NYU STERN Salomon Center for the Study of Financial Institutions

Peliminary Program for the Stern Microstructure Meeting, Friday, May 19, 2017

Supporting funding is provided by NASDAQ OMX through a grant to the Salomon Center at Stern.

Program Committee	Tarun Chordia, Goizueta School, Emory University Joel Hasbrouck, Stern School, NYU Bruce Lehmann, School of Global Policy and Strategy, UCSD Paolo Pasquariello, Ross School, University of Michigan Gideon Saar, Johnson School, Cornell University

The Stern Microstructure Conference is open to everyone with an interest in market microstructure research. The sessions will be held at the Management Education Center, 44 W. 4th St., NYC (near the southeast corner of Washington Square Park). For more complete directions see http://www.stern.nyu.edu/AboutStern/VisitStern/index.htm. There will be a registration desk in the lobby.

Registration Instructions: E-mail salomon@stern.nyu.edu with "SMC2017" in the subject line. Please indicate if you will be joining us for the dinner the night before. Other inquiries: jhasbrou@stern.nyu.edu.

Please note: Hard copies of the papers will not be available at the conference. The schedule below is tentative and subject to revision. (Please don't make travel plans contingent on any particular ordering of the papers.)

Thursday, May 18	
6:30 pm	Dinner (open to all registered conference attendees) *The dinner, breakfast and lunch will be at the school. Registration and check-in will be in the first-floor lobby.*
Friday, May 19	
8:30 am - 9:00	Continental Breakfast
9:00 - 10:00	*A Tale of One Exchange and Two Order Books: Effects of Fragmentation in the Absence of Competition* Alejandro Bernales (University of Chile), Italo Riarte (University of Chile), Satchit Sagade (Goethe University Frankfurt and Research Center SAFE), Marcela Valenzuela (University of Chile), Christian Westheide (University of Mannheim) *Discussant:* Sabrina Buti (Dauphine Université Paris)

10:00 - 11:00	*Toward a Fully Continuous Exchange* Pete Kyle University of Maryland Mina Lee Washington University in St. Louis *Discussant:* Haoxiang Zhu (MIT)
11:00 - 11:15	Break
11:15 - 12:15	*Institutional Rigidities and Bond Returns around Rating Changes* Matthew Spiegel (Yale School of Management), Laura Starks (University of Texas at Austin) *Discussant:* Kumar Venkataraman (Cox School, SMU)
12:15-1:15	Lunch
1:15-2:15	*Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays* Michael Brolley Wilfrid Laurier University David Cimon Bank of Canada *Discussant:* Eric Budish, University of Chicago
2:15-3:15	*A Model of Multi-Frequency Trade* Nicolas Crouzet Northwestern University Ian Dew-Becker Northwestern University Charles Nathanson Northwestern University *Discussant:* Laura Veldkamp, Stern School, NYU
3:15-3:30	Break
3:30-4:30	Secondary Market Trading and the Cost of New Debt Issuance Ryan Davis (University of Alabama Burmingham), David Maslar (University of Tennessee), Brian Roseman (California State University Fullerton) Discussant: Carole Comerton-Forde (University of Melbourne)
4:30	Adjourn

A Tale of One Exchange and Two Order Books: Effects of Fragmentation in the Absence of Competition

Alejandro Bernales* Italo Riarte† Satchit Sagade‡ Marcela Valenzuela§
Christian Westheide¶

First version: August 2016
This version: May 2017

Abstract

Exchanges nowadays routinely operate multiple limit order markets for the same security that are almost identically structured. We study the effects of such fragmentation on market performance using a dynamic model of fragmented markets where agents trade strategically across two identically-organized limit order books. We show that fragmented markets, in equilibrium, offer higher welfare to intermediaries at the expense of investors with intrinsic trading motives, and lower liquidity than consolidated markets. Consistent with our theory, we document improvements in liquidity and lower profits for liquidity providers when Euronext, in 2009, consolidated its order flow for stocks traded across multiple, country-specific, and identically-organized limit order books onto a single order book. Our results suggest that competition in market design, not fragmentation, drives previously documented improvements in market quality when new trading venues emerge; in the absence of such competition, market fragmentation is harmful.

Keywords: Fragmentation, Competition, Liquidity, Price Efficiency

JEL Classification: G10, G12

*University of Chile (DII), abernales@dii.uchile.cl
†University of Chile (DII)
‡Department of Finance and Research Center SAFE, Goethe University Frankfurt, sagade@safe.uni-frankfurt.de
§University of Chile (DII), mvalenzuela@dii.uchile.cl
¶Finance Area, University of Mannheim, and Research Center SAFE, Goethe University Frankfurt, westheide@uni-mannheim.de

For helpful comments and discussions we thank Jonathan Brogaard, Peter Gomber, Jan-Pieter Krahnen, Katya Malinova, Albert Menkveld, Andreas Park, Talis Putnins, Ioanid Rosu, Erik Theissen, and Vincent van Kervel, conference participants at the 2014 Market Microstructure: Confronting Many Viewpoints Conference, 2016 SAFE Market Microstructure Workshop, 2016 CMStatistics Conference, 2016 India Finance

Conference, 2017 Securities Markets: Trends, Risks and Policies Conference, and seminar participants at University of Birmingham, University of Mannheim, University of Manchester, University of Frankfurt, University of Chile, Pontifical Catholic University of Chile. Sagade and Westheide gratefully acknowledge research support from the Research Center SAFE, funded by the State of Hessen initiative for research LOEWE. Valenzuela acknowledges the support of Fondecyt Project No. 11140541 and Instituto Milenio ICM IS130002.

When you split these liquidity pools [...] what happens is that overall volumes tend to go up because the market starts to arbitrage and tries to put the market back together, the value of data goes up. And the whole thing for us turns out to be very good business [...] we don't think it's in the best interest of the market [...]

– Jeffrey Sprecher, Chairman and CEO, Intercontinental Exchange during the Q1 2017 Earnings Call dated 03 May 2017

1. Introduction

Increased fragmentation of trading activity has been one of the most significant changes experienced by equity markets in recent years. Equity markets in the United States, the European Union, and elsewhere have evolved from national/regional stock exchanges being the dominant liquidity pools to a fragmented multi-market environment where a stock now trades on multiple exchanges. These markets have simultaneously also experienced a process of consolidation as a result of national and international mergers of exchanges such that only a small number of operators, each running several exchanges, now compete with one another. For example, in the United States, the three large exchange operators – Intercontinental Exchange, Nasdaq OMX, and BATS – currently operate a total of ten lit equity exchanges. While it is possible that exchange operators allow a certain degree of competition between the different exchanges they own, it appears implausible that such competition would be similar to that between exchanges run by different operators. In most cases, the individual exchanges operated by a single operator employ almost identical rules and use the same technology such that differences between exchanges are minimal. This raises the question as to the effects of fragmentation when competition between venues is absent or minimal.

In this paper, we examine the effects of fragmentation on market performance through a dynamic equilibrium model which characterizes such a multi-market environment. Our

model is set up as a stochastic trading game in which a single asset can be traded in two identically-organized limit order markets. Agents, who are heterogeneous in terms of their intrinsic economic reasons to trade the asset, enter the market following a Poisson process, and make endogenous trading decisions depending on market conditions (e.g. where to submit an order, the type of order, and the limit price). Agents can reenter the market to revise or cancel previously submitted limit orders. They make optimal decisions depending on the state of both limit order books, the stochastically evolving fundamental value of the asset, their private values, and costs of delaying order execution. Limit orders in both order books are independently executed based on price and time priority. By comparing a multi-market environment to a consolidated market setup, we analyze the effects of fragmentation across multiple venues when these venues do not actively compete with each other.

Our model builds on those developed by Goettler et al. (2005, 2009) to characterize a single limit order market. They present a dynamic model in which investors make asynchronous trading decisions based on the prevailing market conditions. We extend their model to describe a fragmented limit order market setting. This is a non-trivial task as the diversity of trading options and trading rules in this setting significantly increases the decision-state space. Furthermore, in contrast to Goettler et al. (2005, 2009), we do not rely on model simplifications to reduce this large state space.

We focus on liquidity, price efficiency, and welfare. In the model, agents endogenously decide whether they provide or consume liquidity. Agents who have an intrinsic motive to trade balance the delay costs associated with submitting limit orders and immediacy costs associated with submitting market orders when determining their optimal strategy. Agents with large absolute private values are more likely to submit market orders because of the proportionally higher expected delay costs. Agents with no intrinsic trading motives generate their profits solely from the trading process. Consequently, they are more patient and hence act as intermediaries by either submitting new limit orders, or sniping mispriced limit orders as in Budish et al. (2015).

In a fragmented environment, agents who provide liquidity submit less aggressive limit orders than in a consolidated market because they can submit an order to one market in order to avoid the time priority of standing limit orders in the second market. This reduction in competition among liquidity providers in a fragmented market translates into higher immediacy costs for liquidity demanding agents.

A comparison of welfare observed in the two different market setups shows that aggregate welfare does not differ markedly between a consolidated and fragmented market. However, the distribution of welfare between the different agent types changes, primarily due to lower price competition in fragmented markets. Agents without any intrinsic trading motive are better off in a fragmented market; their expected payoffs are significantly higher as they obtain better terms of trade. Conversely, fragmented markets are welfare-reducing for agents with exogenous trading motives due to higher costs of obtaining immediacy.

Agents' order submission strategies in fragmented versus consolidated markets have a direct impact on liquidity and price discovery. We find that quoted spread and top-of-book depth are higher in the multi-market environment. We also observe that actual trading costs, proxied using effective spreads, and liquidity providers trading gains, proxied using realized spreads, are lower in a single market setup. At the same time, microstructure noise, defined as the absolute difference between quote midpoint and the fundamental value of the asset, is also higher when markets are fragmented. The above results hold irrespective of whether we measure liquidity and microstructure noise using local or inside quotes. These results also assume exogenous market entry and constant agent populations in both scenarios.

If we were to endogenize market entry of different agent types by allowing them to make entry decisions based on the trade-off between expected trading profits and participation costs, the higher profits in fragmented markets earned by agents without any intrinsic motive should lead to their increased participation. In a computationally simpler alternative, we re-parameterize the model by doubling the number of such agents in the fragmented market and compare its outcomes to those observed under the original parameterization. We find

that quoted bid-ask spreads – albeit lower than in the earlier discussed fragmented market – remain higher than in the single market. Quoted depth in this setup is also highest across the three scenarios. Effective and realized spreads remain higher than in the single market. Conversely, price efficiency improves in this setup because the presence of a higher number of intermediaries leads to prices reacting faster to the arrival of public information. Finally, we obverse an incremental shift in welfare towards agents without intrinsic trading motives when their arrival rate is doubled, while aggregate welfare does not change significantly.

We empirically test the model predictions by examining a unique event in which Euronext, starting 14 January 2009, implemented a single order book per asset for their Paris, Amsterdam, and Brussels markets. Euronext previously operated multiple independent order books for stocks cross-listed on these markets. The event led to a decrease in fragmentation for the affected stocks. Existing empirical studies, such as Foucault and Menkveld (2008), Hengelbrock and Theissen (2009) and Chlistalla and Lutat (2011), examining the effects of new exchange operators entering a market can be viewed as joint tests of fragmentation and competition. This is because the entry of a new market, in addition to increasing fragmentation, also materially alters the competitive environment. The new operator typically attempts to differentiate its platform along critical features such as trading speed, transaction fees, or the ability to execute large blocks. In contrast, the multiple order books operated by Euronext had exactly identical trading protocols before the implementation of a single order book.

The empirical analysis broadly confirms the theoretical results. We find quoted spreads in the consolidated market to be lower by 30% than local spreads in an individual order book before the event. Quoted depth (both local and at the inside quotes) is also higher after consolidation but the results are statistically insignificant. This is consistent with the empirical level of intermediation in fragmented markets being in between the two theoretically modeled scenarios. Consistent with our theoretical results, effective spreads, both measured using local and inside quotes, are smaller after consolidation. Higher competition in the single order book reduces the potential for rent extraction by liquidity providers, resulting

in 35% lower realized spreads after consolidation. Price impact, the other component of the effective spread, in the absence of private information measures the extent of trading at stale prices, and remains unchanged when compared to the price impact based on inside quote midpoints in the fragmented market. Price efficiency, measured using autocorrelations and variance ratios, also improves after consolidation, although the improvements are weakly significant at best.

While we are unable to empirically compute welfare effects, we find that the introduction of a single order book leads to a weakly significant increase in trading volume, This is despite the elimination of arbitrage trades between the multiple Euronext markets, which are responsible for up to 7.8% of the trading volume before the introduction of a single order book. This is likely due to reduced transaction costs allowing more participation by investors with intrinsic trading motives and is consistent with our theoretical results.

Our results contribute to the literature on equity market fragmentation.[1] Early theories on fragmentation such as Mendelson (1987), Pagano (1989), Chowdhry and Nanda (1991) highlight the positive network externalities generated by consolidating trading on a single venue. Harris (1993) argues that fragmentation can emerge as a consequence of real-world frictions and heterogenous trading motives. Even in some of the above models, a consolidated market is no longer the equilibrium outcome when the fragmented markets differ in their absorptive capacity and institutional mechanisms (Pagano, 1989), and when traders are allowed to split their orders over time (Chowdhry and Nanda, 1991). Madhavan (1995) argues that markets fragment only if there is a lack of trade disclosure. Fragmentation in his model benefits dealers and large traders, and increases volatility and price inefficiency. In possibly the most relevant study to today's competitive landscape of equity markets, Foucault and Menkveld (2008) model competition between two limit order books and predict that the entry of a second market increases consolidated depth, and that increased use of smart order routers leads to an increase in liquidity in the entrant market.

[1] See Gomber et al. (2016) for a detailed survey of this literature.

The empirical study closest to our paper is Amihud et al. (2003) who study the reduction in fragmentation on the Tel Aviv Stock Exchange resulting from the exercise of deep in-the-money share warrants and find an increase in stock price and improvement in liquidity. However, their results cannot be extended to modern equity markets because: (i) the stocks and warrants traded periodically in single or multiple batch auctions as opposed to continuously in limit order markets; (ii) the warrant and the underlying stock cannot be considered as perfectly fungible assets such that investors are indifferent between holding the two.

Hengelbrock and Theissen (2009) and Chlistalla and Lutat (2011) analyze the market entry of Turquoise and Chi-X, respectively, in the European markets and find positive effects on liquidity in the main market. Boehmer and Boehmer (2003) and Nguyen et al. (2007) examine the impact of NYSE's entry in the ETF market and also find improvements in different measures of liquidity. Riordan et al. (2010) find that new entrants contribute to the majority of quote-based price discovery for the FTSE100 stocks in the UK. Kohler and von Wyss (2012) and Hellström et al. (2013) find that fragmentation in the Swedish market increases liquidity, for all but large stocks, and price efficiency for all stocks. O'Hara and Ye (2011) analyze overall fragmentation in the US equity markets and find that it is not harmful to market quality. Degryse et al. (2015) and Gresse (2017) differentiate between lit and dark fragmentation and find that the former improves liquidity, but disagree on the effects of the latter.

We contribute to this literature by analyzing the impact of fragmentation across multiple, identically-organized limit order books on market performance. We consider a dynamic model of multiple limit order markets that incorporates several real-world features and allows for more flexible agent behavior as compared to previous models (see for example Mendelson, 1987; Pagano, 1989; Chowdhry and Nanda, 1991; Biais, 1993; Parlour and Seppi, 2003). We provide evidence that fragmentation has detrimental effects on market quality and welfare, benefiting intermediaries at the expense of agents who trade for intrinsic motives.

The remainder of the paper is structured as follows. Section 2 describes the theoretical model central to our analyses. In Section 3, we analyze the theoretical implications of consolidated versus fragmented markets on welfare and market quality. In Section 4 we present the empirical results from the event study. Finally, we conclude in Section 5.

2. Multi-Market Model

2.1 Model Setting

Consider an economy in continuous-time with a single financial asset that is traded on two independent financial markets. The economy is populated by risk-neutral agents trading the asset. Agents arrive sequentially following a Poisson process with intensity λ, and they can use either of the two financial markets to trade the asset. Agents do not cooperate, and they make trading decisions based on a maximization of expected payoffs. Hence, trading activity in the two financial markets reflects a sequential non-cooperative game, where agents make asynchronous decisions by taking into account private reasons to trade the asset, market conditions and the potential strategies employed by other agents arriving in the future.

The two financial markets in the economy, denoted by $m \in \{1, 2\}$, are organized as limit order markets. Agents can submit limit orders and market orders. A limit order is a commitment made by an agent to trade the asset at a price p in the future, where the value of p is decided by the agent at order submission time. A market order is an order to buy or sell immediately at the best available price, where this price is provided by a previously submitted limit order. Hence, a buy (sell) market order submitted by an agent is always matched with a sell (buy) limit order previously submitted by another agent. Agents submitting limit orders are liquidity providers, whereas agents submitting market orders are liquidity consumers.

As in limit order markets found in the real world, the order books are described by a discrete set of prices at which orders can be submitted. The limit order book at time t and in

market m, $L_{m,t}$, is characterized by the set of prices denoted by $\{p_m^i\}_{i=-N_m}^{N_m}$, where $p_m^i < p_m^{i+1}$ and N is a finite number. Let d be the distance between any two consecutive prices, which will be referred to as tick size (i.e. $d = p_m^{i+1} - p_m^i$). The tick size is assumed to be equal for both limit order books. In both limit order books, there is a queue of unexecuted buy or sell limit orders associated with each price. Let $l_{m,t}^i$ be the queue in the limit order market m at time t associated with price p_m^i. A positive (negative) number in $l_{m,t}^i$ denotes the number of buy (sell) unexecuted limit orders, and it represents the depth of the book $L_{m,t}$ at price p_m^i. Thus, in the book $L_{m,t}$ at time t, the best bid price is $B(L_{m,t}) = sup\{p_m^i | l_{m,t}^i > 0\}$ and the best ask price is $A(L_{m,t}) = inf\{p_m^i | l_{m,t}^i < 0\}$. If the order book $L_{m,t}$ is empty at time t on the buy side or on the sell side, $B(L_{m,t}) = -\infty$ or $A(L_{m,t}) = \infty$, respectively. All agents observe both limit order books (i.e. prices and depths at each price) before making any trading decision.

In each market, the limit order book respects price and time priority for the execution of limit orders. In the book $L_{m,t}$, limit orders submitted earlier at the same price p_m^i are executed first, and buy (sell) limit orders at higher (lower) prices have priority in the queue, even if other orders with less competitive prices are submitted earlier. Time and price priority apply independently for each limit order book.[2] The limit order price determines whether an order is a market order: an order to buy (sell) at a price equal to or above (below) the best ask (bid) price is a market order and is executed immediately at the best ask (bid) price.

Agents can monitor both limit order books. However, due to limited cognition, they cannot immediately modify their unexecuted limit orders after a change in market conditions. In that sense, decisions regarding limit order submissions are sticky. Traders re-enter the market to modify unexecuted limit orders according to a Poisson processes with parameter λ_r, which is the same for both markets and is independent of the arrival process.

Agents are heterogeneous in terms of their intrinsic economic motives to trade the asset.

[2] The existence of an order protection rule ensuring price priority across order books does not affect the outcomes of the model.

These motives are reflected in their private values. Each agent has a private value, α, which is known by the agent. α is drawn from the discrete vector $\Psi=\{\alpha_1, \alpha_2, ..., \alpha_g\}$ using a discrete distribution, F_α, where g is a finite integer. Private values reflect the fact that agents would like to trade for various reasons unrelated to the fundamental value of the asset (e.g. hedging needs, tax exposures and/or wealth shocks). They are idiosyncratic and constant for each agent.

Agents face a cost when they cannot immediately trade the asset, which is called a delaying cost. The delaying cost is reflected by a discount rate ρ applied to the agent's payoff (with $0 < \rho < 1$). The value ρ is constant and has the same value whether orders are executed in $L_{1,t}$ or $L_{2,t}$. This delaying cost does not represent the time value of the money. Instead, it reflects opportunity costs and the cost of monitoring the market until an order is executed.

The fundamental value of the asset, v_t, is stochastic and known by agents; its innovations follow an independent Poisson process with parameter λ_v. In case of an innovation, the fundamental value increases or decreases by d, both with an equal probability of 0.5, where d is the tick size of the limit order books.

The heterogeneity of agents (in terms of private values), the delaying costs and the fundamental value of the asset all play an important role in agents' trading behavior. On the one hand, suppose agent x with a positive private value (i.e. $\alpha > 0$) arrives at time t_x. This agent has to be a buyer because she would like to have the asset to obtain the intrinsic benefit given by α. In this case, the agent's expected payoff of trading one share is: $(\alpha+v_{t'}-p)e^{-\rho(t'-t_x)}$, where p is the transaction price, t' is the expected time of the transaction, and $v_{t'}$ is the expected fundamental value of the asset at time t'. Moreover, if the value of α is very high, the agent may also prefer to buy the asset *as soon as possible* in order to avoid a high delaying cost (i.e. the agent has a discount on the level of α given by $(e^{-\rho(t'-t_x)}-1)\alpha$). She may even prefer to buy the asset immediately using a market order. Consequently, an agent with a high positive private value will probably be a liquidity consumer. However,

there is no free lunch for the liquidity consumer. The agent will probably have to pay an immediacy cost that is given by $(v_{t'} - p)^{-\rho(t'-t_x)}$, since it is likely that $v_{t'} - p < 0$. The agent will accept this immediacy cost because she is mainly generating her profits from the large private value, α, rather than from the transaction *per se*.[3]

On the other hand, suppose an agent y with a private value equal to zero (i.e. $\alpha = 0$) arrives at time t_y. This agent needs to find a profitable opportunity purely in the transaction process because she does not obtain any intrinsic economic benefits from trading. Consequently, she is willing to wait until she obtains a good price relative to the fundamental value. Thus, this agent will probably act as a liquidity provider and receive the immediacy cost paid by the liquidity consumer. It is important to note that agents with $\alpha = 0$ are indifferent with respect to taking either side of the market because they can maximize their benefits by either selling or buying (i.e. by respectively maximizing $(p - v_{t''})e^{-\rho(t''-t_y)}$ or $(v_{t''} - p)e^{-\rho(t''-t_y)}$, where t'' is the expected time of the transaction).

Liquidity providers are also affected by the so-called picking-off risk because limit orders can also generate a negative payoff if they are in an unfavorable position relative to the fundamental value. A limit buy (sell) order executed above (below) the fundamental value of the asset generates a negative economic benefit in the transaction. For example, suppose that the agent I with $\alpha = 0$ first arrives at time $t = 0$. Additionally, suppose that this agent has a standing limit buy order at the best bid price, B in market $m = 1$. Suppose that the current time is t^* and v_{t^*} is the current fundamental value of the asset, such that $v_{t^*} > B$. In this case, the agent can make a positive profit if the order is executed immediately at time t^*; this potential profit is given by $(v_{t^*} - B)e^{-\rho t^*}$. Now suppose at time t^{**}, the fundamental value of the asset decreases to level $v_{t^{**}}$, which is below B (i.e. $v_{t^{**}} < B$) and simultaneously agent II with private value $\alpha = 0$ arrives in the market. Since agent I cannot immediately modify her unexecuted limit order, agent II can submit a market sell order, and pick off the limit

[3] A similar example can be explained in the other direction in case of an agent with a negative private value (i.e. $\alpha < 0$) having a preference to sell.

buy order submitted by agent I. Agent II is thus able to generate an instantaneous profit equal to $(B - v_{t^{**}})$ whereas agent I has a negative realized payoff given by $(v_{t^{**}} - B)e^{-\rho t^{**}}$.[4] Consequently, limit buy orders generally have prices below v_t while limit sell orders have prices above v_t. If that were not the case, a newly arriving agent could pick off limit buy (sell) orders above (below) v_t. This also implies that limit orders in unfavorable positions should disappear quickly from both limit order books.

We center each limit order book at the contemporaneous fundamental value of the asset, i.e. by setting $p_m^0 = v_t$. Suppose at time $t = 0$ the fundamental value is v_0, but after a period τ the fundamental value experiences some innovations and its new value is v_τ, with $v_\tau - v_0 = qd$, where q is a positive or negative integer. In this case, we shift both books by q ticks to center them at the new level of the fundamental value v_τ. Thus, we move the queues of existing limit orders in both books to take the relative difference with respect to the new fundamental value into account. This implies that prices of all orders are always relative to the current fundamental value of the asset. This transformation allows us to greatly reduce the dimensionality of the state-space because agents always make decisions in terms of relative prices regarding the fundamental value of the asset.[5]

Each agent can trade one share and has to make three main trading decisions upon arrival: i) to submit an order either to $L_{1,t}$ or $L_{2,t}$; ii) to submit either a buy or a sell order; and iii) to choose the limit price, which implies the decision to submit either a market or a

[4] A similar example, but in opposite direction, can be explained for the cost of being picked off with a limit sell order below the fundamental value of the asset.

[5] It is important to note that under this normalization, we can still observe limit orders being picked-off. For example, suppose that the current time is t and the fundamental value is v_t; hence $p_m^0 = v_t$. Suppose, that the current bid price is $B(L_{m,t}) = p_m^{-1}$ and the ask price is $A(L_{m,t}) = p_m^2$. Subsequently, at time t_{po}, if the fundamental value decreases by twice the amount of the tick size (i.e. $q = -2$), after re-centering the book, the bid and ask prices are $B(L_{m,t_{po}}) = p_m^1$ and $A(L_{m,t_{po}}) = p_m^4$, respectively. Thus, a newly arriving agent can submit a market order against the limit order at the bid price to generate a profit. Subsequently, the limit order at p_m^1 will disappear, and the new bid price will be below the price at the center of the book (i.e. $B(L_{m,t_{po}+\Delta t}) = p_m^0$, where Δt is the time until the limit buy order above the fundamental value is picked-off).

limit order, depending on whether the price is inside or outside the quotes.[6,7] As mentioned above, an agent can re-enter the market and modify her unexecuted limit order. Hence, she has to make the following additional trading decisions after re-entering: i) to keep her unexecuted limit order unchanged or to cancel it; ii) in case of a cancellation, to submit a new order to $L_{1,t}$ or $L_{2,t}$; iii) to choose whether the new order will be a buy or a sell order; and iv) to choose the price of the new order.

The decision to leave the order unchanged has the advantage of maintaining the it's time priority in the respective queue. The negative side of leaving an order in any of the books unchanged is the potential costs agents can incur when the fundamental value of the asset moves in directions that affect the expected payoff. For example, in the case of a reduction in v_t, a limit buy order could be priced too high. This possibility represents an implicit cost of being picked off. Conversely, when the asset value increases, a buy limit order has the risk of waiting for a long period before being executed.

Therefore, agents have to take the possibility of re-entry into account when they make their initial decision after arriving in the economy. Once an agent submits a limit order, she remains part of the trading game until her order is executed; she exits the market forever after trading the asset.

2.2 Agents' Dynamic Maximization Problem and Equilibrium

There is a set of states $s \in \{1, 2, \dots, S\}$ that describes the market conditions in the economy. These market conditions are observed by each agent before making any decision. The state s that an agent observes is described by the contemporaneous limit order books, L_1 and L_2; the agent's private value α; and in the case that the agent previously submitted a limit order

[6] We can include additional shares per agent in the trading decision. However, similarly to Goettler et al. (2009), we assume one share per trader to make the model computationally tractable.

[7] A potential decision to wait outside any of the markets (without submitting an order) is not optimal because there are no transaction fees, submission fees or cancellation fees. An agent can always submit a limit order far away from the fundamental value such that it is unlikely to be executed, but if executed, the potential economic benefit is high.

to any of the books, the status of that order in L_1 or L_2, i.e. its original submission price, its queue priority in the book, and its type (i.e. buy or sell). The fundamental value of the asset, v, is implicitly part of the variables that describe the state s, since agents interpret limit prices relative to the fundamental value. For convenience, we set the arrival time of an agent to zero in the following discussion.

Let $a \in \Theta(s)$ be the agent's potential trading decision, where $\Theta(s)$ is the set of all possible decisions that an agent can take in state s. Suppose that the optimal decision given state s is $\tilde{a} \in \Theta(s)$. Let $\eta(h|\tilde{a},\ s)$ be the probability that an optimally submitted order is executed at time h. The probability $\eta(\cdot)$ depends on future states and potential optimal decisions taken by other agents up to time h. The probability $\eta(0|\tilde{a},\ s)$ is equal to one if the agent submits a market order, while $\eta(h|\tilde{a},\ s)$ converges to zero as the agent submits a limit order further away from the fundamental value. Let $\gamma(v|h)$ be the density function of v at time h, which is exogenous and characterized by the Poisson process of the fundamental value of the asset at rate λ_v. Thus, the expected value of the optimal order submission $\tilde{a} \in \Theta(s)$, if the order is executed prior to the agent's re-entry time h_r, is:

$$\pi(h_r, \tilde{a}, s) = \int_0^{h_r} \int_{-\infty}^{\infty} e^{-\rho h} \left((\alpha + v_h - \tilde{p})\tilde{x}\right) \cdot \gamma(v_h|h) \cdot \eta(h|\tilde{a},\ s) dv_h dh \tag{1}$$

where $\tilde{p}$ and $\tilde{x}$ are components of the optimal decision $\tilde{a}$, in which $\tilde{p}$ is the submission price and $\tilde{x}$ is the order direction indicator (i.e. $\tilde{x} = 1$ if the agent buys and $\tilde{x} = -1$ if the agent sells). The expression $(\alpha + v_h - \tilde{p})\tilde{x}$ is the instantaneous payoff, which is discounted back to the trader's arrival time at rate ρ.

Let $\psi(s_{h_r}|h_r, \tilde{a}, s)$ be the probability that state s_{h_r} is observed by the agent at her re-entry time h_r, given her decision $\tilde{a}$ taken in the previous state s. The probability $\psi(\cdot)$ depends on the states and potential optimal decisions taken by other agents up to time h_r. In addition, let $R(h_r)$ be the cumulative probability distribution of the agent's re-entry time, which is exogenous and described by the Poisson process governing agents' re-entry with rate λ_r.. Thus, the Bellman equation that describes the agent's problem of maximizing her total

expected value, $V(s)$, after arriving in state s is given by:

$$V(s) = \max_{\tilde{a} \in \Theta(s)} \int_0^{\infty} \left[\pi(h_r, \tilde{a}, s) + e^{-\rho h_r} \int_{s_{h_r} \in S} V(s_{h_r}) \cdot \psi(s_{h_r} | h_r, \tilde{a}, \ s) ds_{h_r} \right] dR(h_r) \qquad (2)$$

where S is the set of possible states. The first term is defined in Equation (1), and the second term describes the subsequent payoffs in the case of re-entries.

The intuition for the equilibrium is that each agent behaves optimally by maximizing her expected utility, based on the observed state that describes market conditions (as in Equation (2)). In this sense, optimal decisions are state dependent. They are also Markovian, because the state observed by an agent is a consequence of the previous states and the historical optimal decisions taken in the trading game. We obtain a stationary and symmetric equilibrium, as in Doraszelski and Pakes (2007). In such an equilibrium, optimal decisions are time independent, i.e., they are the same when an agent faces the same state in the present or in the future.

The trading game is also Bayesian in the sense that an agent knows her intrinsic private value to trade (α), but she does not know the private values of other agents that are part of the game. Hence, our solution concept is a Markov perfect Bayesian Equilibrium (see Maskin and Tirole, 2001). In the trading game, there is a state transition process where the probability of arriving in state s_{h_r} from state s is given by $\psi(s_{h_r} | \tilde{a}, \ s, h_r)$.[8] Thus, two conditions must hold in the equilibrium: agents solve equation (2) in each state s, and the market clears.

As mentioned earlier, the state s is defined by the four-tuple ($L_{1,t}$, $L_{2,t}$, α, *status of previous limit order*), where all variables that describe the state are discrete. Moreover, each agent's potential decision a is taken from $\Theta(s)$, which is the set of all possible decisions that can be taken in state s. This set of possible decisions is discrete and finite given the features of the model. Consequently, the state space is countable and the decision space is finite; thus

[8] It is important to note that $\psi(s_{h_r} | \tilde{a}, s, h_r) = \psi(s_{h_r} | s)$, since optimal decisions are state dependent and Markovian, and we focus on a stationary and symmetric equilibrium.

the trading game has a Markov perfect equilibrium (see Rieder, 1979). Despite the fact that the model does not lend itself to a closed-form solution, we check whether the equilibrium is computationally unique by using different initial values.

2.3 Solution approach and model parametrization

Given the large dimension of the state space, we use the Pakes and McGuire (2001) algorithm to compute a stationary and symmetric Markov-perfect equilibrium. The intuition behind the Pakes and McGuire (2001) algorithm is that the trading game by itself can be used, at the beginning, as a learning tool in which agents learn how to behave in each state. At the beginning, we set the initial beliefs about the expected payoffs of potential decisions in each state. Agents take the trading decision that provides the highest expected payoff conditional on the state they observe. Subsequently, agents dynamically update their beliefs by playing the game and observing the realized payoffs of their trading decisions. Thus, the algorithm is based on agents following a learning-by-doing mechanism.

The equilibrium is reached when there is nothing left to learn, i.e., when beliefs about expected payoffs have converged. We apply the same procedure used by Goettler et al. (2009) to determine whether the equilibrium is reached. The Pakes and McGuire (2001) algorithm is able to deal with a large state space because it reaches the equilibrium only on the recurring states class. Once we reach the equilibrium after making agents play in the game for at least 10 billion trading events, we fix the agents' beliefs and simulate a further 600 million events. Therefore, all theoretical results presented in this paper are calculated from the last 600 million simulated events, after the equilibrium has already been reached.

The multi-market model involves a higher level of complexity than a single market setup. First, the state space increases enormously in a multi-market environment, because all combinations of variable values across the two order books have to be considered. Second, in contrast to Goettler et al. (2005, 2009), we do not use model simplifications to reduce the large state space generated by our multimarket model. Goettler et al. (2005) assume that

cancellations are exogenous, and Goettler et al. (2009) reduce the dimension of the state space by using information aggregation (in the spirit of Krusell and Smith, 1998 and Ifrach and Weintraub, 2016). Goettler et al. (2009) also describe the limit order book by only considering the bid and ask prices, the depth at the top of the book, and the cumulative buy and sell depths in the book. We avoid such model simplifications as they may induce the kernel of state variables to be non-Markovian. We instead solve the model by only employing the Pakes and McGuire (2001) algorithm.[9] While parameterizing our model, we use the same market characteristics for both limit order markets. In addition, since our model is an extension of the dynamic model of a single market presented in Goettler et al. (2009), we use the same parameters as in their study.

We set the intensity of the Poisson process followed by the agents' arrivals to one. A unit of time in our model is equal to the average time between new trader arrivals. The intensity of the Poisson process followed by the agents' re-entry is set to 0.25; the intensity of the Poisson process followed by the innovations of the fundamental value is set to 0.125. We set the tick size in both order books to one, and the number of discrete prices available on each side of the order book on both markets to $N_1 = N_2 = 31$. The delaying cost reflected by the rate ρ is set to 0.05. The private value α is drawn from the discrete vector $\Psi = \{-8, -4, 0, 4, 8\}$ using the cumulative probability distribution $F_\alpha = \{0.15, 0.35, 0.65, 0.85, 1.0\}$.[10]

While market entry is exogenous in our model, we posit that, if entry were exogenous, higher profits generated by any agent type in fragmented markets would likely increase their participation.In a computationally simpler alternative, we create an additional parameter configuration by keeping the arrival rates of agents with non-zero private value unchanged

[9] The implementation of the Pakes and McGuire (2001) algorithm, applied to our multi-market model, requires between 600GB and 800GB of RAM, depending on the parameters used. We relied on a high performance computing facility with latest generation processors and 1TB of RAM, which ran over 5-6 weeks to obtain the equilibrium.

[10] As a robustness check, we multiply the following original Goettler et al. (2009) parameters by 0.8 and 1.2: the delaying cost, ρ; the agents' arrival intensity λ; the innovation arrival intensity of the fundamental value, λ_v; and the re-entering intensity λ_r. The results obtained are qualitatively similar to the results presented here.

and doubling the arrival rate of agents with private value equal to zero. In other words, we set the intensity of agent arrival to 1.3 and draw the different agent types from the cumulative distribution $F_\alpha = \{0.15/1.3, 0.35/1.3, 0.95/1.3, 1.15/1.3, 1.0\}$. In addition to the above rationale, this alternative configuration allows to proxy for a second empirical fact observed in real-world markets. It is often the case that liquidity providers are active in multiple limit order books. van Kervel (2015) describes a model of order cancellations in fragmented markets where high-frequency liquidity providers duplicate their orders across multiple order books to improve execution probabilities while simultaneously managing adverse selection risk. A comparison of the different market outcomes across the three (two fragmented and one consolidated) scenarios allows us to highlight potential effects, if any, associated with increased intermediation in fragmented markets.

3. Theoretical Implications

We are interested in examining the theoretical implications of the effects of market fragmentation on trading behavior, welfare, and market quality. To do so, we generate a dataset of trades and order book updates by simulating 10 million events for the following three specifications: (i) a consolidate market with one limit order book; (ii) a fragmented market with two limit order books; and (iii) a fragmented market with two limit order books and twice as many agents with no intrinsic value as the first two specifications. We compute mean levels of the measures of interest under all three market settings.

3.1 Trading Behavior

The order submission strategy determines the price formation of an asset and the liquidity of the market, and as a consequence, it has a direct effect on the welfare of individuals and society. Hence, it is important to analyze how the introduction of a second limit order book affects the trading behavior of agents. We study the trading patterns of agents in single and

fragmented markets. For the latter, we provide results for two scenarios: when the arrival rate of agents without exogenous reasons to trade is the same as in a single market and when the rate is twice as large. Table 1 presents the results.[11,12]

We find that agents submit more aggressive limit orders in a single market compared to a fragmented market. In a single market setting, about 36% of the orders are placed at the best ask price, whereas this is the case for only about 28% of orders in the fragmented market. If the arrival rate of traders without exogenous reasons to trade is doubled, almost 33% of limit order are submitted at the best ask price, probably because of the higher degree of competition among limit order traders in this setup. More aggressive limit orders in a single market compared to a multiple market setting with same arrival rates lead to a higher picking-off risk, i.e., the share of executed limit orders that are picked off, inducing agents to cancel their orders more often. We find that the picking-off risk is indeed lower in a fragmented market. The results in Table 1 indicate that the picking-off risk declines from 21.80% in a single market to 20.82% in multiple markets. When the arrival rate of intermediaries is doubled, the picking-off risk is even lower. Untabulated results reveal that the picking-off risk, in this setting, is higher for each agent type, which is consistent with a higher competition between speculators. However, this measure decreases on average compared to a single setting, because of the higher share of agents of type $\alpha = 0$, who have the lowest picking-off risk. A higher picking-off risk induces agents to cancel their limit order more often, increasing the execution time from her arrival time until the execution of her limit order. Consistent with this intuition, the average number of limit order cancellations per trader is 1.2 in a single order book as compared to 1.01 when there is a second book. We also corroborate that limit orders execute faster in a multi-market setting. The average execution time is 8.61 in a single market, whereas in a fragmented market the time is reduced

[11] As the model is symmetric we focus on the sell side of the market. The results for the buy side of the market are analogous.

[12] We do not report standard errors because the large number of trader arrivals implies that the standard errors on the sample means are sufficiently low such that a difference in means of an order of 10^{-2} is significantly different from zero.

to 7.15 units of time in a fragmented market with the same distribution of agents.

If we double the arrival rate of market-makers, the number of cancelations is 1.58 and the execution time is 13.10, which is also consistent with higher competition of limit orders inducing agents to cancel more often, increasing, in turn, their execution time. The much longer time until execution can be explained by the fact that an overwhelming share of limit order traders in this setup are intermediaries, who are patient traders.

Table 2 shows the proportions of limit orders and market orders submitted by each trader type. We report the distribution of limit orders and market orders for a given trader type. As expected, we find that agents with intrinsic motives to trade (i.e., $|\alpha| \neq 0$) act as liquidity demanders, whereas agents with no intrinsic motives to trade (i.e., $|\alpha| = 0$) act as liquidity suppliers. Almost all of the agents without intrinsic motive to trade (i.e., $|\alpha| = 0$) act as speculators submitting limit orders. Only about 5% of them submit market orders to take advantage of mispriced limit orders. Conversely, about 72% agents with private value $|\alpha| = 8$ submit market orders.

The behavior of agents with private value $|\alpha| = 4$ is in between those of the other types. The choice between limit and market orders does not markedly differ between the single and multi-market setups with the same trader populations. However, differences in order choice between the trader types are more pronounced when we doubled the arrival rates of zero private value agents, as traders with non-zero α use limit orders much less frequently.

Our findings are consistent with the study of Goettler et al. (2009) who examine the trading behavior in a single market setting. They also find that agents with $|\alpha| = 0$ supply liquidity to the market, agents with extreme valuation ($|\alpha| = 8$) are more likely to demand liquidity, and the behavior of agents with $|\alpha| = 4$ is in between that of the more extreme types.

Although our findings reveal that fragmentation does not change the main strategies adopted by traders, it is interesting to notice that, assuming an unchanged population of

traders, agents with private value $|\alpha| = 8$ submit a higher proportion of limit orders when there are two limit order markets. We will show later that market fragmentation leads to wider spreads. As market orders are more expensive in such a setting, some agents with exogenous reasons to trade prefer to submit more limit orders when there are two limit order books. However, when we increase the arrival rate of market makers, the latter appear to crowd out the limit order submissions of other types of traders.

3.2 Market Quality

In this subsection, we compare consolidated and fragmented markets in terms of the major determinants of market quality, i.e., liquidity and price efficiency.

We begin by estimating the effect of market fragmentation on various measures of quoted and traded liquidity. We calculate liquidity measures employing either *local* or *inside quotes*. Local quotes comprise the bid and ask prices of one of the markets whereas inside quotes are combine the highest bid and the lowest ask across the two limit order books.

We measure daily quoted liquidity by time-weighted quoted spreads and time-weighted top-of-book depth. We also report the total number of limit orders waiting to be executed on the sell side of the market. Panel A of Table 3 provides the results. Our theoretical findings indicate that fragmentation by and large impairs liquidity. This is illustrated by wider spreads and lower depth when there are two limit order markets. In particular, both local and inside quoted spreads decrease about 1.04 and 0.34 ticks, respectively, when the market moves from a fragmented to a single market and the arrival rates of all trader types are the same as in the single market. Spreads are also reduced in the single market compared to when the arrival rate of zero private value agents in the fragmented market is twice as large, although the effect is smaller.

Naturally, because of order flow fragmentation between the two markets, fragmented markets also show a decrease in the top-of-book depth. Local top-of-book depth is reduced

by more than 30% in a fragmented market. Inside depth is also lower as compared to the single market. The results change if we double the participation of agents of type $\alpha = 0$: the increased number of liquidity providers leads to a substantial increase in inside depth, and local depth is also slightly higher than in the single market scenario. Thus, our results with respect to quoted liquidity show that spreads are unambiguously smaller in a single market whereas the results for depth are ambiguous.

Improvements in quoted liquidity do not necessarily translate into actual transaction cost savings for traders submitting market orders. Thus, we next compare differences in traded liquidity in single and fragmented markets. We measure traded liquidity by the trade-weighted effective spreads, which capture the actual transaction costs incurred by traders submitting marketable orders. The effective spread is calculated as follows:

$$\text{effective spread} = x_t(p_t - m_t)/m_t, \tag{3}$$

where x_t is +1 for a buyer-initiated order, p_t is the traded price, and m_t is the mid-quote. We further decompose effective spread into realized spread and price impact (adverse selection). The former is calculated as follows:

$$\text{realized spread} = 2x_t(p_t - m_{t+k})/m_t, \tag{4}$$

where k is the number of seconds in the future. As the results are qualitatively similar, we only report the findings for 30 seconds. Finally, price impact is effective spread minus realized spread. The price impact captures the level of information in a trade, whereas the realized spread measures liquidity providers' compensation after accounting for adverse selection losses associated with informed orders. As our model does not contain private information, the price impact measure captures picking-off risk associated with stale limit orders when new (public) information arrives in the market. Just like quoted liquidity, we compute local and inside variants of all three measures using the inside quote midpoints across the two books and local quote midpoints in the order book where a transaction is

executed.

Panel B of Table 3 reports the results. Transaction costs in terms of effective spread and realized spread are higher when there are two limit order books. When the arrival rate of zero private value agents remains the same, effective spreads decrease from 1.80 ticks and 1.45 ticks based on local and inside quotes, respectively, in fragmented markets, to 1.31 ticks in the single market. The differences are even larger if we double the arrival rate of agents of type $\alpha = 0$.

Realized inside and local spreads are higher in the fragmented market by approximately 0.15 ticks with the same population of agents, and higher by about one half of a tick if we double the participation of intermediaries.

Price impact measured relative to local quotes is lower in the single market whereas it is similar when measured relative to inside quotes. This is because, in fragmented markets, a newly arriving trader is more likely to trade in an order book containing a stale quote, leading to a higher local price impact. The inside price impact is smaller because the inside quote midpoint already reflects part of the information. Price impacts are lower if we double the arrival rate of agents of type $\alpha = 0$, likely because the increased arrival rates leads to the exploitation of even mispricing of even small magnitudes. The local price impact in this scenario is still slightly larger than that in the single market, though the inside price impact is substantially smaller.

Finally, we analyze the degree of inefficiency in prices when the market consists of one order book as opposed to multiple ones. If an asset is traded on multiple markets, the degree of price dislocations may be exacerbated *ceteris paribus*, making prices on each book less efficient than they would be if all demand and supply were to meet on a single order book. In the context of our model, the effect of these frictions is measured as the deviation of the quote midpoint from the fundamental value v_t. In Panel C, we present the mean absolute difference between the quote midpoint and the fundamental value. This value changes from 0.67 ticks

in a fragmented market to 0.46 ticks in a single market. The corresponding differences based on inside quotes are in the same direction although the magnitudes are lower. However, in fragmented markets with doubled arrival rate of zero private value agents, microstructure noise is lower than in the single market, suggesting that prices are more efficient in this case. This result is expected as in the absence of private information, a higher number of traders with no intrinsic reasons to trade results in a faster adjustment of quotes when public information arrives. Thus, the degree of mispricing depends on the number of intermediaries in the market and, because their number in real fragmented markets is likely larger but not twice as large as it is in consolidated markets, our model makes no strong predictions about differences in price efficiency between such markets.

3.3 Welfare Analysis

In order to analyze the potential economic benefits per agent and for the whole market, we examine the effect on welfare of both single and fragmented markets. Welfare is measured as the average realized payoff per agent. In addition, we decompose the realized payoffs of investors to analyze the gains and losses from the trading process.

Suppose that an agent with a private value α and delaying discount rate of ρ arrives on the market at time t. She submits an order (i.e., a limit order or a market order) to any of the books at price $\tilde{p}$ with order direction $\tilde{x}$ (i.e., to buy or to sell). Suppose that the agent does not modify the order, and it is finally executed at time t' when the fundamental value is $v_{t'}$ (in the case of a market order $t = t'$). Then the realized payoff of the agents from the order execution is given by:

$$\Pi = e^{-\rho\left(t'-t\right)}\left(\alpha + v_{t'} - \tilde{p}\right)\tilde{x}. \tag{5}$$

We can decompose the agents' payoffs and rewrite (5) as:

$$\begin{aligned} \Pi = Gains\ from\ private\ value + Waiting\ cost\ + Money\ Transfer,\ where \\ Gains\ from\ private\ value = \alpha\tilde{x} \\ Waiting\ cost = (e^{-\rho(t'-t)} - 1)\alpha\tilde{x} \\ Money\ Transfer = e^{-\rho(t'-t)}(v_{t'} - \tilde{p})\tilde{x} \end{aligned} \tag{6}$$

The first term in (6), *gains from private value*, represents the gains obtained directly from the exogenous reasons to trade for each agent, $\alpha\tilde{x}$. Agents initially submitting a limit order do not trade immediately after arriving on the market. and, thus have to wait until they obtain their private values. This waiting process is costly due to the delaying cost ρ. The second term in (6), *waiting cost*, reflects the cost paid by agents in terms of delaying the *gains from private value*.

The realized payoff in (5) results from a transaction in which one agent buys the asset and another agent sells it at a price that may differ from the fundamental value. The third term in (6), *money transfer*, reflects the difference between the fundamental value $v_{t'}$ and the transaction price $\tilde{p}$, and thus the money gained (or lost) in the transaction. It is discounted depending on the arrival time of the trader. In general, the *money transfer* is associated to the immediacy cost incurred when an agent wants to immediately realize her private value. For example, an agent who submits a market order realizes her intrinsic private value without delay. Thus, this trader does not have any *waiting cost*, but she may have to pay a cost for demanding immediacy, which would be reflected in a negative *money transfer*.

Table 4 presents the results. In the first set of columns, we present the results of (5), i.e., the average payoff for each trader type in each market scenario. We find a similar global welfare in the three setups. While the aggregate welfare effects are altogether negligible, the shifts among categories of agents are substantial. Agents with non intrinsic motives to trade (i.e., $|\alpha| = 0$) take more advantages from fragmented markets and, as a consequence, have

higher profits. When we double the arrival rate of agents without intrinsic motive to trade, the welfare for each such agent decreases, but their aggregate welfare is larger than in the other two scenarios. In a single market scenario, agents place more aggressive orders to jump the queue and thus to raise their probability of execution. This fact generates competition on price due to the consolidation of all order flow in a single trading venue. Contrarily, in the presence of multiple markets, as there is no time priority across order books, traders can, with a positive probability, jump ahead of the standing limit orders in one market by submitting an order to the other market. The lack of time priority reduces competition on price such that agents with $|\alpha| = 0$ obtain better terms of trade.

As aggregate welfare effects in the model are not quantitatively meaningful, any interpretation regarding the overall desirability of fragmentation needs to go beyond the model. Market participation in the model is exogenous. In real markets, one would expect that traders endogenously decide about their market entry based on the trade-off between expected trading profits and participation costs. Thus, the higher profit earned by liquidity providers in fragmented markets should lead to an increased participation of this group of traders. If participation in markets is costly - a realistic assumption considering the investments made by intermediaries in modern equity markets - these additional traders incur costs that do not increase aggregate welfare, i.e., the privately optimal decisions are not socially optimal. This suggests there are welfare losses resulting from market fragmentation.

Next we analyze the second and third components of total payoff described in (6). In the next set of columns, we report the waiting cost and money transfer per trader.[13] Agents with intrinsic motives to trade (i.e., $|\alpha| \neq 0$) exhibit a reduction in absolute waiting costs in the fragmented market, even more so if market maker participation is doubled. However, they obtain worse terms of trade, which is reflected in high money transfer costs. For example,

[13]Note that in Table 4, the total money transfer do not add up to zero because they are discounted back to time t and $t' - t$ is different for the trader who submits the market order and the trader who submits the corresponding limit order due to traders' asynchronous arrivals. However, the instantaneous money transfer not discounted back does add up to zero.

agents with private value $|\alpha| = 8$ experience smaller money transfer losses in consolidated markets as compared to fragmented markets (-0.572 ticks versus -0.626 or -0.835 ticks). This is because lower waiting costs do not compensate for the losses associated with money transfer. Finally, agents with $|\alpha| = 0$ obtain higher gains from trading in fragmented markets primarily through higher money transfer gains.

In conclusion, the welfare of agents with non intrinsic motives to trade is increased under the presence of a second limit order book and this gain is to the detriment of traders with non-zero private values, likely because price competition in fragmented markets is less severe. Hence, they pay the cost associated with higher profit for agents with $|\alpha| = 0$ in fragmented markets.

4. Empirical Application

In this section, we test the empirical predictions generated by our model in Section 3.2. We also indirectly address the predictions about welfare described in Section 3.3.[14] To this end, we conduct an event study based on Euronext's decision to implement a single order book per asset for their Paris, Amsterdam, and Brussels markets. Pagano and Padilla (2005) and Nielsson (2009) analyze the effects of integrating trading on Euronext for stocks listed in these three markets.

4.1 Euronext's Institutional Background

We begin by describing Euronext's institutional arrangements leading up to the introduction of the Single Order Book. Euronext was formed in 2000 following a merger of the Paris, Amsterdam and Brussels stock exchanges. In 2002, the Lisbon Stock Exchange became the fourth exchange to merge with Euronext.[15] Stock listings on Euronext pertain to a

[14] We cannot empirically test the predictions from Section 3.1 pertaining to trading behavior due to data limitations.

[15] In 2007, Euronext merged with the NYSE to form NYSE Euronext, which was taken over by Intercontinental Exchange in 2012. In 2014, Euronext was spun off through an IPO.

listing on one or more national markets.[16] Until 13 January 2009, each national listing corresponded to the operation of one limit order book. For example, a stock listed on the Paris market would be traded on the limit order book of Euronext Paris. Firms cross-listed in multiple Euronext markets traded in parallel on multiple Euronext order books, besides other competing markets. On 16 August 2007, the exchange announced its intention to eliminate this arrangement for their Paris, Amsterdam and Brussels markets by unifying all trading in these markets on to a single order book, the so-called "Market of Reference" (MoR). Cross-listed firms had to choose one MoR that continued operating after the implementation of a single order book. This new arrangement was implemented on 14 January 2009.

The existence of multiple order books led to fragmentation of order flow routed to Euronext. As the rules and trading protocols governing the individual order books were identical, the introduction of a single order book decreased fragmentation for the stocks without any corresponding change in the competitive environment. Pagano and Padilla (2005) describe the steps taken by Euronext to standardize its trading protocols and technological platform as the source of the efficiency gains generated through the merger. This is particularly relevant as it allows us to test the isolated effects of fragmentation. Euronext, in its press release announcing the event, made clear that the trading environment remained unchanged: "The Single Order Book will have no impact on the NSC system as the market rules and order book management will remain unchanged [...] In practice, from a trading perspective, Single Order Book implementation simply means the end of order book trading on marketplaces other than the market of reference."[17] Moreover, as it was based on a business decision by Euronext, all multi-listed stocks received the same treatment such that there was no selection bias. Finally, the announcement was made more than one year before the event date in order to allow market participants to adapt and test their trading systems. This eliminates potential concerns about the event date confounding with other

[16] With the implementation of the Markets in Financial Instruments Directive (MiFID), all rules prohibiting trading outside the national markets were repealed such that investors can now trade these stocks in any regulated market.

[17] See Euronext press release dated 14 January 2009.

market events around the same time.[18] Thus, this empirical analysis of a transition from a multi-market environment to a single market setup can be viewed as a natural experiment, allowing us to compare the outcomes with those obtained from our theoretical model.

4.2 Sample Selection

A total of 45 instruments, cross-listed on at least two of the three Euronext markets, are affected by the event. However, we reduce the sample of treated stocks used in our study for several reasons. First, we remove stocks whose primary listing is not on Euronext. These include stocks whose main trading activity takes place in other European markets or in the United States. Second, we eliminate exchange-traded mutual funds because we do not expect their trading activity to be comparable to that of listed firms. Finally, we require that there not only exist multiple Euronext order books before the event, but also that the total share of trading activity on the less active order books is at least equal to 1% of the respective stock's total Euronext trading volume. This reduces the list of instruments to ten. We further exclude one additional stock due to data errors, reducing our final sample to nine stocks.[19] The number of stocks is small due to the unique nature of the event we study. Nonetheless, our sample consists of the whole population of stocks affected by the event, except a subset of stocks which are excluded through objective criteria.

We construct a matched control group of stocks based on stock price and market capitalization obtained from Compustat Global using the distance metric employed by Huang and Stoll (1996), and subsequently, in many other market microstructure studies. Specifically, for each stock in our treatment group, we identify the stock that is its closest match in terms of these two criteria as on the last trading day of 2008 (30 December 2008). The population of stocks from which the control group is constructed comprises all stocks with a primary

[18] Although the original date of implementation was postponed, this was due to technical reasons as opposed to concerns about market conditions. The final implementation date was announced more than 60 days in advance.

[19] One stock in our sample was listed in all three Euronext markets. However, we exclude the least active limit order book as it had a market share of 0.3%.

listing on Euronext not affected by the event.

Davies and Kim (2009) simulate the matching performance of a control group constructed using multiple criteria such as price volatility, trading volume and industry classification, in tests of differences for variables typically used in the microstructure literature, and conclude that one-to-one sampling without replacement based on stock price and market capitalization provides the best results. They also show that results obtained by matching based on the distance metric employed by Huang and Stoll (1996) are similar to those obtained when using the Mahalanobis distance measure.

Using a control group allows us to identify the effects of reduced fragmentation, implicitly controlling for market-wide changes in variables such as liquidity and volatility. It also allows us to control for two additional market-wide changes implemented by Euronext close in time to the introduction of a single order book. First, a harmonized settlement platform known as the Euroclear Settlement for Euronext-zone Securities for all French, Dutch and Belgian stocks was implemented on 19 January 2009. Second, the Universal Trading Platform, having "superior functionality, faster speed and much greater capacity", was introduced on 16 February 2009.[20] These were market-wide events that affected both the control and treatment stocks. Consequently, we can attribute any difference in trading activity and market quality between the two groups exclusively to market consolidation resulting from the introduction of a Single Order Book.

For the purpose of our analysis, we define all days from the beginning of December 2008 to 13 January 2009 as the pre-event period and all days from 26 January 2009 to the end of February 2009 as the post-event period. We exclude all trading days from the event date until the end of the subsequent calendar week in order to eliminate any effect associated with the transition.

[20]See press release titled "NYSE Euronext's European Equities Trading Successfully Migrates to the Universal Trading Platform" dated 17 February 2009.

4.2.1 Data Description and Summary Statistics

We use high-frequency data from Thomson Reuters Tick History between December 2008 and February 2009 for the purpose of our analysis. This data contains trades and order book updates time-stamped with a millisecond resolution. We apply two filters to the data. First, we eliminate all order book updates where the best bid or ask prices are zero or the bid-ask spread is negative. Next, we exclude trades that are executed during the opening and closing auctions as well as trades within the first and last minute of the continuous trading session.

Table 5 describes the characteristics of stocks in the treatment and Huang and Stoll (1996) control group. The average market capitalization across stocks in the treatment (control) group of €4.4 (€4.8) billion and the average stock price of €18.4 (€18.2) are suggestive of high matching quality based on these two variables. The share of the more active venue as a percentage of total Euronext volume across all the days before the event ranges from 54% to 98% across the nine stocks in the treatment group. The simple (volume-weighted) average across all stocks is 78% (62%). This implies that almost 40% of the total Euronext volume was executed on the less active market. The market share of the sole listing Euronext venue for the stocks in the control group is, by construction, 100%. Trading activity when measured in terms of number of trades also provides a similar picture.

4.2.2 Estimation Methodology

In order to test the main implications of our model, we compute several variables capturing the trading activity, liquidity and price efficiency of the stock, as described in Section 3.2. Similar to our numerical results, we calculate both local and inside measures. Unsophisticated investors who choose to trade only on a single order book are likely to select the more active and liquid one. Hence, we compute the local measures for the market having higher trading volume during the pre-event period. The inside measures use the highest bid and the lowest ask across the two limit order books. We estimate a panel difference-in-differences

regression with stock and day fixed effects and standard errors double clustered by stock and day. We estimate the regression for levels and natural logarithms of the variables of interest in order to account for the wide dispersion in the levels of these variables across the stocks in our sample.

4.3 Empirical Results

4.3.1 Quoted Liquidity

We begin by analyzing the effect on quoted spread and top-of-book depth. Table 6 presents the results. Consistent with our theoretical findings, we observe an overall improvement in quoted spreads on Euronext after the introduction of a single order book. The more active of the two Euronext markets experiences a significant reduction in local spreads of 81bps or approximately 30%. The effect on inside spreads depends on the test specification and is statistically insignificant. The absence of a significant improvement in the inside spread can be explained by the fact that in real markets, different from our model, market participants do not always route their orders optimally, i.e. to the market offering the highest bid or lowest ask.[21] Thus, while in the model inside spreads correctly reflects the gains, before adverse selection, that liquidity providers expect to earn, a non-zero probability of traders routing their orders to the market not offering the best price means expected gains earned by liquidity providers are in reality larger. This effect of suboptimal order routing vanishes after consolidation, *ceteris paribus* leading to an increase in inside quoted spreads. Conversely, an increase in price competition among liquidity providers, as predicted by the theory, leads to a decrease in quoted spreads after consolidation. These effects empirically cancel out such that the coefficients for inside spreads appear insignificant.

[21] In European markets, best execution requirements allow brokers to consider other criteria besides price when making order routing decisions. In contrast to the US, European markets also do not have an order protection rule that requires exchanges to re-route orders to venues offering a superior price. Even in the US, communication latencies between geographically dispersed exchanges and exceptions to the order protection rule result in liquidity takers obtaining sub-optimal prices.

We observe a positive though statistically insignificant effect of order flow consolidation on local and inside top-of-book depth. These results lie in between those observed in the simulations with different participation rates of market makers. In other words, they are consistent with an amount of market-making in fragmented markets that is larger than, but less than twice as large as that in a consolidated market. The results for local and inside depth do not markedly differ, which is in contrast to the theoretical predictions where inside depth in the fragmented market is relatively higher. Differences between the tick sizes on Euronext as compared to those in the theory may drive this result. The empirical tick size, relative to the price fluctuation, is substantially smaller than that in the simulations,[22] such that instances with the same best prices offered on the two order books are infrequent, leading to a relatively smaller inside depth than in a large-tick market. Ye (2017) illustrates the negative relationship between flickering quotes and tick size in a single market setup. Extending this argument to fragmented markets, prices across multiple markets will be synchronized less frequently when the tick size is small.

4.3.2 Traded Liquidity

Table 7 presents the results for effective spreads and their decomposition. Effective inside (local) spreads decrease by an economically large 14.5% (37.5%) after the introduction of a single order book, though only the results for local spreads are unambiguously statistically significant. Realized inside spreads significantly decrease by 31.3%, 42.1%, and 45.9%, at the 10-, 30-, and 60-second horizon, respectively. The corresponding decrease in realized local spreads is larger in magnitude and also significant. Local price impacts decrease across all specifications, even though the statistical significance varies. All the above mentioned results on traded liquidity are consistent with our theoretical predictions.

The empirically observed change in inside price impacts differs depending on the empirical

[22]Tick sizes on Euronext during our sample period are smaller than in most international markets. A simulation applying parameters that would closely match Euronext tick sizes is infeasible because the large number of possible prices would lead to a corresponding increase of the state space.

specification and is never significant, whereas our theory tells us the effect should be near-zero or positive. The previously-mentioned fact that liquidity takers empirically sometimes do not trade on the market offering the best price may explain why inside price impacts are not larger in the fragmented market. An order trading against a standing limit order at a price inferior to the lowest ask or highest bid does not mechanically generate an inside price impact even if it executes against the entire limit order, leading to a relatively smaller inside price impact compared to the theoretical predictions. Additionally, the empirical results, in contrast to the model, also capture the effects associated with private information possessed by traders. Lower transaction costs potentially allow traders with small amounts of private information to profitably participate, leading to a decrease in average price impact. The latter channel may cancel out the positive effect of consolidation on price impact predicted by our theory.

4.3.3 Price Efficiency

In our numerical simulation, we examine the price efficiency by measuring the extent to which the mid quote deviates from the fundamental value v_t. Empirically, as we cannot observe the fundamental value, we measure price efficiency using return autocorrelations and variance ratios. Return autocorrelations are measured at 30 second and 5 minute intervals. Variance ratios capture the deviation between long-term and short-term return variance and are calculated as one minus the ratio of long-term and short-term return variance, each scaled by the respective time periods. We calculate variance ratios between 30 second and 5 minute returns variances. As in Boehmer and Kelley (2009), we measure the impact of consolidation on absolute values of both measures because we are interested in departures from a random walk in either direction. The closer these measures are to zero, the more closely does the price path resemble a random walk. Similar to the liquidity measures, we calculate price efficiency based on local and inside quotes.

Table 8 presents the results.[23] The variance ratio becomes closer to one after the implementation of the single order book, though the change is statistically insignificant. The results for autocorrelations also point to improved price efficiency although only results for the 5-minute autocorrelation are significant at the 10% level. These results generally provide evidence for unchanged or higher price efficiency in consolidated markets. When compared to the theoretical results, this appears consistent with a fragmented market containing more but less than twice as much intermediation as a consolidated market.

4.3.4 Trading Volume and Arbitrage

The existence of multiple order books empirically allows market participants to earn arbitrage profits by exploiting occasions of crossed markets, i.e. situations where the bid price on one order book is higher than the ask price on the other. These situations would otherwise be immediately resolved by the adjustment of limit order prices. In other words, such trades do not contribute to an increase in price efficiency, but only result in losses for limit order traders who consequently impose higher trading costs on liquidity seekers. This arbitrage-driven rent extraction may lead to welfare losses if otherwise beneficial trades are crowded out (Foucault et al., 2017; Budish et al., 2015).

We measure trades associated with such "toxic" arbitrage and the resulting costs to market-makers in the empirical data as follows. We start by identifying instances of a crossed order book. Such a situation can arise as a result of new order(s) submitted to either or both order book(s). Next, we identify whether these instances are resolved through a trade, quote-update, or both. This approach is similar to Foucault et al. (2017) who define the resolution through trades as toxic arbitrage if the following two conditions are fulfilled: (i) prices offered in different markets allow aggressive traders to earn a profit by trading against the bid on one market and ask on the other; (ii) they are able to do so because of liquidity providers' slow reaction to new information, rather than them offering attractive prices to manage their

[23] The empirical results based on returns measured at other frequencies are qualitative similar to those reported here and are available upon request.

inventories. Fragmentation is an obvious precondition for such arbitrage trades to occur. For each stock-day, we calculate the number of unique crossed instances, the fraction of a day when inside spreads are on average negative, and the total trading volume contributing to the resolution of a crossed market. Panel A of Table 9 reports the mean values for each stock across all days in the pre-event period. The frequency of unique instances of a crossed market for an average day ranges from 0.3 to 622 across all stocks, with an average value of 124, which corresponds to one instance every four minutes. An average stock has a negative inside spread for 6.4% of the continuous trading session. Finally, 7.8% of the total trading volume on Euronext for an average stock can be attributed to the resolution of instances where the two markets are crossed. Approximately 50% of this, or almost 4% of the total Euronext trading volume, is associated with toxic arbitrage as defined in Foucault et al. (2017).

Since, by construction, arbitrage trades between multiple Euronext order books are eliminated after the introduction of a Single Order Book, trading volume should, everything else equal, be reduced. Panel B of Table 9 shows that the actual change in trading volume is in fact weakly positive. This suggests that the volume transacted by investors with intrinsic motives to trade increases in the consolidated order book. This welfare gain is consistent with our theory. The amount of volume traded by agents with intrinsic reasons to trade is constant in the model because private values are assumed to be sufficiently large such that they never refrain from trading. The increase in trading by such agents suggested by our empirical results indicates that some traders who were earlier crowded out in a less liquid fragmented market, now participate, leading to an overall welfare gain after consolidation.

5. Conclusion

We examine the effects of market fragmentation when competition between markets is non-existent or at best minimal. Such fragmentation is routinely observed after exchange mergers,

when a single exchange operator continues operating multiple order books to trade the same asset post merger. In an attempt to extract synergies from the merger, the operator typically eliminates structural and technological differences across the merging markets resulting in operator-level order flow fragmenting across (nearly) identical limit order books.

Our model allows us to examine the effects on several aspects of market performance such as liquidity, price efficiency, agents' payoffs and overall welfare. As limit order priority is not enforced across markets, fragmentation leads to reduced competition between intermediaries. This results in the deterioration of liquidity in fragmented markets as compared to the consolidated market benchmark. While overall welfare remains largely unchanged under both market setups, the distribution of welfare across the heterogeneous agent types in the model is markedly different. Agents with intrinsic trading motives extract lower payoffs in fragmented markets whereas agents acting as intermediaries are better off in fragmented markets.

These higher intermediation gains should, under conditions of endogenous entry, lead to more intermediaries entering the market. We mimic these conditions by doubling the population of intermediaries in the model while keeping all other market parameters constant. We observe that under these conditions the allocation of trading gains between intermediaries and non-intermediaries shift further in favour of the former while still not altering overall market welfare materially. These results point to fragmentation-induced investment in intermediation capacities, such as high-speed connections required to access the trading systems and real-time data feeds from multiple venues, being socially wasteful.

We empirically test the model implications by investigating the effects of Euronext's decision to introduce a single order book for their Paris, Amsterdam, and Brussels markets. As opposed to existing empirical research on this question which necessarily investigates the joint impact of changes in fragmentation and competition (say, when a new trading center venue the market), this event allows us examine the effects associated with the consolidation of multiple non-competing order books. The empirical analysis broadly confirms the

theoretical predictions related to the effects on liquidity, price efficiency, and market makers' profits. Additionally, we also obtain evidence that trading volume after consolidation does not decrease even though the amount of arbitrage trading in the market mechanically reduces after the event. This suggests that, while the (substantial) revenues generated by modern exchanges' from the sale of market data may decrease after consolidation, improvements in market quality need not come at the expense of reduced trading fees for the exchange operators.

Overall our results suggest that the positive externalities associated with consolidating order flow in a single location (or fewer locations) still exist and are substantial. This is true even in modern electronic limit order markets where the activities of high-frequency traders serve to integrate fragmented order books. The adverse effects of fragmentation are significantly larger for unsophisticated investors who do not possess the technological ability to route their trades to the most advantageous trading center. For such investors consolidation of order flow, at least between non-competing markets, likely results in transaction cost reductions.

Our results also have important policy implications. Regulators may be able to improve the welfare of investors who trade for intrinsic motives by: (i) preventing individual market operators from keeping an artificially high(er) level of order flow fragmentation in the absence of commensurate benefits; and (ii) limiting excessive investment in intermediation capacities necessary to link multiple order books which come at a cost to end investors.

References

Amihud, Y., B. Lauterbach, and H. Mendelson (2003). The value of trading consolidation: evidence from the exercise of warrants. *Journal of Financial and Quantitative Analysis 38*(04), 829–846.

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

Boehmer, B. and E. Boehmer (2003). Trading your neighbor's ETFs: Competition or fragmentation? *Journal of Banking and Finance 27*(9), 1667–1703.

Boehmer, E. and E. K. Kelley (2009). Institutional Investors and the Informational Efficiency of Prices. *Review of Financial Studies 22*(9), 3563–3594.

Budish, E., P. Cramton, and J. Shim (2015). The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response. *Quarterly Journal of Economics 130*(4), 1547–1621.

Chlistalla, M. and M. Lutat (2011). Competition in securities markets: the impact on liquidity. *Financial Markets and Portfolio Management 25*(2), 149–172.

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

Davies, R. J. and S. S. Kim (2009). Using matched samples to test for differences in trade execution costs. *Journal of Financial Markets 12*(2), 173–202.

Degryse, H., F. de Jong, and V. v. Kervel (2015). The Impact of Dark Trading and Visible Fragmentation on Market Quality. *Review of Finance 19*(4), 1587–1622.

Doraszelski, U. and A. Pakes (2007). A framework for applied dynamic analysis in io. *Handbook of industrial organization 3*, 1887–1966.

Foucault, T., R. Kozhan, and W. W. Tham (2017). Toxic Arbitrage. *Review of Financial Studies 30*(4), 1053–1094.

Foucault, T. and A. J. Menkveld (2008). Competition for Order Flow and Smart Order Routing Systems. *Journal of Finance 63*(1), 119–158.

Goettler, R. L., C. A. Parlour, and U. Rajan (2005). Equilibrium in a Dynamic Limit Order Market. *Journal of Finance 60*(5), 2149–2192.

Goettler, R. L., C. A. Parlour, and U. Rajan (2009). Informed traders and limit order markets. *Journal of Financial Economics 93*(1), 67–87.

Gomber, P., S. Sagade, E. Theissen, M. C. Weber, and C. Westheide (2016). Competition Between Equity Markets: A Review Of The Consolidation Versus Fragmentation Debate. *Journal of economic surveys*. forthcoming.

Gresse, C. (2017). Effects of Lit and Dark Market Fragmentation on Liquidity. *Journal of Financial Markets*. forthcoming.

Harris, L. E. (1993). Consolidation, Fragmentation, Segmentation and Regulation. *Financial Markets, Institutions & Instruments 2*(5), 1–28.

Hellström, J., Y. Liu, and T. Sjögren (2013). Time-varying return predictability and equity market fragmentation.

Hengelbrock, J. and E. Theissen (2009). Fourteen at One Blow: The Market Entry of Turquoise. *Available at SSRN 1570646*.

Huang, R. D. and H. R. Stoll (1996). Dealer versus auction markets: A paired comparison of execution costs on NASDAQ and the NYSE. *Journal of Financial Economics 41*(3), 313–357.

Ifrach, B. and G. Y. Weintraub (2016). A framework for dynamic oligopoly in concentrated industries.

Kohler, A. and R. von Wyss (2012). Fragmentation in European Equity Markets and Market Quality – Evidence from the Analysis of Trade-Throughs. Technical Report 1210.

Krusell, P. and A. A. Smith, Jr (1998). Income and wealth heterogeneity in the macroeconomy. *Journal of political Economy 106*(5), 867–896.

Madhavan, A. (1995). Consolidation, fragmentation, and the disclosure of trading information. *Review of Financial Studies 8*(3), 579–603.

Maskin, E. and J. Tirole (2001). Markov perfect equilibrium: I. observable actions. *Journal of Economic Theory 100*(2), 191–219.

Mendelson, H. (1987). Consolidation, Fragmentation, and Market Performance. *Journal of Financial and Quantitative Analysis 22*(02), 189–207.

Nguyen, V., B. F. Van Ness, and R. A. Van Ness (2007). Short- and Long-Term Effects of Multimarket Trading. *Financial Review 42*(3), 349–372.

Nielsson, U. (2009). Stock exchange merger and liquidity: The case of Euronext. *Journal of Financial Markets 12*(2), 229–267.

O'Hara, M. and M. Ye (2011). Is Market Fragmentation Harming Market Quality? *Journal of Financial Economics 100*(3), 459–474.

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

Pagano, M. and J. A. Padilla (2005). Gains from Stock Exchange Integration: The Euronext Evidence. *Working Paper*.

Pakes, A. and P. McGuire (2001). Stochastic algorithms, symmetric markov perfect equilibrium, and the 'curse'of dimensionality. *Econometrica 69*(5), 1261–1281.

Parlour, C. A. and D. J. Seppi (2003). Liquidity-Based Competition for Order Flow. *Review of Financial Studies 16*(2), 301–343.

Rieder, U. (1979). Equilibrium plans for non-zero-sum markov games. *Game theory and related topics*, 91–101.

Riordan, R., A. Storkenmaier, and M. Wagener (2010). Fragmentation, competition and market quality: A post-mifid analysis. *Working Paper*.

van Kervel, V. (2015). Competition for Order Flow with Fast and Slow Traders. *Review of Financial Studies 28*(7), 2094–2127.

Ye, M. (2017). Who Provides Liquidity, and When: An Analysis of Price vs Speed Competition on Liquidity and Welfare.

Table 1. Impact on Trading Behavior

This table shows measures of trader behavior, such as, the percentage of limit orders executed among all limit orders submitted, the probability of being picked-off after submitting a limit order, the number of limit orders submitted per trader, the number of limit order cancellations per trader, the average time between the instant in which a trader arrives and his execution (in time units of our model), the time between the instant in which a trader arrives and the execution of his limit order (in time units of our model) and the probability of submitting a limit sell order at the ask price (i.e., an aggressive limit sell order). All the measures are calculated for a consolidated market and two versions of fragmented market: one with the same distribution of zero private value agents as in a single market and the other one with double arrival rate of zero private value agents. Since the model is symmetric on both sides of the book it is not necessary to also report the probability of submitting a limit buy order at the bid price. The probability of being picked-off is calculated with executed limit orders: we take the number of limit sell (buy) orders that are executed when their execution price is below (above) the fundamental value of the asset, which is divided by all the limit orders executed in the market. All trading behavior measures are determined as mean of 20 million market new entries in equilibrium. Standard errors for all trader behavior measures are small enough since we use a large number of simulated events. The Markov equilibrium is obtained independently for each scenario.

	Single Market	Fragmented Markets	Fragmented Markets Double $\alpha = 0$
Prob. of submitting a limit sell order at the ask price	35.87%	28.45%	32.62%
Execution time of a limit order	8.61	7.15	13.10
Prob. of being picked-off for a limit order	21.80%	20.82%	10.92%
Number of limit order cancelations per trader	1.20	1.01	1.58

Table 2. Impact on Trading Behavior by Agent's Type

This table shows the distribution of limit orders and market orders separated by private value α. The results are reported for a consolidated market and two versions of fragmented market: one with the same distribution of zero private value agents as in a single market and the other one with double arrival rate of zero private value agents. The first three columns show the proportion of limit orders and market orders for a given agents' type α. The next set of columns present how the orders are distributed through the different private values $|\alpha| = 0, 4, 8$. LO denotes limit orders, whereas MO market orders. All trading behavior measures are determined as mean of 20 million market new entries in equilibrium. Standard errors for all trader behavior measures are small enough since we use a large number of simulated events. The Markov equilibrium is obtained independently for each scenario.

$\|\alpha\|$		0	4	8
Single Market	LO	94.6%	68.6%	27.7%
	MO	5.4%	31.4%	72.3%
	Total	100%	100%	100%
Fragmented Market	LO	93.9%	67.8%	29.2%
	MO	6.1%	32.2%	70.8%
	Total	100%	100%	100%
Fragmented Market (Double $\alpha = 0$)	LO	97.7%	34.8%	7.4%
	MO	2.3%	65.2%	92.6%
	Total	100%	100%	100%

Table 3. Impact on Liquidity

Panel A shows the quoted spread and depth for a market containing a single order book and two order books considering arrival rates of zero private value agents being the same and double as in a single market. We present both local and inside liquidity measures. The former refers to measures employing local quotes, i.e., the bid and the ask prices of a local market, whereas the latter refers to liquidity measures using inside quotes, i.e., the highest bid and the lowest ask across the two limit order books. Panel B describes the difference in traded liquidity in consolidated and fragmented markets. We report the level of effective spread, and its decomposition into realized spreads and price impact based on 30 second future quote midpoints. We calculate effective spread as defined in (3) and realized spread as defined in (4). The price impact is then given by the difference. Finally. Panel C presents the difference in price (quote midpoint) efficiency in consolidated and fragmented markets. We report the mean and the standard deviation of the microstructure noise which is defined as the absolute difference between quote midpoint and fundamental value v_t.

Panel A: Quoted Liquidity

	Single Market	Fragmented Market	Fragmented Market Double ($\alpha = 0$)
Quoted Spread: Local	1.565	2.601	2.240
Quoted Spread: Inside	1.565	1.904	1.860
Quoted Depth: Local	1.584	1.082	1.692
Quoted Depth: Inside	1.584	1.445	2.751

Panel B: Traded Liquidity

	Single Market	Fragmented Market	Fragmented Market Double ($\alpha = 0$)
Effective Spread: Local	1.312	1.799	1.862
Effective Spread: Inside	1.312	1.452	1.613
Realized Spread 30: Local	0.865	1.013	1.372
Realized Spread 30: Inside	0.865	1.011	1.371
Price Impact 30: Local	0.441	0.789	0.487
Price Impact 30: Inside	0.441	0.442	0.242

Panel C: Price Efficiency

	Single Market	Fragmented Market	Fragmented Market Double ($\alpha = 0$)
Microstructure Noise Local: Mean $\lvert v_t - p_t \rvert$	0.464	0.670	0.369
Microstructure Noise Inside: Mean $\lvert v_t - p_t \rvert$	0.464	0.570	0.350

Table 4. Decomposition of Welfare by Trader Type

We report the average welfare defined in 5, waiting cost and money transfer defined in 6, all of them per trader differentiated by private value. The first row reports the results for a single market, i.e., a market organized as a single limit order market, whereas the second row for a fragmented market, i.e., a market organized as two limit order markets. Finally, the third row reports the results when the arrival rate of zero private value agents is double. The three measures are reported in ticks and calculated as the mean over 20 million new arrivals in equilibrium. Standard errors for all measures are small enough due to the large number of simulated events and are hence omitted. The Markov equilibrium is obtained independently for each scenario.

	Total	Average welfare per trader				Waiting cost per trader				Money transfer per trader			
	Welfare	Private Value —α—			Total	Private Value —α—			Total	Private Value —α—			Total
	per Period	0	4	8		0	4	8		0	4	8	
Single Market	3.742	0.543	3.510	7.265	3.745	0.000	-0.350	-0.162	-0.189	0.543	-0.140	-0.572	-0.065
Fragmented Market	3.740	0.626	3.479	7.202	3.740	0.000	-0.355	-0.172	-0.193	0.626	-0.166	-0.626	-0.066
Fragmented Market (Double $\alpha = 0$)	3.757	0.485	3.312	7.137	2.890	0.000	-0.127	-0.029	-0.049	0.485	-0.561	-0.835	-0.142

Table 5. Stock Characteristics

This table reports the characteristics of the treatment stocks and the corresponding control stocks generated based on Huang and Stoll (1996). Market Capitalization is the product of shares outstanding and Stock Price as on 31 December 2008, Trading Volume and Number of Trades is the average daily trading volume and number of trades for each stock between 1 December 2008 and 13 January 2009. We also report the market share of the two Euronext order books. Large (Small) order book is the order book with higher (lower) trading volume.

Panel A: Treatment Stocks

	Market Cap	Stock	Trading Volume			Number of Trades		
	€ million	Price (€)	€ '000	% Large	% Small	Count	% Large	% Small
DEXI	3,355	2.9	8,514	54.3%	45.7%	2,843.1	52.6%	47.4%
FOR	2,187	0.9	25,239	71.5%	28.5%	6,613.4	71.7%	28.3%
ISPA	24,985	17.2	180,346	55.6%	44.4%	18,231.8	52.7%	47.3%
UNBP	8,598	104.9	37,687	87.7%	12.3%	4,268.1	86.4%	13.6%
GLPG	80	3.8	183	77.1%	22.9%	67.9	27.9%	72.1%
ONCOB	87	6.6	23	90.7%	9.3%	3.5	73.5%	26.5%
RCUS	193	6.2	119	98.0%	2.0%	68.3	98.5%	1.5%
VRKP	105	20	43	94.4%	5.6%	19.5	94.3%	5.7%
THEB	59	3.5	11	68.4%	31.6%	7.5	77.1%	22.9%
MEAN	4,405	18.4	28,018	77.5%	22.5%	3,569.2	70.5%	29.5%

Panel B: Control Stocks (Huang and Stoll, 1996)

	Market Cap	Stock	Trading Volume			Number of Trades		
	€ million	Price (€)	€ '000	% MoR	% Alternate	Count	% MoR	% Alternate
STM	3,355	4.6	15,157	100.0%	0.0%	2,635.1	100.0%	0.0%
CNAT	3,635	1.3	5,333	100.0%	0.0%	2,265.3	100.0%	0.0%
ABI	25,439	15.9	75,884	100.0%	0.0%	7,245.6	100.0%	0.0%
HRMS	10,652	101	12,193	100.0%	0.0%	1,546.2	100.0%	0.0%
OMT	84	4	15	100.0%	0.0%	9.2	100.0%	0.0%
TAM	81	6.8	29	100.0%	0.0%	23.5	100.0%	0.0%
AMG	184	6.9	2,989	100.0%	0.0%	903.5	100.0%	0.0%
SMTPC	117	20	20	100.0%	0.0%	9.4	100.0%	0.0%
DEVG	63	3.5	156	100.0%	0.0%	111.0	100.0%	0.0%
MEAN	4,846	18.2	12,420	100.0%	0.0%	1,638.8	100.0%	0.0%

Table 6. Empirical Findings: Impact on Quoted Liquidity

This table presents the results on the impact of the introduction of a single order book on quoted liquidity. We calculate quoted spread and depth in single and fragmented markets. We present both local and inside liquidity measures. The former refers to measures employing local quotes, i.e., the bid and the ask prices of a local market, whereas the latter refers to liquidity measures using inside quotes, i.e., the highest bid and the lowest ask across the two limit order books. We estimate a difference-in-difference regression for quoted spread and quoted depth, in level and logarithm, and report the coefficient of the variable interacting the event dummy (which equals one for all days on or after 26 January 2009 and zero otherwise) with the treatment dummy (which equals one for all treatment stocks and zero for all control stocks). We employ stock and day fixed effects and double cluster standard errors by stock and day. In order to calculate local liquidity we choose one of the two order books in the simulated data and the venue with the larger trading volume in the pre-event period in the empirical analysis. Inside liquidity, in fragmented markets, is measured based on the best quotes (highest bid and lowest ask) across the two order books, and in consolidated markets, it is equal to the local liquidity. *, **, *** denote significance at 10%, 5%, and 1%, respectively.

	Treatment	Control	Effect Size	
	Post-Pre	Post-Pre	Levels	Logs
Quoted Spread: Local	−0.542	0.266	−0.808**	−0.322**
Quoted Spread: Inside	−0.056	0.266	−0.323	0.090
Quoted Depth: Local	−878	−5,435	4,589	0.017
Quoted Depth: Inside	−485	−5,435	4,979	0.005

Table 7. Empirical Findings: Impact on Traded Liquidity

This table describes the difference in traded liquidity in consolidated and fragmented markets. We report the impact of the introduction of single order book on traded liquidity. We estimate a difference-in-difference regression for effective spreads, realized spreads, and price impact, in level and logarithm, and report the coefficient of the variable interacting the event dummy (which equals one for all days on or after 26 January 2009 and zero otherwise) with the treatment dummy (which equals one for all treatment stocks and zero for all control stocks). We employ stock and day fixed effects and double cluster standard errors by stock and day. In both panels, we compute local and inside traded liquidity. We measure local liquidity based on quotes on the order books where a a transaction is executed. Inside liquidity, in fragmented markets, is measured based on the inside quotes across the two order books, and in consolidated markets, it is equal to the local liquidity. *, **, *** denote significance at 10%, 5%, and 1%, respectively.

	Treatment	Control	Effect Size	
	Post-Pre	Post-Pre	Levels	Logs
Effective Spread: Local	−1.228	0.151	−1.394*	−0.375**
Effective Spread: Inside	−0.764	0.151	−0.926*	−0.145
Realized Spread 10: Local	−1.115	0.079	−1.211*	−0.517***
Realized Spread 30: Local	−1.110	0.064	−1.190*	−0.643***
Realized Spread 60: Local	−1.109	0.031	−1.153	−0.725***
Realized Spread 10: Inside	−0.803	0.079	−0.894	−0.376***
Realized Spread 30: Inside	−0.817	0.064	−0.893	−0.547***
Realized Spread 60: Inside	−0.796	0.031	−0.836	−0.615***
Price Impact 10: Local	−0.112	0.072	−0.183	−0.164
Price Impact 30: Local	−0.118	0.086	−0.204*	−0.133
Price Impact 60: Local	−0.120	0.121	−0.241**	−0.214
Price Impact 10: Inside	0.040	0.072	−0.032	0.123
Price Impact 30: Inside	0.053	0.086	−0.033	0.161
Price Impact 60: Inside	0.032	0.121	−0.090	0.024

Table 8. Empirical Findings: Impact on Price Efficiency

This table describes the difference in price (quote midpoint) efficiency in consolidated and fragmented markets. We report the impact of the introduction of single order book on price efficiency. We estimate a difference-in-difference regression for absolute values of return autocorrelation measured at 30-second and 5-minute intervals and the variance ratio based on 30-second and 5-minute returns, in level and logarithm, and report the coefficient of the variable interacting the event dummy (which equals one for all days on or after 26 January 2009 and zero otherwise) with the treatment dummy (which equals one for all treatment stocks and zero for all control stocks). We employ stock and day fixed effects and double cluster standard errors by stock and day. In order to calculate local price efficiency measures we choose one of the two order books in the simulated data and the venue with the larger trading volume in the pre-event period in the empirical analysis. Inside price efficiency, in fragmented markets, is measured based on the inside quotes across the two order books, and in consolidated markets, it is equal to the local price efficiency. *, **, *** denote significance at 10%, 5%, and 1%, respectively.

	Treatment	Control	Effect Size	
	Post-Pre	Post-Pre	Levels	Logs
Autocorrelation 30: Inside	−0.007	0.007	−0.015	0.201
Autocorrelation 30: Local	−0.014	0.007	−0.021*	−0.233
Autocorrelation 300: Inside	−0.013	−0.002	−0.011	−1.444*
Autocorrelation 300: Local	−0.022	−0.002	−0.020	−1.545*
Variance Ratio 30/300: Inside	−0.045	−0.017	−0.028	−0.224
Variance Ratio 30/300: Local	−0.052	−0.017	−0.035	−0.206

Table 9. Impact on Trading Volume and Cross Market Arbitrage Analysis
Panel A reports the impact of the introduction of single order book on total Euronext trading volume. We estimate a difference-in-difference regression for the trading volume, in level and logarithm, and report the coefficient of the variable interacting the event dummy (which equals one for all days on or after 26 January 2009 and zero otherwise) with the treatment dummy (which equals one for all treatment stocks and zero for all control stocks). We employ stock and day fixed effects and double cluster standard errors by stock and day. Panel B summarizes the arbitrage opportunities arising on the two Euronext order books during the pre-event period i.e., between 1 December 2008 and 13 January 2009, and their resolution. Section 4.3.4 describes how we identify each abitrage opportunity. Unique Instances are the average daily frequency of arbitrage opportunities on the two order books betwen 08:01 and 16:29, Negative Spread Time is the total amount of time during a trading session when the markets are crossed, Magnitude of Negative Spread is the frequency with which the negative bid-ask spread is equal to one tick, two ticks, three ticks, four ticks, and five or more ticks, and Trading Volume is the average daily volume which can be attributed towards resolution of the arbitrage opportunities. *, **, *** denote significance at 10%, 5%, and 1%, respectively.

Panel A: Arbitrage Analysis

Stock	Unique Instances	Negative Spread Time	Trading Volume
DEXI	138.1	12.8%	1,145,164
FOR	183.8	11.5%	2,361,274
ISPA	622.0	6.0%	14,185,211
UNBP	166.1	3.0%	1,968,394
GLPG	2.4	4.3%	7,558
ONCOB	0.3	1.6%	990
RCUS	0.8	9.5%	9,020
VRKP	1.2	6.7%	4,351
THEB	0.6	2.6%	1,553
MEAN	123.9	6.4%	2,187,057

Panel B: Trading Volume

	Treatment	Control	Effect Size	
	Post-Pre	Post-Pre	Levels	Logs
Total Volume	2,628	-948	3,638*	0.080

Toward a Fully Continuous Exchange

Albert S. Kyle* Jeongmin Lee†

February 27, 2017

Abstract

We propose continuous scaled limit orders to implement Fischer Black's vision of financial markets. By making trading continuous in price, quantity, and time, continuous scaled limit orders eliminate rents high frequency traders earn exploiting artifacts of the current market design. By avoiding time priority, this new order type protects slow traders from being picked off by high frequency traders and makes high frequency traders compete among themselves. All traders, regardless of their technological capacity, can optimally spread trades out over time to minimize adverse price impact. Organized exchanges should move not toward more discreteness but toward a full continuity.

Keywords: Market microstructure, smooth trading, auction design, market design.

*Robert H. Smith School of Business, University of Maryland, College Park, MD 20742, USA; akyle@rhsmith.umd.edu. Kyle has worked as a consultant for various companies, exchanges, and government agencies. He is a non-executive director of a U.S.-based asset management company.

†Olin Business School, Washington University, St. Louis, MO 63130, USA; jlee89@wustl.edu.

About half a century ago, Fischer Black (1971*a*,*b*) made bold predictions about how stock market trading would change if the design of the stock market moved from the human-dominated specialist system to a system in which trading and market-making used computers. He predicted that liquidity would not be supplied cheaply, especially over short periods of time. Realizing that trading large quantities over a short horizon was expensive, customers would spread large trades out over time to reduce temporary price impact costs. He believed an efficient market design could reduce bid-ask spreads on small trades to a vanishingly small level while providing practical ways for large traders to reduce impact by trading gradually over time.

The purpose of this paper is to show how to implement Fischer Black's vision of an efficient market design using a new order type that we call "continuous scaled limit orders." Continuous scaled limit orders eliminate the rents that high frequency traders earn at the expense of other traders and thus also eliminate resulting inefficiencies in today's markets. To illustrate this point, let us first describe how the current markets work.

Since the late 1990s, human beings have been replaced by computerized limit order books. The trading of equities in the U.S. and Europe has in recent decades become dominated by continuous limit order books which handle millions of buy and sell orders each day. A continuous limit order book is, however, hardly continuous. A standard limit order is a message conveying an offer to buy or sell a discrete quantity at a discrete price, where the quantity is an integer multiple of minimum lot size and the price is an integer multiple of a minimum tick size. In most U.S. stocks, the minimum lot size is one share or one hundred shares and the minimum tick size is \$0.01 or one cent per share. A limit order book then processes discrete orders sequentially in the order of their arrivals. Because sending, receiving, and processing messages take time, no trader can trade in continuous time. Thus, a continuous limit order book has elements of discreteness in price, quantity, and time.

In today's markets, high frequency traders who expend real resources to acquire technological advantages earn rents exploiting artifacts of the current market design related to the discreteness of prices, quantities, and time. When several traders want to purchase shares at the same price at the same time, exchanges often allocate trades

based on time priority; the first trader in line to buy or sell at a given price is the first to receive quantities traded at that price. High frequency traders use their speed to take advantage of time priority by placing orders quickly to be the first in the queue. High frequency traders also use their speed to "pick off" slow traders orders by hitting or lifting stale bids or offers before the slow traders can cancel them. Furthermore, today's limit order book requires an allocation rule because discrete prices and quantities prevent the market clearing price from being uniquely defined. The allocation rule provides additional rents high frequency traders can earn from gaming it.

A continuous scaled limit order is a message conveying an offer to buy or sell gradually at a specific trading rate over a specific range of the prices. With such orders, traders' inventories are piecewise differentiable functions of time, with the rates of buying or selling changing when the price changes. Traders can buy at a faster rate when prices fall and sell at a faster rate when prices rise. Continuous scaled limit orders make price, quantity, and time continuous.

With continuous scaled limit orders, all orders are treated symmetrically and executed simultaneously. Because slow traders spread their orders over time, high frequency traders can pick off only a small quantity before slow traders cancel their orders. With the market clearing price uniquely defined, an allocation rule is no longer necessary. This automatically eliminates the rents high frequency traders would have earned from gaming it. More importantly, there is no time priority. The market is no longer the fastest-takes-all. High frequency traders with varying speeds and bandwidths compete with one another. This increased competition among high frequency traders has a broader implication for economic efficiency. Today's market structure encourages arms race among fast traders to become the fastest as emphasized by Harris (2013); Li (2014); Biais, Foucault and Moinas (2015); Budish, Cramton and Shim (2015). Continuous scaled limit orders deter over-competition in technology by increasing competition in trading, which further benefits slow traders who experience price improvements.

Fischer Black was remarkably prescient. Large institutional traders around the world nowadays spread their trading out over time exactly like he said they would. Widespread algorithmic trades are often executed by breaking large intended trades into many small pieces and trading the many small pieces over time. For example, some algorithms try

to achieve the volume-weighted average price ("VWAP") of trades during a day by trading gradually along with the rest of the market. Our proposal for continuous scaled limit orders allows traders to do this without incurring large bandwidth costs for placing, modifying, and canceling thousands of orders throughout the day so that all traders regardless of their technological capacity can implement their trading strategies in an efficient manner.

Theoretical models of dynamic trading are also consistent with traders optimally choosing to trade gradually using continuous scaled limit orders. In the model of Kyle, Obizhaeva and Wang (2017) traders face temporary and permanent price impacts. Because traders have private information, the price moves against the trader, meaning that the price goes up when the trader wants to buy, and the price goes down when the trader wants to sell. Moreover, the extent to which the price moves against the trader increases in the speed with which the trader buys or sells because more urgency signals stronger private information. Therefore, traders smooth their trading over time with optimal trading strategies that almost perfectly map into continuous scaled limit orders.

Such gradual trading directly opposes to the model of Grossman and Miller (1988), in which continuously present market makers must satisfy urgent trading needs of buyers and sellers. In their model, traders demand urgency because they do not take into account their own price impact costs; instead, they trade as perfect competitors. In a one-period model, Kyle and Lee (2017) show that fully strategic traders restrict quantities they trade whenever they face price impacts and may even completely refrain from trading, foregoing gains from trade. This suggests that strategic traders do not demand urgency and choose to trade gradually over time.

We believe that trading with continuous scaled limit orders dominates the current market design. While we cannot prove continuous scaled limit orders are an optimal mechanism, this new order type eliminates rents high frequency traders earn from exploiting the discreteness in today's markets. By allowing all traders to trade gradually without being picked off, continuous scaled limit orders make rapid trading more expensive compared to slower trading. As a result, traders are deterred from acquiring ultra short-term information with little to no social value and are encouraged to pro-

duce more long-term information. Future exchanges should move not toward more discreteness but toward full continuity.

The plan of this paper is as follows. Section 1 describes the difference between continuous scaled limit orders and standard limit orders. Section 2 explains how continuous scaled limit orders benefit long-term traders by eliminating socially counterproductive games high-frequency traders play using their speed to pick off resting limit orders and exploit time and price priority when the tick size is economically meaningful. It also shows how our proposal addresses the efficiency costs of a high-frequency trading arms race better than the proposal of Budish, Cramton and Shim (2015). Section 3 discusses remaining issues such as transparency and trust, execution of market orders, flash crashes, speed bumps, privately arranged trades, minimum resting times, market fragmentation, dark pools, and clock synchronization. Section 4 show that our proposal is deeply grounded in relevant economic theory. Continuous scaled limit orders allow traders to implement with greater message efficiency the gradual trading strategies that they are implementing today.

1 Continuous Scaled Limit Orders

Today's exchanges operate as "continuous limit order books" which process discrete limit orders arriving sequentially in continuous time. Each limit order is a *message* conveying a contractually binding offer to buy or sell a specific quantity at a specific price. The message also includes information about time stamps, the identities of traders, and routing. Traders send messages to exchanges to place, cancel, or modify limit orders. Exchanges log messages and send traders additional messages to confirm receipt of the messages and to update prices and quantities for shares bought or sold. Encryption and decryption of messages is computationally costly. Sending, receiving, and processing messages takes time and consumes real resources such as telecommunications bandwidth and computer processing power.

A continuous limit order book has elements of discreteness with respect to price, quantity, and time. Standard limit orders are *discrete* in both price and quantity in the sense that the price is an integer multiple of a minimum tick size and the quantity is

an integer multiple of minimum lot size. Whether orders are processed one-at-a-time or in batches, continuous limit order books are discrete in time in the sense that finite quantities are exchanged at specific points in time based on the arrival of orders rather than exchanged gradually over time. For example, a standard limit order to buy 100 shares at a price of $40.00 per share will be executed immediately when an order to sell 100 shares at a price of $40.00 arrives; it is not executed at a rate of one share per second over a time period of 100 seconds.

Although messages are sent and received in continuous time, no trader can effectively trade continuously because there are time lags associated with sending, receiving, and processing orders. The degree to which a trader can participate continuously depends on the speed of the trader's technology and is ultimately limited by the speed of light. From a trader's perspective, the market operates more continuously if the trader can send, receive, and process messages at a faster speed than others. A trader who can easily and cheaply send 100 limit orders to buy or sell one share of stock each over a time period of 100 seconds (or milliseconds) can effectively participate more continuously than a trader who cannot do so because it is technologically impractical or too costly. The discreteness of today's continuous limit order books in price, quantity, and time gives faster traders advantages with respect to slower traders.

In this section, we introduce dynamic trading with *continuous scaled limit orders* to achieve continuity in price, quantity, and time. Continuous scaled limit orders are different from standard limit orders in two respects. First, prices and quantities vary continuously. Second, trades are executed continuously over time. Continuous scaled limit orders allow traders to participate continuously while consuming fewer real resources.

We begin by describing how current exchanges work using standard limit orders.

Sequential Auctions of Standard Limit Orders. Currently, exchanges process standard limit orders sequentially in the order in which they arrive. A limit order is a message with three parameters: a buy-sell indicator, a quantity Q, and a price P, where Q

and P are multiples of a minimum lot size and a minimum tick size respectively.[1] In the U.S. market, the stated minimum tick size for most actively traded stocks is currently one cent per share. It was reduced from 1/8 of a dollar (12.5 cents per share) to 1/16 of a dollar (6.25 cents per share) in the late 1990s and reduced again to its current level of one cent per share in 2001. There is also a distinction between "round lots" of 100 shares for most stocks and "odd lots" of fewer than 100 shares. Historically, odd lots have been subject to different order execution and price reporting rules.

A standard buy limit order conveys the message "Buy up to Q shares at a price of P or better." Let X denote the number of shares purchased and let $p(t)$ denote the market clearing price. Then X always satisfies

$$X = \begin{cases} Q & \text{if } p(t) < P, \\ \alpha Q & \text{if } p(t) = P, \\ 0 & \text{if } p(t) > P. \end{cases} \qquad \text{where } \alpha \in [0,1], \tag{1}$$

If the market price $p(t)$ is above the limit price P, nothing is bought; if it is below the market price $p(t)$, the order is fully executed ($X = Q$). If the market clearing price $p(t)$ exactly equals the limit price P, X depends on the rule of assigning market clearing quantities α. Depending on α, the order receives a full execution ($\alpha = 1$), a partial execution ($0 < \alpha < 1$), no executed quantity ($\alpha = 0$).[2]

An allocation rule to determine α is necessary because of discreteness in the limit price and quantity. The market demand schedule calculated from aggregating all buy orders and the market supply schedule calculated by aggregating all sell orders are discontinuous step functions. Although the market demand schedule is weakly downward

[1] An order may also contain additional time parameter T_1 defining the time when the order begins execution. We assume for simplicity that orders are for immediate execution and are good until canceled.

[2] The notation in equation (1) is meant to convey intuition; it is not meant to be mathematically precise. With more formal notation, the quantities Q, P, α, and X would have superscripts indicating the identity of the specific message, which could be mapped to a specific trader. The price $p(t)$ is the same for all traders and changes over time. If a limit order rests in the market for some period of time, then α and X would become functions of time $\alpha(t)$ and $X(t)$. The quantity $X(t)$ would be a monotonically increasing step function of time indicating the cumulative number of shares bought or sold as of time t. The fraction $\alpha(t)$ could be interpreted as the fraction of the remaining quantity $Q - X(t)$ executed at time t.

sloping and the market supply schedule is weakly upward sloping, there may not be a unique point of intersection. Instead, there is typically a pair of best bid and offer prices with excess demand at the best bid and excess supply at the best offer. The exchange typically chooses as the market clearing price the price at which trading volume is maximized. Since there is typically excess supply or demand at this price, some rule is needed to allocate prices and quantities.

Orders are matched according to rules specifying price and time priority. Price priority matches incoming executable limit orders against the lowest sell prices and highest buy prices in the limit order book. When there is more than sufficient quantity at a given price to satisfy an incoming limit order, time priority executes the oldest limit order at the best price first. Traders have strategic incentives to place orders in a manner which exploits price and time priority at the expense of other traders. Obviously, fast traders have an incentive to place orders quickly, to get ahead of other traders in the time priority queue at a given price.

Conceptually, one way to get around the need for an allocation rule is to allow traders to submit orders which are not discontinuous step functions but rather arbitrary weakly monotonic functions which specify quantity demanded or supplied as a function of price. If traders choose continuous upward-sloping supply schedules and continuous downward sloping demand schedules, then there is a unique market clearing price at which the market exactly clears and all traders' quantities demanded and supplied are fulfilled ($\alpha = 1$). This is typically what happens in theoretical models of market equilibrium. In rational expectations model with exponential utility and normally distributed random variables—or models with quadratic storage costs—the demand and supply schedules are linear.

This approach makes limit orders continuous in quantities and prices but not continuous in time by eliminating minimum tick size and minimum lot size. It does not make quantities continuous functions of time. Our approach makes trading continuous in price, quantity, and time. We explain first how to make trading continuous in time, then explain later how to make trading continuous in price and quantity.

Auctions of Continuous Standard Limit Orders. Quantities traded can be made continuous functions of time by adding to each limit order an urgency parameter specifying the maximum rate at which to buy or sell. We define a "continuous standard limit order"as an order which conveys the message, "Buy up to a cumulative total of $Q_{\max}$ shares at a price of $P_{\max}$ or better at maximum rate $U_{\max}$ shares per hour." The quantities $Q_{\max}$ and $U_{\max}$ are multiples of a minimum lot size and $P_{\max}$ is a multiple of a minimum tick size. The speed parameter $U_{\max}$ defines the maximum of the derivative of the trader's inventory as a continuous function of time.[3] The trading speed $U\big(p(t)\big)$ is a function of the the market clearing price $p(t)$ at time t; it is given by

$$U\big(p(t)\big) := \begin{cases} U_{\max} & \text{if } p(t) < P_{\max}, \\ \alpha \cdot U_{\max} & \text{if } p(t) = P_{\max}, \\ 0 & \text{if } p(t) > P_{\max}. \end{cases} \qquad \text{where } \alpha \in [0,1] \tag{2}$$

For an order placed at time t_0 and canceled or filled at time $T_{\max}$, the cumulative quantity executed by time t is given by the integral

$$Q(t) := \int_{t_0}^{t_0+t} U\big(p(\tau)\big)\,\mathrm{d}\tau, \qquad \text{for} \qquad t \in [0, T_{\max}]\,. \tag{3}$$

If the order is canceled at time $T_{\max}$ without being filled, then $Q(T_{\max}) < Q_{\max}$; if the order is filled at time $T_{\max}$, then $Q(T_{\max}) = Q_{\max}$.

When the price is strictly below $P_{\max}$, the trader buys at rate U. When the price is strictly above $P_{\max}$, the inventory does not change, implying $\mathrm{d}Q(t)/\mathrm{d}t = 0$. If the price remains low enough so that that order is executed at maximum rate U, the order will be fully executed exactly after $T_{\max} = Q/U$. If the price fluctuates above and below P, the full execution will take longer than Q/U. If the price stays above P, the order will not be executed. Since $U(p(t))$ changes only when $p(t)$ changes and $p(t)$ changes only when discrete events like order arrivals, executions, and cancelations occur, the cumulative quantity executed $Q(t)$ is a piecewise continuously differentiable function of time. A standard limit order corresponds to $U \to \infty$, which allows the cumulative

[3] We conjecture that future exchanges could develop additional order types which allow $U_{\max}$ to be a function of other market characteristics such as trading volume, price volatility, or "market liquidity".

quantity executed to be a discontinuous step function.

When the market clearing price is exactly equal to the limit price P during order execution, the trader's inventory changes at a rate such that $0 \leq \mathrm{d}Q(t)/\mathrm{d}t \leq U_{\max}$. The exact rate $\alpha U_{\max}$ depends on the rule for allocating market-clearing quantities.[4]

While $Q_{\max}$, and $U_{\max}$ are multiples of minimum lot size, the cumulative quantity traded $Q(t)$ is an arbitrary real number. To settle market clearing quantities, we propose the following approach. Let X denote the net purchases or sales a trader makes, calculated at the end of the day based on full or partial execution of all orders the trader has submitted. The quantity X can be expressed as the sum of an integer portion fraction part ϵ by writing $X = \mathrm{int}(X) + \epsilon$. To clear the fractional part of X, we propose cash-settling the fraction ϵ by buying $1-\epsilon$ shares or selling ϵ shares in a manner such that the expected fractional share traded is approximately zero. This insures that traders have little incentive to game the end-of-day settlement of these fractional shares.

Continuous orders allow traders to slice their orders into small pieces and gradually trade toward their target inventories. As discussed below, economic theory implies that such order shredding is an optimal trading strategy. Nowadays large institutional investors by in the manner implied by theory. They shred large orders into small pieces and trade numerous small quantities more or less continuously throughout the day. Implementing such strategies in today's markets requires sending numerous messages, which is more costly for traders with low technological capacity. Continuous limit orders allow all traders to trade smoothly without being equipped with large bandwidth and processing power. To the extent that price impact depends not only the quantity traded but also the speed with which the same quantity is traded, traders can optimally choose their trading speed by trading off the price impact against the impatience of their trading needs.

Continuous orders do not eliminate the need for an allocation rule which determines the fractional rate of order execution α when there is excess flow demand or supply at the market clearing price. To deal with the possibility that faster traders may

[4]The notation in equations (2) and (3) is also meant to be intuitive, not mathematically rigorous. More formally, the quantities $U_{\max}$, $P_{\max}$, $U(p(t))$, and α should have subscripts indicating the order to which they apply. The quantities $U(p(t))$ and α are functions of time t.

be able to profit at the expense of slower traders by gaming the allocation rule with continuous limit orders, we propose continuous scaled limit orders, which we discuss next.

Market Design with Continuous Scaled Limit Orders. We define a "continuous scaled limit order" as a generalization of a continuous limit order. Instead of one price $P_{\max}$, a continuous scaled limit order conveys the message, "Buy up to $Q_{\max}$ total shares at prices between P_L and P_H at maximum rate $U_{\max}$," where $Q_{\max}$ and $U_{\max}$ are multiples of a minimum lot size and P_L and P_H are multiples of a minimum tick size satisfying $P_L < P_H$ .If $P_L = P_H$, the order corresponds to a continuous (unscaled) limit order. Then the trading speed $U\left(p\left(t\right)\right)$ is a function of the the market clearing price $p\left(t\right)$ given by

$$U\left(p\left(t\right)\right) := \begin{cases} U_{\max} & \text{if } p\left(t\right) < P_L, \\ \left(\frac{P_H - p(t)}{P_H - P_L}\right) U_{\max} & \text{if } P_L \leq p\left(t\right) \leq P_H, \\ 0 & \text{if } p\left(t\right) > P_H. \end{cases} \tag{4}$$

A continuous scaled limit buy order defines a piecewise linear demand schedule according to which the derivative of a trader's inventory $U\left(p\left(t\right)\right)$ is equal to $U_{\max}$ when the price is less than P_L, is equal to zero when the price is greater than P_H, and decreases linearly when the price is between P_L and P_H. The trader's inventory $Q\left(t\right)$ is defined by equation (3).

A set of continuous scaled limit buy orders defines an aggregate flow demand schedule, denoted $D\left(p\right)$, as the sum of the trading speed $U(p)$ of all buy orders. An aggregate demand schedule is the graph of a continuous, weakly monotonically decreasing, piecewise linear function of price p, with possible kinks at integer multiples of the minimum tick size. An aggregate supply schedule, denoted by $S\left(p\right)$, is defined analogously to a demand schedule and is the graph of a continuous, weakly monotonically increasing, piecewise linear function.

Suppose the aggregate demand and supply schedules to intersect at a point where either of the two is not flat. Then the excess demand schedule $D\left(p\right) - S\left(p\right)$ is strictly decreasing in the neighborhood of the intersection, and, thus, there exist P_0 and P_1,

where P_1 is one tick size larger than P_0, such that

$$D(P_0) - S(P_0) \geq 0 \qquad \text{and} \qquad D(P_1) - S(P_1) < 0. \tag{5}$$

Define the relative order imbalance $\omega \in [0,1]$ by

$$\omega := \frac{D(P_0) - S(P_0)}{D(P_0) - S(P_0) - D(P_1) + S(P_1)}. \tag{6}$$

Then the market clearing price $p(t)$ is uniquely defined by

$$p(t) = P_0 + \omega(P_1 - P_0). \tag{7}$$

Intuitively, the price is a weighted average of the two prices P_0 and P_1, with weights $1-\omega$ and ω proportional to the excess demand and supply at these prices.

If the demand and supply schedules intersect at overlapping flat sections, then we adopt the convention that the market clearing price is the midpoint of the overlapping interval. We do not expect this to be the case. Suppose the demand and supply schedules intersect over a horizontal interval. Then each buyer could increase a minuscule amount of demand at the lower price of the interval, forcing the price down. Similarly, each seller could increase a minuscule quantity of supply at the higher price of the interval, forcing the price up. Since a flat demand schedule around the intersection is not an optimal response to a flat supply schedule and vice versa, we expect the demand and supply schedules almost always to intersect at a single point which uniquely defines the market clearing price $p(t)$ as above.

Requiring the price limits P_H and P_L to be multiples of minimum tick size makes both the aggregate demand and supply schedules to be piecewise linear functions of price p with all kinks occurring at integer multiples of the minimum tick size. This feature simplifies algorithmically the calculation of the market clearing price $p(t)$. The aggregate demand schedule and the aggregate supply schedule can both be described as vectors of fixed length, with each vector entry corresponding to the demand or supply at a particular price. The vectors are monotonic in in quantities. This makes it easy to calculate the two prices P_0 and P_1 at which the difference between quantity sup-

plied and quantity demanded changes sign. The price can then be calculated as a real number from , which is an arbitrary real number from equation (7). Given the speed of modern computers, these calculations are nowadays trivial. Since the calculations are performed at the exchange, they do not involve sending and receiving extra messages.

Furthermore, since the price $p(t)$ and thus the trading rates $U\big(p(t)\big)$ are uniquely defined, an allocation rule α is no longer necessary; it does not appear in equation (4). Since the allocation rule is not necessary, traders can accurately infer the quantities they trade from a public feed of prices, or equivalently from P_0, P_1, and ω. Exchanges need not send constant updates of prices and quantities for each fractional share bought on each order. Sending confirmation messages at infrequent time intervals, like one second or one minute, would be sufficient. This conserves bandwidth and computation costs because sending and receiving messages is computationally costly.

With continuous scaled limit orders, a trader is likely to place, modify, and cancel orders much less frequently than with standard limit orders. A continuous scaled limit order automatically implements a strategy to buy patiently over time, as Fischer Black (1971*a*) suggest traders would want to do. The patient strategies which traders use today can be implemented with small number of continuous scaled limit orders rather than a gigantic number of standard limit orders. As we discuss next, such orders not only conserve the real resources needed to operate an organized exchange but also level the playing field between fast and slow traders.

2 Practical Implications for High Frequency Trading

High frequency traders expend real resources to acquire technological advantages over other traders related to lower latency, larger bandwidth, and more processing power. As we discuss in this section, this technological advantage allows fast traders to make profits exploiting artifacts of the current market design related to discreteness of prices, quantities, and time. Although such rents may have great private value, such rents have little to no social value; they are earned at the expense of other traders with less advanced technology.

Slow traders often seek to profit by uncovering long-term information about the

value of assets. This long-term information tends to create a positive externality by giving the market signals about value which can steer resource allocation decisions related to investment and corporate strategy. To the extent that the fast traders increase the trading costs of slow traders, the fast traders discourage production of socially valuable long-term information. Continuous scaled limit orders create long-term social value by reducing the incentives high frequency trader have to engage in a costly technology arms race..

High frequency traders may also perform socially useful services by using their speed to arbitrage prices better and to hold inventories temporarily for short periods of time. Continuous scaled limit orders improve the efficiency with which these services are formed by making it cost effective for slower traders to participate in providing trading services which would otherwise be too technologically expensive for slow traders to provide.

This section first discusses how continuous scaled limit orders eliminate artificial discreteness in price, quantity, and time in the current markets. This not only diminishes the rents earned by fast traders but also changes the nature of competition among fast and faster traders to make the market more competitive. We then compare continuous scaled limit orders to frequent batch auctions proposed by Budish, Cramton and Shim (2015) and random delays proposed by Harris (2013).

2.1 How Fast Traders Earn Rents in Today's Markets.

Fast traders earn rents in today's market by using their speed to process information and submit messages faster. This allows them to profit by arriving early, canceling early, and taking advantage of the allocation rule when there is time and price priority.

To illustrate these ideas, consider a hypothetical stock with a price of about $40.00 per share and volume of about one million shares per day. Suppose the return volatility of the stock is 2.00 percent per day. Thus, a one standard deviation event represents a price change of 2.00 percent of $40.00 or 80 cents per share. This price, share volume, and volatility are typical for a stock just below the median of the S&P 500.

Suppose a portfolio manager desires to buy 10 000 shares of this stock over the course

of one day. Such an order represents one percent of one day's trading volume, a typical amount that an institutional investor might want to trade in one day. Buying 10 000 shares will likely incur significant, unavoidable price impact costs related to adverse selection. Now suppose that the trader submits a 10 000 share limit order and leaves it resting in the market. Such a strategy exposes the order to being exploited by faster traders in several ways. We examine these next.

Arriving Early and Canceling Early. Fast traders can access, process, and act on short-term information than unfolds over short periods of time like fractions of a second. They can learn the price in other markets before others and attempt to take a cross-market arbitrage. Alternatively, fast traders may be able to use public information within a market, such as quantities and prices of active bids and offers, to infer others' trading motives and anticipate their orders to the advantage of the fast traders themselves.

For example, suppose that the institutional investor entered the 10 000 share order in reaction to some fast-unfolding piece of information. Suppose a fast trader entered an order to purchase 4000 shares at the same price based on reacting to the same information. If the fast trader's arrives one microsecond earlier than the slower trader's order, then the fast trader gains time priority in the limit order book. If there are incoming orders to sell 4000 shares at $40.00, the fast traders takes the other side of all 4000 shares because of time priority. If the price rises substantially immediately after these 4000 share finish executing, the fast trader gains all of the benefit from the purchase of 4000 shares and the slower trader gains nothing. The slow trader loses the entire trading opportunity by being one microsecond slower than the fast trader.

If there is an infinitesimal tick size, then the fast trader does not need to be fast to step in front of the 10 000 share order. He can place a limit order to buy at $40.000 000 001 and thereby gain price priority at negligible cost. With this slight modification, the example plays out in the same way.

Now suppose there is no order ahead of the 10 000 order in the time priority queue at $40.00 per share. Suppose new short-term information, observed simultaneously by all traders, suddenly changes the expected future price of the stock from $40.00 per share to $39.80 per share. Fast traders will race to hit the 10 000 share buy order while

the slower buyer simultaneously will try to cancel the 10 000 share order first. The likely outcome is that the fastest trader hits the 10 000 share order before it can be canceled, earning an instantaneous profit of 20 cents per share on 10 000 shares, or $2000. The slow trader, whose order is "picked off," loses $2000.

The expected losses associated with being picked off are proportional to the size of the order, the frequency with which relevant information events occur, and the price movement associated with the events conditional on their occurring. If prices follow a martingale, there are some interesting connections between the frequency of information events and the size of the price movements that result from them. Suppose that 20 cents per share of return standard deviation results from such information events. This corresponds to one information event which results in a 20 cent per share price change. The same 20 cents of standard deviation can also result from 4 events which move prices 10 cents each (since $10 \times \sqrt{4}) = 20$) or 16 events which move prices 5 cents each (since $5 \times \sqrt{16} = 20$). Clearly, holding constant the size of resting limit orders, the total losses to resting limit orders are greater when a given standard deviation of returns volatility is associated with many small information arrivals. Total losses per share of resting limit orders are 20 cents when there is one information event ($1 \times 20 = 20$), 40 cents when there are 4 events ($10 \times 4 = 40$), and 80 cents when there are 16 events ($16 \times 5 = 80$). Since market prices tend to change in very small increments, the presumption must be that costs of being picked off are significant when measured in cents per resting-order share. In the limit as prices follow geometric Brownian motion, leaving a resting limit order of any size continuously in the market, replacing it with a new order every time it is picked off, results in infinite losses on infinite trading volume with infinitesimal losses on each order.

Clearly, this logic suggests that the potential net gains from following a market-making strategy of continuously place limit orders to buy at the bid price and sell at the offer price are going to be greater for a fast trader than a slow trader since the fast trader can more easily avoid losses from being picked off. This logic does not imply that a slow trader who only wants to buy or only wants to sell should never place resting limit orders. A slow trader who wants to trade in one direction must weigh the losses from being picked off against the bid-ask spread costs from placing executable limit or-

ders to sell at the bid price or buy at the offer price. In equilibrium, it is possible that these costs are about the same for slow traders, with slow traders therefore following mixed strategies of sometimes placing executable orders which hit bids and lift offers while other times placing non-executable orders to attempt to buy at the bid or sell at the offer before being picked off. Another possible equilibrium is that high frequency traders are so competitive among themselves and so adept at avoiding being picked off that the bid-ask spread is very tight, due to numerous fast traders competing at the best bid and offer prices, that slower traders always find it optimal to sell at the bid and buy at the offer.

A fast trader may also use the 10000 share buy order as a free "liquidity option" as discussed by Cohen et al. (1978). A fast trader may place another 10000 share buy order at a price of $40.01 per share. Price priority places the face trader at a better position in the queue. Suppose there are some incoming sell orders executable at a price of $40.00 per share. The fast trader's order will begin to execute at a price of $40.01 per share before the slow trader's order executes at all. If the price rises after the fast trader has bought some shares, he makes profits but the slow trader earns nothing. If the price looks like it might fall after the fast trader has bought some shares, he can cancel his order early and place a new order to sell the shares he just bought to the slow trader by hitting his resting order. The fast trader loses only $0.01 per share on up to 10000 shares, or $100. The possible gains if the price rises would likely be much greater, thereby stacking the odds in favor of the fast trader and against the portfolio manager. The portfolio manager's order will likely execute when prices move against him and will likely not execute when prices move in his favor. If negative information arrives before the fast trader has bought any shares at $40.01, he avoids losses by canceling early. The slow trader may limit the losses of fast traders by placing small orders.

To summarize, fast traders earn gains by arriving early to pick off resting orders and canceling early to avoid being picked off. These advantages of arriving early and canceling early do not specifically take advantage of tick size and allocation rules.

Gaming the Allocation Rule with Minimum Tick Size. As discussed in Section 1, the discreteness in the limit price and quantity makes some allocation rule necessary be-

cause multiple combinations of the price and quantity may clear the market. Different allocation rules determine the fractional allocation α in different ways. For example, time priority specifies that before newer orders receive any execution ($\alpha > 0$), older orders must receive full execution ($\alpha = 1$). Instead of time priority, some markets use a "pro rata" or proportional allocation rule according to which all orders receive the same fractional allocation α. Both time priority and pro rata allocation create incentives for gaming which benefit fast traders at the expense of slow traders.

In addition to using their speed to arrive early and pick off resting limit orders or cancel early to avoid being picked off, fast traders can also use their speed to make profits by gaming the allocation rule.

The reason is, essentially, that both the time priority and the pro-rata allocation reward traders from providing liquidity. At first, this might seem fair. Placing large orders before everyone else gives everyone else opportunities to hit the order and thus exposes the trader to being picked off. Not all traders, however, have the same ability to provide liquidity. It is more costly for slow traders to provide liquidity as they are more likely to be picked off. Furthermore, if fast traders can cancel their orders before everyone else can hit them, fast traders do not have to provide any liquidity. Therefore, an allocation rule that results from discrete prices produces additional rents that fast traders can earn at the expense of the rest of the market.

To illustrate how fast traders might game the allocation with a nontrivial tick size, consider the following example. There are two portfolio managers, a buyer and a seller. A buyer wants to buy 10 000 shares and a seller wants to sell 10 000 shares. They both would be happy to trade at a price of \$40.0050. With a one cent tick size, however, the allocation rule must determine the price at the bid of \$40.00 or the ask of \$40.01. Now a fast trader can place orders to sell at the offer price of \$40.01 and to buy at the bid price of \$40.00 as well. It depends on the allocation rule whether the buyer and the seller can trade with each other or not.

If the allocation rule is based on time priority, the fast trader may gain the best position in time priority queue by being at the best bid or offer first. For example, if the market recently changed from being offered at \$40.00 to being bid at \$40.00, this change may have occurred as a result of an incoming executable limit buy order trading against

an existing offer. After this trade occurred, there may have momentarily been no bid or offer at $40.00. If traders realize that a new best bid is likely to be established at $40.00, then fast traders may be the first to establish this bid, thereby obtaining time priority. If there is uncertainty about whether $40.00 is going to be the bid price or the offer price, then slow traders may avoid placing either a buy or sell limit order at this price for fear of being picked off.

With the pro-rata allocation, a fast trader can gain a larger allocation by placing a large order. For example, suppose a fast trader places orders to sell 90000 shares at $40.01 and to buy 90000 shares at $40.00, even though there are only 10000 shares available on the other side of his trades. Now suppose the limit price on the buy order at $40.00 is increased to $40.01. This order will fully execute at a price of $40.01. The pro-rata allocation rule assigns the fast trader 9000 shares while the slow seller will trade only 1000 shares. It is more economically advantageous for the fast trader to submit large orders than a slow trader because the fast trader can cancel orders more quickly to avoid being picked off when conditions change. If the market clearing price falls one tick and begins to bounce back and forth between $39.99 and $40.00, then the fast buyers will cancel their bids at $40.00, leaving the buyer to buy at the new offer price of $40.00. This is, of course, what prevents slow traders from gaming the allocation rule like fast traders in the first place.

By placing arbitrarily gigantic large orders, the fast trader can have almost all of the 10000 shares allocated to him. As prices bounce back and forth between $40.00 and $40.01, the fast traders earned $0.01 in spread profits on each share bought at $40.00 and sold at $40.01. These profits are proportional to the minimum tick size. A large tick size provides economic incentives for fast traders to place large orders at the bid and offer, forcing slow traders to incur a high bid-ask spread cost when they buy or sell.

In sum, fast traders earn rents at the expense of portfolio managers by exploiting the time priority, the price priority, and the minimum tick size. The discreteness in time, price, and quantity in today's exchanges rewards traders who can submit and cancel orders quickly, making the market winner-takes-all, where only the fastest wins.

Message Costs. One way for the portfolio manager to protect himself from fast traders is to buy 10 000 shares gradually over time by placing many small orders, none of which leaves large quantities resting in the market for a significant period. Nowadays large traders shred orders into small pieces, one share each, several price points, change prices as needed to keep close to market. For example, a traders may choose to participate in about one percent of trading volume on a relatively continuous basis. If the trader approximately matches the prices of other traders, he will obtain the Volume-Weighted Average Price (VWAP).

For example, the portfolio might trade 10 000 shares by placing 100 limit orders for 100 shares each, revising the limit prices as necessary to ensure that the orders are executed gradually over the day. Suppose a trader keeps an order close to the market, changing it each time the market moves one tick. If 80 cent standard deviation results from independently distribute price changes of plus or minus one cent, then price changes 6400 times per day, about once every 3–4 seconds. This increases the number of times limit prices on orders need to be changed. Purchasing 10 000 shares may require many tens of thousands of messages.

Sending numerous messages is costly, especially for traders with smaller bandwidth or processing power. When message costs are economically significant, traders face a tradeoff between incurring high message costs and submitting large orders. As a result of this trade-off, they may submit large messages and leave them resting in the market for a longer period of time, expos the orders to being picked off by fast traders. Consistent with the idea that fast traders have lower message costs than slow traders, Kirilenko et al. (Forthcoming) show that high-frequency traders have trades that are half as large (five versus ten contracts) as other traders.

Suppose the stock's daily return volatility of 2.00 percent per day results from the price impact of 100 independently distributed institutional bets of one percent of daily volume each. If prices fluctuate as a result of incoming orders then each bet is expected to move prices about 0.20 percent, or 20 basis points (calculated as $2.00/\sqrt{100} = 0.20$). This price impact of 8 cents per share is the natural, unavoidable price impact associated with order flowing creating return volatility. With suboptimal execution resulting from message costs, the price impact may larger in expectation, perhaps as little as 21

basis points or perhaps as large as 30 basis points or more. Quantifying these costs empirically takes us beyond the scope of this paper.

2.2 How Continuous Scaled Limit Orders Help Slow Traders.

Continuous scaled limit orders dramatically lower the potential rents fast traders earn at the expense of slow traders.

With continuous order types, fast traders do not earn substantial rents from arriving early. There is no longer time priority; all orders are treated symmetrically and executed simultaneously. The reward for placing an order one millisecond early lasts one millisecond. Suppose a portfolio manager submits one continuous scaled limit order to buy 10 000 shares at a price between $40.00 and $40.01 at a maximum rate of one share per second. There are 23 400 seconds of regular hours from 9:30 a.m. to 4:00 p.m. during a trading day. The trader can revise the limit price to keep the order close to the market. The order will be executed in one day if the market price is above $40.01 at least 42.73 percent of the day. When new public information suddenly changes the expected future price of the stock from $40.00 per share to $39.80 per share, the losses associated with being picked off are economically negligible. Since the continuous order buys at a maximum rate of one share per second, the portfolio manager's loss is limited to less than $0.20 if he cancels the order in less than one second. This is far less than losing $2000 when a standard limit order for 10 000 shares is picked off in the same way.

Similarly, the free "liquidity option" provided by slow traders is no longer valuable. When the price looks like it might fall, fast traders may try to liquidate their purchases by hitting the resting limit orders. The number of shares they can liquidate, however, is now much smaller. If the portfolio manager cancels his order within one second, fast traders can sell a maximum of only one share, not 10 000 shares. This eliminates the value of the liquidity option.

Since the market is no longer the fastest-takes-all, slow traders are protected by competition among fast traders. Suppose in the previous example that the slow trader took much longer than one second to cancel his order after the the public information was released. As fast traders race to sell their stocks to the slow trader, the price will

quickly go down. With the improved price, the losses to the slow trader will be much less than \$0.20 per share per second. If the equilibrium price falls \$0.18 per share due to competition among fast traders, the slow trader only loses \$0.02 per share per second.

The increased competition among fast traders has broader implications for economic efficiency. Today's winner-takes-all market structure encourages arms race among fast traders to become the fastest, as emphasized by Harris (2013); Li (2014); Biais, Foucault and Moinas (2015); and Budish, Cramton and Shim (2015). In a sense, fast traders excessively compete on their technology to avoid competition in price. Both over-competition in technology and under-competition in trading can be economically inefficient.

To summarize, continuous scaled limit orders address both inefficiencies. First, by providing a mechanism by which traders can trades gradually without having to send numerous messages, they reduce the rents that fast traders as a whole can earn by picking off slow traders considerably. Second, by removing time priority and treating orders symmetrically, they make fast traders with varying capacities compete with one another, which further reduces the rents that an individual fast trader can earn and the incentives to invest in technology to become the fastest.

Continuous scaled limit orders, unlike continuous (unscaled) limit orders, allow the market clearing price to be continuous even when the limit prices (P_H and P_L) respect the minimum tick sizes, which renders the allocation rule unnecessary and, thus, gaming the allocation rule impossible. Naturally, continuous scaled limit orders eliminate the rents fast traders earn from gaming the allocation rule. To illustrate how this works, suppose there are two portfolio managers, a buyer and a seller, who now can submit continuous scaled limit orders. The buyer places an order to buy $Q_{\max}^{BUY} = 10\,000$ shares between $P_L^{BUY} = \$40.00$ and $P_H^{BUY} = \$40.01$ at maximum rate $U_{\max}^{BUY} = 1$ share per second. The seller places an order to sell $Q_{\max}^{SELL} = 10\,000$ between $P_L^{SELL} = \$40.00$ and $P_H^{SELL} = \$40.01$ shares at maximum rate $U_{\max}^{SELL} = 1$ share per second. If the buyer and the seller are the only traders in the market, then the equilibrium price is the midpoint \$40.0050, and the buyer and seller trader with each other at a rate of 1/2 share per second.

Now suppose a high frequency trader tries to get between the buyer and the seller by

buying between $P_L^{HFT} = \$40.00$ and $P_H^{HFT} = \$40.01$ shares at maximum rate $U_{\max}^{HFT} = 2$ shares per second. Since

$$D(P_0) = 3, \qquad D(P_1) = 0, \qquad S(P_0) = 0, \qquad S(P_1) = 3, \tag{8}$$

we obtain

$$\omega = \frac{D(P_0) - S(P_0)}{D(P_0) - S(P_0) + S(P_1) - D(P_1)} = \frac{3}{4}, \qquad p(t) = (1-\omega)P_0 + \omega P_1 = 40.0075. \tag{9}$$

The higher price reduces the buyer's rate of buying from $U^{BUY} = 0.50$ shares per second to $U^{BUY} = 0.25$ shares per second and raises the seller's rate of selling from $U^{SELL} = 0.50$ shares per second to $U^{SELL} = 0.75$ shares per second. The high frequency trader buys $U^{HFT} = 0.50$ shares per second. As a result of his participation, the high frequency trader drives the price above the midpoint, but does not change the sum of the buyers rate of buying and the sellers rate of selling, which is 1 share per second.

With continuous scaled limit orders, the high frequency trader earns a profit by predicting future prices, not by earning a spread by intermediating trade between the buyer and seller. For example, cross-market arbitrage opportunities may still exist, and high frequency traders may exploit these opportunities. Competition among high frequency traders will make such arbitrage opportunities disappear quickly.

2.3 Comparison with Frequent Batch Auctions.

To reduce the rents that fast traders earn and the resulting arms race among fast traders, Budish, Cramton and Shim (2015) propose frequent batch auctions which match orders at discrete time intervals. Their approach contrasts with our approach in that they propose to make time more discrete while we propose to make time more continuous. Although frequent batch auctions have several desirable properties, frequent batch auctions do not sufficiently address all the perverse incentives that high-frequency traders enjoy in today's markets. Our continuous scaled limit orders fix these problems more robustly.

Frequent batch auctions reduce the costs slow traders incur from being picked off.

Here is the intuition of Budish, Cramton and Shim (2015). Suppose that a super-fast trader can react to changing market conditions in 2 milliseconds, a fast trader can react in 5 milliseconds, and a slow trader (portfolio manager) can react in 50 milliseconds. As before, suppose a slow trader has a limit order to buy 10 000 shares at $40.00 resting in the market. Suppose for now that a batch auction is held each second. If new public information that changes the stock value to $39.80 arrives one millisecond before the next batch auction, even the super-fast trader cannot react fast enough, and the slow trader's order is not picked off. If conditions change 3–4 milliseconds before the next batch auction, the super-fast trader can pick off the resting limit order at the next auction, and the fast high frequency traders' similar orders arrive too late. If conditions change 5–49 milliseconds before the next batch auction, the orders of both the super-fast and the fast traders arrive in time for the auction but the slow trader is unable to cancel. The slow trader may lose less than $2000 or $0.20 per share due to competition among fast traders. If the change occurs between from 50–1000 milliseconds before the auction, the portfolio manager successfully cancels his order.

This logic would seem to suggest that the longer batching interval reduces the losses of slow traders. If the news arrives with constant probability over time, a portfolio manager will be picked off with a probability that corresponds to 50 milliseconds divided by the length of the batching interval. The one-second interval reduces the loss of the portfolio manager by at least about 95 percent, and perhaps more than 99 percent if the competition among fast traders improves the price that the portfolio manager pays.

The logic, however, is incorrect because the order size submitted to auctions depends on the batching interval. Suppose a trader would place a one-share order if batch auctions are held every second. If batch auctions are held every two seconds, the same trader might submit an order for two shares. The theoretical trading models of Vayanos (1999) and Du and Zhu (2017) are consistent with this interpretation. If traders place larger orders in batch auctions, the losses suffered when the order are picked off are proportionally larger as well. Holding batch auctions every two seconds rather than every second may halve the probability of an order being picked off at a given auction, but a doubled order size the doubles the losses conditional on being picked off. Since these two effects cancel, changing the time interval between batch auctions does not change

the expected dollar losses traders suffer from being picked off.

Batch auctions do not resolve the costs of being picked off unless all traders optimally slice their orders and trade gradually. As we discussed earlier, without continuous order types, order shredding requires sending numerous messages, which is especially costly for traders with small bandwidth or processing power. Lee et al. (2004) and Barber et al. (2009) examine trading on the Taiwan Stock Exchange, which had one to two batch auctions every 90 seconds from 1995 to 1999. They show while large institutions smooth out their trading by participating in numerous auctions, individual traders place less frequent orders. Individual traders lose more than two percent of Taiwan's GDP trading stocks. These results are consistent with the interpretation that message costs cause slow traders to place suboptimally few and large orders.

Since frequent batch auctions do not address discreteness in the price, fast traders will still exercise their superior ability to game the allocation rule as they do in today's market standard limit orders. The risks of being picked off by fast traders when new information arrives a few milliseconds before the next auction limit slow traders' capacity to play the same games as fast traders.

Another issue is whether orders not fully executed from previous auctions should have time priority compared to new orders submitted to the current auction. On the one hand, it might be argued that traders who placed their orders in the previous auction should receive priority since they bear the risk of being picked off by other traders who observe the order imbalance and choose not to place their orders in the first place. On the other hand, it might be argued that it is likely that older orders in the limit order book come disproportionately from fast traders because their ability to react more quickly allows them to place large orders. Either way, it is likely that fast traders will be able to earn extra rents by exploiting the auction rules.

Clock synchronization is a major issue with frequent batch auctions. It is technologically difficult for exchanges to synchronize clocks exactly. If one exchange holds its frequent batch auction a millisecond or so earlier than another one, the outcome of the early exchange may be used by super-fast traders to pick off orders on the late exchange. Even with perfectly synchronized clocks, competing exchanges holding simultaneous single-price auctions will likely produce prices consistent with arbitrage opportunities

across the same stock traded on different exchanges and arbitrage opportunities across different assets traded on the same exchange. With continuous scaled limit orders, fast traders eliminate such arbitrage opportunities by submitting multiple offsetting orders. They do not have to wait for the next batch auction.

The last issue is transparency. Real-time pre-trade transparency is inconsistent with the spirit of frequent batch auctions because fast traders can exploit such information. If exchanges broadcast changes to the limit order book in real time, traders will wait until the end of the one-second batch interval before submitting new orders to prevent other traders from being able to react to their order changes, which rewards fast traders. Thus exchanges should not broadcast changes to the limit order book in real time but instead should consider only publishing information about unexecuted orders in the limit order book immediately after batch auctions, if at all. Suspicious traders may still suspect that exchanges will leak information about their orders to other traders. It is, therefore, important that exchanges have mechanisms in place to ensure that some traders do not obtain such information before other traders.

3 Policy Issues Related to Implementation

This section discusses how commonly proposed policies play out with continuous scaled limit orders. We first discuss pre-trade and post-trade transparency. We next discuss policies related to transparency, including competition among exchanges, dark pools, minimum resting times, privately arranged trades, and our proposed solution—quantity speed bumps. Finally, we discuss issues related to flash crashes, including price speed bumps and execution of market orders.

3.1 Transparency

This subsection discusses the issue of transparency with continuous scaled limit orders. Typically pre-trade transparency refers to publicly announcing information about current bid and ask prices, quantities at the best bid and ask, and potentially the quantities bid and ask at prices below or above the best bid and offer. Post-trade transparency

refers to revealing traders the prices and quantities traded in transactions.

With continuous scaled limit orders, these concepts play out differently. Post-trade transparency might consist of revealing trading volume and price, without revealing how many traders are buying and selling. As discussed in Section 1, an allocation rule is unnecessary because the market clearing price is always uniquely determined. Thus, traders can accurately infer the total quantity executed on their orders and the average price paid or received on their orders from the public feed of the market clearing prices. Such straightforward execution of all orders provides full post-trade transparency without exchanges having to send constant updates to all traders.

Pre-trade transparency implies releasing information that traders find useful for constructing optimal strategies. To determine the effect of new buy and sell orders on prices and trading rates, traders need to know the slopes of the aggregate demand and supply schedules around the market clearing price. Using the notation in Section 1, it follows that the minimum actionable pre-trade transparency includes the aggregate demand rates, D_0 and D_1, and supply rates, S_0, and S_1, and the two price points P_0 and P_1 around the market clearing price $p(t)$. These six pieces of data can be used to calculate the slope of the supply schedule $S_1 - S_0$, the slope of the demand schedule $D_0 - D_1$, the relative order imbalance ω in (6), the market clearing price as in equation (7), and the aggregate rate of trading volume

$$v(t) := S(P_0) + \omega(S(P_1) - S(P_0)) = D(P_1) + (1-\omega)(D(P_0) - D(P_1)). \tag{10}$$

The slopes of the supply and demand schedules determine the dynamic depth of the market. Given that the aggregate demand and supply schedules are piecewise linear functions with kinks at multiples of the minimum tick size, traders might want to know the slopes of aggregate demand and supply schedules outside the market clearing price. The exchanges may make public the aggregate demand and supply rates $D(p)$ and $S(p)$ at several integer multiples of the minimum tick size around P_0 and P_1. One argument for disclosing the slopes of the demand and the supply schedules outside the market clearing price is that fast traders can learn this information anyway. Fast traders with large bandwidth can place buy and sell orders away from the market for

brief periods of time, determine urgency away from the market from the execution of these orders over a few milliseconds, then cancel the orders quickly.[5] Determining exactly the price interval over which aggregate demand and supply rates are disclosed is a complex subject which takes us beyond the scope of this paper.

3.2 Market Fragmentation.

Competition Among Exchanges. In today's markets, various exchanges operate simultaneously and compete for trading volume. We believe that continuous scaled limit orders would be widely used in many exchanges. Suppose one exchange offers continuous scaled limit orders and the other standard limit orders. Which exchange will attract the most trading volume? We think the exchange offering continuous scaled limit orders will attract the most volume because its traders will not pay rents to fast traders while conserving bandwidth costs. Consider what happens to a resting limit order when the price suddenly changes. On the one hand, traders on the continuous exchange will pick off the orders on the standard exchange and earn meaningful profits if the size of the resting order is significant. On the other hand, traders on the standard exchange will not make meaningful profits picking off the orders on the exchange offering continuous scaled limit orders.

Dark Pools. Dark pools are trading venues which are not open to all traders and do not have pre-trade transparency. Dark pools exist for many reasons. In the 1990s and earlier, many large block trades were arranged privately off the NYSE exchange floor in the upstairs market. Negotiating trades privately outside the exchange is like participating in a dark pool. Dark pools also exist so that dealers can internalize small order from unsophisticated, uninformed customers. Dark pools also exist to facilitate trading inside the bid-ask spread and to avoid the adverse selection costs incurred when orders are picked off by fast traders.

[5] In a market with standard limit orders, traders need to know the quantities and prices at the best bids and offers. Currently, many exchanges also reveal quantities and prices for supply and demand schedules away from the market clearing price.

We think continuous scaled limit orders on organized open exchanges would dominate dark pools, including privately arranged trades in upstairs dealer markets. Historically, the frequency of large block trades declined after electronic order handling technology improved in the later 1990s, tick size was reduced to $0.01 in 2001, the NYSE specialists became less active in intermediating trades, and order flow dispersed across competing exchanges. Traders instead shredded large orders into tiny pieces which were executed as smaller trades of 100 or 200 shares. Furthermore, continuous scaled limit orders are designed to make gradual execution of large orders more cost effective for institutional traders by eliminating slippage in execution costs due to tick size and allocation rules and by reducing the bandwidth costs of executing large orders with many small trades.

Minimum Resting Time. Dealers have incentives to steer customers to trading venues which benefit the dealers at the expense of their customers. To protect unsophisticated customers from bad execution, we propose a minimum resting time for all dark pools. Dark pools would have to post tentative matched transactions to public scrutiny for some minimum resting time during which any market participant would be allowed to take one side or the other of the transaction, perhaps after offering modest price improvement.[6] For example, if a dark pool matches a 100 share trade at $39.99, this proposed transaction might be exposed to the market for five seconds, during which time the buyer or seller can by displaced by any trader offering price improvement of $0.01. [7]

[6] An alternative to our proposal is SEC regulations which mandate that customer orders be given "best execution" according to a regulatory definition. This approach, however, is unlikely to be optimal in a trading environment with rapid technological change, competing exchanges, and incentives for regulatory arbitrage.

[7] The rule is defined by two parameters: a five-second minimum exposure time and $0.01 minimum price improvement. These parameters might vary with the level of trading activity in the stock, with longer times and greater price improvement required for less actively traded stocks. The two parameter values proposed here are hypothetical. The optimal parameter might be quite different, say 1 second and zero price improvement. The two parameter values should be coordinated so that the free option to trade has little economic value if both sides of the transaction are matched at a market price. The parameters should also mimic the rules for infinitely impatient trades on the exchange offering continuous scaled limit orders.

Privately Arranged Trades. Similar problems arise in privately arranged trades brought to the exchange to be executed in a coordinated manner. Suppose two traders privately negotiate a gigantic trade outside the market. They negotiate a trade for, say, one million shares at $41.00, one entire day's normal trading volume traded at a price $1.00 higher than the prevailing price at the time the trade is negotiated. On an exchange that offers continuous scaled limit orders, two traders might enter continuous scaled limit orders to buy and sell, respectively, one million shares at rates of one billion shares per second at a price range of $39.99 to $41.01. If both orders arrive in the market at about the same time, both orders will fully execute their desired one million shares in one millisecond at a price close to $41.00. By executing such a large quantity so fast, the two traders will likely make it impossible for other traders in the market to participate in the transaction in a meaningful manner.

Such order executions are problematic. Despite its large size, one side of the trade may be a naive and poorly informed customer, perhaps the victim of an unscrupulous intermediary. Even if both the buyer and the seller are sophisticated and well-informed, there is a sense in which they are taking advantage of positive externalities provided by a transparent liquid market while not providing positive externalities to other traders. If all traders were to negotiate all trades privately, there is a danger that markets would be less transparent and less liquid, making all traders worse off.

Solution: Quantity Speed Bumps. Exchanges can deal with this issue by requiring orders of large urgency to take a meaningful amount of time to execute. For example, large urgency might be defined as a level of urgency which would execute one day's trading volume in five minutes, or 200 000 shares per minute for this stock. A meaningful amount of time is enough time for traders with moderately slow technology to submit orders to participate in the transaction. If a slow trader can react in approximately 50 milliseconds, any order which trades at an urgency of 200 000 shares per minute or faster might be required to have a minimum resting time of 5 seconds and not be fully executed in less than 5 seconds.[8] In effect, a minimum resting time for very

[8] If it is possible to execute such an urgent order fully in less than five seconds, either the order could be rejected by the exchange or, alternatively, the urgency of the order reduced so that full execution takes

urgent orders prevents traders from supplying instantaneous liquidity to other traders, which allows any trader with a 50 millisecond response time to participate in at least 99 percent of the time the order is actively in the market. Maintaining a level playing field suggests coordinating this minimum resting time rule with the the rule for crossing privately negotiated trades.[9]

Maker-Taker Pricing. When there is a legally binding minimum tick size, exchanges will engage in strategies of regulatory arbitrage to allow trading at fractional ticks. In the U.S. market, one mechanism for engaging in regulatory arbitrage is called "maker-take pricing." With maker-taker pricing, a trader placing a resting limit order pays a negative transactions fee while the trader placing an executable order pays a higher fee. For example, instead of both the buy- and sell sides to a trade paying a fee of \$0.0002 per share, the nonexecutable order "making" the market incurs a fee of $-\$0.0030$ and the order executable "taking" the market pays a fee of \$0.0034. Either way, the total fees earned by the exchange from matching a buy and a sell are \$0.0004 per share (since $2 \times \$0.0002 = -\$0.0030 + \$0.0034 = \$0.0004$).

This example of maker-taker pricing is economically equivalent to shifting all prices up by \$0.0032 per share, approximately 1/3 of a cent. Not surprisingly, there are also exchanges symmetrically offering "taker-maker" pricing, which has the effect of shifting prices down by approximately 1/3 of a cent. Altogether, the effect of maker-taker and taker-maker pricing is to cut the minimum tick size by a factor of approximately three. Since the best price jumps around from one exchange to another as prices change by fractions of a cent, maker-taker pricing rewards traders with low message costs and high bandwidth at the expense of other traders. For unsophisticated traders, the market becomes less transparent and more confusing, especially if data feeds report market bids and offers in whole cents which do not net out maker-taker fees.

With continuous scaled limit orders, there is no minimum tick size. There is therefore no regulatory arbitrage for maker-take fees to exploit. We believe that continuous

a minimum of five seconds.

[9] Of course, traders might try to violate the spirit of the rule by trading through multiple accounts with undisclosed common ownership or coordination. Such suspicious trading, which would be genuinely highly coincidental if not the result of coordination, should trigger an automatic audit by the exchange.

scaled limit orders would make maker-taker pricing go away.

3.3 Flash Crashes

Continuous scaled limit orders do not automatically prevent flash crashes, during which rapid executions of large orders cause substantial temporary disruptions to prices and volumes. On May 6, 2010, for example, one trader entered a series of orders to sell approximately $4 billion of S&P 500 E-mini futures contracts over a period of about 20 minutes rather than several hours that would have been typical for such a large amount of selling. Subsequently, prices collapsed by more than five percent and then quickly rebounded, as discussed by Kirilenko et al. (Forthcoming). The large seller who caused the flash crash above used an automated algorithm to participate in about 9 percent of trading volume without regard to price and time. The order executed very rapidly because trading volume increased dramatically partly as a result of his trading.

In many cases, extremely rapid selling is likely not an optimal strategy but rather a mistake; the traders who cause flash crashes do not benefit from them economically because they trade at unfavorable prices after the market moves against them. We believe that continuous scaled limit orders focus traders' attention on the time dimension of their orders, and thus would make flash crashes less likely. With continuous scaled limit orders, it is still possible that some traders may disrupt the market by trading large quantities quickly, whether intentionally or unintentionally. As Black (1971*a*) observed, it is a fundamental property of markets that executing large quantities over short periods of time will create adverse price movements.

Price Speed Bumps. To prevent unreasonable prices at times when new public information or extremely urgent orders move prices we propose price speed bumps. The implementation is straightforward. A speed bump begins when the price changes quickly over a short period of time, for example, by more than one cent per second, plus five cents, over any period during the day. Suppose the price has been stable at $40.00 per share for several minutes, at which point a sudden order imbalance makes the tentative market clearing price fall by $0.20 per share to $39.80. Since the maximum immediate

price change allowed is $0.05 per share and $39.95 is well-above the tentative price of $39.80, the speed bump kicks in. The speed bump stays in effect until the minimum price it allows, which falls at the rate of $0.01 per second, generates no excess supply. Excess supply is calculated by hypothetically executing at the minimum allowed price all orders in the market over the time interval that the speed bump is in effect. At the moment when the minimum allowed price generates excess supply, the new market clearing price will be the slightly higher price that clears the market for the entire duration of the speed bump.

This particular structure for a speed bump has several desirable features. First, if the price falls dramatically due to new very short-term information, very slow traders who do not cancel their orders receive price improvement. Second, if a trade with an extreme urgency triggered the price decline, the speed bump protects a naive urgent trader from his price impact by allowing new orders flowing into the market to offer price improvement. Third, the speed bump is hard to game. Suppose a trader places a large urgent order for the purpose of disrupting trading by stopping price formation, then tries to cancel the order before the minimum allowed price ever becomes a market clearing price. Then the cancelation itself is likely to end the speed bump and execute all of his disruptive trades at the worst possible price for him. The rule discourages intentionally disruptive as well as naively disruptive trading.

Market Orders. Nowadays a market order is essentially a limit order with an infinite price for a buy order and a price of $0.01 for a sell order. If a computer receives such an order, and there are no reasonable bids and offers available, the computer may execute the order at an unreasonably high or low price. During the flash crash of May 6, 2010, many market orders for individual stocks were executed at a price of $0.01 even though the stocks traded at prices like $40.00 per share seconds before and seconds after the orders were executed.

The possibility of executions at unreasonable prices suggests that market orders should either not be allowed or, if allowed, should not always be executed immediately at the best available price. We propose to replace a market order with a continuous scaled limit order with an automatic speed designed to achieve good quality execu-

tion over a short amount of human time. For example, a 100 share market order in the $40.00 stock might execute over 100 seconds, buying at a rate of one share per second with limit prices close to the market. Then the limit prices adjust gradually to more aggressive levels only if the execution is unusually slowly because prices are rapidly moving against the order. If a trader wants the more urgent execution of his order, then he could explicitly enter a continuous scaled limit order with the desired speed parameter, in which case the trader has himself to blame if his order creates a sudden temporary distortion in prices.

The way in which market orders are executed has changed over time. With human trading, a human broker would likely execute a market order by asking for bid and ask prices, accept the prices if they were competitive in the sense of being consistent with recent transactions, and ask for prices again if the available bids and offers did not seem reasonable. Asking for prices several times might take several seconds or even a minute or two, depending on the speed of recent trading. Our proposal for market orders resembles the way an honest, competent human broker might have handled market orders in the era of human trading.

4 Discussion of Related Literature and Institutions

A persistent theme in market microstructure concerns whether traders demand to trade immediately as opposed to slowly in the way continuous scaled limit orders are designed to help achieve.

Static Models. In theoretical models, infinite urgency results from assuming that noise trading is exogenous or assuming that traders act like perfect competitors. Under either assumption, a given quantity is traded immediately regardless of price.

In the model of Kyle (1989), informed and uninformed traders submit demand schedules which are downward sloping as a result of imperfect competition and risk aversion. Noise traders mimic infinite urgency by trading an exogenous quantity.

Grossman and Miller (1988) present a model of competitive trading in which market makers are continuously present in the market buy traders with a need to hedge

an inventory shock are not continuously present. If M market makers have the same risk aversion as one trader, the trader hedges the fraction $M/(M+1)$ of his endowment shock. This model does not justify artificially stimulating a demand for immediacy by increasing the tick size. They assume that traders are non-strategic perfect competitors who believe they do not incur price impact costs. In fact, such costs are substantial and induce traders to trade gradually to reduce trading costs.

In the one-period model of Kyle and Lee (2017), informed traders also receive endowment shocks. In contrast to the two models above, all traders are strategic. They show that optimal exercise of monopoly power induces privately informed traders not to demand urgency. Instead, they hedge only a fraction of endowment shocks to market impact. Trading less aggressively because of market power does not reduce the informativeness of prices. Indeed, the opposite is the case; traders trade more aggressively precisely when they have less price impact and their private information is not reflected in prices.

Dynamic Models. In the model of Kyle (1985), noise traders demand to trade exogenous random quantities immediately, and market makers supply immediacy by offering an upward-sloping supply schedule which allows traders to buy or sell significant quantities immediately. The informed trader does not need to trade with urgency because he has monopolistic access to private information which does not decay over time. Since price impact does not depend on time, the informed trader's price impact costs do not depend on how urgently he buys or sells. By trading gradually, the informed trader walks up and down the residual supply schedule like a perfectly discriminating monopolist.

The noise traders, who trade with infinite urgency, do not take advantage of the reduction in price impact costs that would result from trading smoothly. If noise traders were to trade gradually over an arbitrarily short period of time, they would halve their price impact costs. Not doing so essentially implies that noise traders do not take advantage of an arbitrage opportunity. If noise traders were to slow down their trading slightly, so that their inventories were a differentiable function of time rather than a Brownian motion, then noise traders would cut their trading costs in half but the mar-

ket makers would lose money. The equilibrium would collapse and be replaced by something else. What it is replaced with depends on the noise traders' motivations for trading, which might be inventory shocks or private values. Our proposal is designed to implement a trading equilibrium which would result from the natural operation of market forces in a trading environment as free of frictions as possible. In particular, we eliminate frictions associated with minimum tick size, minimum lot size, a costs associated with submitting, modifying, and canceling many orders.

Modeling optimal trading strategies with private information in an equilibrium setting is in principle very complicated. Kyle, Obizhaeva and Wang (2017) consider models of continuous trading on private information, with trade generated by overconfidence or stochastic private values. There is no exogenous demand for immediacy. The assumption of constant absolute risk aversion and normally distributed random variable allow to models to have nearly-closed-form solutions for equilibrium prices, quantities, and trading strategies. Each trader acquires new information continuously and trades on it with the expectation of making a profit. Traders are willing to take the other side of one another's trades because the believe trades of other traders are based on overconfidence or private values. There is an equilibrium in which all traders' trade slowly. Each trader submits a continuous demand schedule to by at a rate linear in price, linear in the trader's inventory, and linear in the trader's private valuation of the asset. The demand schedule defines the derivative of the trader's inventory as a function of the price. These trading strategies map almost perfectly into continuous scaled limit orders.

Vayanos (1999) considers trading model motivated by privately observed endowment shocks in discrete time. Du and Zhu (2017) consider a similar model in which investors receive private information about a liquidating dividend. Instead of holding auctions continuously, both models implement batch auctions by trading take place at discrete points in time. As the period between batch auctions is reduced, traders' expect a more liquid market and expand the quantities they expect to trade. For very frequent batch auctions, the expected quantity traded is approximately proportional to the length of the period between batch auctions.

Similar intuition describes all of these models. Traders trade gradually in order to exercise monopoly power optimally to control trading costs. Less aggressive strategies

reduce market impact costs because the aggressiveness with which a trader buys or sells signals his private information. When trade is motivated by overconfidence, the price reveals an average of traders' valuations immediately. Therefore, price react quickly even though quantities react slowly.

These equilibrium models imply that a finite tick size, a minimum lot size, or discrete batch auctions alter the underlying equilibrium. The models of Vayanos (1999) and Du and Zhu (2017) pay particular attention to the welfare properties of changing the interval between batch auctions. Their models suggest that there may welfare gains associated with moving from continuous batch auctions (equivalent to continuous scaled limit orders) to auctions held at more infrequent intervals (equivalent to non-continuous scaled limit orders). When information arrives almost continuously, the optimal time interval between batch auctions is almost zero.

Institutional Issues. The U.S. Securities and Exchange Commission (SEC) is currently implementing a "tick pilot" to study the effect of increasing the minimum tick size from one cent to five cents. The tick pilot proposal is the opposite of ours since it proposes to increase rather than decrease tick size. The intuition for the tick pilot is that if the bid-ask spread is wider, there will be more quoted instantaneous depth at the best bid and offer; this will allow impatient traders to trade toward their desired inventories faster. In principle, this could be socially desirable if there is demand for immediacy which is not being met due to market failures. The tick pilot disfavors small traders who want to buy or sell fewer shares than available at the best bid or offer. It disfavors poorly informed traders who cannot time their trades based on whether the midpoint of the bid-ask spread is cheap or expensive. It also creates incentives for dealers to route unsophisticated traders' orders to platforms where the dealer will be the opposite side of trades that are unprofitable for their customers.

The tick pilot draws intellectual support from research based on the idea that traders demand immediacy. The idea that market makers provide a risk-sharing service to investors is unrealistic. A typical investor is an asset management company managing billions of dollars in assets with a mandate to bear market risk. Market making firms are nowadays high frequency trading firms which are willing to bear limited risk. For

example, Kirilenko et al. (Forthcoming) found that high frequency traders took maximum net long or short positions of about $250 million during the flash crash; they hold positions on average for two minutes. Baron, Brogaard and Kirilenko (2013) find that high frequency traders earn about $6 per contract (1 basis point) on trades with small traders and about one dollar per contract on trades with institutional investors. Earning 0.1 basis points over two minutes corresponds to earning a return of about 50% for holding the same risk for an entire year. Is it reasonable to assume that an asset manager with tens of billions in assets under management be willing to pay so much for so little?

Duffie (2010) suggests that slow-moving capital results from search frictions with adverse selection. Dealer markets provide an efficient search mechanism when investors do not pay continuous attention, it takes time to search, intermediaries may cause bottlenecks. Our proposal solves the inattention problem by allowing one message to implement a near optimal gradual trading strategy. If all traders are continuously present in the market and can use any trading strategy, they will likely trade gradually over time.

Glosten (1994) argues that a consolidated, competitive limit order book with continuous prices and quantities dominates other types of exchanges. In his one-period model, time is not divisible. This leads to a finite equilibrium bid-ask spread in which very small orders incur a positive cost. We believe that allowing the limit order book to evolve continuously in time will drive the bid-ask spread on infinitesimally small trades to zero. Indeed, this interpretation is almost immediately implied by the models of Kyle, Obizhaeva and Wang (2017), Vayanos (1999), and Du and Zhu (2017).

Kyle and Viswanathan (2008) argue that two goals of a markets are to provide market liquidity and prices conveying economically useful information. Continuous scaled limit orders deter traders from trading on high-frequency information and from exploiting allocation rules to gain time or price priority. By reducing trading costs for traders who acquire long-term information, continuous scaled limit orders both increase market liquidity and allow prices to contain more long-term information.

5 Conclusion

Continuous scaled limit orders make it possible to implement Fischer Black's vision of continuous electronic markets without requiring traders to place enormous quantities of limit orders. Continuous scaled limit orders do not eliminate price impact costs, which are a natural feature of markets in which adverse selection is important. Continuous scaled limit orders dramatically reduce the profits that high frequency traders make by using their speed to exploit time priority, price priority, large tick size. This enhances economic efficiency by reducing incentives to invest in costly technology to win playing a zero-sum game. Other policy ideas to reduce the high-frequency-trading arms race include frequent batch auctions proposed by Budish, Cramton and Shim (2015) and random message processing delays proposed by Harris (2013). Unlike these proposals, continuous scaled limit orders directly address the source of underlying problem, the perverse incentives created by limit order discreteness in price, quantity, and time.

References

Barber, Brad M., Yi-Tsung Lee, Yu-Jane Liu, and Terrance Odean. 2009. "Just How Much Do Individual Investors Lose by Trading?" *Review of Financial Studies*, 22(2): 609–632.

Baron, Matthew, Jonathan Brogaard, and Andrei Kirilenko. 2013. "The Trading Profits of High Frequency Traders." Working Paper.

Biais, Bruno, Thierry Foucault, and Sophie Moinas. 2015. "Equilibrium Fast Trading." *Journal of Financial Economics*, 116(2): 292–313.

Black, Fischer. 1971*a*. "Toward a Fully Automated Exchange, Part I." *Financial Analysts Journal*, 27(6): 29–34.

Black, Fischer. 1971*b*. "Toward a fully automated stock exchange, Part II." *Financial Analysts Journal*, 27(6): 24–28.

Budish, Eric, Peter Cramton, and John Shim. 2015. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." *Quarterly Journal of Economics*, 130(4): 1547–1621.

Cohen, Kalman J., Steven F. Maier, Robert A. Schwartz, and David K. Whitcomb. 1978. "The Returns Generation Process, Returns Variance, and the Effect of Thinness in Securities Markets." *Journal of Finance*, 33(1): 149–167.

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *Journal of Finance*, 65(4): 1237–1267.

Du, Songzi, and Haoxiang Zhu. 2017. "What Is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*, Forthcoming: available at `http://ssrn.com/abstract=2857674`.

Glosten, Lawrence R. 1994. "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, 49(4): 1127–1161.

Grossman, Sanford J., and Merton H. Miller. 1988. "Liquidity and Market Structure." *Journal of Finance*, 43(3): 617–633.

Harris, Larry. 2013. "What to Do About High-Frequency Trading." *Financial Analysts Journal*, March/April: 6–9.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun. Forthcoming. "The Flash Crash: High Frequency Trading in an Electronic Market." *Journal of Finance*.

Kyle, Albert S. 1985. "Continuous Auctions and Insider Trading." *Econometrica*, 53(6): 1315–1335.

Kyle, Albert S. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies*, 56: 317–356.

Kyle, Albert S., and Jeongmin Lee. 2017. "Information and Competition with Symmetry." available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2892141`.

Kyle, Albert S., and S. Viswanathan. 2008. "How to Define Illegal Price Manipulation." *The American Economic Review: Papers and Proceedings*, 98(2): 274–279.

Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang. 2017. "Smooth Trading with Overconfidence and Market Power." *Review of Economic Studies*, Accepted for Publication: available at `http://ssrn.com/abstract=2423207`.

Lee, Yi-Tsung, Yu-Jane Liu, Richard Roll, and Avanidhar Subrahmanyam. 2004. "Order Imbalances and Market Efficiency: Evidence from the Taiwan Stock Exchange." *Journal of Financial and Quantitative Analysis*, 39(2): 327–341.

Li, Wei. 2014. "High Frequency Trading with Speed Hierarchies." available `https://ssrn.com/abstract=2365121` or `http://dx.doi.org/10.2139/ssrn.2365121`.

Vayanos, Dimitri. 1999. "Strategic Trading and Welfare in a Dynamic Market." *Review of Economic Studies*, 66(2): 219–254.

Institutional Rigidities and Bond Returns around Rating Changes

Matthew Spiegel[†]
Laura Starks[‡]

November 19, 2016

†Yale School of Management, P.O. Box 208200, New Haven, CT 06520-8200. Email: matthew.spiegel@yale.edu.

‡McCombs School of Business, University of Texas, Austin, TX 78712, laura.starks@mccombs.utexas.edu.

We would like to thank Catherine Nolan for providing us with numerous institutional details and insights regarding the bond market. Comments from William Goetzmann and Geert Rouwenhorst are gratefully acknowledged. We also with to thank seminar participants at South Florida University, Georgia State University and the Yale finance group's brown bag for helpful comments and suggestions.

Abstract

Corporate bonds face institutional rigidities from the division between investment grade and non-investment grade clientele. Examining how rigidities affect returns requires a methodology that takes the infrequent trading of bonds into account. Using a methodology that modifies the repeat sales method by incorporating bond characteristics, subsequent to a bond rating crossing the investment/non-investment boundary, we find the transaction price for the bond shows significant negative (or positive) abnormal returns over time, followed by a partial recovery. We further show that a structural shift occurred in these reactions after the financial crisis.

Many corporate bond portfolio managers' investment strategies depend critically on bond rating categories, in particular, the categorization between investment grade and non- investment grade (high-yield or junk) bonds. Because many bond portfolio strategies have a targeted benchmark index and specify limits on the proportion of each category in the portfolio, if a bond's rating approaches or crosses the boundary between investment and non-investment grade, the manager often has an imperative or an incentive to sell the position. At the same time, institutions on the other side of the divide should be ready to serve as the counterparty for the sell position, but are not always willing to do so. This institutional rigidity combined with the low levels of liquidity in the bond market can lead to disruptions when a bond's rating changes. In particular, if the rating change pushes a bond across the investment/non-investment grade boundary, the desire by one market segment to sell may not be met equally by a desire by the other market segment to buy, creating price pressure, a result similar to the mutual fund fire sales documented in the equity markets. Such fire sales result in negative abnormal returns on the affected stocks for a period of time, followed by a partial rebound.[1] In this paper we address the question of how institutional rigidities caused by bond ratings influence the bonds' returns and the extent to which they do so. We hypothesize that the rating changes can lead to long-lived price adjustments due to the institutional rigidities, which will then be partially if not completely reversed.

To test our hypotheses we use the Trade Reporting and Compliance Engine, more commonly known as TRACE, which since 2002 has provided publicly available bond transaction data.[2] However, important empirical challenges exist in analyzing corporate bond transactions. The first arises from the low liquidity levels in the markets given that we find most corporate bonds trade less than once a month, if at all. Our hypotheses are focused on short run serially correlated returns, which may be

[1] Fire sales in the stock market, due to mutual fund flows have generated a great deal of interest. Papers by Coval and Stafford (2007), Ali, Wei and Zhou (2011), and Dyakov and Verbeek (2013) among others examine this issue. These papers identify equity fire sales by a pattern of serially correlated returns followed by a partial recovery. We suggest a similar pattern of returns can be induced by bond market institutional rigidities.

[2] See http://www.investopedia.com/terms/t/trace.asp for a brief history of TRACE's development.

difficult to measure with this lack of liquidity. A further empirical challenge is that the bond market illiquidity poses impediments to establishing a benchmark index with which abnormal returns can be measured. We provide a novel approach to the problem by developing an econometric model that combines two techniques from real estate research, a literature that faces similar data issues. The resulting returns are then used to determine whether the institutional rigidities, along with the limited liquidity, play a role in short-term bond returns and whether other aspects of bond portfolio manager trading strategies mitigate or exacerbate the institutional rigidities.

Our methodology is based on the repeat sales regression used in the real estate literature to deal with two characteristics common to the both the bond and the real estate markets, heterogeneous assets that trade infrequently (e.g., Goetzmann, 1992; Francke, 2010; and Peng, 2012). This technique calculates the returns between pairs of transactions on the same asset. Given the infrequency of sales, the resulting returns vary in length and cover different periods of time. To create an equally-weighted index, the returns are then regressed on a set of indicator variables representing the return per period.[3] However, the broad equally-weighted indices usually employed in the repeated real estate sales regressions would be insufficient for corporate bond return analyses because bond returns vary systematically with bond characteristics such as issue size, credit quality, industry, years to maturity and other factors. Although with stocks, such a problem is handled to some degree by estimating a factor model and then adjusting the benchmark returns accordingly, that approach would be impractical with securities that trade as infrequently as do bonds. Consequently, we create a custom index for each bond by estimating a "characteristic-weighted repeat sales index" in which we weight the repeat sales data by characteristic distance.[4] Consider a bond *i* with a vector of characteristics X_{it} as of date *t*. Another bond *j*

[3] For example if a model contains returns for dates 2 through 8 and a house sells on dates 3 and 5 it would have a dummy of one for returns 4 and 5 and zero elsewhere. Standard modifications account for heteroskedasticity in the data.

[4] This modification borrows from a separate strand of the real estate literature on hedonic models (Meese and Wallace (1991)).

in the data set has a characteristic vector X_{jt}. The model then calculates a Euclidean distance between the characteristic vectors and reweights the data accordingly. Thus, data from bonds with characteristics similar to the one in question are given larger weights in the regression model than those further away. Using this technique we generate a set of daily benchmark returns for each bond. These are then subtracted from the observed returns on the bond at issue to generate a set of abnormal returns (AR).

When analyzing stock data, calculated abnormal returns typically span a constant length of time, for example, a day or a month. However, for bonds the irregular time between observed trades means that the estimated index-adjusted ARs span various lengths of time. To account for this a second regression is run on the bond-by-bond ARs to estimate daily ARs for the set of bonds impacted by the event in question. The result is an estimated daily AR around the event date and we test the null hypothesis that the sum of the regression coefficients (CARs) equals zero.

The results from the estimated CARs over the sample period indicate that when a bond is downgraded from investment to non-investment grade, the ARs over the following days approximate −130 basis points (bps), a decrease in price that is both statistically and economically important. However, over the next few weeks the bonds gain back almost half their loss, about 60 bps. Downgrades to bonds that were already non-investment grade do not have as strong a price effect with a post announcement CAR of about −40 bps over the next week, which largely reverses by the end of the second week after the announcement. Bonds in the investment grade category that face downgrades but do not cross the investment/non-investment grade border have little market reaction either immediately or in the weeks that follow.

Thus, the institutional rigidity created by the investment/non-investment grade boundary has an important effect on bond returns, leading to returns that stray, for a time at least, from a random walk. One can rank the negative initial price reactions to a downgrade and subsequent recovery from smallest

to largest: bonds starting and ending as investment grade, bonds starting and ending as non-investment grade and finally bonds that cross from investment to non-investment grade. This pattern is consistent with the hypothesis that the bonds' price pressure from investment grade funds takes some time to ameliorate when they are forced out of an issue and have to wait for non-investment grade demand to come in to take possession of it. However, there are clearly other forces at work as well since the post negative return and recovery reaction is also seen, if to a much smaller degree, for down-graded bonds that are already in the non-investment grade category.

Upgrades in which the bonds move across the investment/non-investment grade boundary have a more muted effect on bond returns than the border-crossing downgrades, but they still have significant returns. Specifically, bonds moving from the noninvestment to investment grade category experience small positive CARs in the first 2 weeks following the announcement and going out 8 weeks the returns are closer to 70 bps. In contrast, bonds that remain in their rating class after an upgrade see little if any change in value even after 8 weeks have passed. Again, the contrast in these patterns is consistent with institutional rigidities having an impact on bond returns. The fact that there is no significant difference in returns for upgrades in the same rating class, but that crossing the border from non-investment grade to investment grade provides significant increases in returns suggests that part of this return comes from the fact that the bonds have to transfer one set of potential investors, non-investment grade funds, to another, the investment grade funds. Since the latter has not had any reason to hold the bonds and thus research them prior to the rating change, it can take them some time to absorb the issue. This kind of friction can produce the return patterns seen in the border-crossing upgrades.

One aspect of the investment grade/non-investment grade rigidities in the financial markets allows us a unique identification when the market segmentation effects are likely to be most severe. Benchmark indices vary over types of bondholders. In particular, investment grade bond mutual funds

tend to use Barclays indices as benchmarks, while non-investment grade bond mutual funds tend to use the Bank of America Merrill Lynch (BofAM) indices. The two indices do not score bonds equivalently or at the same time, which can result in a set of bonds, that for at least a period of time are dropped by one index provider but not included by the other. We dub these observations as "orphan" bonds, the bonds that experience a rating downgrade that drops the bonds out of the investment grade category using Barclays scoring rule, but not under the BofAML rule. As a consequence, these bonds lack a natural constituency since they are not in the benchmark index used by either investment or non-investment grade funds. That is, prior to the downgrade institutional investment grade funds would be the natural holders of these issues. Post announcement they no longer are. At the same time the non-investment grade funds do not have the bonds in their benchmark either. Thus, because of the difficulty in finding buyers, prices for these bonds keep falling as markets attempt to clear. This unique status provides a particularly appropriate test of whether institutional rigidities influence bond returns. We find the evidence supporting our hypotheses to be quite strong as the orphan bonds lose more than 500 bps of their value over the 8 weeks that follow the downgrade announcement.

Our hypotheses and results that rating changes have effects on bond returns run counter to some of the earlier literature on the impact of ratings changes. For example, Weinstein (1977) finds that rating changes follow bond price declines (with a 6-month lag), but that there was no impact after the ratings change, a pattern that has been replicated in numerous subsequent studies. The conclusion from this literature has been that bond rating changes reflect past market performance but correlate weakly, if at all, with future risk-adjusted returns. However, the earlier authors did not have access to bond transaction prices, which only became publicly available in 2002. Consequently, authors accounted for the missing trade data through techniques such as using data from pricing services (e.g., Wansley, Glascock and Clauretie, 1992) or trader quotes (e.g., Warga and Welch, 1993). The problem with such techniques is that market makers and pricing services may quote stale prices on bonds that have not

traded for some time and that are unlikely to do so in the near future. Moreover, quotes are not price commitments and at most are only firm for a trivial volume. For a meaningful lot size, dealers may only feel compelled to produce accurate quotes after the arrival of a bona fide transaction query. This can easily result in what looks like return momentum following a rating change, even if there is none.

Several later studies (Hite and Warga, 1977; May 2010; Ellul, Jotikasthira, and Lundblad 2011) document some effects after a bond downgrade, but they have encountered other problems. First, benchmarking poses a problem for any study seeking to estimate bond returns. One solution has been to use commercial benchmarks as in Hite and Warga (1997). They find some price drift in the month before and after the announcement of a bond downgrade. The abnormal returns in their econometric model are net of a Lehman Brothers index that tracks bonds with a similar maturity and rating level. However, the firms calculating the benchmark returns have the same problem anyone else seeking to analyze bonds have – a lack of pricing data because bonds trade very infrequently.[5]

May (2010) uses the TRACE data to examine bond returns by value-weighting all bonds that traded on days *t* and *t*−1 with the same rating and broad maturity class, thus, avoiding the problem of using non-price estimates. He finds that both downgrades and upgrades impact bond returns in the two-day event window around the change with downgrades having the stronger economic impact. However, since most bonds do not trade even once a week, let alone daily, such a rule drops most issues from the database. Further, since those bonds that do trade frequently tend to be the larger, more liquid, and more recent issues, this obviously skews, and potentially biases, the indices created from them and potentially any empirical analysis based on these indices.

[5] Commercial firms work around this problem by calculating "matrix prices" to fill in for the missing data. However, Warga and Welch (1993) show that these prices often lag the market by a substantial length of time. Furthermore, when looking at rating changes across the investment-noninvestment boundary, the bond in question will be at the edge of any benchmark based on a particular rating class. This may make the benchmark a poor representative of how the bond would have done absent the event in question.

The study of insurance company transactions and prices around bond rating changes by Ellul, Jotikasthira, and Lundblad (2011) is the most similar to our paper. Our paper differs from theirs in a number of ways that add a unique contribution. First, and most importantly, we provide insights into the structural changes in the bond markets that occurred after the financial crisis. Moreover, our methodology of marking to a basket of similar securities rather than marking to model, allows for additional insights. In addition, we have a fuller sample of bonds and their trading because our data contains all bond transactions rather than just transactions with an insurance company on one side. As Bessembinder, Maxwell and Venkataraman (2006) point out, although insurance companies hold a large proportion of corporate bonds (estimated by Schultz (2001) to be about 40%), they only account for 12.5% of corporate bond trading. Thus, we have a much larger sample for examining abnormal returns since their analysis covers 384 bonds while ours covers over 8,000 bonds.[6] Finally, we examine crossing the investment/noninvestment grade boundary from both directions.

A further differentiation of our paper from previous research is that we consider how behavioral regularities in trading such as trend chasing, positive relationship between liquidity and return, and reaching for yield may extend or mitigate the effects from the ratings change. That is, in analyzing the effects of the institutional rigidities, one needs to also account for the effects from these trading behaviors.

The paper is structured as follows: in Section I we discuss the constraints faced by institutions that lead to rigidities in their willingness to hold certain issues. We next provide an overview of the data In Section II and discuss bond liquidity and document how infrequently most issues trade in Section III. In Section IV we contrast the scoring system used by Barclays and BofAML. In Section V we develop the

[6] Other articles have examined bonds around the investment grade/non-investment grade boundary in order to study other issues such as the effects on firms' investments (Chernenko and Sunderam, 2012), the purpose of credit ratings (Bongaerts, Cremers and Goetzmann, 2012), and how index labeling affects the bonds (Chen, Lookman, Schürhoff and Seppi, 2014).

econometric model used to estimate bond ARs and CARs and present the results in Section VI for various rating changes and various sample periods. We provide our conclusions in Section VIII.

I. Institutional Constraints

Corporate bond portfolio managers have a variety of investment goals. Some invest across all bonds, regardless of rating. Many, however, restrict their holdings to either investment grade or high yield issues, resulting in somewhat segmented markets. Whether a bond belongs in one category or another generally depends on the ratings assigned the bond by the rating agencies, the most prominent of which in the U.S. are S&P, Moody's and Fitch.[7] Examples of these restrictions can be found in the prospectuses of corporate bond mutual funds. The Calvert Long-Term Income Fund (ticker CLDAX) states within its "Principal Investment Strategies" section,

> The Fund typically invests at least 65% of its net assets in investment grade, U.S. dollar-denominated debt securities, as assessed at the time of purchase. A debt security is investment grade when assigned a credit quality rating of BBB- or higher by Standard & Poor's Ratings Services ("Standard & Poor's") or an equivalent rating by another nationally recognized statistical rating organization ("NRSRO"), including Moody's Investors Service or Fitch Ratings, or if unrated, considered to be of comparable credit quality by the Fund's Advisor.[8]

In theory, this means the fund can invest in any bond that one of the rating agencies has designated as investment grade and hold some non-investment grade issues as well. However, an added factor that could push bond portfolio managers to focus on investment grade bonds is that these managers are not oblivious to benchmark risk. Holding bonds outside the benchmark imposes significant performance risk. This incentive may be especially strong in the bond market where an individual bond's upside potential is quite limited, although its downside is not. Thus, the benchmark used to assess a bond's portfolio performance may also influence the manager's desire to hold or shun a particular issue. For example,

[7] For expositional clarity the investment, noninvestment and distressed classifications will be referred to as rating categories.
[8] Page 34.

the benchmark for CLDAX is Barclays Long U.S. Credit Index. (Our search through a number of investment grade fund prospectuses shows this is typical.)

The prospectuses of high yield funds also provide similar disclosure regarding the restrictions on their holdings. For example, Calvert's High Yield Bond Fund (ticker CYBAX, CHBCX and CYBYX depending on class) states

> Under normal circumstances, the Fund will invest at least 80% of its net assets (including borrowings for investment purposes) in high yield, high risk bonds, also known as "junk" bonds. The Fund will provide shareholders with at least 60 days' notice before changing this 80% policy. . . . When a corporation issues a bond, it generally submits the security to one or more nationally recognized statistical rating organizations ("NRSROs") such as Moody's Investors Service ("Moody's") or Standard & Poor's Ratings Services ("Standard & Poor's"). These services evaluate the creditworthiness of the issuer and assign a rating, based on their evaluation of the issuer's ability to repay the bond. Bonds with ratings below Baa3 (Moody's) or BBB- (Standard & Poor's) are considered below investment grade and are commonly referred to as junk bonds. Some bonds are not rated at all. The Advisor determines the comparable rating quality of bonds that are not rated. [9]

As with the earlier income fund examples, the prospectus states the fund's benchmark which in this case is the BofA Merrill Lynch High Yield Master II Index. (After reviewing a number of prospectuses this benchmark appears to be the industry standard for high yield funds.)

Barclays and BofA Merrill Lynch seem to run the industry standard benchmarks, making the list of bonds they either do or do not include of upmost importance to numerous fund managers. Since these are benchmarks, the firms publish rules governing when a bond is or is not included. For the Barclay investment grade indices the rule governing inclusion is:

> Securities must be rated investment grade (Baa3/BBB-/BBB- or higher) using the middle rating of Moody's, S&P and Fitch; when a rating from only two agencies is available, the lower is used; when only one agency rates a bond, that rating is used. In cases where explicit bond level ratings may not be available, other sources may be used to classify securities by credit quality:
>
> • Expected ratings at issuance may be used to ensure timely index inclusion or to properly classify split-rated issuers.

[9] Page 26.

• Unrated securities may use an issuer rating for index classification purposes if available. Unrated subordinated securities are included if a subordinated issuer rating is available.

The BofA Merrill Lynch (BofAML) index, however, uses a somewhat different rule based on an average score from S&P, Moody's and Fitch. Table 1 displays the numerical score assigned to each rating. After calculating a score BofAML then rounds out the result, rounding *up* numbers ending in 0.5.[10] For example, a bond has a rating from S&P of BBB2, from Moody's of Baa3 and none from Fitch. Then the score is (9+10)/2 = 9.5 and this is rounded up to 10. In a case like this, the result is identical to what the algorithm used by Barclays produces. However, there are cases where they are not. Consider a bond without a Fitch rating but with scores from S&P of BBB2 and Moody's of Ba1. In this case, the average score is 10 based on the BofAML rule. However, the Barclays algorithm yields an 11 since it takes the lower score when there are just two. While scoring discrepancies like this are not common, they do occur, from which we derive tests of orphan bonds in Section G.

A. Institutional rigidities and bond trades

A primary cause for institutional rigidities in the bond markets is the existence of a boundary based on bond ratings for portfolio managers' investment strategies. When a security drops from a portfolio's benchmark index because of a ratings change, managers have two reasons to sell the issue. First, as noted earlier, bond portfolios, particularly bond mutual funds, typically have clauses restricting the extent to which they can hold securities outside their primary strategic universe. As examples, consider the rules imposed on Oppenheimer's Corporate Bond Fund (OFIAX), T. Rowe Price's Corporate Income Fund (PRPIX) and Calvert's Income Fund (CFICX). These funds are investment grade corporate bond funds that include holding restrictions in their prospectuses. OFIAX limits its high yield holdings to 20% of its portfolio, PRPIX has a 15% limit and CFICX has a 35% limit. High yield funds have similar types of restrictions on their holdings of investment grade bonds. For example, both Oppenheimer's Global High

[10] We thank Preston Peacock from BofAML for providing us with the details regarding how they round scores ending in 0.5.

Yield Fund (OGYAX) and T. Rowe Price's High Yield Fund (PRHYX) limit their investment grade holdings to 20%.[11] Calvert's High Yield Bond Fund (CYBAX) does not have a strict limit, stating that investment grade bonds are "permitted but not a principal investment strategy." Thus, for either type of fund, a bond with a ratings change will move out of the fund's primary investment category, which then adds to the weight of the fund's portfolio that the prospectus limits, giving the fund manager a requirement or an incentive to sell the issue.

Beyond the prospectus limits, the potential benchmark tracking error created by the removal of the bond from the index also provides fund managers an incentive to sell the bond issues that move outside their mandate. For example, suppose a bond constitutes 5 bps of the index and 7 bps of a fund's portfolio, which means relative to the benchmark the fund is long the issue by 2 bps. If the bond is removed from the index, the fund is suddenly long the issue relative to the benchmark by 7 bps, a nontrivial swing that positions the fund from slightly bullish in the issue to very bullish. The fund manager can then achieve a position in line with the benchmark by selling the issue. Moreover, funds that took a relatively bearish position (under 5 bps) may have an even stronger incentive to sell. What was once a short position relative to the index is suddenly long.

B. Hypotheses on the Effects of Rating Changes

In this section we develop hypotheses regarding bond rating changes in two diverse circumstances, (1) the rating remains within the bond's category, that is, before and after the ratings change the bond is either an investment grade or non-investment grade bond; (2) the rating change moves the bond across the investment grade/non-investment grade boundary. Our primary focus revolves around the latter, that is, the effects of institutional rigidities on bond trading during the downgrades and upgrades due to the restrictions regarding the category of bonds a portfolio manager can hold, which should play a major

[11] Although Oppenheimer qualifies this restriction as only applying "under normal market conditions" the fund firm does not further define what this means.

role when a ratings change forces a bond across categories. The institutional rigidities should not affect bond returns when a rating change leaves a bond's broad ratings category unchanged.

Complications for testing the hypotheses arise because of systematic trading effects from microstructure models in which markets have limited liquidity due to an intermediary's inventory concerns.[12] In such models, large buys or sales are spit up over time and lead to serially correlated returns. Prices ultimately overshoot their long run equilibrium value and then partially reverse back. Further complicating the tests are several documented behavioral regularities in institutional investor trading that need to be considered as these regularities may extend or mitigate the trading effects from the ratings change: trend chasing, positive relationship between liquidity and return, and reaching for yield.

Trend Chasing: Evidence shows that when prices increase for a security, some investors engage in trend chasing by adding the security to their portfolio or increasing their current holdings. The opposite effect tends to occur after a price decline. Evidence that institutional investors behave this way is extensive (Grinblatt, Titman and Wermers (1995), Wermers (1999), Badith and Wahal (2002) and Alti, Kaniel and Yoeli (2012)).

With regard to bond funds, the trend-chasing hypothesis suggests prices should overshoot their equilibrium values after a ratings change; since they tend to follow price moves. On the way up trend-chasing funds will want to buy, but the trend chasing should restrict the supply of sellers. The opposite would hold when bonds lose value. The resulting pattern from the trend-chasing hypothesis is that returns should be serially correlated for some time and then partially reverse.

[12] A general discussion of these models can be found in Madhavan's (2000) survey article. A particularly relevant example to the current paper can be found in Keim and Madhavan (1996). In their model a block sale comes through that overwhelms the market's short term liquidity provision. When that happens returns exhibit positive serial correlation over time as the position is worked off. Once the block is exchanged, prices will have generally overshot their equilibrium value and then move in the opposite direction for a time.

Liquidity and Past Returns: A related phenomenon is that higher returns lead to higher future liquidity (Chordia, Roll and Subrahmanyam (2001) and Hameed, Kang and Viswanathan (2010)). If an institution wants to sell a security whose price has recently risen, there should be a smaller temporary price impact than if the same sized sale occurred following a price drop. Thus, for bond upgrades liquidity should increase, which would reduce the degree to which returns are serially correlated. This phenomenon should also reduce or eliminate any tendency for prices to overshoot their equilibrium value. For bond downgrades, the opposite should be true.

Reaching for Yield: Another related hypothesis with behavioral elements is based on the evidence that many institutional investors appear to chase yields. For example, if there are two AA bonds and one has a slightly higher yield, the investor would be more likely to hold the higher yielding one. Evidence in support of this tendency is provided in studies of the trading of insurance companies (Becker and Ivashina (2015) and Merrill, Nadauld and Strahan (2015)) that finds these investors appear to overweight relatively high yielding securities within the rating classes they hold. Thus, if institutional investors are reaching for yield, they would be natural buyers for bonds that move categories on downgrades. Similarly, they would be sellers for bonds that upgrade to a new category. This type of trading behavior should help offset the impact of trend chasers when an upgrade occurs.

The above discussion can be summarized as:

Hypothesis 1: (a) Following a ratings downgrade in which a bond remains in either the investment grade or non-investment grade category, i.e., its initial broad ratings category, trend chasing and liquidity provision suggests we should observe serially correlated returns on the bond and its prices should overshoot their equilibrium value. Reaching for yield will mitigate this. (b) For upgrades in which the bond remains in its ratings category, trend chasing should induce serially correlated returns and prices that overshoot their equilibrium values. Both liquidity provision and reaching for yield should mitigate

this tendency. Overall, serial correlation and overshooting should be more pronounced for downgrades than upgrades. Institutional rigidities should have little effect for bonds that remain in their original ratings category after a change in rating.

When a bond switches from one ratings category to another, trend chasing and liquidity provision are likely to have the same influence on prices that they do when a bond remains within its original broad ratings category. However, this is not true for portfolio managers reaching for yield or those constrained by institutional rigidities. The combination may even exacerbate their individual influences.

When a bond rating switches from investment to non-investment grade, the bond goes from being among the highest yielding bonds in its ratings category to among the lowest. For investment grade funds reaching for yield, these bonds are initially attractive. However, after the downgrade they either have to sell out of the position or have strong incentives to do so because of the benchmark deviation they will face if they continue to hold a bond out of their benchmark. Moreover, among the bonds' potential buyers, the high yield portfolio managers reaching for yield will find these bonds to be particularly unattractive because the bonds will be the lower yielding in their new category. The combination of the investment grade bond portfolio managers wanting to sell and the high yield managers being less interested leads to net selling pressure and thus, may lead to serially correlated returns and prices that overshoot their equilibrium value. In contrast, upgrades that switch a bond's rating category cause it to go from being among the lowest yielding in its category to among the highest. While the bond was initially unattractive to high yield funds that are reaching for yield it is now particularly attractive to investment grade funds that wish to do so.

Hypothesis 2: (a) *When a rating downgrade causes a bond to switch into a new ratings category the institutional rigidity will work in the same direction with the three regularities, which suggests we should observe serially correlated returns and prices that overshoot their equilibrium values. (b) When a ratings*

upgrade causes a bond to switch into a new ratings category, reaching for yield should help reduce the degree to which returns are serially correlated and prices overshoot their equilibrium value. (c) Overall, the greatest degree of serial correlation and price overshooting should occur for downgrades that switch a bond's ratings category.

II. Data

The tick-by-tick bond prices are obtained from the TRACE database for the period beginning July 1, 2002 and ending on June 30, 2015. The reported volume for each transaction is truncated, with non-investment grade bonds being reported as $1 million plus for transactions over $1 million (in par value). The truncation for investment grade bonds is higher; trades in excess of $5 million in par value are listed as 5 million plus.[13] We convert the prices and reported volumes into daily closing prices through the following algorithm. If a bond trades just once during the day we use that trade price as the closing price. If the bond trades multiple times during the day, we use the last trade as the closing price provided the trade volume is large enough to yield a truncated value (an institutional sized trade). Otherwise the closing price we use is derived by computing a size-weighted average of the last three trades in the day.[14]

Bond characteristics are drawn from the Mergent Corporate Bond Securities Database. This database reports a number of bond characteristics including the bond's ratings by the major rating agencies, call schedules, and coupon frequency among others. To be included in the final sample, a bond must be rated by S&P, Moody's or Fitch and also must conform to the following terms: (1) make semi-annual coupon payments, (2) accrue interest on a 360 day year, (3) have USA as the country of domicile,

[13] See the TRACE data guide offered by the Wharton Research Data Services at https://wrds-web.wharton.upenn.edu/wrds/query_forms/variable_documentation.cfm?vendorCode=TRACE&libraryCode=trace&fileCode=trace&id=ascii_rptd_vol_tx.

[14] If there are only two trades, then they are size weighted and averaged to produce a closing price. Trades dated on weekends and bond holidays are dropped from the database.

(4) list its denomination and payments in US dollars (this excludes "Yankee" bonds), (5) have a type of PSTK, PS, EMTN, MBS, TPCS or CCOV and (6) have an industry code below 40.[15]

III. Bond Liquidity

The limited trading in the corporate bond market means that market prices are unavailable to either estimate factor loadings or a bond's current market value. How problematic this is depends on how infrequently a trade takes place. In Table 2 we provide some indication of just how serious this issue is in the corporate bond market. The table displays, by percentile rank, what fraction of days per year a bond trades. (Throughout the paper, "days" refers to trading days, i.e., when the market is open.)

Table 2, summarizes a measure we term "fraction of days traded" (FDT). Because we only observe prices when a bond trades, we do not have precise information on the time at which the bond enters or leaves the market. Consequently, to assess trading frequency we use the following algorithm. A bond *B* is included in year *Y*'s data if it trades in any day during or prior to year *Y* and during or after year *Y*. The FDT in year *Y* for bond *B* then has as its numerator, the number of days bond *B* traded in year *Y*. The denominator contains the number of trading days in year *Y* on or after the first trading day and on or before the last trading day observed for *B* in the entire dataset. Some examples:

- Example 1: A bond trades in 2005 and 2007 but not 2006. FDT(2006) = 0/total trade days in 2006 = 0.
- Example 2: A bond trades on July 11, 2006 and July 12, 2006 but never before or after. FDT(2006) = 2/2 = 1, since total trade days during 2006 between the first and last trade date in the bond is 2.

[15] This excludes bonds issued by foreign agencies, foreign governments, supranationals, the U.S. Treasury, a U.S. Agency, a taxable municipal entity or is in the miscellaneous or unassigned group.

This measure is designed to overstate just how frequently a bond trades during the year. The number of days built into the denominator assumes that once a bond's final trading date is observed the bond leaves the market and can never trade again. Similarly, it assumes a bond is unavailable for trade prior to the first date it appears in TRACE. This is clearly untrue. Unless the bond has been called or matured trading can take place, even if trades do not occur. The point is to provide some intuition regarding how infrequently bonds trade and this measure provides an upper bound on that concept.

Within each year the bonds that are part of that year's sample are ranked by their FDT. Table 2 displays the percentile break points, in percentage terms. Rows represent years for which a full year of data is available. Many of the lower percentile cells contain zeros due to bonds that trade prior to and after the year in that row, but not in that year. The early and late years have entries in all columns due to how end point problems impact the FDT calculation. If a bond trades in 2001 and 2004, it is not included in the 2003 row since the earlier 2001 trade occurred prior to the initial date for the Trace database. Similarly, a bond that trades in 2013 and again in 2016 will not be included in the 2014 data as a 0, since the 2016 observation occurs after the end period for the sample. However, the important point to note is that even with this very conservative measure of trading frequency the median bond trades about 12 or 13 days during the average 254 trading days in a year, which amounts to only about 5% of the available trading days.

Panel B in Table 2 reports the FDT in a different way by measuring time by the number of years since a bond first trades (t_0). The year 1 label is applied to all trades from t_0 to its first anniversary. For example, if a bond first trades on July 11, 2006 all trades in that issue up to July 10, 2007 are aggregated into year 1. The denominator equals all trading days between July 11, 2006 and July 10, 2007. The figures in Panel B across all years shows just how infrequently most issues trade both initially and over time. The median value is just 8.59% during year 1 and falls off to 3.12% by year 4. Even bonds with an FDT score at the 75 percentile, trade on just under 20% of the available trading days in their first year

and by year 4 are down to just under 10%. It is only at the 95% level that the drop off in FDT becomes somewhat less severe over the years. However, even at this level it goes from 37.5% down to 23.5% from years 1 to 4. If the criteria for calling an issue liquid is that it averages close to 1 trade a week, then only bonds in the top 1% of all issues can be said to be liquid past the 5th anniversary of their first trade.

IV. Rating Changes

The hypotheses developed earlier are based on the idea that crossing the boundary between investment and non-investment grade leads to different market reactions than when ratings change within each classification. The drop in a bond's liquidity over time, as indicated in Table 2, suggests rating changes that occur farther from the initial issue date will be accompanied by relatively thin transactions data. We next examine whether this holds in our data by collecting any rating changes that lead a bond to move across the investment/non-investment grade boundary. Panel A of Table 3 reports these numbers by calendar year and Panel B reports them by the number of years since the bond was first rated. Panel A shows a clear variation across years. Prior to 2011, downgrades occur more frequently than upgrades. Clearly, the financial crisis led to a large number of downgrades in 2008 and 2009. From 2011 to 2013 upgrades became more common, with a particular jump in their occurrence in 2013, although in 2014 there were a few more downgrades than upgrades. Panel B of the table shows that many of the rating changes leading to a change in classification to or from investment and non-investment grade occur years after a bond has been issued. Combined with the evidence in Table 2, this occurs after trading in the bond has likely dropped significantly. Thus, estimates of how rating changes impact bond returns will necessitate making up for the lack of daily pricing data.

Given the increased motivation for trading in bonds that cross between investment and non-investment grade, the question arises as to whether the additional trading is sufficient to employ the return estimates commonly used on stock data to the bond data at hand. To check this we tabulate the

FDT over the months and days following a rating change that pushes a bond across categories and report the results in Table 4. In the months prior to a rating change, consistent with previous evidence showing that information precedes bond rating changes, the bonds initially in the investment grade group trade more often than their peers; on about 20% of all days. Those initially in the non-investment grade group trade only about 6% or 7% of the time until the month prior to being upgraded. In the prior month trade nearly doubles in frequency to around 12% or 13%. In the months following a downgrade from investment to non-investment grade a typical bond's FDT score drops significantly. After about 6 months these bonds seem to trade about as often as the non-investment grade bonds that were ultimately upgraded. The reverse is also true of those bonds upgraded from investment to investment grade. Their FDT scores jump. Six months prior to the rating change they have a FDT score of about 6% and after the rating change it goes to about 15% and remains there for at least 6 months. These patterns are consistent with the general observation the bond market liquidity declines with the rating (Han and Zhou (2007) and Chen, Lesmond, and Wei (2007) and Kalimipalli and Nyak (2012)).

V. Estimating Bond Returns

As Section III shows, most bonds trade too infrequently to create factor loading estimates as could be done for stocks. Even creating a benchmark to use as a factor poses a challenge. Real estate is an area in which academics need to deal with heterogeneous assets that trade infrequently. A popular solution has been to employ a repeat sales regression. We create what can be called a distance-weighted bond index by combining the repeat sales algorithm with a kernel estimation model used by Meese and Wallace (1991).[16]

In a traditional repeat sales model, the price p of asset i at dates b (buy) and s (sale) is assumed to follow

[16] Meese and Wallace (1991) used their statistical model to estimate San Francisco housing returns.

$$p_{is} = p_{ib} \prod_{t=b+1}^{s} (1 + r_{mt}) \varepsilon_{it} \quad (1)$$

where r_{mt} is the return on the benchmark portfolio *m* and ε_{it} is a log normal error term. The model assumes no intervening cash flows exist between the two transaction dates. Taking logs, letting *R* = log(1+*r*) and $e = \log(\varepsilon)$ yields

$$\log(p_{is} / p_{ib}) = \sum_{t=b+1}^{s} R_{it} + e_{it}. \quad (2)$$

The model in equation (2) can then be estimated by using dummies equal to 1 if the time period *t* is between the buy and sale dates and zero otherwise. The variance of each observation equals $(s-b)\sigma_e^2$ where σ_e^2 is the variance of *e*. Equation (2) can be estimated via weighted least squares to account for the heteroscedasticity across observations.

Equation (2) is based on the assumption that there are no intervening cash flows between sales.[17] The vast majority of corporate bonds pay coupons semi-annually. (This study drops the few that do not.) Because coupons arrive so infrequently most transactions pairs lack an intervening cash flow and thus satisfy equation (2). The few trading pairs where this is not true have been dropped from the data used to estimate returns in this paper.

The standard repeat sales model works well when the goal is to measure the average return across a broad array of illiquid assets. However, the model may not do as well when dealing with assets that have particular characteristics within the group. In our setting the focus is on bonds that recently transitioned between investment and non-investment grade, which are not representative of the whole

[17] Geltner and Goetzmann (2000) propose a variant of the repeat sales model that can handle transaction pairs with intervening cash flows. As a practical matter, to estimate the model with any reliability there need to be sufficient time 0 data points. The TRACE data lacks that requirement and when we attempted to implement the Geltner and Goetzmann model the design matrix was not numerically invertable.

bond market. In particular, their returns may vary from the general corporate bond market due to characteristics such as maturity, current yield and industry. The repeat sales model can be adapted to this problem by using a variant of the technique suggested by Meese and Wallace (1991), in their case for estimating a hedonic model. Conceptually, we adapt the model to distance weight the rows in (2) to account for how far in characteristic space the observation is from the bond whose return one wants to benchmark.

Define the distance between two observations *i* and *j* with characteristics vectors $X = (x_1 \ldots x_n)$, with *n* characteristics by $D\left[X_{it}, X_{jt}\right]$. In the current application, *D* is a ratio in which the numerator is the Euclidean distance between the two sets of characteristics, where the characteristics *x* are normalized to have unit standard deviations. The denominator is a value that sets *D*<1 for X% of the data. The characteristics we employ are a bond's current yield, days to worst call date and a measure derived from the issuer's 4-digit SIC code, defined as a dummy equal to 1 if two firms are in different 4-digit SIC codes, and 0 if they are in the same one.[18] All characteristics are measured as of the date of the first trade in each repeat sales pair.

Following Meese and Wallace (1991) once distances are calculated the observations are then weighted with the tri-cube function

$$\begin{aligned} W_j &= \left(1 - D_{ij}^3\right)^3, \text{ if } D_{ij} < 1 \\ &= 0 \text{ otherwise.} \end{aligned} \tag{3}$$

After weighting the rows in (2), the return parameters are then estimated via least squares. The resulting estimates are used as the benchmark returns for asset *i*. A bond's abnormal return between dates *b* and *s* is then estimated as

[18] The days to worst call date is the same as days to maturity if the bond is either non-callable or if the yield to worst call date is the maturity date.

$$AR_{i,b,s} = \log\left(p_{bi}/p_{si}\right) - \sum_{t=b+1}^{s} \hat{R}_t \quad (4)$$

where $\hat{R}_t$ is the distance weighted repeat sales estimate of R_t.

In principle, the repeat sales index in equation (2) can be estimated using the entire TRACE database from 2002 to date. However, current computing power makes this approach technologically challenging. The solution used in the analysis that follows is to estimate equation (2) year-by-year using 3 year rolling windows. For example, the benchmark index for 2006 rating changes is created from trade data spanning January 1, 2005 to December 31, 2007. Similarly, the 2007 benchmarks are estimated using data from January 1, 2006 to December 31, 2008.

B. Commercial Benchmarks as an Alternative

Institutional investors typically have their returns compared to the benchmark indices by BofAML and Barclays. While these benchmarks are readily available, using them to estimate the abnormal return to a bond that transitions between investment and non-investment grade seems likely to produce biased results. Relative to an investment grade index, these bonds are rated at the lowest edge of the comparison group and at the upper edge in terms of risk. Relative to a non-investment grade index the opposite is true. It seems unlikely that either index will yield an appropriate return adjustment. Furthermore, default rates are not linear in the ratings. Rather they are convex in their numerical scores, as documented by Emery, et al. (2008) for corporate bonds and Altman and Suggitt (2000) for syndicated loans.

Another option is to use an overall bond index. However, the overall bond market is not equally distributed across rating classes. According to the Securities Industry and Financial Markets Association, high-yield bonds have comprised between 6% and 25% of the overall new issues market. An overall

bond index will therefore skew towards less risk and a higher rating than a bond transitioning between investment and non-investment grade. Finally, there is the issue of index measurement. The index providers also have to deal with the lack of transactions on which to base prices. Their solution is to estimate a bond's value using a spread to Treasuries. For example, the Barclays US Corporate Index Factsheet states, "Most securities in the US Corporate Index are priced using a spread to Treasuries. . . " While this technique is simple, if ratings are sticky estimated price changes will lag the market.[19]

VI. Estimating and Testing Cumulative Abnormal Bond Returns

Estimating equation (4) across bonds produces a list of abnormal bond returns (AR).[20] However, due to the infrequency with which bonds trade these returns cover various spans of time. One option is to assume the ARs are spread evenly between trade dates. For example, if the AR is 100 bps from date 2 to 12 one can assign an AR of 10 to each day. However, this can lead to problems when the time span crosses a date boundary where the suspicion is that ARs before and after differ.

A variation of the repeat sales model can help deal with the fact that the AR calculations cover varying lengths of time. The hypothesis is that the AR on date *t* relative to some event data is R_t. Thus, one can estimate

$$AR_{i,b,s} = \sum_{t=b+1}^{s} R_t + e_t \tag{5}$$

and then calculate the CAR from date t_0 to t_1 as $\sum_{\tau=t_0}^{t_1} \hat{R}_t$, where $\hat{R}_t$ is the estimated value of R_t. A standard significance test can then be conducted as to whether the CAR (sum of the regression coefficients) does or does not equal 0. As in a standard repeat sales regression, return data on the left

[19]Again, see Warga and Welch (1993) and Hite and Warga (1997) for detailed discussions of this issue.
[20] Here, as in the prior discussions returns should be interpreted as log(1+*r*); the estimated value produced by the repeat sales regression.

hand side of (5) will exhibit heteroscedasticity in proportion to the time between sales. A simple weighted regression will correct this and is what this paper uses to estimate the reported $\hat{R}_t$.

As Webb (1988 and 1991) shows estimates from a repeat sales model, like those in equations (2) and (5) suffer from measurement errors that follow an AR(1) process. While this issue disappears in large samples it can be problematic when the data is thin, as it is in some of the tests conducted here. A simple solution is to bootstrap the estimates. The mean return estimates based on the random sampling eliminates the AR(1) measurement error and produce robust standard errors.[21] In what follows, all of the abnormal return estimates using equation (5) and estimates derived from them are based on bootstrapped values. We do not bootstrap the repeats sales estimates from the first stage regression (equation (2)). These negatively serial correlated measurement errors should not qualitatively impact the final estimates from (5), which are the ones of interest.[22]

Once the first stage regression is completed the ARs from it are stacked and estimated against a second repeat sales model. To the degree that the negative serial correlation in the parameter estimates then feeds into the estimated ARs, these should either be completely or close to independent across much of the sample. For example, ARs for a bond in 2005 are calculated with a dataset that is independent of the one used to calculate the ARs for a bond in 2010. Thus, in the second stage regression where the first stage ARs become the dependent variable, the result is just additional noise.

[21] We thank William Goetzmann for this insight and suggestion.

[22] It is also true, that attempting to bootstrap the first stage regression would take current computers months to perform the calculations. Without bootstrapping calculating the custom index and the resulting ARs for each bond in a category (e.g. downgrades within the non-investment grade category) takes about a day.

C. CARs in the Days around a Category Crossing

The first set of tests look at the daily CARs around a rating change that either leaves a bond in the same ratings category or moves it between investment and noninvestment. Day 0 is the date on which the rating change is recorded in the Mergent database. The results are in Table 6.

Table 6 Columns 1 and 2 tabulate the results for rating changes that drop a bond from investment to non-investment grade. The estimated post announcement returns are in the −180 bps range. The results are both economically and statistically significant. There is some evidence of continued negative returns for another day or two after which the price appears to level off with a net loss of approximately 250 bps.

The next set of columns examine the returns to a bond that is downgraded but still remains inside its original ratings category. For downgrades within both the investment and non-investment grade categories returns are relatively small and not persistently significant. Nevertheless, they are uniformly negative indicating that investors in these issues may well suffer losses after a rating downgrade. Within the non-investment grade category it appears that losses within a week of the downgrade come to somewhere between 50 and 70 bps, with some evidence of the start of a recovery by day 10.

On the right side of Table 6 are the returns around upgrades. Upgrades for bonds that cross from noninvestment to investment grade show evidence of positive returns that take a week or two to materialize post announcement. By the end of the second trading week post announcement (day +10) returns are about 50 bps and statistically significant using either the BofAML or Barclays rule. Upgrades for bonds beginning and ending in the non-investment grade category are smaller. There is some evidence using the Barclays rule that the returns are positive for about a week and then mean revert to some degree. A similar pattern appears under the BofAML rule, but there the significance levels are

quite a bit lower. For bonds that are upgraded within the investment grade category the estimated returns are small, of inconsistent sign and lack statistical significance. It is very hard to reject the idea that the announcements have no impact on the market.

The results from Table 6 support the hypothesis that institutional rigidities play an important role in the market reaction when a bond downgrades from the investment to the noninvestment category. For other downgrades that do not move a bond across the investment/noninvestment border, there is little return reaction possibly due to trend chasing affecting the market's overall liquidity and the influence of reaching for yield apparently offset it. With respect to upgrades into the investment grade category, institutional rigidities seem to play a role as well, but to a lesser degree. (Of course, it may be that the rigidities play just as strong a role but are offset by other factors. For example, liquidity, reaching for yield and trend chasing may all combine to produce a reduced overall price impact.) In the other upgrade tests, it may be that the additional liquidity in the investment grade market (see Table 4) is sufficient to offset the impact of a ratings upgrade, while the same may not be true for non-investment grade bonds.

D. CARs Prior to a Rating Change

As noted earlier, numerous studies have found that bond rating changes follow changes in the bonds' market prices. Thus, the ratings change may be expected. We examine this possibility in Table 7 by repeating the analysis in Table 6 for the weeks prior to the rating change. The results are generally consistent with previous studies. Almost across the board, downgrades follow negative returns in excess of 100 bps. For bonds that cross from investment to non-investment grade the negative returns start at least five weeks prior to the downgrade and ultimately total over 150 bps. For downgrades among non-investment grade bonds that remain non-investment grade (i.e. do not transition to distressed) negative returns only precede rating changes by between 1 and 3 weeks and the statistical significance depends

on the scoring rule used. For changes among investment grade bonds that then remain in the investment grade category the negative returns are close to 120 bps but seem to start as much as 6 weeks earlier. [23]

For upgrades the picture is again somewhat mixed. Bonds upgraded from noninvestment to investment and those that start and end in the noninvestment category yield statistically significant positive abnormal returns prior to the rating change. About 70 bps in the former case and 100 in the latter. But, for upgrades that involve bonds that start and end in the investment category there is no economically or statistically significant indication that rating changes follow a string of positive returns.

E. CARs Following a Ratings Change

As demonstrated in the earlier tables the corporate bond market is an illiquid market. As pointed out earlier, microstructure models suggest that this illiquidity can lead prices to overshoot their equilibrium values and then bounce back to some degree in the weeks following the initial price change (Keim and Madhavan (1996)). In Table 8 we address this issue by examining reported CARs that begin at the end of the 10th day following a ratings change, which is the day on which the results reported in Table 6 end.

Table 6 shows that up to day 10 bonds downgraded from investment to non-investment grade experience post announcement returns in the range of −233 bps. Table 8 shows that these same bonds have a partial rebound of 130 or 78 bps depending on the scoring rule used, which is consistent with the Keim and Madhavan model in which price pressure from block sellers and limited liquidity produce serially correlated returns and prices that overshoot their equilibrium value. A similar rebound is observed for downgrades that cause a bond to start and end as non-investment grade. Again, this occurs

[23] We conjecture that the rating agencies are reluctant to downgrade investment grade bonds and require a relatively long negative run of news before doing so. For non-investment grade bonds, the rating agencies may worry about missing an event that leads to default. They therefore require a shorter negative news run prior to downgrading such bonds. Of course, these are just conjectures at this point and we do not pursue them any further in this paper.

in the part of the market where liquidity is likely to be thinnest. In contrast, the bonds that start and end in the investment grade category after a downgrade that showed at most a modest post announcement negative return in Table 6, have little evidence of a price rebound in Table 8.

For upgrades in which a bond crosses from the noninvestment to investment grade category, reports an approximate 50 bps post announcement price reaction in the days following the rating change. Table 8 indicates a continued upward drift for an additional 4 weeks. By trading day +26 returns have increased by another 61 to 78 bps. In contrast, the other rating upgrade groupings show little consistent evidence of systematic price changes. At most, for bonds that are upgraded but start and end in the noninvestment category there may be a small positive post announcement price increase of between 13 and 22 bps after 3 weeks, a gain which then reverses itself by week 5. However, the evidence for this is pretty thin and it would be reasonable to conclude that one cannot dismiss the null hypothesis that the post announcement benchmark adjusted returns are random noise around zero.

F. CARs for Border Bonds that then Cross the Border

Following a ratings change, Table 6 through Table 8 compare bonds that move from an investment or non-investment grade category to another against those that remain in their initial category. However, the potential size of the rating change for bonds that switch categories is quite a bit larger than for bonds that do not. Consider a bond with an initial rating of 1 (AAA). For this bond to continue to be included in the no change investment grade category, it cannot fall below a rating of 10, thus, a change of at most 9 points. For the AAA bond to move from the investment to non-investment grade category, its rating has to change by at least 10 points. In fact, it can change up to 15 points (to 16 points) and still be included in the investment to non-investment grade group. Large rating changes such as this are likely perceived very differently by market participants than a typical downgrade that simply moves a bond by a single rating point. As noted earlier, another problem with a potential large rating change is

that rather than being rated, a ratings agency may simply drop coverage, resulting in measurement error being introduced into the ratings history. That is, rating agencies are not obligated to rate a bond for its entire life. A firm that undergoes a large negative or positive corporate event may induce a rating agency to drop coverage, in which case such a decision will not be included in the database. All one has is the last rating; a rating that is no longer valid. It seems likely that the market is fully aware of this and that the firms creating bond benchmarks adjust their scoring rules accordingly. This measurement error can lead to situations where bond rating changes are underestimated based on the available ratings data.

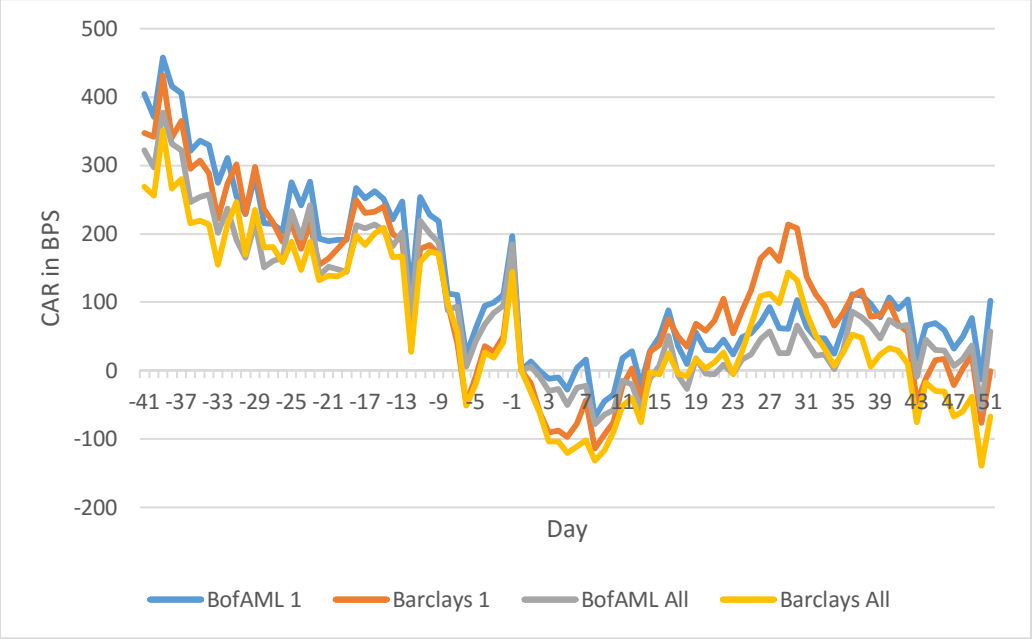


Figure 1: Downgrades that Change a Bond's Classification from Investment to Noninvestment.

Figure 1 and Figure 2 address the issue of whether large rating jumps are driving the prior results. The two graphs compare the CARs for bonds that start near the border of their rating class and then cross versus bonds that simply move from one rating class to another. The legend shows both the scoring system used and the filter used to select bonds. The prefix “BofAML” or “Barclays” indicates the classification scoring system. A suffix of “all” indicates that the CARs include any bond that begins in one

class (investment or noninvestment) and then crosses to the other class. A suffix of "1" indicates that the CARs only include bonds that begin within one point of the border prior to crossing it.

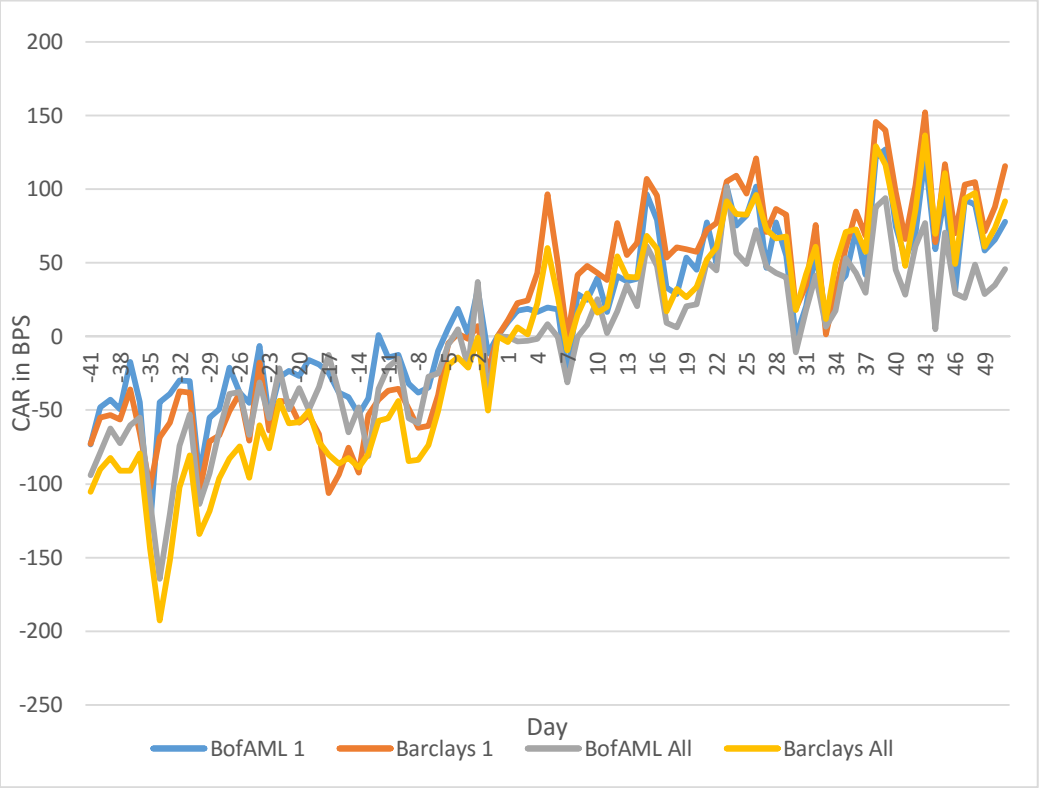


Figure 2: Upgrades that Change a Bond's Classification from Noninvestment to Investment.

While including either all bonds or just those that are near the border prior to crossing makes some difference in the CARs, these do not seem to be material. It is true that including all bonds results in a slightly lower CAR overall. For downgrades, the average CAR differs by 63 bps and 85 bps based on the BofAML and Barclays scoring rules respectively. For upgrades, the average differences are 25 and 22 bps respectively. However, more importantly, in terms of the overall return pattern there seems to be little difference. For both upgrades and downgrades bonds tend to see a price reaction over a number of days following the rating change after which there is some evidence of a recovery (stronger for downgrades than upgrades).

G. Orphan Bonds

As pointed out earlier, because BofAML and Barclays use slightly different classification rules, they do not always lead to the same groupings. In principle, some bonds may be rated investment grade by one and non-investment grade by the other. Recall, that investment grade funds tend to use the Barclays classification rule and investment grade funds the BofAML rule. The result is that some bonds are orphans in that they are non-investment grade via Barclays rule and investment grade via the BofAML rule.[24,25] While orphans are not created very often, they do provide another way to test the hypothesis that institutional frictions lead to liquidity problems. If the frictions arising from a change in rating category impact bond returns, then orphan bonds should be particularly vulnerable. These are bonds that are not part of the benchmarks used by either income or high yield funds. If institutional rigidities are important, this is the group that should be most affected. The current investment grade holders need to sell the issue, but non-investment grade funds have no reason to buy it as it is not in their benchmark either.

Table 5 lists the number of bonds that become orphaned in each year. For these bonds institutional rigidities should play a particularly dramatic role. There is, however, no reason to believe the other three factors listed in Section I.B will impact any differently in this case than in the more general case of a ratings downgrade.

[24] While cases where a rating change leads Barclays to rate the issue investment and BofAML non-investment grade are theoretically possible, the data do not contain any examples where it occurred.

[25] For the orphan bond tables, cases where the BofAML and Barclays rules produce scores that differ by 3 or more are dropped. Occasionally, a bond goes from the higher end of the investment grade scale into either the low end of the non-investment grade scale or even into default. Whether this will be reflected in the Mergent database depends on if all of the agencies issue new ratings in response to the change in the company's fortunes. For example, following a default Moody's may reduce the rating to a D while S&P and Fitch may just stop following the issue. In this case, the database shows the change to the Moody's rating but does not indicate that any change occurred in the S&P or Fitch rating. Whatever the cause, one suspects that such bonds are not really "orphans" and are recognized by all as being either non-investment grade or distressed and are thus dropped from the orphan analysis.

Hypothesis 3: *Orphan bonds should see the greatest degree of negative serial correlation in returns and the greatest degree of price overshooting.*

For fund managers orphan bonds present a unique problem. These bonds leave the investment grade category under the Barclays scoring rule but not that of the BofAML rule. Since investment grade managers use the former and non-investment grade managers the latter, it is not clear what funds will find these issues conformable with their mandate. An advantage of these cases is that they present a unique test of whether institutional rigidities affect bond returns. For these bonds, investment grade funds have an incentive to sell, while their natural counterparties (non-investment grade funds) have little incentive to buy.

Figure 3 displays the CARs for 120 trading days before and after bonds that are orphaned, defined as date 0. The CARs are relative to the announcement date. Blue dots represent CARs that are not significantly different from zero at the 10% level and orange dots CARs that are. In the 60 days prior to a bond being orphaned it loses nearly 500 bps. This is substantially more than the pre rating change losses seen for any of the other downgrades in Table 7. After the downgrade, these bonds lose approximately another 750 bps until about day 45. After that there appears to be an approximately 250 bps recovery until about day 60 after which returns appear to stabilize.

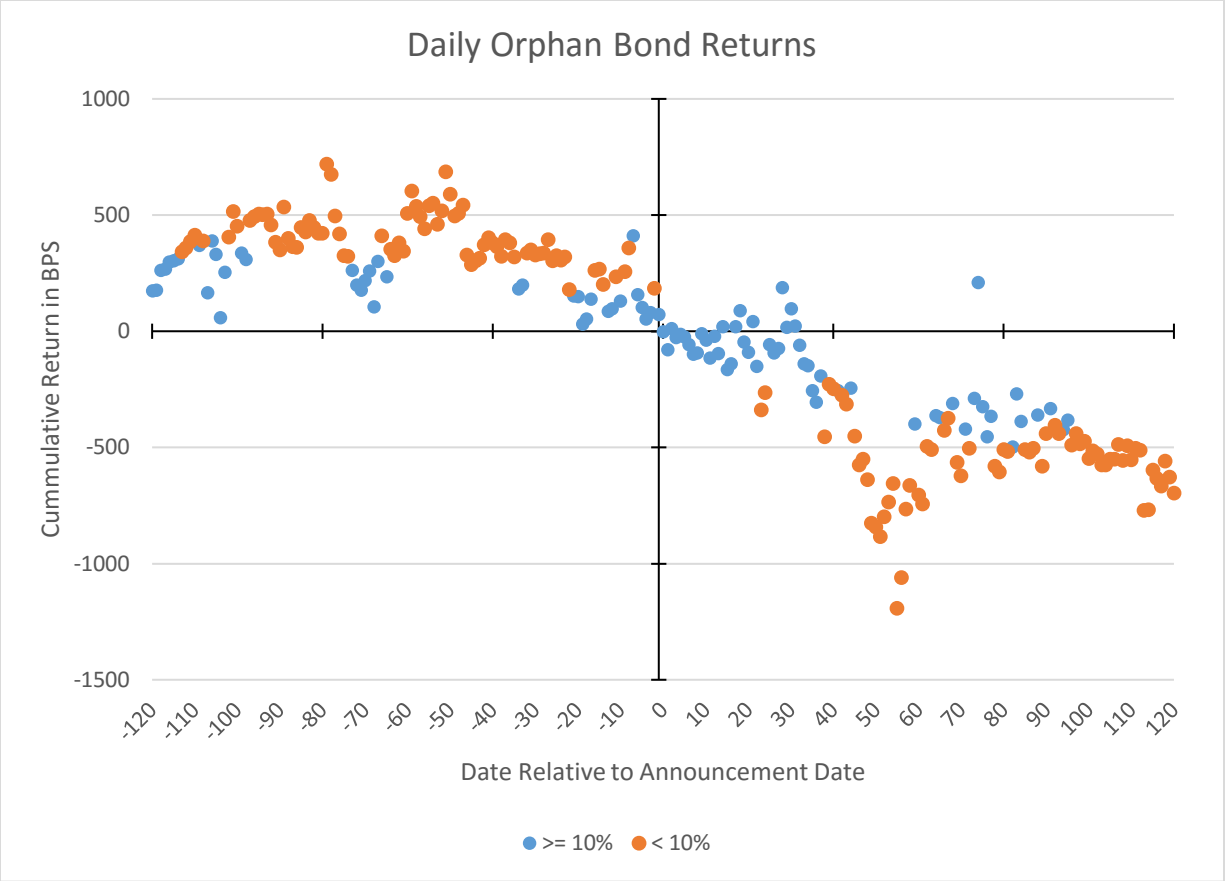


Figure 3: Orphan Bond CARs. Colors represent significance at the 10% level relative to 0.

The post announcement return pattern in Figure 3 is consistent with the institutional rigidity argument. These bonds see a substantial drop in their value while in Barclays investment grade categorization. Once the downgrade goes through the investment grade funds holding the issue are incented to sell it. However, these bonds are still rated investment grade under the BofAML scoring system. That reduces the incentive non-investment grade funds would have to buy the issue. While the investment grade funds may want to eventually sell out of their position, their prospectuses do not typically require that they do so immediately. In a case like this, they may find it optimal to try and hold out as long as they can. Of course, once the price drops far enough it becomes sufficiently attractive that fund managers become willing owners despite the restrictions imposed upon them by their

prospectuses. In a microstructure model like Keim and Madhavan (1996) the result would be long term selling pressure with a recovery. This would result in a pattern like the one in Figure 3.

VII. Changes over Time

In this section we examine how the reaction to bond downgrades around the investment grade/non-investment grade boundary has changed over time given the institutional changes in the bonds markets after the financial crisis of 2007-2008. One aspect of that change has been the change in bond market ratings agencies optimism. Cornaggia, Cornaggia and Hund (2015) argue and present evidence that corporate bonds have optimism in their ratings. Skreta and Veldkamp (2009) argue that this optimism is a result of the pay-by-issuer model.

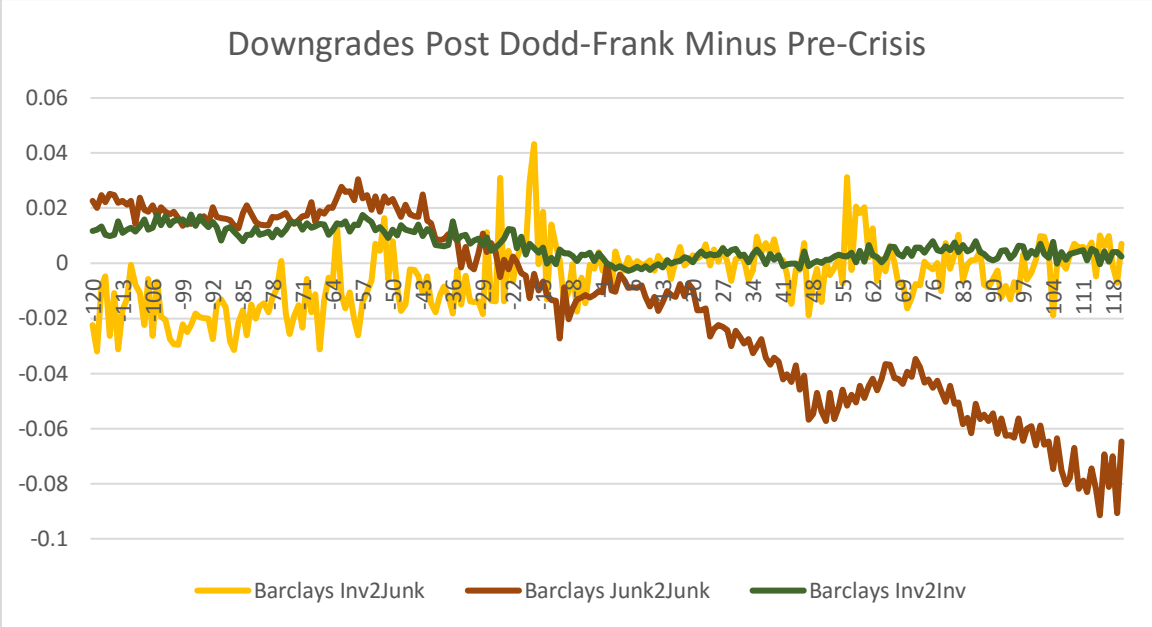


Figure 4: Downgrades post Dodd-Frank minus Pre-crisis

Figure 4, which shows the differences in abnormal returns to downgrades before and after the financial crisis and the passage of the Dodd-Frank law, is consistent with the conjecture that the optimism shown

by ratings agencies has changed since the financial crisis. The figure also suggests that the changes have primarily occurred for the downgrades to high-yield bonds that were already in the high-yield market segment. The downgrades in which a bond still remains an investment grade bond or in which a bond moves from investment grade to non-investment grade have not materially changed. Panel A of Table 3 in our paper also shows that the pattern of downgrades to upgrades changes significantly after the financial crisis. That is, a striking difference exists in the ratio of downgrades relative to upgrades, suggesting again that the rating agencies have changed their issuance of the most optimistic forecasts.

VIII. Conclusion

Many bond funds restrict their holdings to either investment or non-investment grade. When a bond crosses from one category to another, this self-imposed institutional rigidity induces current investors to sell the issue. At the same time, the bond moves into the investment opportunity set for other funds which find that if they so choose, they can now add the bond to their portfolio because it has the moved into their restricted category. Market microstructure models indicate that in an illiquid market with relatively anxious sellers or buyers, security returns will exhibit serial correlation (while the positions are worked off or acquired) followed by a period when prices partially revert. This is the hypothesis tested in this paper.

Given the illiquidity in the corporate bond market, to test this hypothesis we adapt the repeat sales models from the real estate literature in order to estimate abnormal bond returns that arise from institutional rigidities in the market. That model creates a general index about which housing returns vary. Implicitly, this gives each house a factor loading of 1. For bonds, that may be a problematic assumption. To accommodate the factors that may impact benchmark loadings this paper employs a modified repeat sales index. The modification weighs each observation's by the inverse of its distance from the target bond. For each bond this forms a unique distance weighted repeat sales index to use at

its benchmark. Once the vector of abnormal returns are generated around an event, CARs can be estimated and tested for significance using standard regression techniques.

Overall, the empirical results support the institutional rigidity hypothesis in that rating changes that push a bond from one rating category to another lead to return patterns consistent with microstructure theory. Downgrades that cross a bond from investment to noninvestment lead to negative CARs from the announcement day and for a few days afterward. This period is then followed by a partial price rebound. For bonds that see rating changes that leave them in their overall investment or non-investment grade category there is some indication that the bond market's general lack of liquidity leads to prices that overshoot their long run equilibrium value in the non-investment grade market. But it is far more muted than the case where the change drops the security from the investment to noninvestment category. Upgrades show similar patterns to downgrades, but to lesser degree and with returns in the opposite direction.

A unique test of the how institutional rigidities impact bond returns can be seen via an examination of orphan bonds. These bonds undergo rating changes that drop them from investment grade under Barclays scoring rule but not under the BofAML rule. Since investment grade portfolios are typically benchmarked against a Barclays index and non-investment grade portfolios against the BofAML index, there is no natural investment pool for these issues. The results are seen clearly in the data. When a bond becomes an orphan its value drops in the weeks following the rating change and then recovers somewhat. Overall, the loss appears to total close to 5%.

The somewhat artificial designation of a bond as investment or non-investment grade is used by many institutional investors to restrict their holdings. This is obviously done for the convenience of investors and their regulators. But, it also creates an institutional rigidity. When a bond crosses the investment-noninvestment barrier it must also transition from one set of bond funds to another. The

evidence in this paper indicates that this transfer does not take place smoothly. Instead, bond returns exhibit persistent returns for days after the ratings change that are then partially reversed in the weeks that follow. Obviously, this would not happen if the bond market was sufficiently liquid. But it is not. Most bonds trade less than once a month, which makes these transfers potentially very difficult to pull off.

IX. Bibliography

Ali, Ashiq, Kelsey Wei and Yibin Zhou, 2011, "Insider Trading and Option Grant Timing in Response to Fire Sales (and Purchases) of Stocks by Mutual Funds," *Journal of Accounting Research*, 49(3), 595-632.

Alti, Aydogan, Ron Kaniel and Uzi Yoeli, 2012, "Why Do Institutional Investors Chase Return Trends?," *Journal of Financial Intermediation*, 21(4), 694-721.

Altman, Edward and Heather Suggitt, 2000, "Default rates in the syndicated bank loan market: A mortality analysis," *Journal of Banking and Finance*, 24, 229-253.

Badrinath, S.G. and Sunil Wahal, 2002, "Momentum Trading by Institutions," *Journal of Finance*, 57, 2449-2478.

Becker, Bo and Victoria Ivashina, 2015, "Reaching for Yield in the Bond Market," *Journal of Finance*, 70(5), 1863-1902.

Bessembinder, Hendrik, William Maxwell and Kumar Venkataraman, 2006, "Market transparency, liquidity externalities, and institutional trading costs in corporate bonds," *Journal of Financial Economics* 82, 251–288.

Bongaerts, Dion, Martijn Cremers, and William Goetzmann, 2012, "Tiebreaker: Certification and Multiple Credit Ratings," *Journal of Finance* 67, 113-152.

Chen, Long, David Lesmond and Jason Wei, 2007, "Corporate Yield Spreads and Bond Liquidity," *Journal of Finance*, 19-149.

Chen, Zhihua, Aziz Lookman, Norman Schürhoff and Duane Seppi, 2014, "Rating-Based Investment Practices and Bond Market Segmentation," *Review of Asset Pricing Studies* 4, 162-205.

Chernenko, Sergey, and Adi Sunderam, 2012, "The Real Consequences of Market Segmentation," *Review of Financial Studies* 25, 2041-2069.

Chordia, Tarun, Richard Roll and Avanidhar Subrahmanyam, 2001, "Market Liquidity and Trading Activity," *Journal of Finance*, 56(2), 501-530.

Cornaggia, Jess, Kim Cornaggia, and John Hund, 2015, "Credit Ratings across Asset Classes: A Long-term Perspective," Working paper, Georgetown University.

Coval, Joshua and Erik Stafford, 2007, "Asset Fire Sales (and Purchases) in Equity Markets," *Journal of Financial Economics*, 86(2), 479-512.

Dyakov, Teodor and Marno Verbeek, 2013, "Front-Running of Mutual Fund Fire-Sales," *Journal of Banking and Finance*, 37(12), 4931-4942.

Ellul, Andrew, Chotibhak Jotikasthira, and Christian Lundblad, 2011, "Regulatory Pressure and Fire Sales in the Corporate Bond Market," *Journal of Financial Economics* 101, 596-620.

Emery, Kenneth, Sharon Ou, Jennifer Tennant, Frank Kim and Richard Cantor, 2008, "Corporate Default and Recovery Rates, 1920-2007," Moody's Investor Services, https://www.moodys.com/sites/products/DefaultResearch/2007000000474979.pdf.

Francke, Marc, 2010, "Repeat Sales Index for Thin Markets: A Structural Time Series Approach," *Journal of Real Estate Finance and Economics*, 41(1), 24-52.

Geltner, David and William Goetzmann, 2000, "Two Decades of Commercial Property Returns: A Repeated-Measures Regression-Based Version of the NCREIF Index," *Journal of Real Estate Economics and Finance*, 21(1), 5-21.

Goetzmann, William, 1992, "The Accuracy of Real Estate Indices: Repeat Sale Estimators," *Journal of Real Estate Finance and Economics*, 5(1), 5-53.

Grinblatt, Mark, Sheridan Titman and Russ Wermers, 1995, "Momentum investment strategies, portfolio performance, and herding: a study of mutual fund behavior," *American Economic Review*, 85, 1088-1105.

Han, Song and Zhou, Hao, Effects of Liquidity on the Nondefault Component of Corporate Yield Spreads: Evidence from Intraday Transactions Data (March 2008). Available at SSRN: http://ssrn.com/abstract=946802 or http://dx.doi.org/10.2139/ssrn.946802.

Hameed, Allaudeen, Wenjin Kang and S. Vishwanathan, 2010, "Stock Market Declines and Liquidity," *Journal of Finance*, 65(1), 257-293.

Hite, Gailen and Arthur Warga, 1997, "The Effect of Bond-Rating Changes on Bond Price Performance," *Financial Analysts Journal*, 53(3), 35-51.

Investopedia, Trade Reporting and Compliance Engine – TRACE, http://www.investopedia.com/terms/t/trace.asp.

Kalimipalli, Madhu and Subhankar Nyak, 2012, "Idiosyncratic Volatility vs. Liquidity? Evidence from the US Corporate Bond Market," *Journal of Financial Intermediation*, 21, 217-242.

Keim, Donald and Ananth Madhavan, 1996, "The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects," *Review of Financial Studies*, 9(1), 1-36.

Madhavan, Ananth, 2000, "Market Microstructure: A Survey," *Journal of Financial Markets*, 3(3), 205-258.

May, Anthony, 2010, "The Impact of Bond Rating Changes on Corporate Bond Prices: New Evidence from the Over-the-Counter Market," *Journal of Banking and Finance*, 34, 2822-2836.

Meese, Richard and Nancy Wallace, 1991, "Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices," *American Real Estate and Urban Economics Association Journal*, 19(3), 308-332.

Merrill, Craig, Taylor Nadauld and Philip Strahan, 2015, "Final Demand for Structured Finance Securities," working paper Brigham Young University.

Peng, Liang, 2012, "Repeat Sales Regression on Heterogeneous Properties," *Journal of Real Estate Finance and Economics*, 45(3), 804-827.

Schultz, Paul, 2001, "Corporate Bond Trading Costs: A Peek Behind the Curtain," *Journal of Finance* 56, 677–698.

Schultz, Paul, and Sophie Shive, 2016, "Mutual Funds and Bond Market Liquidity," Working Paper, University of Notre Dame.

Securities Industry and Financial Markets Association, Updated 01/19/16, US Corporate Bond Issuance, http://www.sifma.org/research/statistics.aspx.

Skreta, Vasliki, and Laura Veldkamp, 2009, "Ratings Shopping and Asset Complexity: A Theory of Ratings Inflation, *Journal of Monetary Economics* 56, 678-695.

Wansley, James, John Glascock and Terence Clauretie, 1992, "Institutional Bond Pricing and Information Arrival: The Case of Bond Rating Changes," *Journal of Business Finance and Accounting*, 19(5), 733-750.

Warga, Arthur and Ivo Welch, 1993, "Bondholder Losses in Leveraged Buyouts," *Review of Financial Studies*, 6(4), 959-982.

Webb, Cary, 1988, "A Probabilistic Model for Price Levels in Discontinuous Markets," *Measurement in Economics*, ed. By W. Eichhorn, 137-156.

Webb, Cary, 1991, "The Expected Accuracy of a Price Index for Discontinuous Markets," working paper Department of Mathematics, Chicago State University, Chicago IL 60628.

Weinstein, Mark, 1977, "The Effect of a Rating Change Announcement on Bond Price," *Journal of Financial Economics*, 5, 329-350.

Wermers, Russ, 1999, "Mutual Fund Herding and the Impact on Stock Prices," *Journal of Finance*, 54, 581-622.

Table 1: Ratings scale for calculating composite

Scoring system used by BofA Merrill Lynch to determine the index a bond belongs to. Original source: BofA Merrill Lynch Bond Indices; Bloomberg

Numeric	Composite	Moody's	S&P	Fitch
1	AAA	Aaa	AAA	AAA
2	AA1	Aa1	AA+	AA+
3	AA2	Aa2	AA	AA
4	AA3	Aa3	AA-	AA-
5	A1	A1	A+	A+
6	A2	A2	A	A
7	A3	A3	A-	A-
8	BBB1	Baa1	BBB+	BBB+
9	BBB2	Baa2	BBB	BBB
10	BBB3	Baa3	BBB-	BBB-
11	BB1	Ba1	BB+	BB+
12	BB2	Ba2	BB	BB
13	BB3	Ba3	BB-	BB-
14	B1	B1	B+	B+
15	B2	B2	B	B
16	B3	B3	B-	B-
17	CCC1	Caa1	CCC+	CCC+
18	CCC2	Caa2	CCC	CCC
19	CCC3	Caa3	CCC-	CCC-
20	CC	Ca	CC	CC
21	C	C	C	C
22	D	D	DDD-D	

Table 2: Fraction of Days Traded by Year by Bond

Fraction of days traded (FDT) during a one year time period over which a bond trade is observed. The numerator in FDT equals the number of trades in bond *B* during year *Y*. The denominator is the number of trading days in year *Y* between the first and last trade date across all years in bond *B*. Panel A includes a bond in year *Y* if: (1) there is at least one trade in year *Y* or prior to it and (2) at least one trade in year *Y* or after it. Note, a single trade in year *Y* will lead to a bond's inclusion in year *Y*. Also bonds with trade on dates prior to and after but not in year *Y* are included in the year *Y* data. Example: A bond trades in 2005 and 2007 but not 2006. Fraction equals 0/total trade days in 2006 = 0. A bond trades on July 11, 2006 and July 12, 2006 but never before or after. FDT equals 2/2 = 1, since total trade days during 2006 between the first and last trade date in the bond is 2. In Panel B, the same exercise is carried out. But this time Year refers to the number of years since the bond is first observed to trade t_0. Time from t_0 to its first anniversary is labeled 1. The denominator follows the rule for Panel A, with the reference year being the number of years from t_0, again with the first year being from t_0 to its anniversary. Example: A bond first trades on July 11, 2006. It trades again on August 8, 2007 and September 16, 2008. FDT equals 1/trading days between July 12, 2007 and July 11, 2008. If the bond traded on August 8, 2007 but never again traded then FDT equals 1/trading days between July 12, 2007 and August 8, 2007. Row 11+, averages across all years greater than or equal to 11. Displayed is the FDT on a percentage basis (i.e. 100×FDT).

	Percentile						
Year	**5%**	**10%**	**25%**	**50%**	**75%**	**90%**	**95%**
Panel A: FDT by Bond by Calendar Year							
2003	0.85	1.47	3.44	8.52	21.53	38.08	53.59
2004	0.79	1.61	3.92	10.26	22.97	40.50	51.56
2005	0.49	1.19	2.78	7.82	19.05	36.51	47.60
2006	0.40	0.80	2.39	7.94	19.56	36.59	48.21
2007	0.00	0.40	1.98	6.35	17.86	34.61	46.41
2008	0.00	0.40	1.19	4.35	13.44	31.38	45.57
2009	0.00	0.40	1.59	5.16	15.74	32.54	44.34
2010	0.00	0.40	1.19	4.74	15.42	32.00	42.99
2011	0.00	0.40	1.59	5.16	14.68	29.69	40.48
2012	0.00	0.00	1.18	3.70	12.09	27.06	39.92
2013	0.00	0.38	1.52	4.94	14.07	29.28	42.86
2014	0.38	0.77	1.92	5.75	15.33	31.03	42.91
Panel B: FDT by Bond by Year's Since Initial Trade Date							
1	0.40	1.14	3.16	8.59	19.92	37.50	49.80
2	0.38	0.75	1.95	5.40	14.68	30.62	42.44
3	0.00	0.40	1.20	4.11	12.26	26.63	38.94
4	0.00	0.39	1.13	3.12	9.84	23.51	36.11
5	0.00	0.39	0.79	2.77	8.56	20.43	30.56
6	0.00	0.00	0.76	1.92	5.84	16.03	25.18
7	0.00	0.00	0.40	1.53	4.37	12.99	20.76
8	0.00	0.00	0.39	1.18	3.77	11.40	19.32
9	0.00	0.00	0.39	1.18	3.82	9.58	19.63
10	0.00	0.38	0.39	1.52	3.61	9.38	17.29
11+	0.00	0.38	0.55	1.16	3.05	8.78	16.98

Table 3: Changes between Investment and Non-investment grade over Time

Panel A of this table reports for each year the total number of bonds crossing the investment grade and investment grade category using the Bank of America Merrill Lynch (BofAML) or Barclays rule for their indices. Panel B reports the total number of bonds in years since initial rating is defined so that year 1 includes any transition from investment to investment or the reverse occurring within 1 year of the initial rating date.

	Downgrades		Upgrades		Ratio of Downgrades to Upgrades	
Year	BofAML	Barclays	BofAML	Barclays	BofAML	Barclays
Panel A: Calendar Year						
2003	292	312	76	80	3.84	3.90
2004	209	215	135	138	1.55	1.56
2005	1700	2245	216	233	7.87	9.64
2006	688	285	179	155	3.84	1.84
2007	361	433	161	144	2.24	3.01
2008	1604	1572	264	263	6.08	5.98
2009	2361	2467	111	102	21.27	24.19
2010	253	264	137	127	1.85	2.08
2011	152	183	180	176	0.84	1.04
2012	160	157	199	201	0.80	0.78
2013	145	124	352	341	0.41	0.36
2014	125	127	127	101	0.98	1.26

Panel B: Years Since Initial Rating				
	Downgrades		Upgrades	
Year	BofAML	Barclays	BofAML	Barclays
1	1237	1465	266	231
2	1618	1583	312	307
3	1236	1206	257	247
4	864	863	228	225
5	812	874	167	151
6	836	844	136	138
7	371	392	145	139
8	202	233	92	90
9	242	260	101	86
10	221	223	84	94
11+	411	441	349	353

Table 4: Fraction of Days with Trading Around Classification Rating Changes

This table shows the fraction of days in the months or days a bond trades around a rating change to or from investment and investment grade. For months, other than 0, it is the average across bonds of the number of trading days on which the bond traded divided by the number of trading days. For month 0 it is just the day of the rating change. For days it is the fraction of bonds with a trade on that day. Bonds are excluded from a period if there are no trades recorded for them both before and after the period in question. Values are in percentage terms. Standard errors are in square brackets.

Time	Investment Grade to Non-investment grade				Non-investment grade to Investment Grade			
Units	Months		Days		Months		Days	
	BofAML	Barclays	BofAML	Barclays	BofAML	Barclays	BofAML	Barclays
-6	20.87	20.58	11.34	9.33	6.59	6.23	15.80	15.07
	[2.02]	[2.02]	[2.09]	[2.08]	[2.48]	[2.50]	[2.43]	[2.45]
-5	21.34	20.59	10.47	9.67	7.31	6.88	15.09	15.31
	[2.02]	[2.02]	[2.09]	[2.08]	[2.46]	[2.48]	[2.43]	[2.45]
-4	20.44	19.99	10.12	8.81	7.45	6.86	14.39	14.83
	[2.02]	[2.01]	[2.09]	[2.08]	[2.48]	[2.50]	[2.43]	[2.45]
-3	20.48	19.91	11.52	11.57	7.96	7.89	16.27	14.59
	[2.00]	[1.99]	[2.09]	[2.08]	[2.47]	[2.48]	[2.43]	[2.45]
-2	21.89	21.10	13.96	12.44	8.63	8.63	13.92	13.16
	[2.06]	[2.04]	[2.09]	[2.08]	[2.45]	[2.47]	[2.43]	[2.45]
-1	24.34	25.01	14.14	12.78	13.12	12.62	13.21	12.68
	[2.05]	[2.03]	[2.09]	[2.08]	[2.40]	[2.44]	[2.43]	[2.45]
0	21.82	19.00	21.82	19.00	24.53	22.25	24.53	22.25
	[2.09]	[2.08]	[2.09]	[2.08]	[2.43]	[2.45]	[2.43]	[2.45]
1	11.62	10.48	32.29	30.40	15.10	14.87	20.28	20.81
	[2.07]	[2.06]	[2.09]	[2.08]	[2.38]	[2.40]	[2.43]	[2.45]
2	9.66	8.99	27.57	26.60	14.41	13.82	15.33	14.35
	[2.08]	[2.07]	[2.09]	[2.08]	[2.39]	[2.41]	[2.43]	[2.45]
3	7.89	7.55	24.61	24.87	15.80	15.65	15.09	14.11
	[2.10]	[2.09]	[2.09]	[2.08]	[2.39]	[2.41]	[2.43]	[2.45]
4	8.26	8.35	25.13	26.08	15.21	15.15	9.91	8.13
	[2.12]	[2.10]	[2.09]	[2.08]	[2.39]	[2.41]	[2.43]	[2.45]
5	7.79	7.35	21.29	21.07	16.49	16.44	11.56	11.24
	[2.12]	[2.10]	[2.09]	[2.08]	[2.40]	[2.42]	[2.43]	[2.45]
6	8.21	7.80	22.51	24.87	16.54	16.50	8.96	9.33
	[2.11]	[2.10]	[2.09]	[2.08]	[2.42]	[2.45]	[2.43]	[2.45]

Table 5: Orphans by Year

This table shows the number of orphan bonds in the sample each year. An orphan bond is defined as one with a BofAML score of less than or equal to 10 and a Barclays score of greater than or equal to 11. Bonds are only included if the absolute difference between the two scores is less than or equal to the value in Abs(Diff). The upgrades are bonds that having a rating change that causes them to go from BofAML investment grade group to its investment grade group, while remaining in the Barclays investment grade group. Downgrades are bonds that have a rating change that moves them from the Barclays investment grade category to Barclays investment grade category while remaining in the BofAML investment grade category.

	Upgrades		Downgrades	
Abs(Diff)	1	2	1	2
2003	8	16	37	47
2004	14	25	68	71
2005	24	38	155	769
2006	42	52	36	51
2007	31	35	52	60
2008	32	38	103	124
2009	13	17	366	456
2010	30	33	72	81
2011	37	41	28	37
2012	17	24	35	39
2013	32	41	8	10
2014	31	33	36	36

Table 6: Day 0 to +10 Cumulative Abnormal Returns for Bonds Experiencing a Rating Change

This table shows daily bond returns following a rating change. Investment grade ("Inv.") (defined as having a score between 1 and 10) and noninvestment grade ("Junk") (a score of 11 to 16) are based on the rating aggregation method used by either BofAML or Barclays ("Barc"). Bonds with a rating of 17+ are defined as distressed. Column headers indicate downgrades or upgrades for bonds from one rating group to another. Columns "Inv. To Inv." Indicate a rating change for an investment grade bond that remains investment grade after the rating change. Columns "Junk to Junk" indicate the same for bonds initially rated non-investment grade. Finally, "Inv to Junk" and "Junk to Inv." Indicate bonds that switch class after a rating change. Displayed returns by day are the benchmark adjusted cumulative abnormal returns (CAR) from the end of day 0 to the end of day 10 in basis points. Return estimates are based on the mean bootstrapped values repeated 500 times. Day 0 is the date on which a bond's numerical rating score changes. Days are measured in trading, not calendar, days. In the row *X* to *Y* the numbers represent borders. It should be read as pre rating change score greater than or equal to *X* and post rating score less than or equal to *Y*. Bootstrapped *p*-values against a null of 0 is below in square brackets. Key: ***=1%, **=5% and *=10%.

	Downgrades						**Upgrades**					
	Inv. to Junk		**Junk to Junk**		**Inv. to Inv.**		**Junk to Inv,**		**Junk to Junk**		**Inv. to Inv.**	
Day	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**
0	-194.938	-157.632	-15.616	-35.410	-12.379	-12.505	13.635	19.435	13.749	10.321	15.418	12.500
	[0.00]***	[0.00]***	[24.60]	[7.20]*	[11.60]	[13.60]	[29.00]	[22.20]	[7.20]*	[11.20]	[26.40]	[32.40]
1	-184.572	-175.765	-29.816	-40.633	-11.806	-15.474	23.375	30.903	21.719	23.613	-3.319	-6.350
	[0.00]***	[0.00]***	[15.60]	[4.80]**	[16.60]	[12.80]	[24.40]	[16.60]	[1.80]**	[2.20]**	[45.40]	[44.80]
2	-197.277	-218.186	-40.715	-47.246	-17.060	-22.713	29.976	43.337	17.801	25.808	10.087	7.015
	[0.00]***	[0.00]***	[10.20]	[2.60]**	[12.00]	[7.20]*	[20.00]	[8.00]*	[9.40]*	[4.00]**	[31.40]	[36.40]
3	-209.022	-252.310	-55.861	-46.606	-23.158	-26.848	30.626	44.680	18.373	21.743	23.016	22.286
	[0.20]***	[0.00]***	[4.40]**	[4.20]**	[5.00]**	[5.80]*	[16.20]	[7.60]*	[8.00]*	[6.80]*	[15.40]	[23.00]
4	-207.288	-246.507	-66.675	-45.971	-28.594	-16.974	28.715	64.707	8.509	27.182	3.649	-12.017
	[0.00]***	[0.00]***	[3.20]**	[4.60]**	[2.40]**	[16.00]	[20.00]	[5.40]*	[25.20]	[2.80]**	[43.20]	[38.60]
5	-227.234	-256.209	-60.003	-39.044	-34.688	-33.761	31.607	114.154	10.238	25.764	-15.235	-19.551
	[0.20]***	[0.00]***	[7.60]*	[9.40]*	[2.00]**	[2.20]**	[20.60]	[16.20]	[17.00]	[1.60]**	[28.80]	[28.40]
6	-192.621	-235.745	-74.061	-57.773	-33.781	-34.023	32.410	71.889	4.271	24.746	14.018	11.550
	[0.60]***	[0.00]***	[2.60]**	[5.00]**	[2.40]**	[2.00]**	[16.80]	[3.00]**	[34.40]	[2.40]**	[28.60]	[34.00]
7	-182.391	-203.713	-73.559	-51.926	-31.870	-29.754	-12.395	19.601	17.687	41.681	3.567	-6.695
	[0.40]***	[0.00]***	[3.40]**	[7.00]*	[3.80]**	[4.60]**	[38.40]	[34.00]	[9.20]*	[0.00]***	[43.80]	[45.40]
8	-263.868	-273.247	-52.757	-39.172	-30.965	-29.555	43.390	62.157	22.706	34.847	-4.617	-12.207
	[0.00]***	[0.00]***	[12.00]	[14.40]	[5.20]*	[5.80]*	[14.80]	[7.80]*	[6.00]*	[1.20]**	[44.20]	[38.40]
9	-241.856	-252.054	-58.927	-40.121	-33.884	-30.324	35.633	67.880	11.264	20.671	-14.880	-30.196
	[0.00]***	[0.00]***	[8.60]*	[15.00]	[6.20]*	[10.00]*	[13.80]	[2.00]**	[20.80]	[9.40]*	[32.80]	[24.00]
10	-234.680	-233.550	-50.261	-9.872	-18.867	-16.836	53.064	62.461	15.408	23.480	4.215	-0.184
	[0.00]***	[0.00]***	[12.80]	[40.00]	[15.20]	[18.40]	[6.60]*	[4.80]**	[13.00]	[6.00]*	[43.00]	[50.60]

Table 7: Weekly Abnormal Returns prior to a Rating Change

This table shows weekly bond returns preceding a rating change. Investment grade ("Inv.") (defined as having a score between 1 and 10) and noninvestment grade ("Junk") (a score of 11 to 16) are based on the rating aggregation method used by either BofAML or Barclays ("Barc"). Bonds with a rating of 17+ are defined as distressed. Column headers indicate downgrades or upgrades for bonds from one rating group to another. Columns "Inv. To Inv." Indicate a rating change for an investment grade bond that remains investment grade after the rating change. Columns "Junk to Junk" indicate the same for bonds initially rated non-investment grade. Finally, "Inv to Junk" and "Junk to Inv." Indicate bonds that switch class after a rating change. Displayed returns by day are the benchmark adjusted cumulative abnormal returns (CAR) from the end of day −42 to the end of day −1 in basis points. Return estimates are based on the mean bootstrapped value repeated 500 times. Day 0 is the date on which a bond's numerical rating score changes. Days are measured in trading, not calendar, days. In the row *X* to *Y* the numbers represent borders. It should be read as pre rating change score greater than or equal to *X* and post rating score less than or equal to *Y*. Bootstrapped *p*-values against a null of 0 is below in square brackets. Key: ***=1%, **=5% and *=10%.

	Downgrades						**Upgrades**					
	Inv. to Junk		**Junk to Junk**		**Inv. to Inv.**		**Junk to Inv,**		**Junk to Junk**		**Inv. to Inv.**	
Day	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**
-41	38.067	41.387	-15.095	-11.543	2.923	7.749	9.990	6.000	23.433	23.361	1.951	-5.590
	[21.00]	[10.20]	[18.40]	[25.00]	[43.00]	[28.80]	[29.80]	[35.60]	[0.00]***	[0.20]***	[46.20]	[39.80]
-36	-45.125	-13.052	1.620	-4.642	1.518	-13.058	42.290	12.392	40.361	44.360	-4.217	-10.094
	[14.00]	[39.00]	[45.00]	[41.80]	[47.20]	[28.40]	[5.00]**	[34.00]	[0.20]***	[0.00]***	[44.20]	[38.20]
-31	-113.882	-2.892	-39.855	-78.027	-40.525	-64.014	57.983	42.151	31.631	51.038	24.269	24.432
	[6.20]*	[47.80]	[20.40]	[4.80]**	[3.40]**	[0.40]***	[3.00]**	[9.60]*	[2.60]**	[0.20]***	[15.80]	[19.60]
-26	-160.872	-115.271	-1.163	-6.020	-78.660	-82.646	49.378	42.054	44.186	67.468	32.157	21.517
	[3.00]**	[4.00]**	[48.20]	[42.60]	[0.00]***	[0.00]***	[4.20]**	[4.80]**	[1.20]**	[0.00]***	[12.40]	[26.00]
-21	-174.456	-143.735	-48.537	-34.210	-66.138	-67.399	65.019	38.481	82.121	86.395	20.581	23.414
	[0.60]***	[2.00]**	[13.80]	[23.60]	[0.20]***	[0.60]***	[3.40]**	[13.00]	[0.00]***	[0.00]***	[22.80]	[22.60]
-16	-103.037	-75.971	-70.308	-6.622	-99.753	-106.587	44.957	-12.862	83.488	87.734	43.294	41.156
	[9.20]*	[17.80]	[8.80]*	[44.20]	[0.00]***	[0.00]***	[10.20]	[37.80]	[0.00]***	[0.00]***	[5.80]*	[8.40]*
-11	-105.845	-136.582	-126.372	-26.323	-91.627	-104.333	72.646	43.581	86.458	80.015	0.790	-1.742
	[11.60]	[6.60]*	[0.80]***	[31.80]	[0.00]***	[0.00]***	[0.80]***	[6.00]*	[0.00]***	[0.20]***	[48.80]	[46.00]
-6	-338.419	-366.826	-130.892	-39.717	-91.634	-104.603	77.580	40.072	126.386	130.624	-5.862	-6.342
	[0.00]***	[0.00]***	[0.80]***	[22.20]	[0.00]***	[0.00]***	[0.60]***	[9.00]*	[0.00]***	[0.00]***	[40.00]	[39.40]
-1	-168.812	-158.603	-179.588	-69.953	-120.926	-121.672	77.107	61.394	116.214	112.005	17.036	19.316
	[5.60]*	[4.40]**	[0.00]***	[12.60]	[0.00]***	[0.00]***	[2.60]**	[5.20]*	[0.00]***	[0.00]***	[34.00]	[38.40]

Table 8: Weekly Cumulative Abnormal Returns following a Rating Change

This table shows weekly bond returns following a rating change. Investment grade ("Inv.") (a score between 1 and 10) and noninvestment grade ("Junk") (a score of 11 to 16) are based on the rating aggregation method used by either BofAML or Barclays ("Barc"). Bonds with a rating of 17+ are defined as distressed. Column headers indicate downgrades or upgrades for bonds from one rating group to another. Columns "Inv. To Inv." Indicate a rating change for an investment grade bond that remains investment grade after the rating change. Columns "Junk to Junk" indicate the same for bonds initially rated non-investment grade. Finally, "Inv to Junk" and "Junk to Inv." Indicate bonds that switch class post rating change. Displayed returns are the benchmark adjusted cumulative abnormal returns (CAR) from the end of day 10 to the end of day 51 in basis points. Return estimates are based on the mean bootstrapped value repeated 500 times. Day 0 is the date on which a bond's numerical rating score changes. Days are measured in trading, not calendar, days. In the row *X* to *Y* the numbers represent borders. It should be read as pre rating change score greater than or equal to *X* and post rating score less than or equal to *Y*. Bootstrapped *p*-values against a null of 0 is below in square brackets. Key: ***=1%, **=5% and *=10%.

	Downgrades						**Upgrades**					
	Inv. to Junk		**Junk to Junk**		**Inv. to Inv.**		**Junk to Inv,**		**Junk to Junk**		**Inv. to Inv.**	
Day	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**	**BofAML**	**Barc**
11	52.521	54.073	-40.873	-16.187	-11.624	-16.490	-24.123	-2.610	-2.347	-7.305	9.069	15.499
	[1.00]***	[0.60]***	[0.40]***	[24.40]	[12.20]	[2.80]**	[16.00]	[46.60]	[39.00]	[24.20]	[34.60]	[22.20]
16	118.016	152.258	-43.811	-47.570	-1.567	-7.267	39.926	52.912	19.759	19.370	23.805	34.790
	[1.60]**	[0.00]***	[5.60]*	[16.00]	[46.20]	[30.20]	[6.80]*	[2.60]**	[5.40]*	[5.20]*	[29.20]	[23.80]
21	58.562	149.228	-39.756	-84.984	-11.059	-11.283	36.116	29.944	16.209	10.861	1.779	-3.287
	[17.20]	[1.20]**	[11.20]	[4.80]**	[22.20]	[22.40]	[15.80]	[14.20]	[14.20]	[24.40]	[45.80]	[45.00]
26	99.791	241.368	-15.553	-99.237	-7.807	-15.778	61.086	78.571	24.868	10.758	-2.114	-18.408
	[11.00]	[0.20]***	[35.60]	[8.80]*	[27.20]	[16.40]	[7.00]*	[2.40]**	[8.00]*	[26.40]	[44.60]	[22.80]
31	90.579	213.793	15.051	-32.809	8.938	-7.848	-20.389	-6.412	18.899	9.458	-18.820	-29.257
	[14.00]	[0.00]***	[38.00]	[38.40]	[31.20]	[34.40]	[30.20]	[42.60]	[16.00]	[34.80]	[19.60]	[12.80]
36	145.725	188.416	21.810	-36.189	-8.377	-20.374	31.377	45.082	-0.184	-19.016	-16.215	-11.854
	[7.00]*	[2.20]**	[32.00]	[34.00]	[34.20]	[17.60]	[30.80]	[21.00]	[47.60]	[18.20]	[24.40]	[31.20]
41	118.666	142.396	48.547	-8.893	-2.508	-25.743	11.146	25.435	-17.491	-25.199	1.467	-13.270
	[10.40]	[7.40]*	[16.80]	[49.60]	[44.80]	[14.40]	[42.00]	[30.00]	[22.60]	[17.20]	[46.40]	[35.40]
46	87.101	94.085	94.394	44.718	-19.431	-36.708	-7.093	29.980	-12.081	-19.108	-37.520	-56.980
	[18.40]	[18.20]	[5.20]*	[29.60]	[18.40]	[6.60]*	[46.40]	[31.40]	[31.60]	[22.60]	[13.40]	[6.60]*
51	130.523	78.511	143.570	52.294	-2.255	-12.872	37.780	75.669	-0.755	-10.317	7.367	12.231
	[17.60]	[28.00]	[1.40]**	[26.20]	[47.00]	[30.60]	[18.80]	[6.40]*	[50.60]	[34.20]	[39.20]	[34.40]

Order Flow Segmentation, Liquidity and Price Discovery: The Role of Latency Delays[*]

Michael Brolley[†] David Cimon[‡]
Wilfrid Laurier University Bank of Canada

May 11, 2017

— preliminary draft —

Abstract

Latency delays—known as "speed bumps"—are an intentional slowing of order flow by exchanges. Supporters contend that delays protect market makers from high-frequency arbitrage, while opponents warn that delays promote "quote fading" by market makers. We construct a model of informed trading in a fragmented market, where one market operates a conventional order book, and the other imposes a latency delay on market orders. We show that informed investors migrate to the conventional exchange, widening the quoted spread; the quoted spread narrows at the delayed exchange. The overall market quality impact depends on the nature of the delay: "short" latency delays lead to improved trading costs for liquidity investors, but worsening price discovery; sufficiently "long" delays improve both.

[*]The authors would like to thank Corey Garriott, Andriy Shkilko, and Adrian Walton for valuable discussions. Michael Brolley gratefully acknowledges the financial support from the Social Sciences and Humanities Research Council. The views of the authors are not necessarily those of the Bank of Canada. All errors are our own.

[†]Email: mbrolley@wlu.ca; web: http://www.mikerostructure.com.

[‡]Email: dcimon@bank-banque-canada.ca; web: https://sites.google.com/site/dcimon.

"I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues."

—SEC Chair, Mary Jo White, June 5, 2014

Liquidity suppliers prefer to transact against uninformed traders. These uninformed traders are valuable, as they are unlikely to move the market against market makers. Many exchanges, competing for scarce order flow, have specialized to attract these uninformed liquidity demanders. Inverse pricing, dark trading and retail order segmentation facilities have all been studied as ways in which exchanges try to draw these traders from other markets, in part by advertising their market design as a way to disincentivize informed traders from also participating. Recently, some exchanges have imposed latency delays—so-called "speed bumps"—as yet another way of segmenting away retail order flow. Measured on the order of milliseconds or microseconds, latency delays impose a time delay between an order's receipt at the exchange, and its execution.[1] Exchanges advertise latency delays as means of protecting market makers from adverse selection at the hands of high frequency traders (HFTs) acting on extremely short-horizon information; the savings are then passed on through a narrower spread.[2]

As with any market structure change, latency delays have not been without controversy. Beyond the comments from proponents, who tout improved market quality, other market participants have suggested that delays in order execution create an uneven playing field by allowing market makers to "fade" quotes ahead of large orders.[3] Quote "fading" refers to a market maker's ability to revise their quotes after an order is received, but not yet filled. By fading quotes, market makers execute incoming orders at less favourable prices than at the time of initial submission. Indeed, existing evidence from the academic community suggests

[1] A description of the mechanics behind latency delays is available in the appendix.

[2] For one example see "Regulators Protect High-Frequency Traders, Ignore Investors" in Forbes: `http://www.forbes.com/sites/jaredmeyer/2016/02/23/sec-should-stand-up-for-small-investors/\#1c96d49a1ec6`

[3] For one example see "Canada's New Market Model Conundrum" by Doug Clark at ITG: `http://www.itg.com/marketing/ITG_WP_Clark_Alpah_Conundrum_20150914.pdf`

that, not only do latency delays allow market makers to withdraw liquidity, but they may harm liquidity at other markets, by concentrating retail order flow at a single venue. (Chen, Foley, Goldstein, and Ruf 2016)

To resolve these competing explanations, we construct a static, three-period model of sequential trading. In our model, trading occurs in fragmented markets, where one exchange imposes a latency delay. We model traders who are aware of the possibility of an information event, which occurs with some probability in the second period. This information event can be interpreted in two ways, both as a scheduled event such as an earnings announcement or as a fleeting arbitrage opportunity. In the first case, traders are aware of an announcement at fixed point in the second period, while in the second case, traders are aware that arbitrage opportunities become public knowledge at a fixed point in the second period.[4]

In the first period, traders can choose to submit orders to one of two exchanges. One is a standard exchange, which executes orders immediately in the first period. The second is a latency-delayed exchange, which randomly executes orders either immediately in the first period, or after the information event becomes public in the second period.

To differentiate the effects of the delay on different traders, we model two types of traders, uninformed liquidity traders, and informed speculators. Liquidity traders arrive at the market with a need to trade, and have the choice between either submitting an order immediately, or waiting until after the information event. A liquidity trader who chooses to submit an order before the information event may send the order to either the open exchange, which executes instantly, or the speed bump, which delays the order with some probability. Liquidity traders who delay their order risk paying a form of delay cost, should the market move against them. This delay cost represents the need for these traders to seek additional capital, should prices become worse.

Alternatively, if a speculator arrives, they have the option for paying to become informed before the announcement. Similar to liquidity traders, speculators have the option to either

[4]The latter interpretation is similar in many respects to Budish, Cramton, and Shim (2015), who document the fleeting nature of arbitrage opportunities between New York and Chicago.

execute their order immediately at the non-delayed exchange, or submit their order to the delayed-exchange, and risk having the information event arrive before their order executes.

We relate the "length" of a latency delay to the probability that a known (or expected) information event occurs between the trader's order submission, and its execution. In this way, we assume that the delayed exchange imposes a delay that is randomly drawn within a fixed interval, such that private information becomes public within the latency delay with some (expected) probability. Our notion of a latency delay variation has two interpretations, a "longer" delay implies that either: i) the distribution of the random delay widens, or, ii) the time before the public announcement has been reduced.

We show that as the length of the latency delay increases, informed investors migrate away from the latency delayed-exchange. We use this migration to define our results in terms of a "segmentation point". The segmentation point is the delay length at which the speculator with the highest marginal utility for the delayed exchange migrates away from the delayed exchange. As the delay length increases towards the segmentation point, uninformed investor participation at the delayed exchange increases, reaching a maximum at the segmentation point. At the non-delayed exchange, there is a net migration of informed investors, and a net emigration of uninformed investors. The result is a wider quoted spread at the non-delayed exchange; moreover, some speculators who would acquire information in a setting with no delayed market, choose not to become informed.

Once the segmentation point is reached, further increases in the delay create very different results. Spreads at the latency-delayed exchange improve no further, as all informed traders have left the exchange, while uninformed traders continue to incur larger delays. As a result, uninformed traders begin to return to the non-delayed exchange improving bid-ask spreads at the non-delayed exchange and increasing informed trader participation. For a delay of sufficient length, the non-delayed exchange reverts to the conditions present before a latency delay was imposed, while the latency-delayed exchange contains only uninformed traders who previously did not participate in the market prior to the resolution of the information

event.

We make several empirical predictions regarding latency delays. First, we predict that initial prices should improve at the delayed-exchange while they should be worse at standard exchanges. Second, we predict market segmentation effects between exchanges. Liquidity trader participation should increase following the introduction of a delay, and their trading should be concentrated at this exchange. Informed participation should fall following the introduction of a delay, and their trading should be concentrated at standard exchange. Finally, we show that, latency delays have ambiguous effects on price discovery, depending on the length of the delay. Particularly, small latency delays decrease price discovery measures, while simultaneously increasing spreads at other exchanges.

Related Literature. While there is little existing literature on the topic of latency delays, the factors which have led to their creation have been well documented. The first group of relevant literature studies high frequency trading, and its effects on markets. Predatory high frequency trading is generally cited as the rationale for the use of speed bumps and, as such, is essential to understanding their purpose. The second group of literature covers both the drivers and effects of market fragmentation. As a means for exchanges to differentiate themselves, speed bumps can be discussed within this general trend of market fragmentation and competition between exchanges.

As latency delays are on the order of milliseconds or less, market makers who are able to make use of them in a strategic manner are inherently high frequency traders. Several studies of high frequency market makers have shown that they can improve liquidity (Brogaard and Garriott 2015, Brogaard, Hagströmer, Nordén, and Riordan 2015, Subrahmanyam and Zheng 2015). However, work on high frequency liquidity demanders finds that they may increase price efficiency (Carrion 2013) but also increase transaction costs (Chakrabarty, Jain, Shkilko, and Sokolov 2014). High frequency traders have also been shown to improve price discovery through both liquidity supply (Brogaard, Hendershott, and Riordan 2015, Conrad, Wahal, and Xiang 2015) and demand (Brogaard, Hendershott, and Riordan 2014).

Proponents argue that latency delays can curb "predatory" behaviours by high frequency traders, such as inter-market arbitrage. However, critics have suggested that latency delays may also lead to quote fading. Latza, Marsh, and Payne (2014) do not find evidence of predatory quote fading behaviour by HFTs, while Malinova and Park (2016) find that it does occur.[5] Our model confirms some forms of quote fading found in the empirical literature. While we do not allow market makers to fade quotes arbitrarily, we model market makers who may fade quotes in response to new information on the underlying asset value. We show that the ability to update quotes before an order arrives may allow market makers to quote at better initial prices.

Theoretically, the role of HFTs has been studied in a variety of contexts including their role in market-making (Jovanovic and Menkveld 2011), arbitrage (Wah and Wellman 2013), and the incorporation of new information (Biais, Foucault, and Moinas 2015).[6] Menkveld and Zoican (2016) model the effects of known latency within a single exchange, versus latency in reaching the exchange, a friction similar to an intentional latency delay. We complement the existing theoretical work on HFTs by modeling both intentional, randomized delays within exchanges as well as investor migration between exchanges, based on these delays. Further to previous literature, investors base their exchange choice not only with whether other market participants are delayed, but also on whether a delay at one exchange will remove their informational advantage.

The topic of market segmentation is not new within the academic literature. Existing empirical work has found that fragmented markets can have improved liquidity (Foucault and Menkveld 2008) and efficiency (Ye and O'Hara 2011). Additional work by Kwan, Masulis, and McInish (2015) and Gomber, Sagade, Theissen, Weber, and Westheide (2016) studies the use of both dark trading, and other mechanisms, in order to attract order flow.[7] As

[5]Related work by Ye, Yao, and Gai (2013) find evidence of a different behaviour known as quote "stuffing", which we do not address in this paper

[6]A further survey is topics surrounding HFT is present in both Angel, Harris, and Spatt (2011) and O'Hara (2015).

[7]Further theoretical work by Baldauf and Mollner (2016) shows that the net effects of increased fragmentation are ambiguous for liquidity suppliers.

latency delays are another means of attracting order flow, our work confirms the concept of segmentation and suggests additional avenues for empirical market segmentation work.

Existing theoretical work studies the choice of market based on fees (Colliard and Foucault 2012), dark liquidity (Zhu 2014), and the profitability of financial intermediaries (Cimon 2016). We extend existing work by modeling market segmentation based on differences in speed. Taken together with these earlier contributions, our work helps complete the set of factors which may influence market choice by financial system participants.

The closest work to ours is Chen, Foley, Goldstein, and Ruf (2016) who empirically study the introduction of a speed bump on TSX Alpha, a Canadian trading venue. They find that, following the introduction of a speed bump, total volume on the affected exchange decreases. High frequency traders provided a greater proportion of liquidity, compared to non-high frequency traders when the speed bump was in place. Adverse selection on the affected exchange also decreased. For all other exchanges, informed trading increased, leading to wider quoted and effective spreads.

The paper proceeds as follows. Section 1 outlines the model. Section 2 presents a benchmark model of two identical (fragmented) markets with no latency delay, and then extends it to consider the case where one exchange may impose a latency delay on incoming orders. In Section 3, we present empirical and policy predictions. Section 4 concludes.

1 The Model

Security. There is a single risky security with an unknown random payoff v that is equal to $v_0 - \sigma$ or $v_0 + \sigma$, with equal probability, where $\sigma \in (0, 1)$. The security is available for trading at $t = 1$ and $t = 2$. The security's value is publicly announced at $t = 2$ before trading begins. The asset is liquidated at $t = 3$.

Market Organization. There are two exchanges, Fast and Slow, that operate as displayed limit order books: posted limit orders display their quotes to all market participants. Market

orders sent to Exchange Fast fill immediately upon receipt, whereas exchange Slow fills market orders with a random delay. With probability $\delta \in (0,1)$, an order sent to exchange Slow is delayed until $t = 2$, and filled after the announcement of v.[8] Otherwise, the order is filled immediately.

There are two interpretations for this type of latency delay. First, a latency delay of this type can reflect a setting where investors expect an incoming information event (a scheduled announcement), though some investors may not be informed about its direction and magnitude. Alternatively, this type of latency delay can reflect the presence of fleeting arbitrage opportunities at other markets. Speculators who acquire information can be viewed as acquiring the necessary technology to exploit these opportunities. The random nature of the speed bump then represents the fact that, with a delay of any length, speculators may no longer be the first to trade.

Exchange Market Maker. A competitive market maker supplies buy and sell limit orders to both exchanges before investors submit their orders at $t = 1$ and $t = 2$. The market maker is risk-neutral, and receives only the public information, v_0, about the security's fundamental value. The market maker has a zero latency, permitting them to place (and update) limit orders to both exchanges at the beginning of periods $t = 1$ and $t = 2$, before investors place their orders. At $t = 2$, upon the announcement of v, the market maker updates their $t = 1$ limit orders to the public value, v.

The exogenous separation of market makers matches an important feature of latency-delayed venues. In general, orders are delayed, with the exception of orders used for market making purposes. On some venues, this consists of orders pegged at or near the midpoint, while on others it consists of large orders, above a certain size, providing liquidity. Thus, it is generally insufficient to merely submit a limit order to bypass the delay.

Investors. There is a unit mass of risk-neutral investors. At $t = 0$, an investor arrives at the market to trade a single unit of the security. The investor is either a speculator with

[8] A random delay is similar in nature to the latency delay imposed by TMX Alpha, a Canadian trading venue. TMX Alpha delays orders by a random time period of 1-3ms.

probability $\mu > 0$, or an uninformed investor endowed with liquidity needs. Upon arrival, a speculator receives an information acquisition cost γ_i that is distributed uniformly on $[0, 1]$. Speculators may pay γ_i at $t = 0$ to perfectly learn the random payoff v. We refer to those that acquire information as "informed investors", and their mass is denoted $\mu_I \in (0, \mu)$.

With probability $(1 - \mu)$, a liquidity investor arrives. Liquidity investors have no private information, but are endowed with a liquidity need that motivates them to trade. They also pay an additional cost to trade following an adverse price movement that is proportional to the innovation, $c_i = k\lambda_i\sigma$, where $k \in (0, \infty)$. λ_i is a private scaling parameter of the innovation that is distributed uniformly on $[0, 1]$. This cost represents the cost uninformed investors pay to acquire additional capital to trade when the price moves away from them. As this represents a re-capitalization cost, liquidity investors pay this cost only if the price moves against them, not if it moves in their favour. [9] The uninformed investor also pays a constant delay cost $K \in (\sigma, \infty)$ if they elect not to trade. Liquidity investors are buyers or sellers with equal probability.

An investor i may submit a single market order at $t = 1$ or $t = 2$, or not trade. Investors place orders to maximize (expected) profits. Finally, the structure of the model is known to all market participants. The model timeline is illustrated in Figure 1.

Investor Payoffs. The expected payoff to an investor who submits a buy order at $t = 1$ is given by their knowledge of the true value of v, minus the price paid, any information acquisition or delay costs incurred. We denote liquidity investors as L, and informed investors as I. The expected payoffs to investor $i \in \{I, L\}$ submitting an order to exchange $j \in \{\mathsf{Fast}, \mathsf{Slow}\}$ are given by:

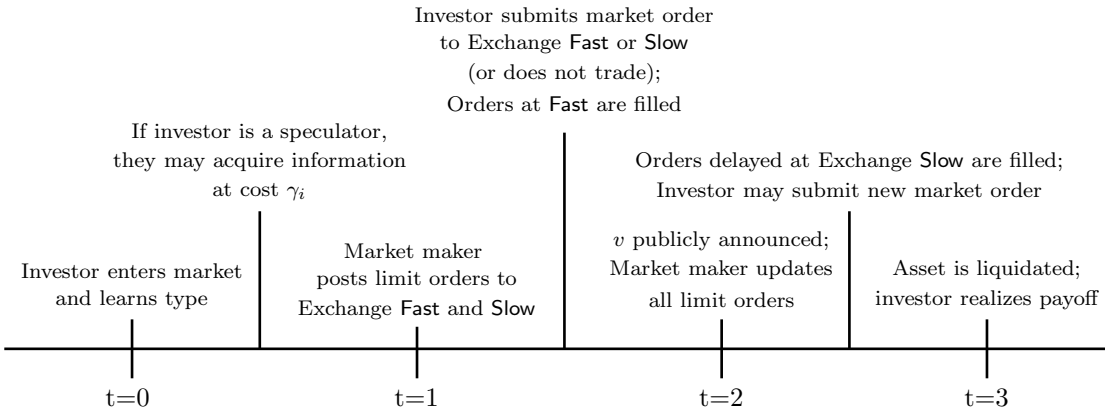
$$\pi_I(\gamma_i;\ \text{Buy at t=1}) = v - \mathsf{E}[\mathsf{ask}_1^j \mid \text{submit at exchange J}] - \gamma_i \tag{1}$$

$$\pi_L(c_i;\ \text{Buy at t=1}) = v_0 - \mathsf{E}[\mathsf{ask}_1^j \mid \text{submit to exchange J}] - \mathsf{Pr}(\text{order delay}) \times \frac{c_i}{2} \tag{2}$$

[9] We concede that a price movement can occur in a beneficial direction, and that the investor could earn a reinvestment return on the proceeds. We assume that the recapitalization cost exceeds the reinvestment return, and as such, normalize the reinvestment return to zero.

Figure 1: Model Timeline

This figure illustrates the timing of events upon the arrival of an investor at $t = 0$, until their payoff is realized at $t = 3$. Speculators face information acquisition costs γ_i, and liquidity investors face delay cost c_i.



The scaling factor of 1/2 in the delay cost of π_L reflects the fact that the asymmetric cost is only incurred if the price moves away from the liquidity investor, which occurs with probability 1/2. An investor i who submits a buy order at period $t = 2$ (or elects not to trade) has a payoff of $-\gamma_i$ if informed (speculators have payoff zero); uninformed investors earn a payoff to not trading of $-K < -\sigma$. Seller payoffs are similarly defined.

2 Equilibrium

In this section, we present two versions of our model: first, we outline a benchmark case where both exchanges are identical (no latency delay), and then subsequently compare our results to a model where Exchange Slow imposes a latency delay.

We search for a perfect Bayesian equilibrium in which the market maker chooses a quoting strategy such that they earn zero expected profits at each venue, and investors choose order submission strategies that maximize their profits. We also search for equilibria where investors use both exchanges. We study the impact of a processing delay at one exchange by comparing it to a case where both exchanges do not impose a processing delay; we refer to this as the *benchmark case*. In effect, a market with two identical exchanges is equivalent

to a single competitive exchange. Because the set-up of our model is symmetric for buyers and sellers, we focus our attention to the decisions of buyers, without loss of generality.

2.1 Identical Fragmented Markets (No Latency Delay)

In the exposition that follows, although both exchanges fill orders without delay, we continue to denote them as Exchange Fast and Slow, to maintain consistency in notation. If both exchanges impose no processing delay ($\delta = 0$), then investors' payoffs simplify considerably. Because any orders submitted to either exchange will be filled at the posted quote, investors who submit orders suffer no risk of the quote updating adversely. Speculator and liquidity investor payoffs to trading on an Exchange j are reduced to:

$$\pi_I(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^j - \gamma_i \tag{3}$$

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^j \tag{4}$$

Note that because a market buy order is filled immediately at the posted quote, the expected profit for a liquidity investor who submits a market buy order at $t = 1$ does not consider c_i directly; instead, the cost of c_i is considered when choosing whether to trade at $t = 1$, or wait until uncertainty is resolved at $t = 2$ (for which they pay c_i).

Given an expectation of investors' order submission strategies, the market maker populates the limit order books at exchanges Fast and Slow. The market maker quotes competitively, setting the ask (and bid) prices at $t = 1$ on exchange Fast and Slow—which we denote $\mathsf{ask}_1^{\mathsf{Fast}}$ and $\mathsf{ask}_1^{\mathsf{Slow}}$, respectively—to account for the expected adverse selection of an incoming buy order:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at Exchange } \mathsf{Fast}] \tag{5}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at Exchange } \mathsf{Slow}] \tag{6}$$

Prices $\mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ are analogously determined through symmetry of buyers and sellers.

At period $t = 2$, v is announced, and the market maker updates their buy orders on both exchanges to $\mathsf{ask}_2^{\mathsf{Fast}} = \mathsf{ask}_2^{\mathsf{Slow}} = \mathsf{bid}_2^{\mathsf{Fast}} = \mathsf{bid}_2^{\mathsf{Slow}} = v$.

Each investor makes two decisions: whether to participate in the market at $t = 1$ (or at all), and if they participate, to which exchange should they submit an order. A speculator receives their information acquisition cost γ_i at $t = 0$, and weighs it against the expected profit of becoming informed. If they acquire information, they subsequently decide to which exchange they will submit an order. Similarly, liquidity investors receive their delay cost c_i at $t = 0$, and choose whether to delay trading to $t = 2$. If they decide to trade at $t = 1$, they choose to which exchange they submit an order.

We characterize these decisions via backward induction. At $t = 2$, speculators (informed and otherwise) have no information advantage, and thus their expected profit is zero. Liquidity investors who did not submit an order at $t = 1$ submit an order to either exchange at $t = 2$ and pay cost c_i. It is always optimal for a liquidity investor to submit an order at $t = 2$, as the cost to abstaining, $K > \max\{c_i\}$.

At $t = 1$, speculators who do not acquire information at $t = 0$ do not trade. If a speculator has chosen to acquire knowledge of v, the now-informed investor knows that delaying until period $t = 2$ is unprofitable, so they choose the optimal exchange to which they submit their order. We denote the probability with which an informed investor submits an order to Exchange Fast as $\beta \in (0, 1)$; they submit an order to Exchange Slow otherwise. Because γ_i only dictates the decision to acquire information, and doesn't factor directly into the venue choice, informed investors use a mixed strategy in β such that they earn a equal payoff at both exchanges. Similarly, a liquidity investor that chooses to trade in $t = 1$ finds that their venue choice is not directly impacted by c_i; they also submit orders to both venues with a mixed strategy, where we denote probability of submitting an order to Exchange Fast as

$\alpha \in (0,1)$, and Exchange Slow otherwise. Buyers' order choice indifference conditions are:

$$\text{Informed Buyer: } \left\{\beta \mid \pi_I^{\mathsf{Fast}}(\text{Buy t=1}) = \pi_I^{\mathsf{Slow}}(\text{Buy t=1}) \iff \mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{7}$$

$$\text{Liquidity Buyer: } \left\{\alpha \mid \pi_L^{\mathsf{Fast}}(\text{Buy t=1}) = \pi_L^{\mathsf{Slow}}(\text{Buy t=1}) \iff \mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{8}$$

We note here that, in the absence of direct impacts by γ_i and c_i, the sole determinant of venue choice for buyers are the ask prices (and similarly bid prices for sellers). If quotes are not equal across both exchanges, then (α, β) cannot be an equilibrium, as there would be migration from the high-priced exchange to the lower priced exchange until prices across both exchanges equate.

Given the venue choice strategies for informed and liquidity investors, the ask prices quoted by the market maker at $t = 1$ can now be characterized as:

$$\mathsf{ask}_1^{\mathsf{Fast}} = v_0 + \frac{\mathsf{Pr}(\text{informed trade at } \mathsf{Fast})}{\mathsf{Pr}(\text{trade at } \mathsf{Fast})} \cdot \sigma \tag{9}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = v_0 + \frac{\mathsf{Pr}(\text{informed trade at } \mathsf{Slow})}{\mathsf{Pr}(\text{trade at } \mathsf{Slow})} \cdot \sigma \tag{10}$$

Liquidity investors are buyers or sellers with equal probability, so only half of liquidity investors who choose to participate in the market at $t = 1$ will buy, independent of the realization of v. Sell prices $\mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ are similarly characterized.

Given α and β, investors make participation decisions at $t = 0$ that characterize the measure of speculators, μ_I and the measure of liquidity investors that participate before $t = 2$, which we denote $\Pr(c_i \geq \underline{c})$. That is, all investors with $c_i \geq \underline{c}$ face a large enough cost of delay c_i, such that they trade prior to period $t = 2$. Speculators receive γ_i in period $t = 0$, and decide whether paying their information acquisition cost is profitable. The mass of speculators that choose to acquire information determines μ_I. To find μ_I, we find the value of γ_i at which a speculator is indifferent to acquiring information and not trading. This is

equal to γ_i such that a speculator earns a zero expected profit from becoming informed:

$$\bar{\gamma} = \max\left\{v - \mathsf{ask}_1^{\mathsf{Fast}}, v - \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{11}$$

Hence, any speculator with $\gamma_i \leq \bar{\gamma}$ will acquire information, and the mass of informed investors at $t = 1$ is equal to: $\mu_I = \mu \times \mathsf{Pr}\left(\gamma_i \leq \bar{\gamma}\right)$. Similarly, we characterize the measure of liquidity investors that participate in the market at $t = 1$, $\Pr(c_i \geq \underline{c})$ by:

$$\underline{c} = \min\left\{v_0 - \mathsf{ask}_1^{\mathsf{Fast}}, v_0 - \mathsf{ask}_1^{\mathsf{Slow}}\right\} \tag{12}$$

Therefore, any liquidity investors with a delay cost greater than $\underline{c}$ choose to trade at $t = 1$. The probability that such a liquidity investor arrives is given by $(1 - \mu) \times \mathsf{Pr}\left(c_i \geq \underline{c}\right)$.

An equilibrium in our model is characterized by: (i) investor participation measures, μ_I and $(1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)$; (ii) investor venue strategies, α and β, and; (iii) market maker quotes at $t = 1$ for each exchange $j \in \{\mathsf{Fast}, \mathsf{Slow}\}$, ask_1^j and bid_1^j. These values solve the venue choice indifference equations (7)-(8), the market maker quoting strategy (9)-(10), and the investor participation conditions, (11)-(12).

Theorem 1 (Identical Fragmented Markets) *Let $\delta = 0$. Then for any $\beta \in (0, 1)$, there exists a unique equilibrium consisting of participation constraints $\mu_I \in (0, \mu)$, $\underline{c} \in [0, \frac{k\sigma}{2}]$ that solve (11)-(12), prices $\mathsf{ask}_1^{\mathsf{Fast}}, \mathsf{ask}_1^{\mathsf{Slow}}, \mathsf{bid}_1^{\mathsf{Fast}}$ and $\mathsf{bid}_1^{\mathsf{Slow}}$ that satisfy (9)-(10), and $\alpha \in (0, 1)$ that solves (7)-(8) such that $\beta = \alpha$.*

Theorem 1 illustrates that, in equilibrium, identical fragmented markets may co-exist, and moreover, they need not attract the same level of order flow, despite offering identical prices. For example, in an equilibrium where $(\alpha, \beta) = (3/4, 3/4)$, Exchange Fast receives three times the order flow of Exchange B, but because $\alpha = \beta$, these probabilities cancel out of the pricing equations (9)-(10), ensuring that the ask (and bid) prices of Exchange Fast and Slow are equal. We summarize this in the Corollary below.

Corollary 1 (Equilibrium Prices) *In equilibrium, ask and bid prices at* $t = 1$ *are equal to* $\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}} = v_0 + \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}(c_i \geq \underline{c})} \cdot \sigma$ *and* $\mathsf{bid}_1^{\mathsf{Fast}} = \mathsf{bid}_1^{\mathsf{Slow}} = v_0 - \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}(c_i \geq \underline{c})} \cdot \sigma$

In what follows, we define the identical fragmented market formulation of our model ($\delta = 0$) as the benchmark case. We denote the equilibrium solutions with the superscript BM (i.e., $\mathsf{ask}^{\mathsf{BM}}$, $\mathsf{bid}^{\mathsf{BM}}$).

2.2 Slow Exchange Imposes a Latency Delay

In this section, we examine the case where Exchange Slow fills investor orders with a random processing delay, such that orders sent to Exchange Slow are filled before $t = 2$ with probability $\delta \in (0, 1)$. The processing delta impacts payoffs to informed and liquidity investors differently. Informed investors face payoffs to Exchange Fast and Slow:

$$\pi_I^{\mathsf{Fast}}(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^{\mathsf{Fast}} - \gamma_i \tag{13}$$

$$\pi_I^{\mathsf{Slow}}(\gamma_i; \text{Buy at t=1}) = v - (1-\delta) \times \mathsf{ask}_1^{\mathsf{Slow}} - \delta \cdot v - \gamma_i \tag{14}$$

By submitting an order to Exchange Slow, informed investors face the possibility of losing their informational advantage. Liquidity do not know v, however, so their expectation of what the announcement of the true value will be is always v_0, and thus the processing delay does not impact their expectation of the future value when buying. Instead, the uncertainty about the outcome of the price manifests in an asymmetrical cost to trading, c_i, that they incur if the price moves in the direction of their desired trade ($v > \mathsf{ask}_1^{\mathsf{Slow}}$). The payoffs to liquidity investors then simplify to:

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^{\mathsf{Fast}} \tag{15}$$

$$\pi_L(c_i; \text{Buy at t=1}) = (1-\delta)(v_0 - \mathsf{ask}_1^{\mathsf{Slow}}) - \delta \cdot \frac{k\lambda_i}{2} \times \sigma \tag{16}$$

Taking this into account, the market maker sets its prices at $t = 1$ in the following way:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \frac{\beta\mu_I}{\beta\mu_I + \mathsf{Pr}(\text{uninformed trade at } \mathsf{Fast})} \cdot \sigma \tag{17}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + \mathsf{Pr}(\text{uninformed trade at } \mathsf{Slow})} \cdot \sigma \tag{18}$$

In period $t = 2$, the value v is publicly announced, so the market maker updates its prices to $\mathsf{ask}_2^{\mathsf{Fast}} = \mathsf{ask}_2^{\mathsf{Slow}} = v$.

When Exchange Slow imposes a processing delay, investors weigh the cost of trading on Exchange Fast immediately, against possibility of a) losing (all or part of) their information if they are informed, or b) paying a higher cost to acquire capital to complete their trade if they are a liquidity investor. A investor's order placement strategy has two equilibrium conditions: i) an indifference condition (IC) between orders to Exchange Fast and Slow, and ii) a participation constraint (PC). For a speculator, the participation constraint PC_I is the maximum information acquisition costs γ_i that lead a speculator to become an informed investor. Then, conditional on participation, the indifference condition IC_I represents the value of β such that an informed investor is indifferent to submitting an order to or B. These conditions are written as:

$$\text{IC}_I\text{: } \delta\sigma = \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] - (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \tag{19}$$

$$\text{PC}_I\text{: } \mu_I = \mu \Pr(\gamma_i \leq \max\{\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}], (1-\delta)(\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}])\}) \tag{20}$$

Liquidity investors face two similar conditions. Their participation constraint PC_L describes the scaling of their delay costs $\underline{\lambda}$ at which they are indifferent to trading in $t = 1$ and waiting until $t = 2$. Then, conditional on participating, their indifference condition IC_L describes the value of $\bar{\lambda}$ such that a liquidity investor is indifferent to submitting an order to either

exchange. We write these conditions below.

$$\text{IC}_L\text{: } \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] = (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] + \delta \cdot \frac{k\bar{\lambda}}{2} \times \sigma \tag{21}$$

$$\text{PC}_L\text{: } \underline{\lambda} = \min\left\{\frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}]}{k\sigma}, \frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]}{k\sigma}\right\} \tag{22}$$

Finally, an equilibrium is characterized by values k such that it is sufficiently costly to delay until $t = 2$ for at least some investors (i.e, $k > \underline{k} > 0$).

Lemma 1 (Costly Delay) *In any equilibrium that satisfies conditions (19)-(22), $k > 2$.*

We can now describe our equilibrium. An equilibrium in a model with a processing delay is characterized by: (i) Ask prices (17) and (18) (and similar bid prices) set by the market maker at exchanges A and B, respectively, such that they earn zero expected profit in expectation; (ii) a solution to the speculator's optimization problem, (19)-(20) and; (iii) a solution to the liquidity investor's optimization problem, (21)-(20). By solving this system, we arrive at the following theorem.

Theorem 2 (Existence and Uniqueness) *Let $k > 2$. For $\delta \in (0, 1)$, there exist unique values μ_I, $\underline{\lambda}, \bar{\lambda}$, β, and prices $\mathsf{ask}_1^{\mathsf{Fast}}, \mathsf{ask}_1^{\mathsf{Slow}}$ given by (17)-(18) that solve equations (19)-(22).*

The nature of the equilibrium depends on the parametrization of the latency delay and can take several forms. For a delay of sufficiently small size, market makers at the delayed exchange are not offered sufficient protection from informed trades. For these delays, both types of traders continue to trade at the delayed-exchange. However, there exists an inflection point, further discussed below, where the delay becomes sufficiently large that informed traders withdraw their flow from the delayed-exchange. For these larger delays, market makers are able to offer vastly improved prices on the latency-delayed exchange, drawing order flow only uninformed traders.

One interpretation of the latency delay is in the context of statistical arbitrage. Instead of interpreting the announcement event as an earnings announcement, it can instead be viewed

as the time with which market makers become aware of arbitrage opportunities. This is similar in many respects to Budish, Cramton, and Shim (2015), who document the fleeting nature of arbitrage opportunities between New York and Chicago. When viewed in this sense, a "short" speed bump is one which is similar in length to the lifespan of actionable arbitrage opportunities. Similarly, a "long" speed bump is one which delays orders sufficiently, such that statistical arbitrage is generally not possible.

3 Empirical Predictions and Policy Implications

We investigate the impact of a latency delay on measures of market quality and price discovery. When Exchange Slow imposes a latency delay, investors who submit an order to Exchange Slow at $t = 1$ face the possibility that private news may become public (i.e., the market maker will update their limit orders) before their order is filled. The latency delay impacts speculators and liquidity investors differently. Speculators do not benefit from a latency delay directly, as a latency delay increases the probability that they may lose their private information advantage, if they trade at Exchange Slow. Hence, ceteris paribus, they prefer an exchange that will execute their order immediately. A liquidity investor's preference, however, depends on their individual costs to delay. Those that have sufficiently low delay costs are impacted more by the price of the order than the possibility of delay, and hence, they may prefer an exchange with a latency delay, if the price offered is at a sufficient discount. Because speculators and liquidity investors' motives are not perfectly correlated, the introduction of a latency delay segments the order flow of the two investor groups, to varying degrees.

The degree of order flow segmentation depends on the parameters of the speed bump. A speed bump is not driven by the magnitude of the delay alone, but the likelihood that a delay of a given length would lead an investors' order to fill after private information becomes public, and hence face updated limit orders. In our model, the latency delay δ takes on this

interpretation. We identify a latency delay δ^*—which we refer to as the "segmentation point"—as the delay such that for all $\delta \geq \delta^*$, no informed traders submit orders to the delayed exchange ($\beta = 1$). Moreover, if no informed traders submit orders to Exchange Slow, then it must be that in equilibrium, $\mathsf{ask}_1^{\mathsf{Slow}} = 0$. Thus, because the cost of trading on exchange Slow is bounded above by the cost of delay, it must be that all uninformed investors participate in the market at $t = 1$ ($\underline{\lambda} = 0$). Given these solutions, we solve equations (19)-(22) for δ^*, yielding the equation:

$$\delta^*(k, \mu, \sigma) = \frac{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} - (1-\mu)(1-\frac{2}{k})}{\sqrt{(1-\mu)^2(1-\frac{2}{k})^2 + (1-\mu)(1-\frac{2}{k})\mu\sigma} + (1-\mu)(1-\frac{2}{k})} \tag{23}$$

We use δ^* to characterize our results on order flow segmentation in Proposition 1 below.

Proposition 1 (Order Flow Segmentation) *Relative to the benchmark value at $\delta = 0$, if Exchange* Slow *imposes a delay $\delta \in (0, 1)$, then:*

- *for $\delta \leq \delta^*$, informed trading on Exchange* Slow *falls ($\beta \downarrow$), and the measure of liquidity investors who submit orders only at $t = 2$ declines ($\underline{\lambda} \downarrow$).*
- *for $\delta > \delta^*$, informed trading concentrates on Exchange* Fast *($\beta = 1$), and all liquidity investors submit orders at $t = 1$ ($\underline{\lambda} = 0$). Moreover, liquidity trading on Exchange* Fast *increases ($\bar{\lambda} \downarrow$).*

While we find that $\beta = 1$ for all $\delta > \delta^*$, we do not predict full order flow segmentation of informed and uninformed investors, as uninformed investors whose delay costs are large enough ($\lambda_i \geq \bar{\lambda}$) still use Exchange Fast. The relationship between the value of δ and the participation of both investor types is shown in Figure 2.

Order flow segmentation represents one of the reasons why latency delays are often advertised by exchanges. Proponents argue that delays are a means of protecting liquidity suppliers from informed investors. Empirically speaking, existing work supports this fact and finds that that exchanges with latency delays have lower informed trading and higher

participation by uninformed orders (Chen, Foley, Goldstein, and Ruf 2016). We show that, for a sufficiently long delay, informed traders do optimally avoid these exchanges altogether, allowing liquidity suppliers to quote a near-zero spread for uninformed investors.

The existence of the latency delay implies that with some probability an order submitted to Exchange Slow will be delayed until after a public information announcement about the security being traded. Thus, the market maker is afforded the opportunity to update their limit orders before the delayed order arrives, allowing them to potentially avoid being adversely selected. Because the potential of updated quotes is equally costly to all informed investors, but not all liquidity investors, it is natural to hypothesize that quoted spreads would differ across exchanges Fast and Slow. Our model yields the following prediction on quoted spread behaviour between Exchanges Fast and Slow, given a latency delay, $\delta \in (0, 1)$.

Proposition 2 (Quoted Spreads) *For $\delta \in (0, 1)$ quoted spreads are narrower for Exchange* Slow ($\mathsf{ask}^{\mathsf{Slow}} \leq \mathsf{ask}^{\mathsf{BM}}$) *and wider at Exchange* Fast ($\mathsf{ask}^{\mathsf{Fast}} \geq \mathsf{ask}^{\mathsf{BM}}$)*, when compared to the benchmark case. For $\delta < \delta^*$, the spread widens at Exchange* Fast *as δ increases, while for $\delta > \delta^*$, the spread narrows at Exchange* Fast *as δ increases.*

While the market maker may have the opportunity to update their quotes before an informed trade clears the latency delay, they face additional costs at the non-delayed exchange. Informed traders concentrate at the non-delayed exchange, increasing adverse selection costs and forcing the market maker to quote worse prices than in the benchmark case. We illustrate the impact of δ on quoted spreads in Figure 4. Proposition 2 reflects the empirical results in Chen, Foley, Goldstein, and Ruf (2016), who find that spreads improve on the exchange with the latency delay, and become worse elsewhere.

An improvement in quotes at Exchange Slow is correlated with our result on order segmentation (Proposition 1): the migration of informed traders to Slow leads to an increase in market participation at $t = 1$ by liquidity investors. To analyze this order segmentation, we define total order submissions (OS) as the probability that an investor who enters, submits

an order at $t = 1$:

$$\mathsf{OS}_{t=1} = \mu\bar{\gamma} + (1-\mu) \times (1-\underline{\lambda}) \tag{24}$$

We then determine from Equation 25 how much of total order submissions at $t = 1$ are expected to result in trades before $t = 2$, our measure for trading volume before $t = 2$.

$$\mathsf{OS}_{t=1} = \mu\bar{\gamma} \times (\beta + (1-\beta)(1-\delta)) + (1-\mu) \times ((1-\bar{\lambda}) + (1-\delta)(\bar{\lambda} - \underline{\lambda})) \tag{25}$$

The right panel of Figure 3 shows that, as liquidity investors increase their participation, the migration of informed traders to Exchange Fast and the resulting increase in quoted spreads at Exchange Fast lead to a decline in informed trader participation, net of which our model predicts an increase in aggregate order submissions. This increase does not lead to an increase in total trading volume, however, as the increase in liquidity investor participation occurs primarily at Exchange Slow, orders at which, fill before $t = 2$ with probability $1 - \delta$. We summarize this result below.

Proposition 3 (Total Volume and Participation) *Relative to the benchmark value at $\delta = 0$, if Exchange* Slow *imposes a delay $\delta \in (0, 1)$, then liquidity investor participation improves ($\underline{\lambda} \downarrow$), and information acquisition falls ($\bar{\gamma} \downarrow$). Moreover, total market order submission at $t = 1$ increases, but expected trading volume prior to $t = 2$ declines.*

The latency delay affects incentives for both liquidity investors and informed investors. For liquidity investors, the improved prices offered by the market maker increases participation. As more liquidity investors enter the market and select the latency-delayed exchange, the market maker probability of adverse selection falls, further improving prices. For informed investors, the latency delay creates a disincentive for information acquisition. As δ increases towards δ^*, the proportion of liquidity investors to informed traders on the non-delayed exchange decreases, increasing spreads and decreasing total participation by informed investors. Moreover, a rise in δ improves the likelihood that an informed trader loses their information advantage if they trade on exchange Slow.

If the delay is sufficiently long, however, ($\delta = \delta^*$), all informed traders segregate to the non-delayed exchange, and all liquidity investors participate before $t = 1$. At this point, that any longer delay cannot improve the adverse selection costs on Exchange Slow, as these costs are already zero. Then, it must be that an increase in the delay probability beyond δ^* can only increase the probability that a liquidity investor pays their delay cost, which must be greater than $\mathsf{ask}_1^{\mathsf{Slow}} = 0$. Thus, for any $\delta > \delta^*$, liquidity investors must migrate from Exchange Slow to Exchange Fast (see Figure 2). For a sufficiently long delay, both informed traders and liquidity traders at the non-delayed exchange revert to the case where no delayed-exchange exists.

In comparison to the benchmark case, we find that the presence of a delayed exchange unequivocally reduces information acquisition by informed investors (and their subsequent market participation). We examine whether this fall in information acquisition arising from the presence of a delayed exchange contributes positively or negatively the price discovery process. In our framework, we define a measure of price discovery as the fraction of trades prior to the announcement of v that can be attributed to informed trades (that is, the permanent price impact of a trade).

$$\mathsf{Price\ Discovery} = \mu\bar{\gamma} \times (\beta \cdot \mathsf{ask}_1^{\mathsf{Fast}} + (1-\beta)(1-\delta) \cdot \mathsf{ask}_1^{\mathsf{Slow}}) \tag{26}$$

An informed investor's contribution to permanent price impact has three components: i) the probability of information acquisition, ii) the likelihood of a trade by an informed investor, and iii) the quote they hit (i.e. their price impact). From Proposition 3, we know that μ_I is lower for any $\delta \in (0, 1)$ when compared to the benchmark case, so the presence of a delayed exchange reduces permanent price impact under component (i). The impact of (ii) and (iii) are more nuanced, however. For $\delta < \delta^*$, the probability of trading at $t = 1$ for informed investors falls for those participating on Exchange Slow, and the quoted spread narrows. The countervailing force to this is that informed investors migrate their participation toward Exchange Fast, where trading before $t = 1$ is guaranteed *and* the quoted

spread is widening. For small δ, the reduction in informed investor volume and tightening of the quoted spread dominates, but for sufficiently large delays $\delta > \hat{\delta} >> \delta^*$ where informed trading is concentrated entirely on Exchange Fast, the latter dominates, and price discovery improves above that of the benchmark case.

Numerical Observation 1 (Price Discovery) *Relative to $\delta = 0$, there exists a unique $\hat{\delta} > \delta^*$, such that for all $\delta < \hat{\delta}$, average price movement attributed to informed trades (permanent price impact) at $t = 1$ worsens. For any $\delta > \hat{\delta}$, price discovery improves.*

An additional consequence of the latency delay is a change in pre-announcement price discovery, as shown in Figure 4. While price-discovery decreases for shorter delays, sufficiently long delays concentrate traders at the exchange with no delay and may improve price discovery measures. Unlike the previous results in this paper, which represent a transfer between liquidity traders and informed investors, the change price discovery information represents a cost imposed on the market by the delayed exchange. This prediction is somewhat at odds with the empirical results of Chen, Foley, Goldstein, and Ruf (2016), as we predict that price discovery may improve following the introduction of some forms of latency delay.

Curiously, we find that with sufficiently short delays, price discovery falls, but spread widen for informed traders. Here, markets lose benefits from price discovery, while informed traders continue to pay higher trading costs. Combined, these two changes represent a cost imposed on other exchanges from the introduction of a latency delay. This form of equilibrium is counter to conventional results, where increased price discovery results in wider spreads, and decreased price discovery allows market makers to quote narrower spreads.

Because of the ambiguous effects on price discovery, the effects on liquidity investors are also not definitive. We examine whether this effect has a positive transfer to liquidity investors via a reduction in trading costs paid on *average* (across liquidity investors of all delay cost types). We write this measure in the following way:

$$\mathsf{ATC} = \int_{\bar{\lambda}}^{1} \mathsf{ask}_1^{\mathsf{Fast}} \mathrm{d}\lambda + \int_{\underline{\lambda}}^{\bar{\lambda}} \mathsf{ask}_1^{\mathsf{Slow}} + \frac{k\sigma}{2}\lambda \mathrm{d}\lambda + \int_{0}^{\underline{\lambda}} \frac{k\sigma}{2}\lambda \mathrm{d}\lambda \tag{27}$$

We now examine how ATC is impacted by the introduction of an exchange with a latency delay, δ. Our result is presented graphically in Figure 4.

Numerical Observation 2 (Liquidity Investor Trading Costs) *There exist unique* $\underline{\delta}$ *and* $\overline{\delta}$ *such that* $0 < \underline{\delta} < \overline{\delta} < 1$ *where liquidity investors:*

- *pay lower average costs if* $\delta < \underline{\delta}$ *or* $\delta > \overline{\delta}$ *relative to* $\delta = 0$.
- *pay higher average costs if* $\delta \in [\underline{\delta}, \overline{\delta}]$ *relative to* $\delta = 0$.

Despite the fact that more liquidity investors participate in the market pre-announcement, the average delay costs borne by those traders increases. This, seemingly contradictory behaviour is a result of new liquidity traders submitting orders in $t = 1$, rather than delaying until $t = 2$. Without the latency delay, liquidity traders with the lowest delay cost are those who choose not to enter the market, and delay trading until the final period. With the latency delay, these low delay cost traders enter the market and trade on the delayed-exchange. For liquidity traders already in the market, an increase in the delay time increases the optimality of trading on the delayed exchange. While these traders are offered better prices, they incur higher delay costs, which increase their total cost of trading. Broadly speaking, traders who begin to enter the market at $t = 1$ as a result of the latency delay are made better off, while many of those who were already in the market are made worse off.

4 Conclusion

Latency delays have been a topic of controversy since their introduction. Proponents contend that they improve liquidity for uninformed investors via narrower spreads, while opponents claim that the liquidity improvement is illusory: the "improved" quotes may fade before they are ever hit. We construct a model of latency delays in order to disentangle potential effects from their introduction.

We find that many of the effects from latency delays depend on the length of the delay. Specifically, we define a "segmentation point", which is the shortest length of a latency

delay such that all informed traders cluster on the non-delayed exchange. As the length of a latency delay increases towards this point, the crowding of informed traders at the non-delayed exchange widens its bid-ask spread. Concurrently, more liquidity traders migrate to the delayed exchange, narrowing the its quoted spread, and increasing its total order flow.

Once the delay increases past the segmentation point, results change drastically. The spread at the latency-delayed exchange holds constant, and liquidity traders begin migrating to the non-delayed exchanges. This migration improves bid-ask spreads at non-delayed exchanges, and encourages more informed traders to (re-)enter the market. Finally, for sufficiently long latency delays, non-delayed markets are identical to the case with no delays, while the delayed markets contain only liquidity traders who did not trade in market with no delayed exchange.

Our model makes several empirical predictions. We predict that, following the introduction of a delay, quoted spreads should improve at the delayed exchange, while worsening at the standard exchanges. We also predict that the presence of a delayed exchange improves liquidity investor participation, and that informed trading should cluster on the non-delayed exchange. Our model also offers several predictions for policy makers. First, we find that the introduction of a delayed exchange can impact other exchanges. Other exchanges are likely to see an increased concentration of informed order flow and a withdrawal of retail order flow. Market makers on these exchanges may require additional protection, or they may withdraw from markets or quote at much worse prices. Alternatively, the delayed exchanges are particularly attractive to uninformed traders. This may create the need for special attention by regulators who may be concerned about protecting retail investors and non-professional market participants. Finally, sufficiently-short latency delays may create a loss in price discovery, combined with an increase in spreads at non-delayed exchanges. This combination represents a cost imposed on other markets from a delayed-exchange. Our model shows that, as with many market structure phenomena, policy makers must take a nuanced view to changes involving latency delays.

References

Angel, James J, Lawrence E Harris, and Chester S Spatt, 2011, Equity trading in the 21st century, *Quarterly Journal of Finance* 1, 1–53.

Baldauf, Markus, and Joshua Mollner, 2016, Trading in fragmented markets, *Available at SSRN*.

Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2015, Equilibrium fast trading, *Journal of Financial Economics* 116, 292–313.

Brogaard, Jonathan, and Corey Garriott, 2015, High-frequency trading competition, .

Brogaard, Jonathan, Björn Hagströmer, Lars Nordén, and Ryan Riordan, 2015, Trading fast and slow: Colocation and liquidity, *Review of Financial Studies* 28, 3407–3443.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High-frequency trading and price discovery, *Review of Financial Studies* 27, 2267–2306.

——— , 2015, Price discovery without trading: Evidence from limit orders, *Available at SSRN 2655927*.

Budish, Eric, Peter Cramton, and John Shim, 2015, The high-frequency trading arms race: Frequent batch auctions as a market design response, *Quarterly Journal of Economics* 130, 1547–1621.

Carrion, Allen, 2013, Very fast money: High-frequency trading on the nasdaq, *Journal of Financial Markets* 16, 680–711.

Chakrabarty, Bidisha, Pankaj K Jain, Andriy Shkilko, and Konstantin Sokolov, 2014, Speed of market access and market quality: Evidence from the sec naked access ban, *Available at SSRN 2328231*.

Chen, Haoming, Sean Foley, Michael A Goldstein, and Thomas Ruf, 2016, The value of a millisecond: Harnessing information in fast, fragmented markets, .

Cimon, David A, 2016, Broker routing decisions in limit order markets, *Available at SSRN*.

Colliard, Jean-Edouard, and Thierry Foucault, 2012, Trading fees and efficiency in limit order markets, *Review of Financial Studies* 25, 3389–3421.

Conrad, Jennifer, Sunil Wahal, and Jin Xiang, 2015, High-frequency quoting, trading, and the efficiency of prices, *Journal of Financial Economics* 116, 271–291.

Foucault, Thierry, and Albert J Menkveld, 2008, Competition for order flow and smart order routing systems, *Journal of Finance* 63, 119–158.

Gomber, Peter, Satchit Sagade, Erik Theissen, Moritz Christian Weber, and Christian Westheide, 2016, Spoilt for choice: Order routing decisions in fragmented equity markets, .

Jovanovic, Boyan, and Albert J Menkveld, 2011, Middlemen in limit order markets, *Western finance association (WFA)*.

Kwan, Amy, Ronald Masulis, and Thomas H McInish, 2015, Trading rules, competition for order flow and market fragmentation, *Journal of Financial Economics* 115, 330–348.

Latza, Torben, Ian W Marsh, and Richard Payne, 2014, Fast aggressive trading, *Available at SSRN 2542184*.

Malinova, Katya, and Andreas Park, 2016, modern market makers, *Retrieved from* ***http://firn.org.au/wp-content/uploads/2016/05/Modern-Market-Makers-Park-Malinova.pdf*** *, October 26, 2016*.

Menkveld, Albert J, and Marius A Zoican, 2016, Need for speed? exchange latency and liquidity, *Review of Financial Studies (Forthcoming)*.

O'Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257–270.

Subrahmanyam, Avanidhar, and Hui Zheng, 2015, Limit order placement by high-frequency traders, *Available at SSRN 2688418*.

Wah, Elaine, and Michael P Wellman, 2013, Latency arbitrage, market fragmentation, and efficiency: a two-market model, in *Proceedings of the fourteenth ACM conference on Electronic commerce* pp. 855–872. ACM.

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

Ye, Mao, Chen Yao, and Jiading Gai, 2013, The externalities of high frequency trading, *Available at SSRN 2066839*.

Zhu, Haoxiang, 2014, Do dark pools harm price discovery?, *Review of Financial Studies* 27, 747–789.

A Appendix

In the appendix, we include a description of the mechanics underlying latency delays, all proofs and figures not presented in-text.

A.1 Latency Delays.

Broadly speaking, latency delays are means by which an exchange imposes a delay on some or all of their incoming orders. Despite being a relatively new feature offered by exchanges, many varieties of latency delay exist.

The most well known type of latency delay is that of IEX in the United States. This delay, sometimes referred to as the magic shoebox, indiscriminately slows down all orders entering the exchange by 350 microseconds. This alone would not prevent multi-market strategies, as traders could simply send their orders to the delayed exchange in advance. However, markets such as IEX generally allow traders to post pegged orders, which move instantaneously in response to external factors Since these pegged orders move instantaneously if trading occurs on other exchanges, market makers using these orders are offered some protection from multi-market trading strategies.

The pegged orders at IEX are available in multiple forms, but the one most relevant to this paper is what is called the "discretionary peg". This order type uses a known algorithm to determine if a price movement is likely, a behaviour IEX refers to as a "crumbling quote".[10] If IEX determines that the quote in a particular security is likely to move, it automatically reprices orders placed at "discretionary pegs", without the 350 microsecond delay.

A second type of delay allows some forms of liquidity supplying orders to simply bypass the delay. These limit orders often have a minimum size, or price improvement requirement, which differentiates them from a conventional limit order. By allowing some orders to bypass the latency delays, market makers who use these orders are able to update their quotes in

[10]Complete documentation is available in the IEX Rule Book, Section 11.190 (g), available here: `https://www.iextrading.com/docs/Investors\%20Exchange\%20Rule\%20Book.pdf`

response to trading on other venues. If the delay is calibrated correctly, this updating can occur before the same liquidity demanding orders bypass the latency delay. Critics contend that these delays also potentially allows market makers to fade their quotes, removing liquidity before any large order reaches the exchange.

This form of latency delay is used on the Canadian exchange TSX Alpha. In the case of TSX Alpha, orders entering the exchange are delayed by a period of 1 to 3 milliseconds before reaching the order book. A special order type, a limit order referred to as a "post only" order, is able to bypass this delay. Unlike a conventional limit order, the "post only" order also contains a minimum size requirement based on the price of the security. These sizes range from 100 shares for high priced to 20,000 shares for lower priced securities.[11]

Finally, a third type of latency delay explicitly classifies traders into two groups. Some traders are affected by the delay, and have their orders held up for a fixed period of time. Other traders are simply not affected and trade as normal. Unlike the other two types of delays which rely on order types, this form requires the explicit division of traders by the exchange into two types. This is used on the Canadian exchange Aequitas NEO, which divides traders into Latency Sensitive Traders, who are affected by the speed bump, and non-Latency Sensitive Traders, who are not.[12] Those are are deemed to be "latency sensitive" are subjected to a randomized delay of between 3 to 9 milliseconds.

[11]Complete documentation is available on the TMX Group website, here: `https://www.tsx.com/trading/tsx-alpha-exchange/order-types-and-features/order-types`

[12]The factors underlying this determination are outlined in Section 1.01 of the Aequitas Neo rule book, available here: `https://aequitasneoexchange.com/media/176022/aequitas-neo-trading-policies-march-13-2017.pdf`

A.2 Proofs

Proof Sketch (Theorem 1).

Investors who choose to buy at $t = 1$ at Exchange j have profit functions given by:

$$\pi_I(\gamma_i; \text{Buy at t=1}) = v - \mathsf{ask}_1^j - \gamma_i \tag{28}$$

$$\pi_L(c_i; \text{Buy at t=1}) = v_0 - \mathsf{ask}_1^j \tag{29}$$

Because exchanges are identical in their operation, it must be that in any equilibrium, their ask and bid prices are identical. These prices are given by the following:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \frac{\beta\mu_I}{\beta\mu_I + (1-\mu)\alpha\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{30}$$

$$\mathsf{ask}_1^{\mathsf{Slow}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + (1-\mu)(1-\alpha)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{31}$$

We then solve $\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{ask}_1^{\mathsf{Slow}}$ for $(\alpha, \beta) \in (0,1)^2$, for all μ_I and $\underline{c}$:

$$\mathsf{ask}_1^{\mathsf{Fast}} = \mathsf{E}[v \mid \text{Buy at } \mathsf{Fast}] = \mathsf{E}[v \mid \text{Buy at } \mathsf{Slow}] = \mathsf{ask}_1^{\mathsf{Slow}} \tag{32}$$

$$\iff \frac{\beta\mu_I}{\beta\mu_I + (1-\mu)\alpha\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma = \frac{(1-\beta)\mu_I}{(1-\beta)\mu_I + (1-\mu)(1-\alpha)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \cdot \sigma \tag{33}$$

$$\iff \beta(1-\alpha) = (1-\beta)\alpha \Rightarrow \beta = \alpha \tag{34}$$

Given equilibrium prices in (30) and (31), we then solve for μ_I and $\underline{c}$. To solve for μ_I, we solve the equation:

$$\mu_I = \mu \times \mathsf{Pr}(\gamma_i \leq \min\left\{v - \mathsf{ask}_1^{\mathsf{Fast}}, v - \mathsf{ask}_1^{\mathsf{Slow}}\right\}) \tag{35}$$

$$\Rightarrow \bar{\gamma} - (v - \mathsf{ask}_1^{\mathsf{Fast}}) = 0 \tag{36}$$

where the simplification in (36) arises from the fact that the ask prices at Exchanges Fast and Slow are identical in equilibrium. We then show that there exists a unique $\bar{\gamma} \in [0,1]$

that solves (36). Given this $\bar{\gamma}$, $\mu_i = \mu \times \bar{\gamma}$ exists, and is unique.

$$\bar{\gamma} = 0 : 0 - (v - 0) < 0 \tag{37}$$

$$\bar{\gamma} = 1 : 1 - \sigma \left(1 - \frac{\mu}{\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)} \right) > 0 \tag{38}$$

where (38) is positive because $\sigma < 1$. Then differentiate equation (36) by $\bar{\gamma}$:

$$\frac{\partial}{\partial \bar{\gamma}}(\bar{\gamma} - (v - \mathsf{ask}_1^{\mathsf{Fast}})) = 1 + \sigma \left(\frac{(1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right)}{(\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right))^2} \right) > 0 \tag{39}$$

for all $\underline{c}$. Then, to show there exists a unique $\underline{c}$, consider the participation constraint for liquidity investors, $\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}} \geq 0$:

$$\underline{c} = 0 : 0 - \frac{\mu_I}{\mu_I + (1-\mu)\mathsf{Pr}\left(c_i \geq 0\right)} \cdot \sigma < 0 \tag{40}$$

$$\underline{c} = 1 : 1 - \sigma > 0 \tag{41}$$

where (41) is positive because $\sigma < 1$. Then differentiate $\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}} \geq 0$ by $\underline{c}$:

$$\frac{\partial}{\partial \underline{c}}(\underline{c} - \mathsf{ask}_1^{\mathsf{Fast}}) = 1 + \sigma \left(\frac{(1-\mu)\mu_i}{(\mu + (1-\mu)\mathsf{Pr}\left(c_i \geq \underline{c}\right))^2} \right) > 0 \tag{42}$$

Thus, a unique equilibrium exists for all $\beta = \alpha \in (0,1)^2$. ■

Proof (Lemma 1). For an equilibrium to exist, we require that liquidity investors will trade before $t = 1$ for a non-zero measure of λ_i on *both exchanges*. To ensure this, a sufficient condition is that the scaling of the cost of delaying trade, k, must be large enough, to entice investors with the largest valuations ($\lambda_i \geq 1 - \epsilon$, for ϵ arbitrarily close to zero) to trade at an exchange that posts the widest possible spread, equal to 2σ. Then, k must satisfy:

$$\frac{k(1-\epsilon)\sigma}{2} > \sigma \iff k > \frac{2}{1-\epsilon} > 2$$

Hence, in any equilibrium where investors use both exchanges, $k > 2$. ■

Proof (Theorem 2). The proof of Theorem 2 proceeds similarly to Theorem 1, except

that we solve the liquidity investor constraints for $\bar{\lambda}$ and $\underline{\lambda}$, instead of $\underline{c}$ and α.

There are three equilibrium cases, defined through the (mixed) strategies of speculators: $\beta = 0$, $\beta(0,1)$, and $\beta = 1$.

Speculators use only Exchange Slow ($\beta = 0$): In this part, we show that no equilibrium exists for $\beta = 0$. To do so, we consider the informed investor's incentive compatibility constraint, evaluated at $\beta = 0$.

$$\text{IC}_I\text{: } \sigma - 0 - (1-\delta)(\sigma - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})}) = \delta\sigma + \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})} > 0 \tag{43}$$

Moreover, because $\mathsf{ask}^{\mathsf{Fast}} = 0$, then $\bar{\gamma} < 1$, implying that informed investors would always have an incentive to deviate to the fast exchange.

Speculators use both exchanges ($\beta \in (0,1)$): We now solve the following system of equations for $\bar{\lambda}, \underline{\lambda}, \bar{\gamma}$ and β, using the method as in the proof of Theorem 1.

$$\text{IC}_I\text{: } \delta\sigma = \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] - (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \tag{44}$$

$$\text{PC}_I\text{: } \mu_I = \mu \Pr(\gamma_i \leq \max\{\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}], (1-\delta)(\sigma - \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}])\}) \tag{45}$$

$$\text{IC}_L\text{: } \mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}] = (1-\delta)\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] + \delta \cdot \frac{k\bar{\lambda}}{2} \times \sigma \tag{46}$$

$$\text{PC}_L\text{: } \underline{\lambda} = \min\left\{\frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Fast}]}{k\sigma}, \frac{2\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]}{k\sigma}\right\} \tag{47}$$

We write (44)-(47) explicitly as:

$$\text{IC}_I\text{: } 1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\left(1 - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})}\right) = 0 \tag{48}$$

$$\text{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})}\right) = 0 \tag{49}$$

$$\text{IC}_L\text{: } \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0 \tag{50}$$

$$\text{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} - \frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} = 0 \tag{51}$$

We first show that, for all $(\bar{\lambda}, \underline{\lambda}, \bar{\gamma}) \in (0,1)^3$, there is a unique $\beta^* \in (0,1)$ that solves (44).

$$\text{IC}_I \mid_{\beta=0} : \delta\sigma + (1-\delta)\frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(\bar{\lambda} - \underline{\lambda})} > 0 \tag{52}$$

$$\text{IC}_I \mid_{\beta=1} : \delta\sigma - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} < 0, \ \forall \delta < \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} = \delta^* \tag{53}$$

Thus, for all $\delta < \delta^*$, there exists a $\beta \in (0,1)$ by the intermediate value theorem that satisfies (48). To show that β^* is unique, we differentiate (48) with respect to β.

$$\frac{\partial}{\partial\beta}(\text{IC}_I) = -\frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda}))^2} < 0 \tag{54}$$

Thus, β^* is unique for all $(\bar{\lambda}, \underline{\lambda}, \bar{\gamma}) \in (0,1)^3$.

We then rearrange (48) to:

$$\delta = \frac{\mu\bar{\gamma}\beta}{\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda})} - (1-\delta)\frac{\mu\bar{\gamma}(1-\beta)}{\mu\bar{\gamma}(1-\beta) + (1-\mu)(\bar{\lambda} - \underline{\lambda})} \tag{55}$$

Equation (55) can then be substituted into (50) and simplified to yield an expression for $\bar{\lambda}$:

$$\bar{\lambda} = \frac{2}{k} \tag{56}$$

Next, we use equation (48) and (49) to solve for $\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}]$, in terms of δ, σ and $\bar{\gamma}$, which we substitute into equation (51):

$$\underline{\lambda} = \frac{2}{k}\left(1 - \frac{\bar{\gamma}}{\sigma(1-\delta)}\right) \tag{57}$$

Then, because the right-hand side equals $\frac{2}{k}\mathsf{E}[\sigma \mid \text{Buy at } \mathsf{Slow}] \in (0,1)$, and $\bar{\lambda} = \frac{2}{k}$, $\underline{\lambda}^*$ exists and is unique for all $\bar{\gamma} \in (0,1)$. Lastly, we show that there exists a unique $\bar{\gamma}^*$ that solves (20), given $\bar{\lambda}^*(\bar{\gamma}), \underline{\lambda}^*(\bar{\gamma})$, and $\beta^*(\bar{\gamma})$.

First, we show that $\bar{\gamma}^* \in [0,1]$ exists, by appealing to the intermediate value theorem:

$$\text{PC}_I \mid_{\bar{\gamma}=0} : 0 - \sigma < 0 \tag{58}$$

$$\text{PC}_I \mid_{\bar{\gamma}=1} : 1 - \sigma(1 - \frac{\mu\beta}{\mu\beta + (1-\mu)(1-\bar{\lambda})}) > 0 \tag{59}$$

where (59) holds by the fact that $\sigma < 1$. Thus, $\bar{\gamma}^* \in (0,1)$ exists. To show that $\bar{\gamma}^*$ is unique, we differentiate (20) by $\bar{\gamma}$:

$$\frac{\partial}{\partial \bar{\gamma}}(\mathrm{PC}_I) = \sigma \frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \beta}{\partial \bar{\gamma}} \cdot \frac{\mu\bar{\gamma}(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \bar{\lambda}}{\partial \bar{\gamma}} \cdot \frac{\mu\bar{\gamma}(1-\mu)\beta}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} < 0 \tag{60}$$

Where the third term is zero by the fact that $\frac{\partial \bar{\lambda}}{\partial \bar{\gamma}} = 0$. Now all we need to show is that $\frac{\partial \beta}{\partial \bar{\gamma}} \geq 0$. If we differentiate (19) by $\bar{\gamma}$, and solve for $\frac{\partial \beta}{\partial \bar{\gamma}}$, we find:

$$\frac{\partial IC_I}{\partial \bar{\gamma}} = -\frac{\mu(1-\mu)(1-\bar{\lambda}) + \frac{\partial \beta}{\partial \bar{\gamma}}\mu\beta(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - (1-\delta)\frac{k}{2} \cdot \frac{\partial \underline{\lambda}}{\partial \bar{\gamma}} = 0 \tag{61}$$

$$\iff \frac{\partial \beta}{\partial \bar{\gamma}} = -\frac{\frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda}))^2} + (1-\delta)\frac{k}{2} \cdot \frac{\partial \underline{\lambda}}{\partial \bar{\gamma}}}{\frac{\mu\beta(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta+(1-\mu)(1-\bar{\lambda}))^2}} > 0 \tag{62}$$

where (62) is positive by the fact that the partial derivative of $\underline{\lambda}$ with respect to $\bar{\gamma}$ is:

$$\frac{\partial \underline{\lambda}}{\partial \bar{\gamma}} = -\frac{2}{\sigma k(1-\delta)} < 0$$

which implies that:

$$\frac{\mu(1-\mu)(1-\bar{\lambda})}{(\mu\bar{\gamma}\beta + (1-\mu)(1-\bar{\lambda}))^2} - \frac{1}{\sigma} < 0$$

Hence, $\bar{\gamma}^*$ is unique.

Speculators use only Exchange Fast ($\beta = 1$): Lastly, we solve equations (44)-(47) for the case where $\beta =$. Inputting $\beta = 1$, we have:

$$\mathrm{IC}_I\text{: } \delta - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} \geq 0 \tag{63}$$

$$\mathrm{PC}_I\text{: } \bar{\gamma} - \sigma\left(1 - \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})}\right) = 0 \tag{64}$$

$$\mathrm{IC}_L\text{: } \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda})} - \frac{\delta k\bar{\lambda}}{2} = 0 \tag{65}$$

$$\mathrm{PC}_L\text{: } \frac{\delta k\underline{\lambda}}{2} = 0 \tag{66}$$

Equation (63) pins down the relation between β and δ: for all $\delta \geq \frac{\mu\bar{\gamma}}{\mu\bar{\gamma}+(1-\mu)(1-\lambda)}$, $\beta^* = 1$.

Moreover, by inspection, we see that $\underline{\lambda}^* = 0$. To prove the existence of a unique $\bar{\gamma}$, we solve equation (64) for $\bar{\gamma}$:

$$\bar{\gamma}^* = \frac{\sqrt{(1-\mu)^2(1-\bar{\lambda})^2 + (1-\mu)(1-\bar{\lambda})\mu\sigma} - (1-\mu)(1-\bar{\lambda})}{2\mu} \tag{67}$$

By inspection, $\bar{\gamma}^*$ exists and is unique as long as the limit $\mu \to 0$ exists, and is in the interval [0,1]. To calculate this limit, we need to apply L'Hôpital's Rule.

$$\lim_{\mu \to 0} \left(\frac{\frac{\partial}{\partial \mu}\left(\sqrt{(1-\mu)^2(1-\bar{\lambda})^2 + (1-\mu)(1-\bar{\lambda})\mu\sigma} - (1-\mu)(1-\bar{\lambda})\right)}{\frac{\partial}{\partial \mu}(2\mu)} \right) = \frac{\bar{\lambda} + \sigma}{4} \in [0,1] \tag{68}$$

Lastly, we show that there exists a unique $\bar{\lambda} \in [0,1]$ that solves (65).

$$\text{IC}_L \mid_{\bar{\lambda}=0} : \frac{\mu\bar{\gamma}}{\mu\bar{\gamma} + (1-\mu)} - 0 > 0 \tag{69}$$

$$\text{IC}_L \mid_{\bar{\lambda}=1} : 1 - \frac{k}{2} < 0 \tag{70}$$

Thus, $\bar{\lambda}^*$ exists. To show that it is unique, we differentiate (65) with respect to $\bar{\lambda}$:

$$\frac{\partial}{\partial \bar{\lambda}}(IC_L) = \frac{\mu\bar{\gamma}(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} + \frac{\partial \bar{\gamma}}{\partial \bar{\lambda}} \cdot \frac{\mu(1-\bar{\lambda})(1-\mu)}{(\mu\bar{\gamma} + (1-\mu)(1-\bar{\lambda}))^2} - \frac{\delta k}{2} < 0 \tag{71}$$

Since $\bar{\lambda} \leq 2/k$, the following holds.

Thus, a unique equilibrium exists for $\left\{\beta, \bar{\lambda}, \underline{\lambda}, \bar{\gamma}\right\} = \left\{1, \bar{\lambda}^*, 0, \bar{\gamma}^*\right\}$ ■

Figure 2: Market Participation by Investor Type

The left panel below depicts the unconditional probabilities of a speculator's action prior to $t = 2$ (β), as a function of the latency delay δ. The right panel illustrates the market participation choices of liquidity investors, as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

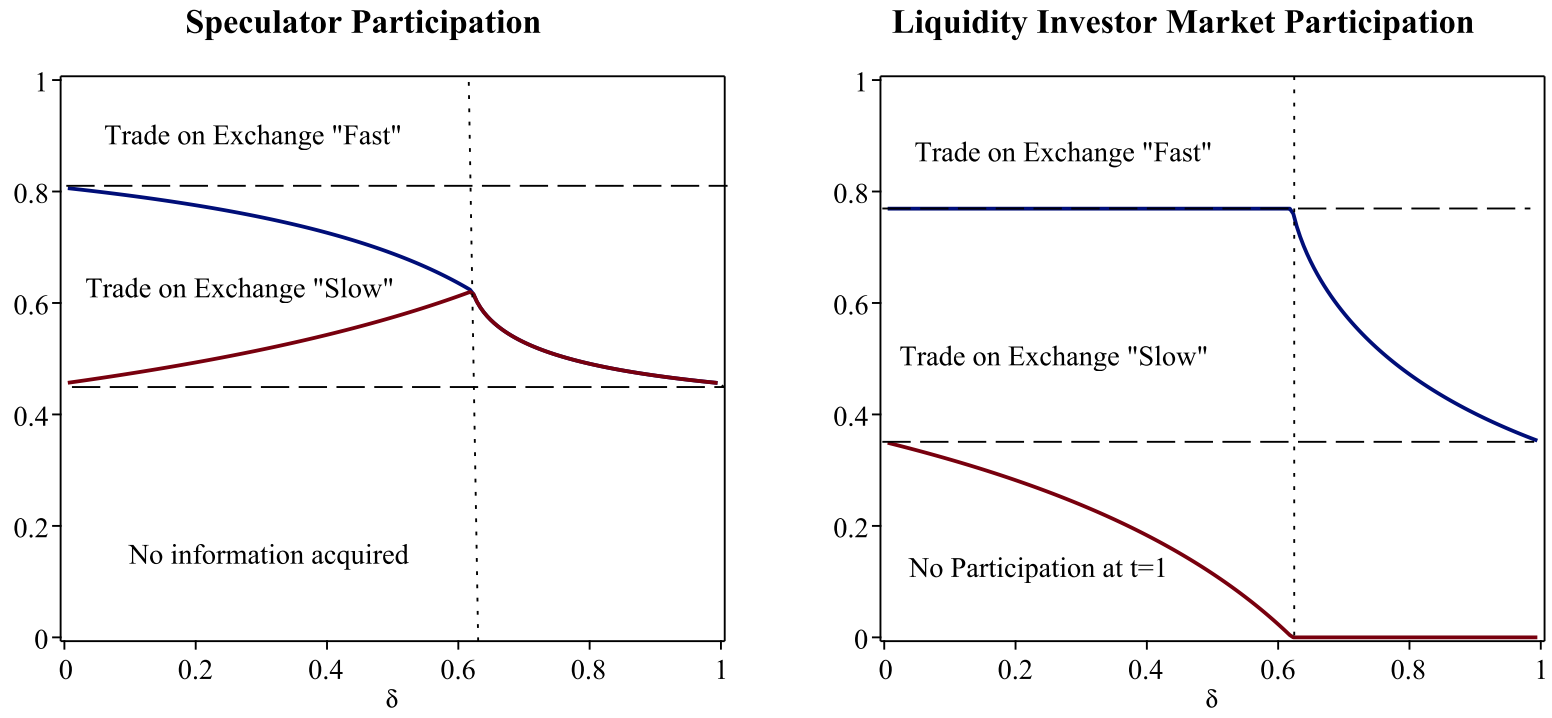


Figure 3: Order Submissions, Trades, and Market Participation

The left panel below depicts total orders submitted and trades executed pre-announcement (prior to $t = 2$), as a function of the Exchange Slow latency delay δ. The right panel illustrates market participation by speculators (μ_I) and liquidity investors (μ_L), as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

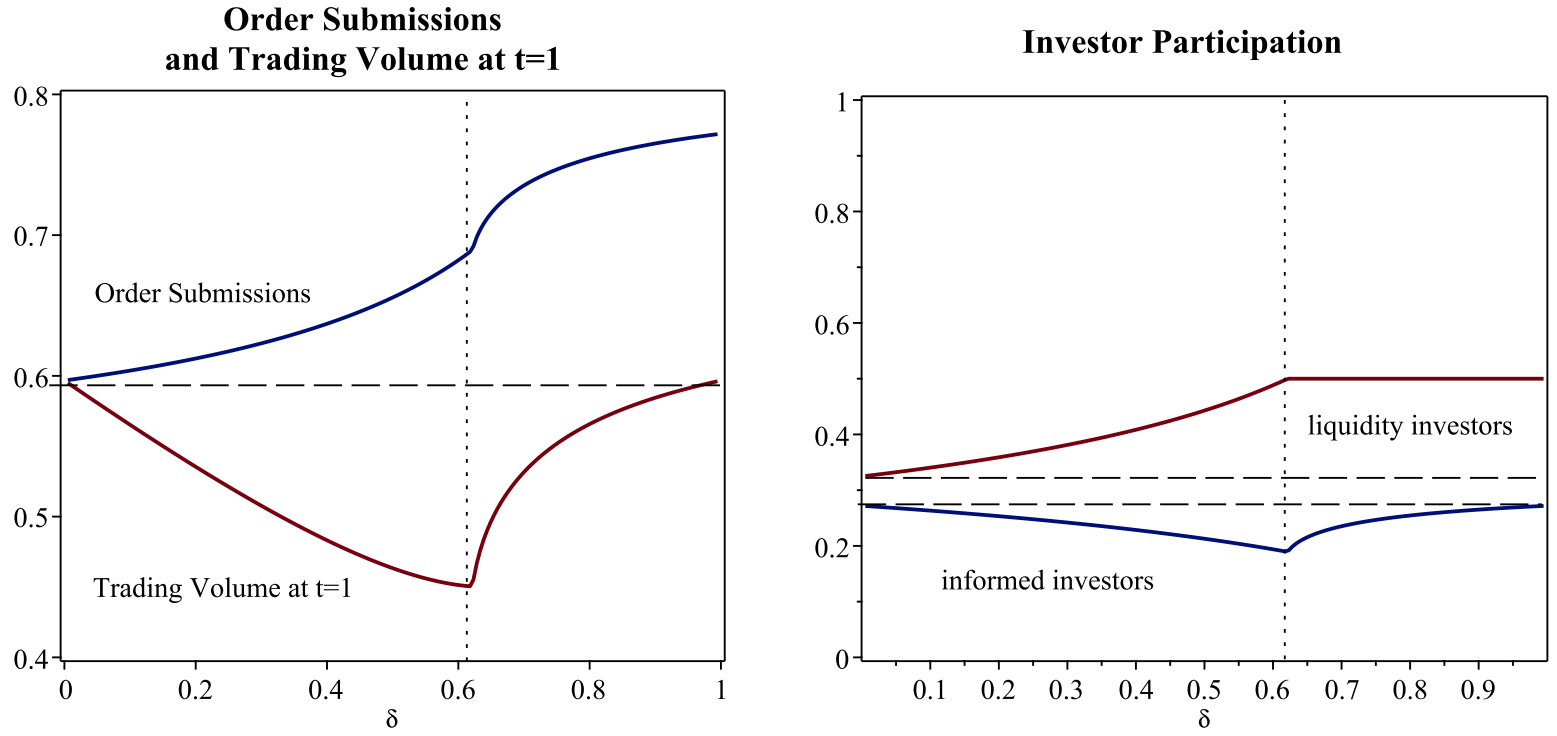


Figure 4: Quoted Spreads and Price Discovery

The left panel below presents the quoted half-spreads for exchanges Fast and Slow at $t = 1$, as a function of the latency delay δ. The right panel depicts price discovery pre-announcement, which we measure as average price movement attributed to informed trades prior to $t = 2$ (the announcement date of v), as a function of the Exchange Slow latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.

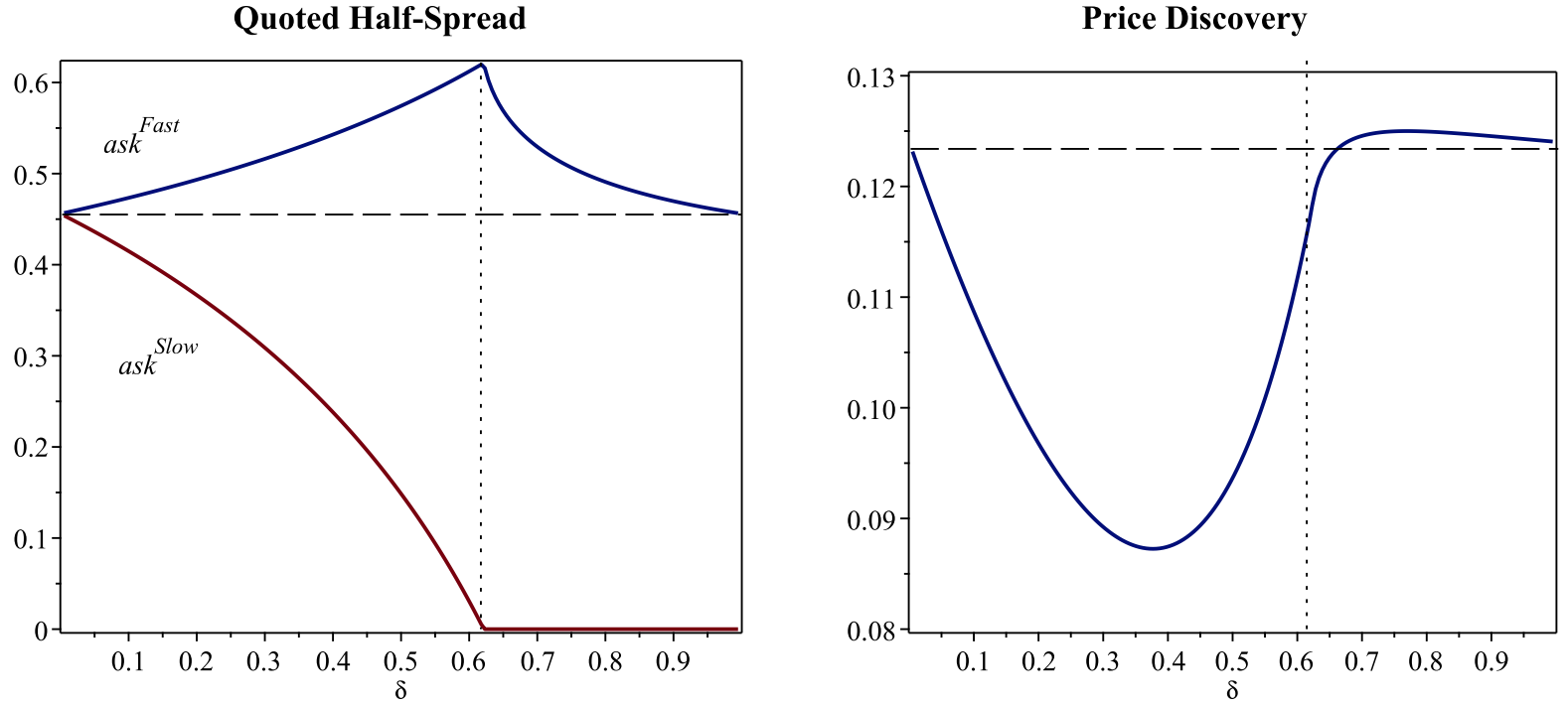
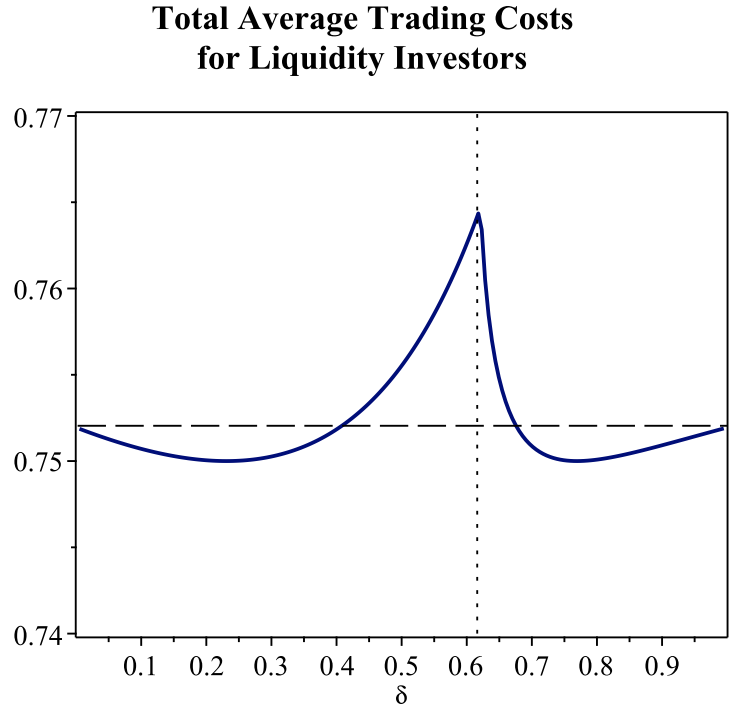
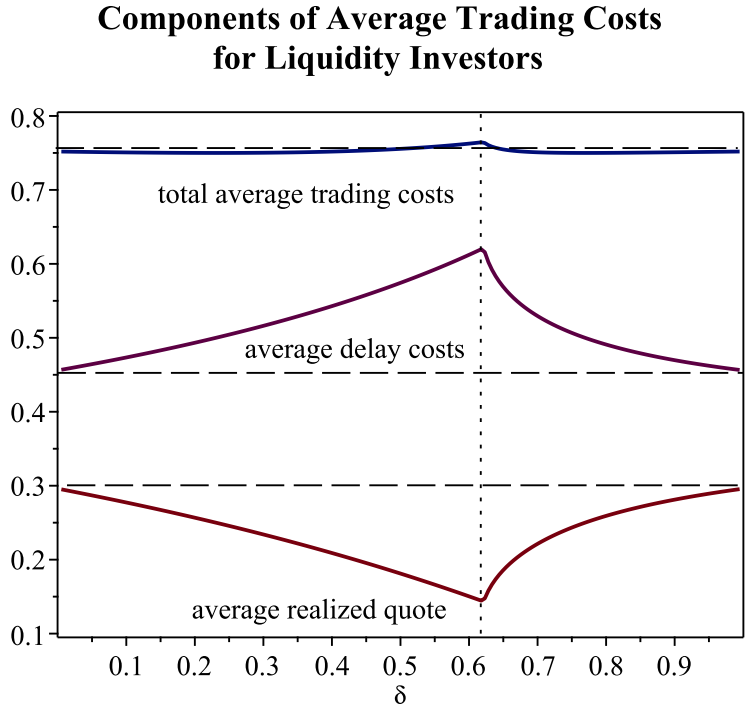


Figure 5: Liquidity Investor Trading Costs

The left panel below illustrates the average trading costs paid by a liquidity investor who enters the market at $t = 0$. In the right panel, we present the trading costs due to delay and the trading costs due to realized quotes separately, as well as the aggregation (from the left panel). We present these costs as a function of the latency delay δ. A vertical dashed line marks δ^*: for all $\delta > \delta^*$, informed investors use only Exchange Fast. Horizontal dashed lines mark values for the benchmark case. Parameter $\mu = 0.5$ and $k = 2.6$. Results for other values of μ and k are qualitatively similar.





A Model of Multi-Frequency Trade*

Nicolas Crouzet, Ian Dew-Becker, and Charles G. Nathanson

March 5, 2017

Abstract

We develop a noisy rational expectations model of financial trade featuring investors who acquire information and trade at a range of different frequencies. In the model, a restriction on high-frequency trading affects efficiency of prices at high frequencies, but leaves low-frequency efficiency unaffected. In a particular equilibrium of the model, traders specialize into trading at individual frequencies. We show that high- and low-frequency investors coexist, trade with each other, and make money from each other. The model matches numerous basic features of financial markets: investors endogenously specialize into strategies distinguished by frequency; volume is disproportionately driven by high-frequency traders; and the portfolio holdings of informed investors forecast returns at the same frequencies as those at which they trade.

Investors in financial markets follow many different strategies, including value investing, technical analysis, macro strategies, and algorithmic trading. These strategies differ in two salient ways. First, they require investors to learn about different aspects of asset prices; market-makers or algorithmic traders care more about the high-frequency movements of prices, while value investing puts more emphasis on their slow-moving features. These investors all understand that their information sets may not overlap, and yet they trade with each other, presumably making some money in the process. Second, these strategies differ in the frequency at which they require investors to trade, or equivalently the rate at which they turn over their positions.

This paper proposes an equilibrium model in which investors endogenously specialize in acquiring information and trading at different frequencies. There is a single fundamentals process, and a continuum of investors who trade forward contracts on the fundamental. These investors also learn about different aspects of asset dynamics. An example of the fundamentals process is the spot price of oil: investors are able to acquire information that tells them about the future path of oil prices, allowing them to potentially earn profits on the forward contracts. As is common elsewhere, in order to grease the wheels of the market, we assume that investors trade against an exogenous flow of demand for forward contracts that fluctuates stochastically over time.

We show that in such a model, there exists a natural (though not necessarily unique) equilibrium in which individual investors endogenously choose to focus on specific frequencies of the

*Crouzet: Northwestern University. Dew-Becker: Northwestern University and NBER. Nathanson: Northwestern University. We appreciate helpful comments from Stijn van Nieuwerburgh.

fundamentals. Some investors learn about low-frequency aspects of oil prices in the sense that they get a signal about their average path over, say, a period of decades, while others learn about higher-frequency behavior, receiving a signal about how oil prices vary from day to day or month to month. This occurs despite the fact that the learning technology is fully general, and in no way tilts investors towards frequency specialization ex-ante.

Given attention allocation – what aspect of fundamentals investors choose to learn about – in equilibrium we show that their positions fluctuate at the frequency at which they receive signals. That is, investors who learn about long-run fundamentals hold positions in forward contracts that fluctuate slowly over time, whereas those who do high-frequency research have positions that vary at high frequencies. So we have a model in which people choose to learn about high- or low-frequency aspects of fundamentals, and that learning causes them to endogenously become high- or low-frequency traders.

While there is other research on investors who trade at different frequencies, that work typically endows investors with investment horizons that differ exogenously.[1] In our setting, all investors have the same objective, maximizing utility over identical horizons. We view this as an important restriction in our setting because it is obviously not the case that people who trade at high frequencies, e.g. turning over their portfolios once per day, really have investment horizons of only 24 hours. Rather, all investors want to maximize the same utility function over wealth, they just go about it in different ways.

What is particularly interesting about the equilibrium that we obtain is that it is not the case that the informed investors trade only with the exogenous demand (i.e. liquidity traders). In fact, high- and low-frequency traders trade with each other. The simple reason is that a high-frequency trader cannot distinguish uninformative demand shocks from the orders of informed low-frequency traders (and vice versa). So in periods when fundamentals are persistently strong, low-frequency traders tend to hold persistently more forward contracts than high-frequency traders and earn profits from them. Similarly, if there is a very transitory increase in fundamentals, the high-frequency traders tend take advantage and earn profits while the low-frequency traders lose, as they ignore the temporary trading opportunity. In that sense, then, everybody is a noise trader sometimes, and they all understand that, but they still participate and make money on average.

The model has a number of predictions for observable features of financial markets. First, as we have already discussed, it predicts that there are traders who can be distinguished by the frequencies at which their asset holdings change over time, and they do research about fundamentals at the same frequencies. So we obtain endogenous high- and low-frequency traders with a specific prediction for how research aligns with trade.

The model also matches salient facts about differences in volume across investors. We can

[1] See, e.g., Amihud and Mendelson (1986), who assume that investors are forced to sell after random periods of time; Hopenhayn and Werner (1996), who assume that investors vary in their rates of pure time preference, and Defusco, Nathanson, and Zwick (2016) who assume that there are sets of investors who are exogenously forced to sell at determinstic horizons that vary across groups. Turley (2012), like us, studies a setting in which investors endogenously choose to learn about high- or low-frequency information, though he studies only a two-period case.

very easily show analytically that high-frequency investors account for a fraction of aggregate volume that is out of proportion to their fraction of total asset holdings. An interesting implication of that result is that incorporating trading costs into the model can have substantial effects on optimal information acquisition strategies. Since high-frequency trade requires paying much larger transaction costs than low-frequency trade, any trading costs cause prices to naturally be less efficient and for there to be less liquidity at high than low frequencies.[2] Moreover, as transaction costs fall, we expect to see a shift towards higher-frequency trade and for prices to become more efficient at high frequencies (see also Turley (2012)).

The model is fundamentally about differences in information across investors. People obtain information in order to make money, and so their asset holdings in general should forecast returns. We see that both in the model and in the data.[3] But different investors' holdings do not forecast returns in the same way. The frequency at which an investor's portfolio holdings forecasts returns is the same as the frequency at which they trade: high-frequency investors' positions forecast returns at very short horizons, while buy-and-hold investors' portfolios forecast returns over much longer periods.

The idea that an investors' asset holdings should forecast returns over a period related to how long those assets will be held is perhaps not surprising. Studies of the holdings of mutual funds and other institutional investors typically examine returns over a period of perhaps 3–12 months. At the other extreme, Brogaard, Hendershott, and Riordan (2014) show that the holdings of high-frequency traders (as defined by NASDAQ) forecast returns over periods of 1–5 seconds – a horizon 7 orders of magnitude smaller than a calendar quarter.

To empirically test our model, we provide novel evidence on the relationship between turnover and asset return predictability. Using form 13F data on institutional asset holdings, we first show that asset turnover within funds is highly persistent over time, suggesting that it is a salient feature of investor strategies. Next, after confirming past results that institutional holdings predict returns, we show that the predictive power of the holdings of high-turnover funds decays much more quickly than those of low-frequency funds, consistent with the model.

Finally, we use the model to study the effects of a policy that restricts high-frequency trade. Such a policy has the obvious effect of reducing the informativeness of prices at high frequencies, but it has no effect at low frequencies. The practical implication is that while prices at any single moment contain less information than without the policy, moving averages of prices remain almost equally informative about moving averages of dividends. So to the extent that economic decisions are made based on an average of prices over time, rather than a price at a single moment, the model implies that restricting high-frequency trade will not reduce the information available for those decisions.

[2] Gârleanu and Pedersen (2013) also discuss how high-frequency information is less valuable in the presence of trading costs, while Dávlia and Parlatore (2016) study in a related setting how trading costs can affect information acquisition, but without our focus on differences across frequencies.

[3] See, e.g., the literature on the predictive power of mutual fund and institutional investor asset holdings for future returns, such as Carhart (1997) and Yan and Zhang (2009), among many others.

To summarize, we develop a model that matches a number of major features of trade in financial markets: investors can be distinguished by the frequencies at which they trade; volume is accounted for by high-frequency traders; and the holdings of investors forecast returns at horizons similar to their holding periods. The model can then be used to analyze the effects of restricting trade at specific frequencies.

In general models with dynamic trade are extremely difficult to solve; solutions typically require some kind of restriction, such as to a very narrow class of driving processes (e.g. AR(1) and Ornstein–Uhlenbeck processes studied in Wang (1993, 1994) and He and Wang (1995)), or to only two or three period horizons. We allow for a very long investment horizon and place only technical constraints on the fundamental processes driving the model and obtain fully analytic solutions. The sacrifice that we make is that information sets are fixed on date 0 – investors only obtain signals once. The model should be thought of as essentially a stationary equilibrium: it gives a steady-state description of trade, volume, and returns. It is not well suited to studying how investors and markets respond to shocks to information sets.

The major advantage of our particular information structure is that it allows us to take a long-horizon dynamic model and solve it as a series of parallel scalar problems. In particular, solving our model is only marginally more difficult than solving a standard single-period/single asset noisy rational expectations model – it reduces to a parallel set of such equilibria. The paper thus has useful methodological contributions for analyzing models of trade over time.

This paper builds on a growing recent literature that tries to understand optimal information acquisition in financial markets. The most important building blocks are the models of van Nieuwerburgh and Veldkamp (2010) and Kacperczyk, van Nieuwerburgh, and Veldkamp (2016) in that we use a highly similar information and market structure and build on their results on optimal information acquisition (their reverse water-filling solution, in particular). Those papers themselves build critically on work by Grossman and Stiglitz (1980), Hellwig (1980), Diamond and Verrecchia (1981), and Admati (1985) on rational expectations equilibria. More recently, research has tried to understand the effects on the equilibria developed in those earlier papers of various limits on information gathering ability (e.g. Banerjee and Green (2015) and Dávila and Parlatore (2016)).

There is also a literature on price dynamics in rational expectations equilibria, though it is relatively small given how difficult dynamic models are to solve. In particular, a series of papers by Wang (1993, 1994) and He and Wang (1995) study the implications of dynamic equilibria on prices and volume. Those papers are based around AR(1) or Ornstein–Uhlenbeck-type dynamics to maintain tractability (see also Wachter (2002)), whereas we study a setting in which the various exogenous time series may follow processes with minimally constrained autocorrelations. Furthermore, we focus on how investment strategies differ across investors, whereas those papers focus on symmetric strategies. A number of papers also study overlapping generations models, which can help alleviate some of the difficulties with dynamic trade.[4]

There is also a large literature on disagreement in financial markets. In addition to the above

[4]See Spiegel (1998), Watanabe (2008), and Banerjee (2011), among others.

work, see, e.g., Townsend (1983), Basak (2005), Hong and Stein (2007), and Banerjee and Kremer (2010), who focus, like us, on dynamics. In our setting, disagreement arises not just because agents receive signals that have random errors, but also because their signals have different relationships with fundamentals. High-frequency and low-frequency investors will often disagree about the price path of the asset over time because they learn about different characteristics of fundamentals – the path over the next few minutes, say, versus the path over the next several years.

Our desire to develop a model that can match salient features of the cross-section of investment strategies follows from a large empirical literature that documents the behavior of many different types of investors and how it affects the aggregate behavior of financial markets. Chen, Jegadeesh, and Wermers (2000), Gompers and Metrick (2001), Nagel (2005), Griffin and Xu (2009), Yan and Zhang (2009), and Brogaard, Hendershott, and Riordan (2014), for example, study the behavior of institutional investors and how their holdings relate to asset returns. Turley (2012), Bai, Philippon, and Savov (2016), and Weller (2016) study how price informativeness has changed over time and how it is affected by trading costs and the number of investors who trade at different frequencies.

Finally, our work is related to a small literature that studies the properties of asset returns and portfolio choice in the frequency domain including Bandi and Tamoni (2014), Chinco and Ye (2016), Chaudhuri and Lo (2016), and Dew-Becker and Giglio (2016).

The remainder of the paper is organized as follows. Section 1 describes the basic environment, and we solve for optimal information acquisition in section 2. Section 3 examines the implications of the model for the behavior of individual investors in a setting that features investors who specialize in trade at a particular frequency. Section 4 presents empirical evidence on the behavior of institutions consistent with out model of specialization. Finally, section 5 presents our key results on the effects of restrictions on high-frequency trade on return volatility and price efficiency at different frequencies, and section 6 concludes.

1 Asset market equilibrium

We begin by describing the basic market structure and the asset market equilibrium. This section introduces the description of trading strategies in terms of frequencies and shows how the frequency transformation makes multi-period investment a purely scalar problem. the problem is solved from the perspective of date 0.

1.1 Market structure

Time is denoted by $t \in \{-1, 0, 1, ..., T\}$, with T even, and we will focus on cases in which T may be treated as large. There is a fundamentals process D_t that investors make bets on with realizations on all dates except -1 and 0. The time series process is stacked into a vector $D \equiv [D_1, D_2, ..., D_T]'$ (variables without subscripts denote vectors) and is distributed as

$$D \sim N(0, \Sigma_D). \tag{1}$$

The fundamentals process is assumed to be stationary, meaning that it has constant unconditional autocovariances. Stationarity implies that Σ_D is Toeplitz (all diagonals are constant), and we further assume that the eigenvalues of Σ_D are finite and bounded away from zero.[5]

On date 0, there is a market for forward claims on D_t for all $t > 0$. A unit mass of investors indexed by $i \in [0, 1]$ meets on date 0 and commits to a set of trades of futures contracts maturing on all dates. P_t denotes the price of a claim to the fundamental D_t.

There is an exogenous supply of futures, Z, which is distributed as

$$Z \sim N(0, \Sigma_Z). \tag{2}$$

Z_t may be thought of as either exogenous liquidity demand or noise trading. The time series process for supply is also assumed to be stationary. For markets to clear, the net demand of the investors for the fundamental on date t must equal Z_t,[6]

$$\forall t : \int_i Q_{i,t} di = Z_t, \tag{3}$$

where $Q_{i,t}$ is the number of date-t forward claims agent i buys.

A concrete example of a potential process D_t is the price of crude oil: oil prices follow some stochastic process and investors trade futures on oil at many maturities. D_t can also be interpreted as the dividend on a stock, in which case P_t is the price of a forward claim on a single dividend.

1.1.1 Modeling equities

While the concept of a futures market on the fundamentals will be a useful analytic tool, we can also obviously price portfolios of futures. We model equity as a claim to the stream of fundamentals over time. To purchase such a claim, one would enter into futures contracts for the fundamental on each date $t + j$. Since futures contracts specify that money only changes hands at maturity, the money that must be set aside on date t for a futures contract that expires at $t + j$ is $P_{t+j}R^{-j}$, where R is the discount rate (which is assumed here to be a constant). The date-t cost of a claim to the entire future stream of fundamentals is therefore

$$P_t^{equity} \equiv \sum_{j=1}^{T-t} R^{-j} P_{t+j} \tag{4}$$

Holding any given combination of futures claims on the fundamental D is therefore also equivalent to holding futures contracts on equity claims (i.e. committing to a trading strategy in equities at prices that are agreed on at date 0). Any desired set of exposures to fundamentals over time

[5] The analysis is similar if a transformation of D_t (e.g. its first difference) is stationary. See appendix section A.

[6] It is also possible to assume that there is an exogenous downward-sloping supply curve of the fundamental that shifts stochastically over time; our results go through similarly. This case is treated as part of the analysis of appendix 6.

can be obtained either through purchases of futures or through suitable trading strategies for the equity claim (assuming prices can be committed to or that they are predetermined, which will be the case in our equilibrium).

Our analysis of pricing will focus on futures as they will give the most direct analog to past work. When we discuss volume and trading costs, though, we will take advantage of the equity-based implementation.

1.2 Information structure

The realization of the time series of fundamentals, $\{D_t\}_{t=1}^T$, can be thought of as a single draw from a multivariate normal distribution. Investors are able to acquire signals about that realization. The signals are a collection $\{Y_{i,t}\}_{t=1}^T$ *observed on date 0* with

$$Y_{i,t} = D_t + \varepsilon_{i,t},\ \varepsilon_i \sim N\left(0, \Sigma_i\right), \tag{5}$$

Information sets in the model are fixed on date 0. Through $Y_{i,t}$, investors can learn about fundamentals potentially arbitrarily far into the future. $\varepsilon_{i,t}$ is a stationary error process in the sense that $cov\left(\varepsilon_{i,t}, \varepsilon_{i,t+j}\right)$ depends on j but not t (again, Σ_i is Toeplitz).

The information structure here is obviously stylized. One interpretation is that we are collapsing to date 0 all the realizations of a stationary process. That is, agents have a machine that gives them signals about fundamentals plus an error, and that machine reports all of its output on date 0. The information structure is meant to generate two important features in the model. First, obviously on any particular date agents can choose to learn about fundamentals on more than just a single date in the future – they can potentially get information about fundamentals in many different periods (e.g. next quarter vs. over the next five years). Second, by restricting $\varepsilon_{i,t}$ to be "stationary", we are forcing agents to choose a fixed policy for information. They build a machine (or a research department) that, rather than yielding information about only a single date, returns information about the entire fundamentals stream over time in a way that places no particular emphasis on any single date.

In choosing Σ_i agents will have two choices to make. First, they will be able to choose how informative their signals are by choosing the variance of ε_i. Second, though, they will be able to choose how accurate the signals are about fundamentals over different horizons. some choices for Σ_i will yield signals that are informative about transitory variation in fundamentals, while others will yield signals that are more useful for forecasting trends.

1.3 Investment objective

All trading decisions are made on date 0. Investors choose demands $\{Q_{i,t}\}_{t=1}^T$ conditional on their observed signals, $\{Y_{i,t}\}_{t=1}^T$, and the set of futures prices, $\{P_t\}_{t=1}^T$. That is, as in past work, agents submit to a central auctioneer demand curves that condition on prices.

We assume that investors have mean-variance utility over cumulative excess returns. Investor i's objective is

$$U_{0,i} = \max_{\{Q_{i,t}\}} E_{0,i}\left[T^{-1}\sum_{t=1}^{T} Q_{i,t}\left(D_t - P_t\right)\right] - \left(\rho T\right)^{-1} Var_{0,i}\left[\sum_{t=1}^{T} Q_{i,t}\left(D_t - P_t\right)\right], \tag{6}$$

where $E_{0,i}$ is the expectation operator conditional on agent i's date-0 information set, $\{P, Y_i\}$. $Var_{0,i}$ is the variance operator conditional on $\{P, Y_i\}$. ρ is risk-bearing capacity per unit of time.

The assumption that all plans are made on date 0 only restricts information sets in a very specific way: it means that investors are not able to condition demand on the *realized* history of fundamentals. That is, it is not free to condition on the history of D_t, even when that history has already been realized. Instead, what agents must condition on is noisy signals about fundamentals, $Y_{i,t}$.

An important implication of that assumption is that agents have no desire to change their investment choices after date 0 since they receive no further information. Agents' trading strategies can thus be equivalently implemented either through a set of purchases of futures contracts or through a dynamic plan for trading the equity claim, as described above.

We interpret the objective as representing a target that an institutional investor might have. Rather than aiming to maximize the discounted sum of returns, as a person who consumes out of wealth might, the investors we study maximize a measure of their performance. The objective can be thought of as representing CARA or quadratic preferences over the sum of excess returns, so it would appear if a manager were paid on date T a fee proportional to total excess returns up to that time. Bhattacharya and Pfleiderer (1985) and Stoughton (1993) also argue that a quadratic contract (which would induce mean-variance preferences) can appear optimally in delegated investment problems. The important characteristic of (6) is that it yields a stationary problem in the sense that there is no discounting to make returns in some periods more important than others.

Finally, note that all investors have the same investment horizon. We show in appendix F that the investment horizon as defined here by T has no effect on information choices in the model – two investors with different T will be equally likely to be high- or low-frequency investors. The simplest way to confirm that fact is to simply note, when we obtain the equilibrium strategies, that T has no effect on the type of information that investors optimally obtain.

1.4 Equilibrium

Conditional on the information choices of the agents – that is, taking the set of Σ_i (which may differ across agents) as given – we study a standard asset market equilibrium.

Definition 1 *An asset market equilibrium is a set of demand functions, $Q_i\left(P, Y_i\right)$, and a vector of prices, P, such that investors maximize utility, $U_{0,i}$, and all markets clear, $\int_i Q_{i,t} di = Z_t \ \forall t$.*

The equilibrium concept is that Grossman and Stiglitz (1980), Hellwig (1980), Diamond and Verrecchia (1981), and Admati (1985). Investors submit demand curves for each futures contract to a Walrasian auctioneer who selects equilibrium prices to clear all markets.

The structure is in fact mathematically that of Admati (1985), who studies investment across a set of assets that might represent stocks in different companies, and the solution from that paper applies directly here. Here we are considering investment across a set of futures contracts that represent claims on some fundamentals process across different dates. We simply rotate the Admati (1985) structure from a cross-section to a time series.

1.5 Trading frequencies

This paper is fundamentally about the behavior of markets at different frequencies, so we need a rigorous concept of what frequencies are. We use the fact that fluctuations at different frequencies represent an (asymptotic) orthogonal decomposition of any time series.

Define a set of vectors of cosines and sines at the fundamental frequencies $\omega_j = 2\pi j/T$ for $j \in \{0, 1, ..., T/2\}$

$$c_j \equiv \sqrt{2/T}\left[\cos\left(2\pi j\left(t-1\right)/T\right); t = 1, 2, ..., T\right]' \tag{7}$$

$$s_{j'} \equiv \sqrt{2/T}\left[\sin\left(2\pi j\left(t-1\right)/T\right); t = 1, 2, ..., T\right]'. \tag{8}$$

A cycle at frequency ω_j has an associated wavelength $2\pi/\omega_j$. $\omega_0 = 0$ thus corresponds to an infinite wavelength, or a permanent shock (a constant vector). ω_1 corresponds to a cycle that lasts as long as the sample – c_1 is a single cycle of a cosine. $\omega_{T/2} = \pi$, the highest frequency, corresponds to a cycle that lasts two periods, so that $c_{T/2}$ oscillates between $\pm\sqrt{2/T}$.

The frequency-domain counterpart to the vector of fundamentals, D, is then

$$d = \Lambda' D \tag{9}$$

$$\text{where } \Lambda \equiv \left[c_0/\sqrt{2}, c_1, ..., c_{T/2}/\sqrt{2}, s_1, s_2, ..., s_{T/2-1}\right], \tag{10}$$

we use the notation $d_j = c_j' D$ and $d_{j'} = s_j' D$ to refer to fundamentals at particular frequencies. When the distinction is necessary, we use the notation j to refer to a frequency associated with a cosine transform and j' to refer to one with a sine transform. In what follows, lower-case letters denote frequency-domain objects. Note that Λ is orthonormal with $\Lambda' = \Lambda^{-1}$.

Since d is a linear function of D, it can be thought of as a vector of payoffs on portfolios of futures given by Λ – portfolios with weights on D_t that fluctuate over time as sines and cosines.

For our purposes, the key feature of Λ is that it approximately diagonalizes *all* Toeplitz matrices and thus orthogonalizes stationary time series.[7]

[7] This is a textbook result that appears in many forms, e.g. Shumway and Stoffer (2011). Brillinger (1981) and Shao and Wu (2007) give similar statements under weaker conditions.

Definition 2 *f_X is the* spectrum *of X with elements $f_{X,j}$, defined as*

$$f_{X,j} \equiv \sigma_{X,0} + 2\sum_{s=1}^{T-1} \sigma_{X,j} \cos(\omega_j s) \tag{11}$$

$$f_X \equiv \left[f_{X,0}, f_{X,1}, ...f_{X,T/2}, f_{X,1}, f_{X,2}, ..., f_{X,T/2-1}\right]'. \tag{12}$$

Lemma 1 *For a stationary time series $X_t \sim N(0, \Sigma_X)$ with autocovariances $\sigma_{X,j} \equiv cov(X_t, X_{t-j})$,*

$$x \equiv \Lambda' X \Rightarrow N(0, diag(f_X)) \tag{13}$$

where $\Rightarrow$ denotes convergence in the sense that

$$\left|\Lambda' \Sigma_X \Lambda - diag(f_X)\right| \leq c_X T^{-1/2} \tag{14}$$

for a constant c_X and for all T.[8] *$diag(f_X)$ is a matrix with the vector f_X on its main diagonal and zero elsewhere.*

Proof. This is a textbook result (e.g. Brockwell and Davis (1991)). See appendix B for a derivation specific to our case. ■

For any finite horizon, the matrix Λ does not exactly diagonalize the covariance matrix of D. But as T grows, the error induced by ignoring the off-diagonal elements of the covariance matrix $\Lambda' D$ becomes negligible (it is of order $T^{-1/2}$), and x is well approximated as a vector of independent random variables.[9] The spectrum of X, f_X, measures the variance in X coming from fluctuations at each frequency. It also represents an approximation to the eigenvalues of Σ_X.[10]

To see why this lemma is useful, consider the vector of fundamentals in the frequency domain, $d = \Lambda' D$. Given that $D \sim N(0, \Sigma_D)$, where Σ_D is Toeplitz, we have

$$\Lambda' D = d \Rightarrow N(0, diag(f_D)). \tag{15}$$

Λ thus approximately diagonalizes the matrix Σ_D, meaning that the elements of d – the fluctuations in fundamentals at different frequencies (with both sines and cosines) are jointly asymptotically independent. Moreover, the same matrix Λ asymptotically diagonalizes the covariance matrix of *any* stationary process. That result will allow us to massively simplify the study of investment over

[8] Two technical points may be noted here. First, as a technical matter, the spectrum f_X must be extended as T grows. A simple way to do that is to suppose that there is a true process for X with a spectrum that is a continuous function f_X, and in any finite sample of length T, there is then an associated spectrum $f_{X,T}$ defined in (11). The second point is that the constant c_X is then a function of that true spectrum f_X; the appendix elaborates on that fact.

[9] For all the stationary processes studied in the paper, we assume that the autocovariances are summable in the sense that $\sum_{r=1}^{\infty} |j\sigma_{X,j}|$ is finite (which holds for finite-order stationary ARMA processes, for example).

[10] f_X represents an approximation to the eigenvalues only in the sense that $\Lambda' \Sigma_X \Lambda \approx diag(f_X)$. Providing a sense in which f_X is actually close to the true eigenvalues of Σ_X is a subtler problem that we do not address here. The specific result in lemma 1 is all that we actually need for our results.

many horizons. It says that set of orthogonal factors underlying all stationary processes is (nearly) the same.

1.6 Market equilibrium in the frequency domain

The approximate diagonalization induced by Λ allows us to solve the model through a series of parallel scalar problems that can be easily solved by hand. Using the asymptotic approximation that d and z are independent across frequencies (and across sines and cosines), we obtain the following frequency-by-frequency solution to the asset market equilibrium.[11]

Solution 1 *Under the approximations* $d \sim N(0, diag(f_D))$ *and* $z \sim N(0, diag(f_Z))$*, the prices of the frequency-specific portfolios,* p_j*, satisfy, for all* j, j'

$$p_j = a_{1,j} d_j - a_{2,j} z_j \tag{16}$$

$$a_{1,j} \equiv 1 - \frac{f_{D,j}^{-1}}{\left(\rho f_{avg,j}^{-1}\right)^2 f_{Z,j}^{-1} + f_{avg,j}^{-1} + f_{D,j}^{-1}} \tag{17}$$

$$a_{2,j} \equiv \frac{a_{1,j}}{\rho f_{avg,j}^{-1}} \tag{18}$$

$$\text{where } f_{avg,j}^{-1} \equiv \int_i f_{i,j}^{-1} di \tag{19}$$

where p_j*,* d_j*, and* z_j *represent the frequency-*j *components of prices, fundamentals, and supply, respectively.* f_i *is the spectrum of the matrix* Σ_i*. See appendix C for the derivation.*

The price of the frequency-j portfolio depends only on fundamentals and supply at that frequency. As usual, the informativeness of prices, through $a_{1,j}$, is increasing in the precision of the signals that investors obtain, while the impact of supply on prices is decreasing in signal precision and risk tolerance. The frequency domain analog to the usual demand function is

$$q_{i,j} = \rho \frac{E[d_j - p_j \mid y_{i,j}, p_j]}{Var[d_j - p_j \mid y_{i,j}, p_j]}. \tag{20}$$

These solutions for the prices and demands are the standard results for scalar markets. What is novel here is that the choice problem refers to trades over time. p_j is the price of a portfolio whose exposure to fundamentals fluctuates over time at frequency $2\pi j/T$. Both prices and demands at frequency j depend only on signals and supply at frequency j – the problem is completely separable across frequencies.

The appendix shows that the frequency domain solution provides a close approximation to the true solution in the time domain. Specifically, the true time domain solution from Admati (1985)

[11] A simple way to see where this solution comes from is to note that, under the asymptotic approximation, $\Lambda' A_1 \Lambda$ and $\Lambda' A_2 \Lambda$ from the Admati solution can be written purely in terms of diagonal matrices, for which addition, multiplication, and inversion are simply scalar operations on the main diagonal.

(with no approximations) can be written as

$$P = A_1 D - A_2 Z \tag{21}$$

for a pair of matrices A_1 and A_2 defined in the appendix that are complicated matrix functions of Σ_Z, Σ_D, and the precisions of the signals agents obtain.

Proposition 1 *The difference between calculating the prices directly in the time domain using the Admati (1985) solution in the time domain and rotating the frequency domain solution back into the time domain is small in the sense that*

$$\begin{aligned} \left|A_1 - \Lambda diag\left(a_1\right)\Lambda'\right| &\leq c_1 T^{-1/2} & (22)\\ \left|A_2 - \Lambda diag\left(a_2\right)\Lambda'\right| &\leq c_2 T^{-1/2} & (23) \end{aligned}$$

for constants c_1 and c_2. Furthermore, while prices and demands are stochastic, the time- and frequency-domain solutions are related through an even stronger result

$$\begin{aligned} e_{\max}\left[Var\left(\Lambda p - P\right)\right] &\leq c_P T^{-1/2} & (24)\\ e_{\max}\left[Var\left(\Lambda q_i - Q_i\right)\right] &\leq c_Q T^{-1/2} & (25) \end{aligned}$$

where the operator $e_{\max}[\cdot]$ denotes the maximum eigenvalue of a matrix (that is, the operator norm), for constants c_P and c_Q.

In other words, among portfolios whose squared weights sum to 1, the maximum variance of the pricing and demand errors – the difference between the truth from the time domain solution and the frequency-domain approximation that assumes that Λ diagonalizes the covariance matrices – is of order $T^{-1/2}$ (that is, the bound holds for *any* portfolio of futures, not just the frequency- or time-domain claims). We note also that these are not limiting results – they are true for all T.

Result 1 shows that for large T, the standard time-domain solution for stationary time series processes becomes arbitrarily close to a simple set of parallel scalar problems in the frequency domain. The time domain solution is obtained from the frequency domain solution by premultiplying by Λ.

2 Optimal information choice

We now model a constraint on information acquisition and characterize optimal strategies. The objective, constraint, and solution are drawn from van Nieuwerburgh and Veldkamp (2009) and Kacperczyk, van Nieuwerburgh, and Veldkamp (2016; KvNV). Our analysis follows theirs closely, except that we are studying a time-series model and a frequency transformation. Whereas KvNV study a symmetric equilibrium in which all investors follow the same information acquisition strategy, we will subsequently argue for the relevance of a separating equilibrium in our setting.

2.1 Objective

Following KvNV, we assume that investors choose information to maximize the expectation of their mean-variance objective (6) subject to a linear constraint on total precision:

$$\max_{\{f_{i,j}\}} E_{-1}\left[U_{i,0} \mid \Sigma_i^{-1}\right] \text{ such that } T^{-1} tr\left(\Sigma_i^{-1}\right) \leq \bar{f}^{-1} \tag{26}$$

where E_{-1} is the expectation operator on date -1, i.e. prior to the realization of signals and prices (as distinguished from $E_{i,0}$, which conditions on P and Y_i). The trace function $tr\left(\Sigma_i^{-1}\right)$ measures the total cost of acquiring a private signal with precision matrix Σ_i^{-1} and is subject to the bound $\bar{f}^{-1}$.[12] This cost function is also equal to the sum of the eigenvalues of the precision matrix. Since the eigenvalues represent the precisions of the orthogonalized signals, it can be thought of as measuring the total precision of the independent parts of the signals. Moreover, since the trace operator is invariant under rotations, this measure of information is invariant to the domain of analysis, time or frequency.[13] That is,

$$T^{-1} tr\left(\Sigma_i^{-1}\right) = T^{-1} \sum_{j,j'} f_{i,j}^{-1} \tag{27}$$

The information constraint is linear in the frequency-specific precisions. Investors also face the constraint that $f_{i,j} = f_{i,j'}$, which ensures that the variance matrix of ε_i is symmetric and Toeplitz.[14]

The appendix shows that, given the optimal demands, an agent's expected utility is linear in the precision they obtain at each frequency.

Lemma 2 *Under the frequency domain representation, when informed investors optimize, each investor's expected utility may be written as a function of their own precisions, $f_{i,j}^{-1}$, and the average across other investors, $f_{avg,j}^{-1} \equiv \int_i f_{i,j}^{-1} di$, with*

$$E_{-1}\left[U_{0,i} \mid \{f_{i,j}\}\right] = \frac{1}{2} T^{-1} \sum_{j,j'} \lambda_j\left(f_{avg,j}^{-1}\right) f_{i,j}^{-1} + constants \tag{28}$$

$\lambda_j(x)$ is a function determining the marginal benefit of information at each frequency with the properties $\lambda_j(x) > 0$ and $\lambda_j'(x) < 0$ for all $x \geq 0$. The fact that $\lambda_j' < 0$ says that the marginal benefit to an investor of allocating attention to frequency j is decreasing in the amount of attention that other investors allocate to that frequency – attention decisions are strategic substitutes. If

[12] Our main analysis considers the case where signals about fundamentals are costly but investors can condition on prices freely. Appendix I considers a case where it is costly to condition expectations on prices and shows that model's predictions results go through similarly with the caveat that investors never choose to become informed about prices, as in Kacperczyk, van Nieuwerburgh, and Veldkamp (2016).

[13] This result relies on the approximation $\Sigma_i \approx \Lambda' diag(f_i) \Lambda$.

[14] KvNV show that the solution of the optimal attention allocation problem (26) are identical if one assumes that the cost of information is measured by the entropy of the investor's signals, which corresponds to the function $\ln\left|\Sigma_i^{-1}\right| \approx \sum_{j,j'} \log f_{i,j}^{-1}$. The key feature of the two cost functions is that they are non-convex in precision.

$f_{avg,j}^{-1}$, the average precision of the signals obtained by other agents, is high, then prices are already efficient at frequency j, so there is little benefit to an investor from learning about that frequency.

The frequency-domain transformation is what allows us to write utility as a simple sum across frequencies. An investor's utility depends additively on the amount of information that they obtain at each frequency. In the time domain, utility is a complicated function of matrices.

2.2 Characterizing the optimum

The critical feature of (28) is that expected utility is linear in the set of precisions that agent i chooses, $\left\{f_{i,j}^{-1}\right\}$. Since the both the objective (28) and the constraint (27) are linear in the choice variables, it immediately follows that agents either allocate all attention to a single frequency, or that they are indifferent between allocating attention across some subset of the frequencies. We then obtain the following solution for attention allocation.

Solution 2 *Information is allocated so that*

$$f_{avg,j}^{-1} = \begin{cases} \lambda_j^{-1}\left(\bar{\lambda}\right) & \text{if } \lambda_j\left(0\right) \geq \bar{\lambda} \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

where $\bar{\lambda}$ *is obtained as the solution to*

$$T^{-1} \sum_{j,j':\lambda_j^{-1}\left(\bar{\lambda}\right)>0} \lambda_j^{-1}\left(\bar{\lambda}\right) = \bar{f}^{-1}. \tag{30}$$

This is the reverse water-filling solution from KvNV. While it may appear mathematically complicated, the intuition is simple: investors allocate attention to signals in such a way that the marginal benefit is equalized to the extent possible across frequencies. It is impossible to allocate negative attention, though, so if the marginal benefit of paying attention to a particular frequency, $\lambda_j\left(0\right)$, is below the cutoff $\bar{\lambda}$, then $f_{i,j}^{-1} = 0$ there for all investors.

The intuition is easiest to develop graphically. Figure 1 plots the functions $\lambda_j\left(0\right)$ and $\lambda_j\left(f_{avg,j}^{-1}\right)$ across frequencies ω_j, where

$$\lambda_j\left(0\right) = f_{D,j}\left(1 + \rho^{-2} f_{D,j} f_{Z,j}\right). \tag{31}$$

The initial marginal benefit of allocating attention is increasing in the amount of fundamental information and the volatility of supply.

The details of the calibration are reported in appendix J. What is important here is simply that $\lambda_j\left(0\right)$ has peaks at low, middle, and high frequencies. Those are the frequencies at which D_t or Z_t is more volatile, so there is more information to potentially be gathered and a larger reward for doing so. For a given value of $\bar{f}^{-1}$, $\lambda_j\left(f_{avg,j}^{-1}\right)$ is a flat line for all j such that $\lambda_j\left(0\right) \geq \bar{\lambda}$. Those are the frequencies that investors learn about. The term "reverse water filling" refers here to the idea that the curve $\lambda_j\left(0\right)$ is inverted and one pours water into it. $\bar{\lambda}$ is then the level of the water's

surface.[15] As the information constraint is relaxed, $\bar{\lambda}$ falls and potentially more frequencies receive attention.

Given the calibration, we see that there are investors acquiring information in three disconnected ranges of frequencies. At the places where $\lambda_j(0)$ is farther above $\bar{\lambda}$, there is more information acquisition, whereas the locations where $\lambda_j(0) = \bar{\lambda}$ are marginal in the sense that they are the next to receive attention if $\bar{\lambda}$ falls.

Another way to interpret the results is to observe the following

Result 1 *The return at frequency j has variance*

$$Var[r_j] = \lambda_j\left(f_{avg,j}^{-1}\right) \tag{32}$$

$$where\ r_j \equiv d_j - p_j. \tag{33}$$

The marginal benefit of acquiring information at a particular frequency is exactly equal to the unconditional variance of returns at that frequency. When returns have high variance, there are potentially large profits to be earned from acquiring information. When returns have zero variance, on the other hand, prices are already perfectly informative, so there is no reason to study fundamentals at such a frequency. So agents desire to learn at the frequencies where returns are most volatile.

The solution derived here characterizes aggregate information acquisition – the sum of the precisions obtained by all the agents at each frequency – but it does not describe exactly what strategy each agent follows; and in fact there are infinitely many strategies for individual investors consistent with the aggregate solution. We now examine one particular solution that leads to the existence of traders who can be characterized by their trading frequencies.

3 Specialization

3.1 The separating equilibrium

Given the assumptions we have made so far, the only restrictions on information allocation are those that ensure that the information allocation condition (29) holds. There are numerous equilibria with that characteristic, though. KvNV focus on the symmetric equilibrium in which all investors allocate their attention in proportion to $f_{avg,j}^{-1}$ at each frequency. There are also many asymmetric and mixed-strategy equilibria.

Since one of our goals is to understand the potential existence and behavior of high-and low-frequency treaders, we now focus on equilibria in which all investors learn about only a single frequency. Specifically, we assume that for every agent i, there is a frequency j_i^* such that $f_{i,j_i^*}^{-1} =$

[15] Again, each frequency (except 0 and π) has an associated sine and cosine. The same amount of precision is required to be allocated to both the sine and cosine at each frequency.

$\bar{f}^{-1}/2$, and $f_{i,j}^{-1} = 0$ for all other j

$$f_{i,j} = f_{i,j'} = \begin{cases} \bar{f}^{-1}/2 \text{ if } j = j_i^* \text{ and } j_i^* \notin \{0, T/2\} \\ \bar{f}^{-1} \text{ if } j = j_i^* \text{ and } j_i^* \in \{0, T/2\} \\ 0 \text{ otherwise} \end{cases} \tag{34}$$

($\bar{f}^{-1}$ is divided by 2 in the first case because the agent pays attention to both the sine and the cosine at frequency j_i^*). Specialization here means that agents obtain information at a single frequency and are uninformed at all other frequencies.

We offer two potential explanations for why specialization would be natural. First, it could be the case that people must pay a fixed cost for each frequency that they learn about. That is, just starting to learn about some aspect of fundamentals might be costly. In that case people would naturally choose to learn about only a single frequency, since all frequencies pay the same marginal benefit, so learning about more than one frequency requires an extra payment of the fixed cost with zero benefit.

Another motivation for specialization is that people might simply have different natural aptitudes or desires for learning about fluctuations at different frequencies. The appendix develops a simple example of such a case that can generate specialization. The basic idea is that if the preference for learning about particular frequencies is sufficiently small (i.e. it is second-order or in a sense lexicographic) then the equilibrium described in the previous section still holds, but with each investor focusing their attention on only a single frequency.

Now obviously in reality nobody learns about just a single aspect of the world. It is also not the case, though, that everybody learns about everything. We focus here on the case with specialization as it is consistent with the evidence discussed above and with new results presented below on the wide divergences in behavior and research across investors.

3.2 Specialization model predictions

We now examine the implications of the model with specialization for the behavior of individual investors, obtaining the following results:

1. Investors can be distinguished by the frequencies at which their portfolio positions fluctuate, and those fluctuations match the frequencies at which they obtain information.

2. The average volume accounted for by an investor is proportional to the frequency at which they trade. In the presence of quadratic trading costs, costs can be linearly decomposed across frequencies and are quadratic in frequency.

3. Investors' positions are correlated with returns most strongly at the frequency they learn about.

4. Investors earn money from liquidity provision, they earn money from trading at the frequency at which they are informed, and they *lose* money to other investors from trading at frequencies at which they are uninformed.

3.2.1 Fluctuations in positions

Result 2 *Investor i's demand at frequency j is*

$$q_{i,j} = z_j + \rho\left[\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right) r_j + f_{i,j}^{-1}\tilde{\varepsilon}_{i,j}\right] \tag{35}$$

where $\tilde{\varepsilon}_{i,j}$ is equal to the jth column of Λ multiplied by ε_i, i.e. the noise in investor i's signal at frequency j, and r_j is the realized return on the jth frequency portfolio.

Investor i's demand depends on three terms. z_j is the stochastic supply at frequency j. Each investor is equally willing to absorb supply, so they all take equal fractions, giving them a common component z_j.

The second term, $\rho\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right) r_j$ reflects investor i's information. At the frequency that investor i pays attention to, $f_{i,j}^{-1} - f_{avg,j}^{-1}$ is positive, so investor i's demand covaries positively with returns at that frequency. That is, investors who learn about low-frequency dynamics hold portfolios that are long when returns are high over long periods, while high-frequency investors hold portfolios that covary positively with transitory fluctuations in returns. At the other frequencies, where investor i does not pay attention, $f_{i,j}^{-1} = 0$, so the investor's demand actually covaries slightly negatively with returns, holding z_j fixed.

The third term, $\rho f_{i,j}^{-1}\tilde{\varepsilon}_{i,j}$ is the idiosyncratic part of demand that is due to the random error in the signal that agent i receives. Note that the standard deviation of $f_{i,j}^{-1}\tilde{\varepsilon}_{i,j}$ is equal to $f_{i,j}^{-1/2}$, so these errors are equal to zero at the frequencies that the investor ignores (i.e. all but one).

When the number of active frequencies (i.e. with $f_{avg,j}^{-1} > 0$) is large, $f_{avg,j}^{-1}$ becomes small relative to $\bar{f}^{-1}$. That means that the term $\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right)$ is close to zero at all frequencies except for the one that the agent pays attention to, j_i^*. Since $\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right) \approx 0$ for all other frequencies, we have

$$Q_{i,t} \approx Z_t + \cos\left(\omega_{j_i^*} t/T\right)\left(\bar{f}^{-1} r_{j_i^*} + \tilde{\varepsilon}_j\right) + \sin\left(\omega_{j_i^*} t/T\right)\left(\bar{f}^{-1} r_{j_i^{*\prime}} + \tilde{\varepsilon}_{j'}\right) \tag{36}$$

Investor i's demand on date t thus is approximately equal to supply on that date plus a multiple the part of returns depending on frequency $\omega_{j_i^*}$, $r_{j_i^*}$ and $r_{j_i^{*\prime}}$, plus an error. The second line shows that what is really going on is that investor i's information can be thought of as a signal about returns interacted with a cosine and a sine.

The important feature of equations (35) and (36) is that they show that each agent's position is equal to Z_t plus fluctuations that come primarily at the frequency that they pay attention to.[16] That is, if some agent allocates all attention to frequency $\omega_{j_i^*}$, then their relative position, $Q_{i,t} - Z_t$,

[16] More formally, the variance of $Q_{i,t} - Z_t$ can be decomposed as $Var\left(Q_{i,t} - Z_t\right) = \sum_j \left(\rho^2\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right)^2 f_{R,j} + f_{i,j}^{-1}\right)$. Now consider a simple case where there are N frequencies that receive equal allocations of information. Furthermore, denote the spectrum of returns as $f_{R,j}$. Then we have

$$\lim_{N\to\infty} Var\left(Q_{i,t} - Z_t\right) = \rho^2 \bar{f}^{-2} f_{R,j_i^*} + f_{i,j}^{-1} \tag{37}$$

which shows that $Q_{i,t} - Z_t$ is driven by fluctuations at a single frequency.

fluctuates over time at frequency $\omega_{j_i^*}$. This can be seen by noting that the sum of a sine and a cosine at frequency $\omega_{j_i^*}$, even with different coefficients, remains a cosine that fluctuates at frequency $\omega_{j_i^*}$, just shifted by a constant. Specifically,

$$Q_{i,t} \approx Z_t + \sqrt{\begin{array}{c}\left(\bar{f}^{-1} r_{j_i^*} + \tilde{\varepsilon}_{j_i^*}\right)^2 \\ + \left(\bar{f}^{-1} r_{j_i^{*\prime}} + \tilde{\varepsilon}_{j_i^{*\prime}}\right)^2\end{array}} \cos\left(\omega_{j_i^*} t/T + C_{i,j}\right) \tag{38}$$

where $C_{i,j}$ is a function of $\left(\bar{f}^{-1} r_j + \tilde{\varepsilon}_j\right)$ and $\left(\bar{f}^{-1} r_{j'} + \tilde{\varepsilon}_{j'}\right)$. So agent i's excess demand is approximately a cosine with a random translation and amplitude.

As a numerical example, figure 2 plots a hypothetical history for a particular agent's position $Q_{i,t}$ in the same calibration that we studied above. We see that $Q_{i,t}$ looks like a sinusoid with noise added; the noise is from the Z_t term in (38). The noise in the agent's signal, $\tilde{\varepsilon}_{j_i^*}$ and $\tilde{\varepsilon}_{j_i^{*\prime}}$, simply changes the amplitude and translation of the cosine in (38).

So equations (35) and (36) deliver our first two basic results for the behavior of individual specialized investors: the investors can be distinguished by the frequencies at which their asset holdings fluctuate, and those frequencies are linked to the type of information that they acquire. The first result, that there are traders at different frequencies, is essentially obtained by design: it follows from the assumption that agents specialize across frequencies. Nevertheless, the finding is interesting for its novelty in a theoretical setting.

The fact that the frequency of trading is related to information acquisition, while not surprising, is certainly not obtained by assumption. In past work, different trading behavior has sometimes been obtained by simply assuming that different agents have different exogenously specified trading horizons. In our case, any investor can potentially trade at any frequency. That choice is entirely endogenous – investors are not forced to trade any particular frequency by assumption (the assumption is that they gather information at a single frequency).

The reason that buy-and-hold investors in our model buy and hold is that they have persistent low-frequency information about fundamentals – they have signals that fundamentals will be strong or weak over long time spans. Similarly, high-frequency investors have transitory high-frequency information. So the model provides a testable prediction that we should observe investors doing research about asset return dynamics that aligns in terms of frequency or time horizon with their average holding periods.

3.2.2 How do investors earn money?

Investors earn returns in the model through two basic mechanisms: providing liquidity and trading on private signals. We can see from the results on demand above that the liquidity function is spread equally across investors. The effects of private information are more interesting.

Result 3 *Investor i's expected profits (which are also equal to the covariance of their positions with*

returns) are

$$E\left[Q_i'R\right] = \sum_j E\left[q_{i,j}r_j\right] \tag{39}$$

$$= \sum_j E\left[z_j r_j\right] + \rho\left(\bar{f}^{-1} - f_{avg,j_i^*}^{-1}\right) f_{R,j_i^*} - \sum_{j\neq j_i^*} \rho f_{avg,j}^{-1} f_{R,j} \tag{40}$$

where the spectrum of returns is (from (32))

$$f_{R,j} = \max\left(\bar{\lambda}, \lambda_j\left(0\right)\right) \tag{41}$$

The first term on the right-hand side is the contribution from each investor's liquidity provision. The second term is the positive covariance of the investor's holdings with returns at the frequency they are informed about. In informed investors have demands that covary positively with returns at a particular frequency, then the investors who are uninformed about that frequency must have demands that covary *negatively* with returns (after accounting for $E\left[z_j r_j\right]$). That is the third term above: there is a negative contribution to the correlation of the investor's demands with returns from the frequencies they do not pay attention to.

It is not the case that trading from frequencies $j \neq j_i^*$ is unprofitable. Investors still earn profits from liquidity provision. It is just the case that some of their profits at those frequencies are taken by investors who are more informed. In some sense, this result is inevitable. The total profits that the informed investors earn as a group come from trading with liquidity demand. If an investor earns more money by becoming informed at some frequency, that must come at the cost of other investors.

Now since the information allocation we find is an optimum, obviously investors must be in some sense comfortable with the losses we see here. Intuitively, the slight trading losses they bear at frequencies other than j_i^* are offset by their gains at j_i^*. But obviously any trading that informed investors do that is not related to exogenous supply must ultimately come at the cost of other informed investors.

So the model has the feature that high-frequency investors earn money at high frequencies, but they lose money at lower frequencies relative to other investors. Low-frequency investors might know that oil prices are on a long-term downward trend. In such a situation, the high-frequency investors can still earn profits by betting on day-to-day movements in oil prices, but they will lose money to those who understand that prices are generally drifting down. Similarly, low-frequency investors will tend to lose out at high frequencies by, for example, failing to trade at precisely the right time, buying slightly too high and selling slightly too low compared to where they would if they had high-frequency information.

3.2.3 Volume and trading costs

We study volume in the representation of the model in terms of equity holdings. Recall that equity is modeled as a discounted claim to dividends on all future dates. An investor's position $Q_{i,t}$ can be acquired either by holding $Q_{i,t}$ units of forwards or $Q_{i,t}$ units of equity. In modeling volume, we consider trading in equity. Using equity to measure volume ensures that a person who has position that does not change between dates t and $t+1$ ($Q_{i,t} = Q_{i,t+1}$) induces no trade volume, whereas if we assumed that every forward position required volume, then each investor's contribution would be $|Q_{i,t}|$ each date, meaning, unrealistically, that buy-and-hold investors would contribute constantly to volume.

The equity volume contributed by investor i is

$$V_{i,t} = |\Delta Q_{i,t}| \tag{42}$$

$$\text{where } \Delta Q_{i,t} \equiv Q_{i,t} - Q_{i,t-1} \tag{43}$$

Recall that investors' positions can be written as functions of cosines and sines. The appendix derives the following result for volume for each investor.

Result 4 *The volume induced by investor i, $|\Delta Q_{i,t}|$, may be approximated as*

$$|\Delta Q_{i,t}| \approx |\Delta Z_t| + \omega_{j_i^*} \bar{f}^{-1} \rho \left| \begin{array}{c} \sin\left(\omega_{j_i^*} t\right)\left(r_{j_i^*} + \tilde{\varepsilon}_{i,j_i^*}\right) \\ +\cos\left(\omega_{j_i^*} t\right)\left(r_{j_i^{*\prime}} + \tilde{\varepsilon}_{i,j_i^{*\prime}}\right) \end{array} \right| \tag{44}$$

and has expectation

$$E\left[|\Delta Q_{i,t}|\right] - E\left[|\Delta Z_t|\right] \approx \omega_{j_i^*} \sqrt{\frac{2}{\pi}} \rho \left(\bar{f}^{-1} \bar{\lambda} + 1/2\right). \tag{45}$$

The approximations converge to true equalities as $T \to \infty$.

So we find that agent i's contribution to volume depends on the volume induced by exogenous supply and also the magnitude of returns at frequency $\omega_{j_i^*}$ ($\bar{\lambda}$).

Agent i's contribution to aggregate volume is also exactly proportional to the frequency they allocate attention to, $\omega_{j_i^*}$. High-frequency investors contribute relatively more to aggregate volume because they have portfolios than change most rapidly. An investor at the very lowest frequency, $\omega = 0$, contributes zero volume beyond that induced by exogenous supply, since their position is nearly constant. Investors at $\omega = \pi$, on the other hand, contribute the maximum possible volume as they approximately turn over their entire portfolios in each period.

Not surprisingly, it is also straightforward to show that high-frequency investors typically will face larger trading costs. As an example, consider quadratic trading costs proportional to $\sum_{t=2}^{T} \left(Q_{i,t} - Q_{i,t-1}\right)^2$. The appendix shows that trading costs can, just like volume, be decomposed across frequencies.

Result 5 *The quadratic variation in an investor's position can be approximated (with convergence as $T \to \infty$) by*

$$\sum_{t=2}^{T} \left(Q_{i,t} - Q_{i,t-1}\right)^2 = \sum_{j,j'} \left(2\pi j\right)^2 T^{-1} q_{i,j}^2. \tag{46}$$

The quadratic trading costs associated with a given demand vector Q_i can be written as a simple sum across frequencies. Trading costs are proportional to the frequency squared. It is thus immediately apparent from our frequency-domain analysis that changes in the magnitude of trading costs have the largest effects on the highest frequencies.

4 Institutional portfolio turnover and return forecasting

In this section, we demonstrate empirically that investment funds specialize in the frequency at which they trade, and we show that the portfolio holdings of high turnover funds forecast returns at relatively shorter horizons than those of lower turnover funds.

4.1 Data

We obtain data on institutional asset holdings from SEC form 13F. These forms list the identities and quantities of securities held by each institution at the end of the filing quarter.[17] The data cover the period 1980–2015. Data on monthly stock returns is taken from CRSP and is aggregated to a quarterly frequency with delisting returns included. We obtain data on the risk-free rate, market return, and Fama–French (1993) factors from Kenneth French's website.

4.2 Fund specialization

Yan and Zhang (2009) define the churn rate $c_{i,t}$ of institution i in quarter t as

$$c_{i,t} \equiv \frac{\min\left(\sum\limits_{s|\Delta S_{i,s,t}>0} P_{s,t}\Delta S_{i,s,t}, \sum\limits_{s|\Delta S_{i,s,t}\leq 0} P_{s,t}\left|\Delta S_{i,s,t}\right|\right)}{\frac{1}{2}\sum\limits_{s} P_{s,t-1}S_{i,s,t-1} + \frac{1}{2}\sum\limits_{s} P_{s,t}S_{i,s,t}}, \tag{47}$$

where $P_{s,t}$ is the price and $S_{i,s,t}$ is the number of shares of stock S held by institution i at the end of quarter t. The churn rate is equal to the minimum of net purchases and sales during quarter t as a fraction of the institution's average value over the two quarters, and it is used to measure the turnover of each institution's portfolio. Due to the presence of the minimum operator, institutions

[17] Institutions are required to report only their holdings of 13(f) securities, a category defined by the SEC that includes exchange-traded equities and some securities that can be converted to equity. Only institutions holding more than \$100,000,000 in 13(f) securities at the end of the quarter must file form 13F, and each institution is required to report only securities for which its holdings exceed \$200,000 or 10,000 shares. Gompers and Metrick (2001) provide more information on these filings. We use Thompson Reuters's database of these filings, which includes the price of each security at the filing date.

must both buy and sell large fractions of their portfolios to register a high churn rate. The mean churn rate is 0.12 and the standard deviation is 0.14, indicating a high degree of right skewness as the minimum churn rate is zero.

If institutions specialize in the rate at which they trade, then the churn rate should be persistent over time within institutions. Figure 5 plots the sample autocorrelations $\text{corr}(c_{i,t}, c_{i,t-\Delta t})$ for $\Delta t \geq 2$. The churn rate strongly persists over time, with an autocorrelation of 0.51 over 10 years and 0.21 over 30 years.[18] We also find that institution fixed effects (δ_i) account for 65 percent of the variance in the churn rate in the regression $c_{i,t} = \delta_i + \varepsilon_{i,t}$, where $\varepsilon_{i,t}$ a residual.

4.3 Fund performance

To separate institutions according to their trading frequency, at each quarter t we divide institutions into deciles, denoted d, based on the mean of $c_{i,t}$ over the previous five years.[19] For simplicity, we restrict attention to top and bottom deciles ($d = 10$ and $d = 1$). Table 1 lists several institutions in these extreme deciles during the most recent quarter in our data. The top decile contains several well-known quantitative and high-frequency trading firms, whereas the bottom contains endowments and insurance companies.

Table 1: Institutions in the top and bottom deciles of churn rate in the fourth quarter of 2015

Top decile	Bottom decile
Arrowstreet Capital	Berkshire Hathaway
Citadel	Bill & Melinda Gates Foundation
Dynamic Capital Management	Lilly Endowment
Ellington Management Group	Longview Asset Management
Quantlab	MetLife
Renaissance Technologies	New York State Teachers' Retirement System
Soros Fund Management	University of Notre Dame
Virtu Financial	University of Chicago

At the beginning of quarter t, we average the portfolio holdings of all the funds in each decile d at the end of quarter $t-1$ (with equal weight on each fund) and then track the returns on that aggregate decile-level portfolio over subsequent quarters, reinvesting proceeds from any delistings in the remaining stocks in the portfolio according to their value weights at that time. The return during quarter t of the decile d portfolio formed in quarter $t-k$ is denoted $r_{d,k,t}$.

We measure the performance of each portfolio by its alpha,

$$r_{d,k,t} - r_t^f = \alpha_{d,k} + \beta_{d,k} F_t + \varepsilon_{d,k,t}, \tag{48}$$

[18] Institution identifiers can be reassigned over time in the 13F data, leading to measurement error that biases the longer-term autocorrelations towards zero.

[19] We restrict attention to institutions for which the $t+1$ return on some of their holdings appears in CRSP, as these are the only institutions that can be analyzed in our return regressions.

F_t is a vector of market risk factors; $F_t = r_t^m - r_t^f$ in the CAPM specification and $F_t = \begin{pmatrix} r_t^m - r_t^f & r_t^{smb} & r_t^{hml} \end{pmatrix}'$ in the Fama-French specification (r_t^{smb} and r_t^{hml} are returns on the SMB and HML portfolios, respectively). We focus on returns over the first two years after portfolio formation by estimating (48) only for $k \leq 8$.

Figure 6 displays alphas in the two specifications. For simplicity, we compare the alphas in the first quarter ($\alpha_{d,1}$) to those in the following seven quarters, $\alpha_{d,>1} \equiv \frac{1}{7}\sum_{j=2}^{8} \alpha_{d,j}$. The holdings of high-turnover funds out-perform more during the first quarter, while those of low-turnover funds out-perform more during subsequent quarters. The difference in differences $(\alpha_{10,1} - \alpha_{10,>1}) - (\alpha_{1,1} - \alpha_{1,>1})$, which measures the relative out-performance of high churn holdings versus low churn holdings at short versus long horizons, is equal to 0.005 in both specifications and is significant at the 5 percent level.[20] So, consistent with the model, high-turnover funds hold stocks that outperform relatively more in the short-run, while low-turnover funds hold assets that display more persistent outperformance.

5 The effects of eliminating high-frequency trade

Recently there has been interest in policies that might restrict high-frequency trading. Some of those policies are aimed at investors who trade at the very highest frequencies (such as the CFTC's recently proposed Regulation AT; see CFTC (2016)). But there are also proposals to discourage even portfolio turnover at the monthly or annual level.[21] There are two ways to interpret such policies. One would be that regulators might impose a tax on trading, which would simply represent a transaction cost. Such a regulation would obviously have the strongest effects on high-frequency traders (that result can be derived in an extension to the present setting with trading costs), but it would ultimately affect all trading strategies. A more targeted policy would be one that simply outlawed following a trading strategy in which positions fluctuate at frequencies above some bound. Such a policy is straightforward to analyze in our framework.

We show in this section that a policy that restricts high-frequency trading by professional investors (as opposed to liquidity traders) reduces liquidity and price informativeness and increases return volatility at high frequencies. At the frequencies not targeted by the policy, though, price informativeness is, if anything, increased.

5.1 The policy

If elimination of high-frequency trading means that investors cannot hold portfolios with components that fluctuate rapidly, it means that those investors are restricted to setting $q_{i,j} = 0$ for ω_j in

[20] Yan and Zhang (2009) similarly find that the fraction of the outstanding shares of a stock held by high-churn institutions predicts subsequent returns, while the fraction held by low-churn institutions does not.

[21] The US tax code, for example, encourages holding assets for at least a year through the higher tax rates on short-term capital gains. There have been recent proposals to further expand such policies (a plan to create a schedule of capital gains tax rates that declines over a period of six years was attributed to Hillary Clinton during the 2016 US Presidential election; see Auxier et al. (2016)).

the relevant frequency range. But when $q_{i,j}$ must equal zero at some set of frequencies, obviously no trader will allocate attention to those frequencies.

In cases where the sophisticated investors must set $q_j = 0$, there can obviously be no equilibrium since liquidity demand is perfectly inelastic. We therefore consider a simple extension to the baseline model where the exogenous supply curve is elastic,

$$Z_t = \tilde{Z}_t + kP_t \tag{49}$$

$$z_j = \tilde{z}_j + kp_j \tag{50}$$

where $\tilde{Z}_t$ is the exogenous supply process and k is the response of supply to prices.

The appendix solves this extended version of the model. We obtain the same water filling equilibrium. For this section, what is most important is the volatility of returns.

For the frequencies that are restricted

$$\text{for } q_{i,j} = 0\text{: } p_j^{restricted} = -k^{-1}\tilde{z}_j. \tag{51}$$

That is, prices at the restricted frequencies are now completely uninformative, depending only on supply, with no relationship with fundamentals. Moreover, the market is completely illiquid in the sense that when exogenous supply increases, there is no change in trade – prices just move so that trade remains at zero. In other words, prices equilibrate the market instead of quantities.

5.2 Return volatility

Result 6 *Given an information policy $f_{avg,j}^{-1}$, the variance of returns at frequency j, when trade is unrestricted (i.e. in the benchmark model from above), is*

$$f_R(\omega_j) \equiv Var(r_j) \tag{52}$$

$$= \lambda_j\left(f_{avg,j}^{-1}\right) \tag{53}$$

$$= \min\left(\bar{\lambda}, \lambda_j(0)\right). \tag{54}$$

where

$$\lambda_j(0) = f_{D,j} + \frac{f_{\tilde{Z},j}}{\left(k + \rho f_{D,j}^{-1}\right)^2} \tag{55}$$

Recall that f_R is also the spectrum of returns, $R_t \equiv D_t - P_t$.

So the spectrum of returns inherits exactly the water-filling property of the marginal benefits of information. In the context of our benchmark calibration, the spectrum of returns is exactly the $\lambda_j\left(f_{avg,j}^{-1}\right)$ curve plotted in figure 1.

That result does not apply when the sophisticated investors are restricted from trading, though.

Result 7 *The variance of returns at any restricted frequency, where $q_{i,j}$ must equal zero, is*

$$f_{R,j}^{restricted} = f_{D,j} + \frac{f_{\tilde{Z},j}}{k^2} \tag{56}$$

and

$$f_{R,j}^{restricted} > \lambda_j(0). \tag{57}$$

The volatility of returns at a restricted frequency is higher than it would be if the sophisticated investors were allowed to trade, even if they gathered no information. Intuitively, when the active investors have risk-bearing capacity ($\rho > 0$), they can absorb some of the exogenous supply. The greater is the risk-bearing capacity, the smaller is the effect of supply volatility on return volatility.

We examine the quantitative implications of restricting trade in the context of the calibration used above. The top panel of figure 3 plots $f_{avg,j}^{-1}$ in the restricted and unrestricted scenarios. The restriction is that investors are not allowed to trade at frequencies above $\omega = 3$ (cycle lengths shorter than 2.1 periods). We see, then, that no information is acquired at those frequencies. That means, though, that investors can allocate their attention elsewhere, so we observe more information acquisition at other frequencies.

The bottom panel of figure 3 plots return volatility in the two regimes and also when investors can trade at all frequencies, but they are just restricted from gathering information at high frequencies. At high frequencies, we see that the restrictive policy has two separate effects that both strongly affect return volatility. First, when investors can trade but do not gather information, there is a jump in return volatility at the frequencies where f_{avg}^{-1} is constrained to zero (up to $\lambda_j(0)$). But under the full restriction, where they cannot trade at all, we see that the effect on return volatility is much larger, due to the reduced risk bearing capacity. At the unrestricted frequencies, return volatility actually weakly declines, again due to the fact that more attention is allocated to those frequencies.

Restricting sophisticated investors (such as dealers or proprietary trading firms) from trading at high frequencies in this model can thus substantially raise asset return volatility at high frequencies – it can lead to, for example, large minute-to-minute fluctuations in prices and returns. Sophisticated traders typically play a role of smoothing prices over time, essentially intermediating between excess inelastic demand in one minute and excess inelastic supply in the next. When they are restricted from holding positions that fluctuate from minute to minute, they can no longer provide that intermediation service. Such behavior has no impact on low-frequency volatility in prices, though. Even when there is no high-frequency trading, changes in average prices between one year and the next are essentially unaffected.

5.3 Price informativeness and efficiency

The fact that the sophisticated investors choose to allocate no attention to high frequencies under the trading restriction obviously has implications for price efficiency there. To see precisely how, we

measure price informativeness as the precision of a person's prediction of fundamentals conditional on observing prices only, $Var\left(D_t \mid P\right)$. In the frequency domain we have

$$\begin{aligned} \bar{\tau}_j &\equiv Var\left(d_j \mid p\right)^{-1} && (58) \\ &= \left(\rho f_{avg,j}^{-1}\right)^2 f_{Z,j}^{-1} + f_{D,j}^{-1} && (59) \end{aligned}$$

price-based precision, $\bar{\tau}_j$ is higher at frequencies where there is less fundamental uncertainty ($f_{D,j}^{-1}$ is lower), where there is less variation in liquidity demand ($f_{Z,j}^{-1}$ is lower) or where investors acquire more information ($f_{avj,j}^{-1}$ is higher). So when trading strategies are restricted and $f_{avg,j}^{-1}$ endogenously falls to zero at the restricted frequencies, price informativeness clearly falls. In fact, when $f_{avg,j}^{-1} = 0$, the ex-post precision at each frequency is exactly $f_{D,j}^{-1}$, which is the prior precision; prices contain no information. The decline in informativeness happens, though, only at the restricted frequencies.[22]

Result 8 *If trade is restricted at some set of frequencies, prices become (weakly) less informative at those frequencies ($\bar{\tau}_j$ falls) but informativeness is unaffected or increased at all other frequencies.*

5.3.1 Informativeness for moving averages of D_t

If a person is making decisions based on estimates of fundamentals from prices and they are worried that prices are contaminated by high-frequency noise, a natural response would be to examine an average of fundamentals and prices over time. For averages of fundamentals, we have the following:

Result 9 *The variance of an estimate of the average of fundamentals over dates t to $t+n-1$ conditional on observing the vector of prices, P, is*

$$\begin{aligned} Var\left(\frac{1}{n}\sum_{m=0}^{n-1} D_{t+n} \mid P\right) &= \frac{1}{nT}\sum_{j,j'} F_n\left(\omega_j\right)\bar{\tau}_j^{-1} && (60) \\ \text{where } F_n\left(\omega_j\right) &\equiv \frac{1}{n}\frac{1-\cos\left(n\omega_j\right)}{1-\cos\left(\omega_j\right)} && (61) \end{aligned}$$

and $\bar{\tau}_j$ is defined in (58).

F_n is the Fejér kernel. $F_1 = 1$, and as n rises, the mass of the Fejér kernel migrates towards the origin. That is, it places progressively less mass on high frequencies and more on low frequencies (it always integrates to 1). Specifically,

$$\begin{aligned} \frac{1}{T}\sum_{j,j'} F_n\left(\omega_j\right) &= 1 && (62) \\ \lim_{n\to\infty} F_n\left(\omega\right) &= 0 \text{ for all } \omega \neq 0 && (63) \end{aligned}$$

[22] To see that result in the time domain, the appendix shows that $Var\left(D_t \mid P\right) = \frac{1}{T}\sum_{j,j'}\bar{\tau}_j^{-1}$. The variance of an estimate of fundamentals conditional on prices at a particular date is equal to the average of the variances across all frequencies. So when uncertainty, $\bar{\tau}_j^{-1}$, rises at some set of frequencies, the informativeness of prices for fundamentals on every date falls by an equal amount.

The total weight allocated across the frequencies always sums to 1, and as n rises, the mass becomes allocated eventually purely to frequencies local to zero.

This result shows that the informativeness of prices for moving averages of fundamentals places relatively more weight on low- than high-frequency informativeness. So even if prices have little or no information at high frequencies – $\bar{\tau}_j$ is small for large j, there need not be any degradation of information about averages of fundamentals over multiple periods, as they depend primarily on precision at lower frequencies (smaller values of j).

The top panel of figure 4 plots the Fejér kernel, F_n, for a range of values of n. One can see that even with $n = 2$, the weight allocated to frequencies above the cutoff of $\omega = 3$ that we use in the example in figure 3 is close to zero. As n rises higher, the weight falls towards zero at a progressively wider range of frequencies. Equation (60) therefore shows that while a reduction in precision at high frequencies due to trading restrictions will reduce the informativeness of prices about fundamentals on any single date, it has quantitatively small effects on the informativeness of prices for fundamentals over longer periods.

Moving averages of fundamentals depend less on the precise high-frequency details of the world, so when high-frequency information is reduced, we would not expect to see a reduction in the informativeness of prices for moving averages. More concretely, going back to our example of oil futures, when high-frequency trade is not allowed, prices become noisier, making it more difficult to obtain an accurate forecast of the spot price of oil at some specific moment in the future. If one is interested in the average of spot oil prices over a year, on the other hand, then we would expect futures prices to remain informative, even when high-frequency trade is restricted.

5.3.2 How should one forecast D_t conditional on prices?

As an alternative to estimating an average of fundamentals over some number of periods, one's goal might alternatively be to specifically forecast fundamentals on just one date. In that case, we see that the effect of high-frequency trade restrictions is to cause investors to focus on averages of prices over multiple periods. That is, to forecast the spot oil price in one particular month, one might use an average of futures prices over neighboring months.

In general, the expectation of the full time series of fundamentals, D, is

$$E\left[D \mid P\right] = \Lambda diag\left(\phi_j\left(f_{avg,j}\right)\right)\Lambda' P \tag{64}$$

ϕ_j is a function of only $f_{avg,j}$ and the behavior of fundamentals and liquidity demand at frequency j with $\phi_j\left(0\right) = 0$. Intuitively, this equation says that dividends are obtained from prices using a filter: prices are transformed into the frequency domain ($\Lambda' P$), a filter is applied that depends on the informativeness of prices at each frequency ($\phi_j\left(f_{avg,j}\right)$), and then D is obtained by transforming back into the time domain.

The effect of eliminating information – setting $f_{avg,j}^{-1} = 0$ – at high frequencies is then simple to analyze. The frequency domain step sets to zero the weight on any frequencies at which there

is no information. That is, in obtaining the expectation of dividends, one first applies a low-pass filter to prices (e.g. Christiano and Fitzgerald (2003)).

Result 10 *When information acquisition is set to zero for frequencies above a cutoff $\bar{j}$, so that $f_{avg,j}^{restricted} = f_{avg,j} 1\{j \leq \bar{j}\}$ (where $1\{\cdot\}$ is the indicator function) the expectation of fundamentals conditional on prices is*

$$E[D \mid P] = \Lambda diag\left(\phi_j\left(f_{avg,j}\right)\right) diag\left(1\{j \leq \bar{j}\}\right) \Lambda' P \tag{65}$$

Specifically, $E[D_t \mid P]$ is equal to the t'th row of $\Lambda diag\left(\phi_j\left(f_{avg,j}\right)\right) diag\left(1\{j \leq \bar{j}\}\right) \Lambda'$ multiplied by the vector of prices, P.

Under the restriction on information acquisition (which, as we saw above, happens when investors may not trade at frequencies above $\bar{j}$), expectations are now calculated by first applying a filter to prices that eliminates high-frequency fluctuations (that is, it sets to zero all components of the price vector P that are spanned by high-frequency ($j \geq \bar{j}$) sines and cosines). That filter is implemented by multiplying prices by $diag\left(1\{j \leq \bar{j}\}\right) \Lambda'$, which eliminates the high-frequency components. Intuitively, since those components are completely devoid of information, they should be filtered out before estimating fundamentals.

A simple benchmark case is where the variance of fundamentals and liquidity demand is constant across frequencies, so that $f_{avg,j}$ is also constant across frequencies in the absence of trading restrictions and $diag\left(\phi_j\left(f_{avg,j}\right)\right)$ is a multiple of the identity matrix. The question, then, is how investors estimate fundamentals on some date t based on the history of prices.

When there is no restriction on prices and $\phi_j\left(f_{avg,j}\right)$ is constant, we see immediately that $E[D \mid P] \propto P$, so that the expectation of fundamentals on date t depends only on the price on date t (for any t).

When trade is restricted, we have $E[D \mid P] \propto \Lambda diag\left(1\{j \leq \bar{j}\}\right) \Lambda' P$. The bottom panel of figure 4 therefore plots a representative row of $\Lambda diag\left(1\{j \leq \bar{j}\}\right) \Lambda'$ for different values of $\bar{j}$ (the interior rows are all highly similar; the boundaries induce some differences). We see that with the trading restriction, the estimate of fundamentals on date t now depends on prices on t and its neighbors. Moreover, the smaller is $\bar{j}$ – the more frequencies that are restricted – the wider is the set of weights applied to prices. Intuitively, then, figure 4 confirms the natural result that when prices are less informative at high frequencies, the simple response is to estimate fundamentals based on a moving average of prices.

5.3.3 Summary

In the end, this section shows that the model has two key predictions for the effects of restrictions on high-frequency trade. First, at the frequencies at which trade is restricted, price informativeness falls and return volatility rises (due to both information effects and liquidity effects). Second, though, price informativeness at low frequencies is, if anything, *improved* by the policy. So if a

manager is making investment decisions based on fundamentals only at a particular moment, then that decision will be hindered by the policy since prices now have more noise. But if decisions are made based on averages of fundamentals over longer periods, e.g. over a year, then the model predicts that there need not be adverse consequences. If anything, low-frequency price informativeness may increase as investors reallocate attention to lower frequencies.

6 Conclusion

This paper develops a model in which there are many different investors who all trade at different frequencies. Investors in real-world markets follow countless strategies that are associated with rates of turnover that differ by multiple orders of magnitude. We show that in fact it is entirely natural that investors would differentiate along the dimension of investment frequency.

It has been standard in the literature for decades to focus on factors or principal components when studying the cross-section of asset returns. For stationary time series, the analog to factors or principal components is the set of fluctuations at different frequencies. So just as it seems natural for investors to focus on particular factors in the cross-section of returns, e.g. value stocks, a particular industry, or a particular commodity, it is also natural for investors to focus on fluctuations in fundamentals at a particular frequency, like quarters, business cycles, or decades.

Such an attention allocation problem can be solved using a combination of standard tools from time series econometrics and the literature on equilibria in financial markets. We show that the model fits a wide range of basic stylized facts about financial markets: investors can be distinguished by turnover rates; trading frequencies align with research frequencies; volume is driven primarily by high-frequency traders; and the positions of informed traders forecast returns at a horizon similar to their holding period.

Since the model has a rigorous concept of what being a high- or low-frequency entails, it is particularly useful for studying the effects of regulatory policies that would restrict trade at certain frequencies, whether by outlawing it or by simply making it more costly. We find that not only do such policies reduce the informativeness of prices at those frequencies, they also reduce liquidity and increase return volatility. In fact, return volatility will in general be raised even above where it would be in the complete absence of information, since eliminating active traders from the market removes their risk-bearing capacity.

At this point, the primary drawback of the model in our view is that it is not fully dynamic. In a certain sense we have to assume that investors do not update information sets over time. While that simplification does not interfere with the model's ability to match a wider range of basic facts about financial markets, a simple desire for realism suggests that incorporating dynamic learning is an obvious next step.

References

Admati, Anat R, "A Noisy Rational Expectations Equilibrium for Multi-Asset Securities Markets," *Econometrica*, 1985, pp. 629–657.

Amihud, Yakov and Haim Mendelson, "Asset Pricing and the Bid-Ask Spread," *Journal of Financial Economics*, 1986, *17* (2), 223–249.

Auxier, Richard, Len Burman, Jim Nunns, Ben Page, and Jeff Rohaly, "An Updated Analysis of Hillary Clinton's Tax Proposals," October 2016.

Bandi, Federico and Andrea Tamoni, "Business-cycle consumption risk and asset prices," *Working Paper*, 2014.

Banerjee, Snehal, "Learning from Prices and the Dispersion in Beliefs," *Review of Financial Studies*, 2011, *24* (9), 3025–3068.

____ **and Brett Green**, "Signal or noise? Uncertainty and learning about whether other traders are informed," *Journal of Financial Economics*, 2015, *117* (2), 398–423.

____ **and Ilan Kremer**, "Disagreement and learning: Dynamic patterns of trade," *The Journal of Finance*, 2010, *65* (4), 1269–1302.

Basak, Suleyman, "Asset pricing with heterogeneous beliefs," *Journal of Banking & Finance*, 2005, *29* (11), 2849–2881.

Bhattacharya, Sudipto and Paul Pfleiderer, "Delegated Portfolio Management," *Journal of Economic Theory*, 1985, *36* (1), 1–25.

Brillinger, David R., *Time Series: Data Analysis and Theory*, McGraw Hill, 1981.

Brockwell, Peter J. and Richard A. Davis, *Time Series: Theory and Methods*, Springer, 1991.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, "High-Frequency Trading and Price Discovery," *Review of Financial Studies*, 2014, *27* (8), 2267–2306.

Carhart, Mark M., "On Persistence in Mutual Fund Performance," *Journal of Finance*, 1997, *52* (1), 57–82.

Chaudhuri, Shomesh E. and Andrew W. Lo, "Spectral Portfolio Theory," 2016. Working paper.

Chen, Hsiu-Lang, Narasimhan Jegadeesh, and Russ Wermers, "The value of active mutual fund management: An examination of the stockholdings and trades of fund managers," *Journal of Financial and quantitative Analysis*, 2000, *35* (03), 343–368.

Chinco, Alexander and Mao Ye, "Investment-Horizon Spillovers: Evidence From Decomposing Trading-Volume Variance," 2016. Working paper.

Christiano, Lawrence J. and Terry J. Fitzgerald, "The Band Pass Filter," *International Economic Review*, 2003, *44(2)*, 435–465.

Commission, Commodity Futures Trading, "Regulation Automated Trading; Supplemental notice of proposed rulemaking," in "Federal Register" 2016.

Dávila, Eduardo and Cecilia Parlatore, "Trading Costs and Informational Efficiency," 2016. Working paper.

DeFusco, Anthony A., Charles G. Nathanson, and Eric Zwick, "Speculative Dynamics of Prices and Volume," 2016. Working paper.

Dew-Becker, Ian and Stefano Giglio, "Asset Pricing in the Frequency Domain: Theory and Empirics," *Review of Financial Studies*, 2016. Forthcoming.

Diamond, Douglas W and Robert E Verrecchia, "Information aggregation in a noisy rational expectations economy," *Journal of Financial Economics*, 1981, *9* (3), 221–235.

Fama, Eugene F. and Kenneth R. French, "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 1993, *33* (1), 3–56.

Gârleanu, Nicolae and Lasse Heje Pedersen, "Dynamic Trading with Predictable Returns and Transaction Costs," *The Journal of Finance*, 2013, *68* (6), 2309–2340.

Gompers, Paul A and Andrew Metrick, "Institutional Investors and Equity Prices," *Quarterly Journal of Economics*, 2001, pp. 229–259.

Griffin, John M and Jin Xu, "How smart are the smart guys? A unique view from hedge fund stock holdings," *Review of Financial Studies*, 2009, *22* (7), 2531–2570.

Grossman, Sanford J. and Joseph Stiglitz, "On the Impossibility of InfoInformation Efficient Markets," *American Economic Review*, 1980, *70*, 393–408.

He, Hua and Jiang Wang, "Differential information and dynamic behavior of stock trading volume," *Review of Financial Studies*, 1995, *8* (4), 919–972.

Hellwig, Martin F, "On the aggregation of information in competitive markets," *Journal of Economic Theory*, 1980, *22* (3), 477–498.

Hong, Harrison and Jeremy C Stein, "Disagreement and the stock market," *The Journal of Economic Perspectives*, 2007, *21* (2), 109–128.

Hopenhayn, Hugo A and Ingrid M Werner, "Information, Liquidity, and Asset Trading in a Random Matching Game," *Journal of Economic Theory*, 1996, *68* (2), 349–379.

Kacperczyk, Marcin, Stijn Van Nieuwerburgh, and Laura Veldkamp, "Rational attention allocation over the business cycle," Technical Report, National Bureau of Economic Research 2016.

Nagel, Stefan, "Short sales, institutional investors and the cross-section of stock returns," *Journal of Financial Economics*, 2005, *78* (2), 277–309.

Shao, Xiaofeng and Wei Biao Wu, "Asymptotic Spectral Theory for Nonlinear Time Series," *The Annals of Statistics*, 2007, *35* (4), 1773–1801.

Shumway, Robert H. and David T. Stoffer, *Time Series Analysis and Its Applications*, New York: Springer, 2011.

Spiegel, Matthew, "Stock price volatility in a multiple security overlapping generations model," *Review of Financial Studies*, 1998, *11* (2), 419–447.

Stoughton, Neal M, "Moral Hazard and the Portfolio Management Problem," *The Journal of Finance*, 1993, *48* (5), 2009–2028.

Townsend, Robert M, "Forecasting the forecasts of others," *The Journal of Political Economy*, 1983, pp. 546–588.

Turley, Robert, "Informative Prices and the Cost of Capital Markets," 2012. Working paper.

Wachter, Jessica A, "Portfolio and consumption decisions under mean-reverting returns: An exact solution for complete markets," *Journal of financial and quantitative analysis*, 2002, *37* (01), 63–91.

Wang, Jiang, "A model of intertemporal asset prices under asymmetric information," *The Review of Economic Studies*, 1993, *60* (2), 249–282.

___ , "A model of competitive stock trading volume," *Journal of political Economy*, 1994, pp. 127–168.

Watanabe, Masahiro, "Price volatility and investor behavior in an overlapping generations model with information asymmetry," *The Journal of Finance*, 2008, *63* (1), 229–272.

Weller, Brian, "Efficient Prices at Any Cost: Does Algorithmic Trading Deter Information Acquisition?," 2016. Working paper.

Yan, Xuemin Sterling and Zhe Zhang, "Institutional Investors and Equity Returns: Are Short-Term Institutions Better Informed?," *Review of Financial Studies*, 2009, *22* (2), 893–924.

A Non-stationary fundamentals

If fundamentals are non-stationary, e.g. if D_t has a unit root, then Σ_D is no longer Toeplitz and our results do not hold. In that case, we assume that D_0 is known by all agents and that the distribution of $\Delta D_t \equiv D_t - D_{t-1}$ is known, with covariance matrix $\Sigma_{\Delta D}$. Then the entire problem can simply be rescaled by defining $\tilde{P}_t \equiv P_t - D_{t-1}$, so that

$$\begin{aligned} R_t &= D_t - P_t \quad (66) \\ &= \Delta D_t - \tilde{P}_t \quad (67) \end{aligned}$$

Our analysis then applies to $\tilde{P}_t$ and ΔD_t, with $Q_{i,t}$ continuing to represent the number of forward contracts on D_t that agent i buys. That is, we are allowing agents to condition demand $Q_{i,t}$ not just on signals and prices, but also the level of D_{t-1}, simply through differencing.

B Proof of lemma 1

Gray (2006) shows that for any circulant matrix (a matrix where row n is equal to row $n-1$ circularly shifted right by one column, and thus one that is uniquely determined by its top row), the discrete Fourier basis, $u_j = [\exp(i\omega_j t), t = 0, ..., T-1]'$ for $j \in \{0, ..., T-1\}$, is the set of eigenvectors.

Let Σ be a symmetric Toeplitz matrix with top row $[\sigma_0, \sigma_1, ..., \sigma_{T-1}]$. Define the function $circ(x)$ to be a circulant matrix with σ_{circ} as its top row. Define a vector σ

$$\sigma \equiv [\sigma_0, \sigma_1 + \sigma_{T-1}, \sigma_2 + \sigma_{T-2}, ..., \sigma_{T-2} + \sigma_2, \sigma_{T-1} + \sigma_1]' \quad (68)$$

Following Rao (2016), we "approximate" Σ by the circulant matrix $\Sigma_{circ} \equiv circ(\sigma_{circ})$. Since Σ_{circ} is symmetrical, one may observe that its eigenvalues repeat in the sense that $u_j' \Sigma_{circ} = u_{T-j}' \Sigma_{circ}$ for $0 < j < T$. Since pairs of eigenvectors with matched eigenvalues can be linearly combined to yield alternative eigenvectors, it immediately follows that the matrix Λ from the main text contains a full set of eigenvectors for Σ_{circ}. The associated eigenvalues are

$$f_{\Sigma_{circ}}(\omega_j) = \sigma_0 + 2 \sum_{t=1}^{T-1} \sigma_t cos(\omega_j t) \quad (69)$$

We can write this relationship more compactly as:

$$\begin{aligned} \Sigma_{circ} \Lambda &= \Lambda f_\Sigma \\ \Lambda' \Sigma_{circ} \Lambda &= f_\Sigma \end{aligned}$$

where the $T \times T$ diagonal matrix f_Σ is given by:

$$f_\Sigma = diag\left(f_\Sigma(\omega_0), f_\Sigma(\omega_1), ..., f_\Sigma\left(\omega_{\frac{T}{2}}\right), f_\Sigma(\omega_1), f_\Sigma(\omega_2), ..., f_\Sigma\left(\omega_{\frac{T}{2}-1}\right)\right)'.$$

The approximate diagonalization of the matrix Σ consists in writing:

$$\begin{aligned} \Lambda'\Sigma\Lambda &= f_\Sigma + R_\Sigma \\ \text{where } R_\Sigma &\equiv \Lambda'(\Sigma - \Sigma_{circ})\Lambda \end{aligned}$$

By direct inspection of the elements of $\Sigma - \Sigma_{circ}$, one may see that the m, n element of R_Σ, denoted $R_\Sigma^{m,n}$ satisfies (defining λ_m to be the mth column of Λ and $\lambda_{m,n}$ to be its m, n element)

$$\begin{aligned} R_\Sigma^{m,n} &\equiv \lambda_m'(\Sigma - \Sigma_{circ})\lambda_n && (70) \\ &= \sum_{i=1}^{T}\sum_{j=1}^{T} \lambda_{m,i}\lambda_{n,j}(\Sigma - \Sigma_{circ})^{m,n} && (71) \\ &\leq \sum_{i=1}^{T}\sum_{j=1}^{T} \frac{2}{T}(\Sigma - \Sigma_{circ})^{m,n} && (72) \\ &\leq \frac{4}{T}\sum_{j=1}^{T-1} j|\sigma_j| && (73) \end{aligned}$$

where $(\Sigma - \Sigma_{circ})^{m,n}$ is the m, n element of $(\Sigma - \Sigma_{circ})$. So R_Σ is bounded elementwise by a term of order T^{-1}. One may show that the weak norm satisfies $|\cdot| \leq \sqrt{T}|\cdot|_{\max}$, where $|\cdot|_{\max}$ denotes the elementwise max norm, which thus yields the result that $|\Lambda\Sigma\Lambda' - diag(f_\Sigma)| \leq bT^{-1/2}$ for some b.

B.1 Convergence in distribution and $\bar{O}$ bounds

Define the notation $\Rightarrow$ to mean that $\Lambda X \Rightarrow N\left(0, \hat{\Sigma}_X\right)$ if $\Lambda X \sim N(0, \Sigma_X)$ and $\left|\hat{\Sigma}_X - \Sigma_X\right| = \bar{O}\left(T^{-1/2}\right)$.

The notation $\bar{O}$ indicates

$$|A - B| = \bar{O}\left(T^{-1/2}\right) \Longleftrightarrow |A - B| \leq bT^{-1/2} \quad (74)$$

for some constant b and for all T. This is a stronger statement than typical big-O notation in that it holds for all T, as opposed to holding only for some sufficiently large T.

Trigonometric transforms of stationary time series converge in distribution under more general conditions. See Shumway and Stoffer (2011), Brillinger (1981), and Shao and Wu (2007).

C Derivation of solution 1

Since the optimization is entirely separable across frequencies (confirmed below), we can solve everything in scalar terms. To save notation, we suppress the j subscripts indicating frequencies in this section when they are not necessary for clarity. So in this section f_D, for example, is a scalar representing the spectral density of fundamentals at some arbitrary frequency.

C.1 Statistical inference

We guess that prices take the form

$$p = a_1 d - a_2 z \tag{75}$$

The joint distribution of fundamentals, signals, and prices is then

$$\begin{bmatrix} d \\ y_i \\ p \end{bmatrix} \sim N\left(0, \begin{bmatrix} f_D & f_D & a_1 f_D \\ f_D & f_D + f_i & a_1 f_D \\ a_1 f_D & a_1 f_D & a_1^2 f_D + a_2^2 f_Z \end{bmatrix}\right) \tag{76}$$

The expectation of fundamentals conditional on the signal and price is

$$\begin{aligned} E\left[d \mid y_i, p\right] &= \begin{bmatrix} f_D & a_1 f_D \end{bmatrix} \begin{bmatrix} f_D + f_i & a_1 f_D \\ a_1 f_D & a_1^2 f_D + a_2^2 f_Z \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ p \end{bmatrix} && (77) \\ &= [1, a_1] \begin{bmatrix} 1 + f_i f_D^{-1} & a_1 \\ a_1 & a_1^2 + a_2^2 f_Z f_D^{-1} \end{bmatrix}^{-1} \begin{bmatrix} y_i \\ p \end{bmatrix} && (78) \end{aligned}$$

and the variance satisfies

$$\begin{aligned} \tau_i &\equiv Var\left[d \mid y_i, p\right]^{-1} = f_D^{-1} \left(1 - \begin{bmatrix} 1 & a_1 \end{bmatrix} \begin{bmatrix} 1 + f_i f_D^{-1} & a_1 \\ a_1 & a_1^2 + a_2^2 f_Z f_D^{-1} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ a_1 \end{bmatrix}\right)^{-1} && (79) \\ &= \frac{a_1^2}{a_2^2} f_Z^{-1} + f_i^{-1} + f_D^{-1} && (80) \end{aligned}$$

We use the notation τ to denote a posterior precision, while f^{-1} denotes a prior precision of one of the basic variables of the model. The above then implies that

$$E\left[d \mid y_i, p\right] = \tau_i^{-1} \left(f_i^{-1} y_i + \frac{a_1}{a_2^2} f_Z^{-1} p\right) \tag{81}$$

C.2 Demand and equilibrium

The agent's utility function is (where variables without subscripts here indicate vectors),

$$\begin{aligned} U_i &= \max_{\{Q_{i,t}\}} \rho^{-1} E_{0,i}\left[T^{-1} Q_i'\left(D-P\right)\right] - \rho^{-2} Var_{0,i}\left[T^{-1/2} Q_i'\left(D-P\right)\right] && (82)\\ &= \max_{\{Q_{i,t}\}} \rho^{-1} E_{0,i}\left[T^{-1} q_i'\left(d-p\right)\right] - \rho^{-2} Var_{0,i}\left[T^{-1/2} q_i'\left(d-p\right)\right] && (83)\\ &= \max_{\{Q_{i,t}\}} \rho^{-1} T^{-1} \sum_{j=0}^{T-1} q_{i,j} E_{0,i}\left[\left(d_j - p_j\right)\right] - \rho^{-2} T^{-1} \sum_{j=0}^{T-1} q_{i,j}^2 Var_{0,i}\left[d_j - p_j\right] && (84) \end{aligned}$$

where the last line follows by imposing the asymptotic independence of d across frequencies (we analyze the error induced by that approximation below). The utility function is thus entirely separable across frequencies, with the optimization problem for each $q_{i,j}$ independent from all others.

Taking the first-order condition associated with the last line above for a single frequency, we obtain

$$\begin{aligned} q_i &= \rho \tau_i E\left[d - p \mid y_i, p\right] \\ &= \rho_i \left(f_i^{-1} y_i + \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i \right) p \right) \end{aligned}$$

Summing up all demands and inserting the guess for the price yields

$$\begin{aligned} z &= \int_i \rho \left(f_i^{-1} y_i + \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i \right) \left(a_1 d - a_2 z \right) \right) di && (85)\\ &= \int_i \rho \left(f_i^{-1} d + \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i \right) \left(a_1 d - a_2 z \right) \right) di && (86) \end{aligned}$$

Where the second line uses the law of large numbers. Matching coefficients then yields

$$\begin{aligned} \int_i \rho \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i \right) di &= -a_2^{-1} && (87)\\ \int_i \rho f_i^{-1} + \rho \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i \right) a_1 di &= 0 && (88) \end{aligned}$$

and therefore

$$\rho \int_i f_i^{-1} di = \frac{a_1}{a_2} \tag{89}$$

Inserting the expression for τ_i into (87) yields

$$a_1 = \frac{\frac{a_1}{a_2} + \rho \left(\frac{a_1}{a_2} \right)^2 f_Z^{-1}}{\rho \left(\int_i f_i^{-1} di + f_D^{-1} + \left(\frac{a_1}{a_2} \right)^2 f_Z^{-1} \right)} \tag{90}$$

Now define aggregate precision to be

$$f_{avg}^{-1} \equiv \int_i f_i^{-1} di \tag{91}$$

We then have

$$\begin{aligned} \tau_i &= \frac{a_1^2}{a_2^2} f_Z^{-1} + f_i^{-1} + f_D^{-1} && (92) \\ \tau_{avg} &\equiv \int \tau_i di = \left(\rho f_{avg}^{-1}\right)^2 f_Z^{-1} + f_{avg}^{-1} + f_D^{-1} && (93) \end{aligned}$$

$$\begin{aligned} a_1 &= \tau_{avg}^{-1}\left(f_{avg}^{-1} + \left(\rho f_{avg}^{-1}\right)^2 f_Z^{-1}\right) = 1 - \frac{f_D^{-1}}{\tau_{avg}} = \frac{\tau_{avg} - f_D^{-1}}{\tau_{avg}} && (94) \\ a_2 &= \frac{a_1}{\rho f_{avg}^{-1}} && (95) \end{aligned}$$

C.3 Proof of proposition 1

In the time domain, the solution from Admati (1985) is

$$\begin{aligned} P &= A_1 D - A_2 Z && (96) \\ A_1 &\equiv I - S_{avg}^{-1}\Sigma_D^{-1} && (97) \\ A_2 &\equiv \rho^{-1} A_1 \Sigma_{avg} && (98) \end{aligned}$$

Standard properties of norms yield the following result. If $|A - B| = \bar{O}\left(T^{-1/2}\right)$ and $|C - D| = \bar{O}\left(T^{-1/2}\right)$, then

$$\begin{aligned} |cA - cB| &= \bar{O}\left(T^{-1/2}\right) && (99) \\ \left|A^{-1} - B^{-1}\right| &= \bar{O}\left(T^{-1/2}\right) && (100) \\ |(A + C) - (B + D)| &= \bar{O}\left(T^{-1/2}\right) && (101) \\ |AC - BD| &= \bar{O}\left(T^{-1/2}\right) && (102) \end{aligned}$$

In other words, convergence in weak norm carries through under addition, multiplication, and inversion. Since A_1 is a function of Toeplitz matrices using those operations, it follows that $|\Lambda' A_1 \Lambda - diag\,(a_1)| = \bar{O}\left(T^{-1/2}\right)$, and the same holds for A_2.

For the variance of prices, we define

$$\begin{aligned} R_1 &= A_1 - \Lambda diag\,(a_1)\,\Lambda' && (103) \\ R_2 &= A_2 - \Lambda diag\,(a_2)\,\Lambda' && (104) \end{aligned}$$

$$
\begin{aligned}
|Var\left[P-\Lambda p\right]| &\leq \left|R_1\Sigma_D R_1'\right| + \left|R_2\Sigma_Z R_2'\right| && (105)\\
&\leq |R_1\Sigma_D|\,|R_1| + |R_2\Sigma_Z|\,|R_2| && (106)\\
&\leq \|\Sigma_D\|\,|R_1|^2 + \|\Sigma_Z\|\,|R_2|^2 && (107)\\
&\leq K\left(|R_1|^2 + |R_2|^2\right) && (108)
\end{aligned}
$$

The first line follows from the triangle inequality; the second line comes from the sub-multiplicativity of the weak norm; the third line uses the fact that, as indicated by Gray (2006), for any two square matrices G, H, $||GH||_2 \leq \|G\|\,|H|$; and the last line follows from the assumption that the eigenvalues of Σ_D and Σ_Z are bounded by some K.

Since the weak norm is invariant under unitary transformations,

$$
|R_1| = \left|\Lambda' R_1 \Lambda\right| = \left|\Lambda' A_i \Lambda - diag\left(a_1\right)\right| \quad , \quad i = 1, 2.
$$

Therefore,

$$
\begin{aligned}
|Var\left[P-\Lambda P\right]| &\leq K\left(\left|\Lambda' A_1 \Lambda - diag\left(a_1\right)\right|^2 + \left|\Lambda' A_2 \Lambda - diag\left(a_2\right)\right|^2\right) && (109)\\
&= \bar{O}\left(\frac{1}{T}\right) && (110)
\end{aligned}
$$

Since $\|\cdot\| \leq \sqrt{T}\,|\cdot|$, $\|Var\left[P^c - P\right]\| = \bar{O}\left(T^{-1/2}\right)$.

D Proof of lemma 2

Inserting the optimal value of $q_{i,j}$ into the utility function, we obtain

$$
E_{-1}\left[U_{i,0}\right] \equiv \frac{1}{2} E\left[T^{-1}\sum_{j=0}^{T-1}\tau_{i,j} E\left[d_j - p_j \mid y_{i,j}, p_j\right]^2\right] \tag{111}
$$

$U_{i,0}$ is utility conditional on an observed set of signals and prices. $E_{-1}\left[U_{i,0}\right]$ is then the expectation taken over the distributions of prices and signals.

$Var\left[E\left[d_j - p_j \mid y_{i,j}, p_j\right]\right]$ is the variance of the part of the return on portfolio j explained by $y_{i,j}$ and p_j, while $\tau_{i,j}^{-1}$ is the residual variance. The law of total variance says

$$
Var\left[d_j - p_j\right] = Var\left[E\left[d_j - p_j \mid y_{i,j}, p_j\right]\right] + E\left[Var\left[d_j - p_j \mid y_{i,j}, p_j\right]\right] \tag{112}
$$

where the second term on the right-hand side is just $\tau_{i,j}^{-1}$ and the first term is $E\left[E\left[d_j - p_j \mid y_{i,j}, p_j\right]^2\right]$ since everything has zero mean. The unconditional variance of returns is

$$
Var\left[d_j - p_j\right] = \left(1 - a_{1,j}\right)^2 f_{D,j} + \frac{a_{1,j}^2}{\left(\rho f_{avg,j}^{-1}\right)^2} f_{Z,j} \tag{113}
$$

So then

$$E_{-1}\left[U_{i,0}\right]=\frac{1}{2}T^{-1}\sum_{j=0}^{T-1}\left(\left(1-a_{1,j}\right)^{2}f_{D,j}+\frac{a_{1,j}^{2}}{\left(\rho f_{avg}^{-1}\right)^{2}}f_{Z,j}\right)\tau_{i,j}-\frac{1}{2} \tag{114}$$

We thus obtain the result that agent i's expected utility is linear in the precision of the signals that they receive (since $\tau_{i,j}$ is linear in $f_{i,j}^{-1}$).

Furthermore,

$$\begin{aligned}\left(1-a_{1,j}\right)^{2}f_{D,j}+\frac{a_{1,j}^{2}}{\left(\rho f_{avg,j}^{-1}\right)^{2}}f_{Z,j} &= \tau_{avg,j}^{-2}f_{D,j}^{-1}+\rho^{-2}f_{avg,j}^{2}\tau_{avg,j}^{-2}\left(\tau_{avg,j}-f_{D,j}^{-1}\right)^{2}f_{Z,j} &(115)\\ &= \tau_{avg,j}^{-2}f_{D,j}^{-1}+\rho^{-2}\tau_{avg,j}^{-2}\left(\rho^{2}f_{avg}^{-1}f_{Z}^{-1}+1\right)^{2}f_{Z,j} &(116)\\ &= \tau_{avg,j}^{-2}\left(f_{D,j}^{-1}+\rho^{-2}\left(\rho^{2}f_{avg}^{-1}f_{Z}^{-1}+1\right)^{2}f_{Z,j}\right) &(117)\end{aligned}$$

So

$$E_{-1}\left[U_{i,0}\right]=\frac{1}{2}T^{-1}\sum_{j=0}^{T-1}\tau_{avg,j}^{-2}\left(f_{D,j}^{-1}+\rho^{-2}\left(\rho^{2}f_{avg}^{-1}f_{Z}^{-1}+1\right)^{2}f_{Z,j}\right)\left(\left(\rho f_{avg,j}^{-1}\right)^{2}f_{Z,j}^{-1}+f_{i,j}^{-1}+f_{D,j}^{-1}\right)-\frac{1}{2} \tag{118}$$

E Derivation of solution 2

Investors allocate attention, $f_{i,j}^{-1}$, to maximize $E_{-1}\left[U_{i,0}\right]$ subject to the constraint

$$\sum_{j,j'}f_{i,j}^{-1}\leq\bar{f}^{-1} \tag{119}$$

and that $f_{i,j}^{-1}=f_{i,j'}^{-1}$. Since the investors are maximizing a linear objective subject to a linear constraint, the optimal policy is clearly to allocate attention $f_{i,j}^{-1}$ only to the frequencies j at which the marginal benefit is equal to the maximum available marginal benefit.

Define the function λ_j

$$\lambda_{j}\left(x\right)\equiv\left(\left(\rho x\right)^{2}f_{Z,j}^{-1}+x+f_{D,j}^{-1}\right)^{-2}\left(f_{D,j}^{-1}+\rho^{-2}\left(\rho^{2}xf_{Z,j}^{-1}+1\right)^{2}f_{Z,j}\right)$$

then $\lambda_j\left(f_{avg,j}^{-1}\right)$ is the marginal benefit from attention to frequency j. Note that $d\lambda_j\left(x\right)/dx<0$. In equilibrium, then, there is a number $\bar{\lambda}$ such that

$$\lambda_{j}\left(f_{avg,j}^{-1}\right)\leq\bar{\lambda}\text{ for all }j \tag{120}$$

Now define $\mathcal{J}\left(\bar{\lambda}\right)$ to be the set of frequencies j such that $\lambda_j^{-1}\left(\bar{\lambda}\right)>0$.[23] That is the set of

[23] Technically, it is the set of frequencies for which $\lambda_j^{-1}\left(\min\left(\bar{\lambda},\lambda_j\left(0\right)\right)\right)>0$.

frequencies for which there is positive attention.

For any frequency that investors allocate attention to,

$$f_{avg,j}^{-1} = \lambda_j^{-1}\left(\bar{\lambda}\right) \tag{121}$$

$$f_{avg,j}^{-1} = \int f_{i,j}^{-1} di \tag{122}$$

Now

$$\sum_{j,j' \in \mathcal{J}} \int f_{i,j}^{-1} di = \int \sum_{j,j' \in \mathcal{J}} f_{i,j}^{-1} di \tag{123}$$

$$= \int \bar{f}^{-1} di = \bar{f}^{-1} \tag{124}$$

So then

$$\sum_{j,j' \in \mathcal{J}\left(\bar{\lambda}\right)} \lambda_j^{-1}\left(\bar{\lambda}\right) = \sum_{j,j' \in \mathcal{J}\left(\bar{\lambda}\right)} f_{avg,j}^{-1} = \bar{f}^{-1} \tag{125}$$

So $\bar{\lambda}$ is obtained by solving $\sum_{j,j' \in \mathcal{J}\left(\bar{\lambda}\right)} \lambda_j^{-1}\left(\bar{\lambda}\right) = \bar{f}^{-1}$.

F Time horizon and investment

At first glance, the assumption of mean-variance utility over cumulative returns over a long period of time ($T \to \infty$) may appear to give investors an incentive to primarily worry about long-horizon performance, whereas a small value of T would make investors more concerned about short-term performance. In the present setting, that intuition is not correct – the $T \to \infty$ limit determines how detailed investment strategies may be, rather than incentivizing certain types of strategies.

The easiest way to see why the time horizon controls only the detail of the investment strategies is to consider settings in which T is a power of 2. If $T = 2^k$, then the set of fundamental frequencies is

$$\left\{2\pi j/2^k\right\}_{j=0}^{2^{k-1}} \tag{126}$$

For $T = 2^{k-1}$, the set of frequencies is

$$\left\{2\pi j/2^{k-1}\right\}_{j=0}^{2^{k-2}} = \left\{2\pi\left(2j\right)/2^k\right\}_{j=0}^{2^{k-2}} \tag{127}$$

That is, when T falls from 2^k to 2^{k-1}, the effect is to simply eliminate alternate frequencies. Changing T does not change the lowest or highest available frequencies (which are always 0 and π, respectively). It just discretizes the $[0, \pi]$ interval more coarsely; or, equivalently, it means that the matrix Λ is constructed from a smaller set of basis vectors.

When T is smaller – there are fewer available basis functions – Q and its frequency domain analog $q \equiv \Lambda' Q$ have fewer degrees of freedom and hence must be less detailed. So the effect

of a small value of T is to make it more difficult for an investor to *isolate* particularly high- or low-frequency fluctuations in fundamentals (or any other narrow frequency range). But in no way does T cause the investor's portfolio to depend more on one set of frequencies than another. While we take $T \to \infty$, we will see that the model's separating equilibrium features investors who trade at both arbitrarily low and high frequencies, and T has no effect on the distribution of investors across frequencies.

G Proofs of specialization model predictions

G.1 Results 2 and 3

$$q_i = \rho\left(f_i^{-1} y_i + \left(\frac{a_1}{a_2^2} f_Z^{-1} - \tau_i\right) p\right)$$

The coefficient on $\tilde{\varepsilon}_i$ is f_i^{-1}. Straightforward but tedious algebra confirms that the coefficient on d is

$$\rho\left(f_{avg}^{-1} - f_i^{-1}\right)(a_1 - 1)$$

The coefficient on z is

$$1 + \rho\left(f_i^{-1} - f_{avg}^{-1}\right) a_2$$

We thus have

$$q_i = \rho\left(f_{avg}^{-1} - f_i^{-1}\right)(a_1 - 1)\, d + \left(1 + \rho\left(f_i^{-1} - f_{avg}^{-1}\right)\right) a_2 z \tag{128}$$

Now note that

$$r = (1 - a_1)\, d + d_2 z \tag{129}$$

So then

$$q_i = \rho\left(f_i^{-1} - f_{avg}^{-1}\right) r + \rho\tilde{\varepsilon}_i + z \tag{130}$$

The result on the covariance then follows trivially.

G.2 Result 4

Approximating first differences with derivatives, we obtain

$$\Delta Q_{i,t} - \Delta Z_t \approx -\sum_{j=0}^{T/2} \frac{2\pi j}{T} \begin{bmatrix} \sin\left(2\pi j t/T\right)\left(\rho\left[\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right) r_j + f_{i,j}^{-1}\tilde{\varepsilon}_{i,j}\right]\right) \\ +\cos\left(2\pi j t/T\right)\left(\rho\left[\left(f_{i,j}^{-1} - f_{avg,j}^{-1}\right) r_{j'} + f_{i,j}^{-1}\tilde{\varepsilon}_{i,j'}\right]\right) \end{bmatrix} \tag{131}$$

where the approximation becomes a true equality as $T \to \infty$. Now if we furthermore use the approximations $f_{i,j_i^*}^{-1} - f_{avg,j_i^*}^{-1} \approx \bar{f}^{-1}/2$ and suppose that the exogenous supply process is small enough that it rarely causes a trader's demand to change signs, then we have

$$|\Delta Q_{i,t}| \approx |\Delta Z_t| + \omega_{j_i^*} \bar{f}^{-1} \rho \left| \begin{array}{c} \sin\left(\omega_{j_i^*} t\right)\left(r_{j_i^*} + \tilde{\varepsilon}_{i,j_i^*}\right) \\ +\cos\left(\omega_{j_i^*} t\right)\left(r_{j_i^{*\prime}} + \tilde{\varepsilon}_{i,j_i^{*\prime}}\right) \end{array} \right|. \tag{132}$$

G.3 Result 5

$$QV\{q_j\} \equiv \sum_{t=2}^{T}(Q_{i,t}-Q_{i,t-1})^2 \approx \sum_{t=2}^{T}\left(\sum_j \frac{2\pi j}{T}\begin{bmatrix} q_j \sin(2\pi jt/T) \\ +q_{j'}\cos(2\pi jt/T)\end{bmatrix}\right)^2 \tag{133}$$

$$= \sum_{t=2}^{T}\sum_{j,k}\left(\frac{2\pi}{T}\right)^2 jk\begin{bmatrix} q_j \sin(2\pi jt/T)\, q_k \sin(2\pi kt/T) + q_{j'}\cos(2\pi jt/T)\, q_k \sin(2\pi kt/T) \\ q_j \sin(2\pi jt/T)\, q_{k'}\cos(2\pi kt/T) + q_{j'}\cos(2\pi jt/T)\, q_{k'}\cos(2\pi kt/T)\end{bmatrix} \tag{134}$$

$$\approx \sum_{j,j'}(2\pi j)^2 T^{-1} q_j^2 \tag{135}$$

where the equality in the first line is approximate in assuming that $\cos(2\pi jt/T) - \cos(2\pi j(t-1)/T) \approx \frac{2\pi j}{T}\sin(2\pi jt/T)$ and the same for the differences in the sines. The third line uses the fact that sines of unequal frequencies are orthogonal (it is approximate because $t=1$ is not included in the sum inserts the integral for $\sin^2$ and $\cos^2$, rather than the exact finite sums. All the approximations here are accurate for large T.

H Proofs of trading restriction results

H.1 Results 6 and 7

If trade by the investors is not allowed at certain frequencies, then obviously markets cannot clear at those frequencies when supply is inelastic. In this section we therefore first solve the model for the case with an upward sloping supply curve and then analyze the effect of eliminating trade on asset prices and returns.

H.1.1 Equilibrium with elastic supply

We now assume that there is exogenous supply on each date of

$$Z_t = \tilde{Z}_t + kP_t \tag{136}$$

where k is a constant determining the slope of the supply curve. One could imagine allowing k to differ across frequencies, which would be equivalent to allowing supply to depend on prices on multiple dates (intuitively, maybe supply increases by more when prices have been persistently high than when they are just temporarily high). Here, though, we simply leave k constant across frequencies. Multiplying by Λ' yields

$$z_j = \tilde{z}_j + kp_j \tag{137}$$

Solving the inference problem as before, we obtain

$$\tau_i \equiv Var\left[d \mid y_i, p\right]^{-1} \tag{138}$$
$$= \frac{a_1^2}{a_2^2} f_{\tilde{Z}}^{-1} + f_i^{-1} + f_D^{-1} \tag{139}$$

and

$$E\left[d \mid y_i, p\right] = \tau_i^{-1}\left(f_i^{-1} y_i + \frac{a_1}{a_2^2} f_{\tilde{Z}}^{-1} p\right) \tag{140}$$

H.1.2 Demand and equilibrium

The investors' demand curves are again

$$q_i = \rho_i\left(f_i^{-1} y_i + \left(\frac{a_1}{a_2^2} f_{\tilde{Z}}^{-1} - \tau_i\right) p\right)$$

Summing up all demands and inserting the guess for the price process yields

$$\tilde{z} + k\left(a_1 d - a_2 \tilde{z}\right) = \int_i \rho\left(f_i^{-1} d + \left(\frac{a_1}{a_2^2} f_{\tilde{Z}}^{-1} - \tau_i\right)\left(a_1 d - a_2 \tilde{z}\right)\right) di \tag{141}$$

Matching coefficients yields

$$\int_i \rho\left(\frac{a_1}{a_2^2} f_{\tilde{Z}}^{-1} - \tau_i\right) di = -a_2^{-1}\left(1 - k a_2\right) \tag{142}$$
$$\int_i \rho f_i^{-1} + \rho\left(\frac{a_1}{a_2^2} f_{\tilde{Z}}^{-1} - \tau_i\right) a_1 di = k a_1 \tag{143}$$

Combining those two equations, we have

$$\rho f_{avg}^{-1} = a_1\left(k + a_2^{-1}\left(1 - k a_2\right)\right) \tag{144}$$
$$= \frac{a_1}{a_2} \tag{145}$$

$$a_1 = \frac{f_{avg}^{-1} + \left(\rho f_{avg}^{-1}\right)^2 f_{\tilde{Z}}^{-1}}{\tau_{avg} + \rho^{-1} k} \tag{146}$$
$$= \frac{\tau_{avg} - f_D^{-1}}{\tau_{avg} + \rho^{-1} k} \tag{147}$$
$$a_2 = \frac{a_1}{\rho f_{avg}^{-1}} \tag{148}$$

H.1.3 Utility

As before, the contribution to optimized utility from frequency j is

$$\left((1-a_{1,j})^2 f_{D,j} + \frac{a_{1,j}^2}{\left(\rho f_{avg}^{-1}\right)^2} f_{Z,j} \right) \tau_{i,j} \tag{149}$$

Furthermore,

$$\begin{aligned}
(1-a_{1,j})^2 f_{D,j} + \frac{a_{1,j}^2}{\left(\rho f_{avg,j}^{-1}\right)^2} f_{\tilde{Z},j} &= \left(\frac{\rho^{-1}k + f_D^{-1}}{\tau_{avg} + \rho^{-1}k} \right)^2 f_{D,j} + \rho^{-2} f_{avg,j}^2 \left(\frac{\tau_{avg} - f_D^{-1}}{\tau_{avg} + \rho^{-1}k} \right)^2 f_{\tilde{Z},j} \\
&= \left(\tau_{avg} + \rho^{-1}k\right)^{-2} \left(\left(\rho^{-1}k + f_D^{-1}\right)^2 f_{D,j} + \rho^{-2} f_{avg,j}^2 \left(\left(\rho f_{avg}^{-1}\right)^2 f_{\tilde{Z}}^{-1} + f_{avg}^{-1} \right)^2 f_{\tilde{Z},} \right. \\
&= \left(\tau_{avg} + \rho^{-1}k\right)^{-2} \left(\left(\rho^{-1}k + f_D^{-1}\right)^2 f_{D,j} + \rho^{-2} \left(\rho^2 f_{avg}^{-1} f_{\tilde{Z}}^{-1} + 1 \right)^2 f_{\tilde{Z},j} \right)
\end{aligned}$$

So

$$E_{-1}\left[U_{i,0}\right] = \frac{1}{2} T^{-1} \sum_{j=0}^{T-1} \left[\begin{array}{c} \left(\tau_{avg,j} + \rho^{-1}k\right)^{-2} \left(\left(\rho^{-1}k + f_{D,j}^{-1}\right)^2 f_{D,j} + \rho^{-2} \left(\rho^2 f_{avg,j}^{-1} f_{\tilde{Z},j}^{-1} + 1 \right)^2 f_{\tilde{Z},j} \right) \\ \times \left(\left(\rho f_{avg,j}^{-1}\right)^2 f_{Z,j}^{-1} + f_{i,j}^{-1} + f_{D,j}^{-1} \right) \end{array} \right] - \frac{1}{2} \tag{150}$$

When there are no active investors and just exogenous supply, we have

$$\begin{aligned}
0 &= \tilde{z}_j + k p_j && (151) \\
p_j &= -k^{-1} \tilde{z}_j && (152) \\
r_j &= d_j - k^{-1} \tilde{z}_j && (153)
\end{aligned}$$

We then have

$$\begin{aligned}
f_R &= f_D + \frac{f_Z}{k^2} && (154) \\
f_{R,0} &= f_{D,j} + \frac{f_{\tilde{Z},j}}{\left(k + \rho f_{D_j}^{-1}\right)^2} && (155)
\end{aligned}$$

H.2 Result 9

We have

$$D \mid Y, P \sim N\left(\bar{D}, \Lambda diag\left(\tau_0^{-1}\right) \Lambda'\right) \tag{156}$$

where τ_0 is a vector of frequency-specific precisions conditional on prices. Now consider some average over D, $F'D$, where F is a column vector. Then

$$\begin{aligned} Var\left(D_t\right) &= 1_t'\Lambda diag\left(\tau_0^{-1}\right)\Lambda' 1_t && (157)\\ &= \left(\Lambda' 1_t\right)' diag\left(\tau_0^{-1}\right)\left(\Lambda' 1_t\right) && (158)\\ &= \sum_{j,j'} \lambda_{t,j}^2 \tau_{0,j}^{-1} && (159)\\ &= \lambda_{t,0}^2\tau_{0,0}^{-1} + \lambda_{t,T/2}^2\tau_{0,0}^{-1} + \sum_{n=1}^{T/2-1}\left(\lambda_{t,n}^2 + \lambda_{t,n'}^2\right)\tau_{0,n}^{-1} && (160)\end{aligned}$$

where 1_t is a vector equal to 1 in its tth element and zero elsewhere and $\lambda_{t,j}$ is the jth trigonometric transform evaluated at t, with

$$\begin{aligned} \lambda_{t,j} &= \sqrt{2/T}\cos\left(2\pi j\left(t-1\right)/T\right) && (161)\\ \lambda_{t,j'} &= \sqrt{2/T}\sin\left(2\pi j\left(t-1\right)/T\right) && (162)\\ \lambda_{t,0} &= \sqrt{1/T} && (163)\\ \lambda_{t,T/2} &= \sqrt{1/T}\cos\left(\pi\left(t-1\right)\right) = \sqrt{1/T}\left(-1\right)^{t-1} && (164)\end{aligned}$$

More generally, then

$$\begin{aligned} Var\left(\frac{1}{s}\sum_{m=0}^{s-1} D_{t+m}\right) &= \frac{1}{s^2}\left(\sum_{m=0}^{s-1} 1_{t+m}\right)'\Lambda diag\left(\tau_0^{-1}\right)\Lambda'\left(\sum_{m=0}^{s-1} 1_{t+m}\right) && (165)\\ &= \frac{1}{s^2}\left(\sum_{m=0}^{s-1}\lambda_{t+m,0}\right)^2\tau_{0,0}^{-1} + \frac{1}{s^2}\left(\sum_{m=0}^{s-1}\lambda_{t+m,T/2}\right)^2\tau_{0,T/2}^{-1} && (166)\\ &\quad + \frac{1}{s^2}\sum_{n=1}^{T/2-1}\left[\left(\sum_{m=0}^{s-1}\lambda_{t+m,n}\right)^2 + \left(\sum_{m=0}^{s-1}\lambda_{t+m,n'}\right)^2\right]\tau_{0,n}^{-1} && (167)\end{aligned}$$

where $\tau_{0,n}$ is the frequency-n element of τ_0. For $0 < n < T/2$

$$\left(\sum_{m=0}^{s-1}\lambda_{t+m,n}\right)^2 + \left(\sum_{m=0}^{s-1}\lambda_{t+m,n}\right)^2 = \sum_{m=0}^{s-1}\sum_{k=0}^{s-1}\frac{2}{T}\left[\begin{array}{c}\cos\left(2\pi n\left(t+m-1\right)/T\right)\cos\left(2\pi n\left(t+k-1\right)/T\right)\\ +\sin\left(2\pi n\left(t+m-1\right)/T\right)\sin\left(2\pi n\left(t+k-1\right)/T\right)\end{array}\right] \quad (168)$$

Now note that

$$2\cos\left(x\right)\cos\left(y\right) + 2\sin\left(x\right)\sin\left(y\right) = 2\cos\left(x-y\right) \quad (169)$$

So we have

$$\left(\sum_{m=0}^{s-1}\lambda_{t+m,n}\right)^2+\left(\sum_{m=0}^{s-1}\lambda_{t+m,n}\right)^2 = \frac{2}{T}\sum_{m=0}^{s-1}\sum_{k=0}^{s-1}\cos\left(\frac{2\pi n}{T}(m-k)\right) \quad (170)$$

$$= 2\frac{s}{T}\sum_{m=-(s-1)}^{s-1}\frac{s-|m|}{s}\cos\left(\frac{2\pi n}{T}m\right) \quad (171)$$

$$= 2\frac{s}{T}F_s\left(\frac{2\pi n}{T}\right) \quad (172)$$

$$= \frac{2}{T}\frac{1-\cos\left(s\frac{2\pi n}{T}\right)}{1-\cos\left(\frac{2\pi n}{T}\right)} \quad (173)$$

where F_s denotes the sth-order Fejér kernel. Note that when $s=T$, the above immediately reduces to zero, since $\cos(2\pi n)=0$. That is the desired result, as an average over all dates should be unaffected by fluctuations at any frequency except zero.

For $n=0$,

$$\left(\sum_{m=0}^{s-1}f_{t+m,0}\right)^2 = \left(\sum_{m=0}^{s-1}\sqrt{1/T}\right)^2 \quad (174)$$

$$= \left(s\frac{1}{T^{1/2}}\right)^2 \quad (175)$$

$$= \frac{s}{T}F_s(0) \quad (176)$$

Since $F_s(0)=s$ (technically, this holds as a limit: $\lim_{x\to 0}F_s(x)=s$).

For $n=T/2$,

$$\left(\sum_{m=0}^{s-1}f_{t+m,T/2}\right)^2 = \frac{1}{T}\left(\sum_{m=1}^{s}(-1)^m\right)^2 = \begin{cases}\frac{1}{T}\text{ for odd }s\\ 0\text{ otherwise}\end{cases} \quad (177)$$

$$= \frac{s}{T}\frac{1}{s}\left(\frac{\sin(s\pi/2)}{\sin(\pi/2)}\right)^2 = \frac{s}{T}F_s(\pi) \quad (178)$$

So we finally have that

$$Var\left(\frac{1}{s}\sum_{m=0}^{s-1}D_{t+m}\right)=\frac{1}{sT}\sum_{j,j'}F_s(\omega_j)\tau_{0,j}^{-1} \quad (179)$$

I Costly learning about prices

I.1 Generic result: no learning from prices

Lemma 3 *Assume that learning from prices is costly. At that at time -1, if agent i decides to infer information from prices, then their capacity constraint is:*

$$Tr(f_i^{-1} + f_P^{-1}) \leq \bar{f}^{-1},$$

where f_P^{-1} is inverse of the variance-covariance matrix of signals contained in prices, and f_i^{-1} is the variance-covariance of the private signals of agent i. On the other hand, if agent i decides not to infer information from prices, then his capacity constraint is:

$$Tr(f_i^{-1}) \leq \bar{f}^{-1}.$$

Then, agents always prefer not to learn from prices.

Proof. If agent i has decided not to learn from prices, then at time 0, their posterior distribution over d is:

$$\begin{aligned} d|y_i \quad &\sim N\left(\mu(y_i), \tau_i^{-1}\right) \\ \tau_i^{NP} \quad &= f_D^{-1} + f_i^{-1} \\ \mu(y_i) \quad &= (\tau_i^{NP})^{-1} f_i^{-1} y_i \end{aligned} \tag{180}$$

Agent i still observes prices; their first-order condition leads to the demand schedule:

$$q_i = \rho \tau_i^{NP} \left(\mu(y_i) - p\right).$$

His time-0 utility is:

$$U_{0,i}^{NP}(y_i; p) \quad = \tfrac{1}{2T} \left(\mu(y_i) - p\right)' \tau_i^{NP} \left(\mu(y_i) - p\right). \tag{181}$$

Since τ_i^{NP} is symmetric, this implies:

$$E_{-1,i}\left[U_{0,i}^{NP}\right] \quad = \tfrac{1}{2T} tr(\tau_i^{NP} V_i^{NP}) + \tfrac{1}{2T} (\mu_i^{NP})' \tau_i^{NP} \mu_i^{NP}, \tag{182}$$

where as before:

$$\begin{aligned} \mu_i^{NP} \quad &= E_{-1,i}\left[\mu(y_i) - p\right] \\ V_i^{NP} \quad &= Var_{-1,i}\left[\mu(y_i) - p\right] \end{aligned} \tag{183}$$

As before, because all fundamentals are mean 0, $\mu_i = 0$. Moreover, by the law of total variance:

$$V_i = \underbrace{Var_{-1}\left[d - p\right]}_{\equiv V_{-1}} - (\tau_i^{NP})^{-1}$$

Therefore,

$$\begin{aligned} E_{-1,i}\left[U_{0,i}^{NP}\right] &= \tfrac{1}{2T} tr(\tau_i^{NP} V_i) \\ &= \tfrac{1}{2T} tr(\tau_i^{NP} V_{-1}) - \tfrac{1}{2T} tr(I) \\ &= \tfrac{1}{2T} tr\left(f_D^{-1} V_{-1}\right) - \tfrac{1}{2T} tr(I) + \tfrac{1}{2T} tr(f_i^{-1} V_{-1}) \end{aligned} \tag{184}$$

The time-(-1) attention allocation problem of such an agent is therefore:

$$\begin{aligned} U_{-1,i}^{NP}\left(f_{avg}^{-1}\right) &= -\tfrac{1}{2} + \tfrac{1}{2T} tr\left(f_D^{-1} V_{-1}\right) + \quad \frac{1}{2T} \max_{f_i^{-1}} \quad tr(f_i^{-1} V_{-1}) \\ \text{s.t.} \qquad & f_{i,j}^{-1} \geq 0 \quad \forall j \in [0, ..., T-1] \\ & tr(f_i^{-1}) \leq \bar{f}^{-1} \end{aligned} \tag{185}$$

For an agent who does learn from prices (but shares the other agent's ex-ante distribution over p and d, summarized by V_{-1}), the attention allocation problem has already been derived; it is given by:

$$\begin{aligned} U_{-1,i}\left(f_{avg}^{-1}\right) &= -\tfrac{1}{2} + \tfrac{1}{2T} tr\left(\left(f_D^{-1} + f_P^{-1}\right) V_{-1}\right) + \quad \frac{1}{2T} \max_{f_i^{-1}} \quad tr(f_i^{-1} V_{-1}) \\ \text{s.t.} \qquad & f_{i,j}^{-1} \geq 0 \quad \forall j \in [0, ..., T-1] \\ & tr(f_i^{-1} + f_P^{-1}) \leq \bar{f}^{-1} \end{aligned} \tag{186}$$

Since f_i^{-1} is diagonal, $f_i^{-1} \to tr(f_i^{-1} V_{-1})$ can be thought of as a linear map on R^T. By the Riesz representation theorem, there is $\lambda \in R^T$ such that $\forall f_i^{-1}$, $tr(f_i^{-1} V_{-1}) = \sum_{j=0}^{T-1} f_{i,j}^{-1} \lambda_j$. Let $\tilde{\lambda}$ denote the element-wise maximum of λ. Note, in particular, that:

$$tr(f_P^{-1} V_{-1}) = \sum_{j=0}^{T-1} f_{P,j}^{-1} \lambda_j.$$

Moreover, after optimization, not learning through prices yields utility:

$$U_{-1,i}^{NP}\left(f_{avg}^{-1}\right) = -\frac{1}{2} + \frac{1}{2T} tr\left(f_D^{-1} V_{-1}\right) + \frac{1}{2T} \tilde{\lambda} \bar{f}^{-1}.$$

Learning through prices yields utility:

$$U_{-1,i}\left(f_{avg}^{-1}\right) = -\frac{1}{2} + \frac{1}{2T} tr\left(\left(f_D^{-1} + f_P^{-1}\right) V_{-1}\right) + \frac{1}{2T} \tilde{\lambda}\left(\bar{f}^{-1} - tr(f_P^{-1})\right)$$

The difference between the two is:

$$\begin{aligned} U_{-1,i}^{NP}\left(f_{avg}^{-1}\right) - U_{-1,i}\left(f_{avg}^{-1}\right) &= \tfrac{1}{2T} \tilde{\lambda} tr(f_P^{-1}) - \tfrac{1}{2T} tr\left(f_P^{-1} V_{-1}\right) \\ &= \tfrac{1}{2T} \tilde{\lambda} tr\left(f_P^{-1}\right) - \tfrac{1}{2T} \textstyle\sum_{j=0}^{T-1} f_{P,j}^{-1} \lambda_j \\ &\geq 0 \end{aligned} \tag{187}$$

Therefore, the agent always prefer not to learn from prices. ■

I.2 The equilibrium when agents do not learn about prices

Guess:

$$p = a_3 d - a_4 z$$

with a_3, a_4 diagonal matrices of size $T \times T$. Straightforward derivations lead to:

$$\begin{aligned} a_3 &= I - (\tau_{avg} + kI)^{-1} \left(f_D^{-1} + kI\right) \\ &= (\tau_{avg} + kI)^{-1} f_{avg}^{-1} \\ a_4 &= \tfrac{1}{\rho} a_3 f_{avg} \\ &= \tfrac{1}{\rho} (\tau_{avg} + kI)^{-1} \\ \tau_{avg} &= f_{avg}^{-1} + f_D^{-1} \\ \tau_i &= f_i^{-1} + f_D^{-1} \end{aligned} \tag{188}$$

Moreover, expected utility is given by:

$$\begin{aligned} E_{-1,i}\left[U_{0,i}^{NP}\right] &= \mathcal{C}^{NP} + \tfrac{1}{2T} tr(V_{-1}^{NP} f_i^{-1}) \\ V_{-1}^{NP} &= f_D \left((I + k f_D)^2 + \tfrac{f_Z f_D}{\rho^2}\right) (I + k f_D + f_D f_{avg}^{-1})^{-2} \\ \mathcal{C}^{NP} &= \tfrac{1}{2T} tr\left(f_D^{-1} V_{-1}^{NP}\right) - \tfrac{1}{2} \end{aligned} \tag{189}$$

J Calibration

$\bar{f}^{-1} = 0.01$

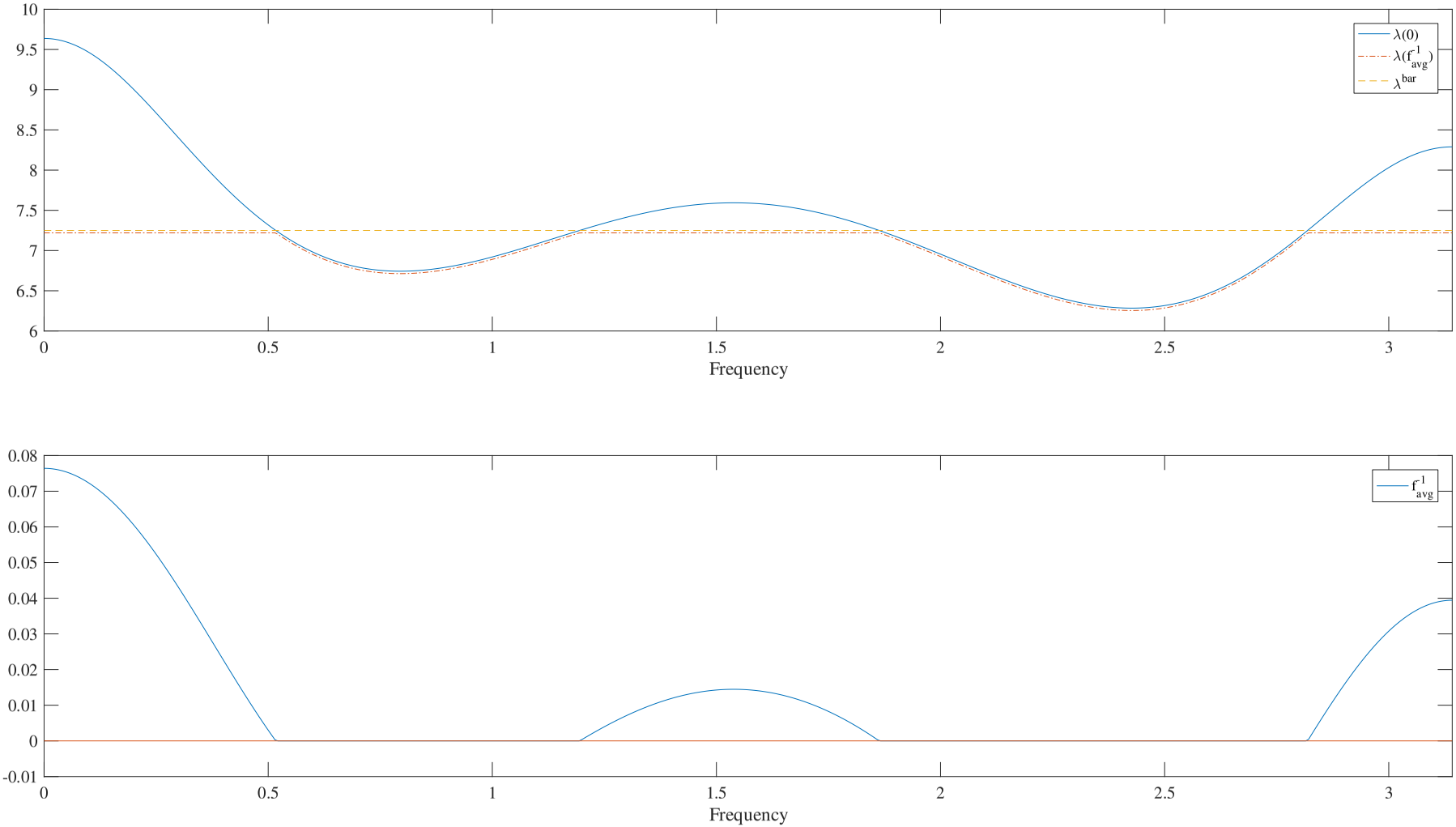
$T = 1000$

$f_D(\omega) = \frac{1}{4}\left|1 - \frac{1}{2}e^{i\omega}\right|^{-2} + 1 - .55\cos(2\omega) + \frac{7}{16}\left|1 + \frac{1}{2}e^{i\omega}\right|^{-2}$

$f_Z(\omega) = \frac{1}{10}\left|1 - \frac{1}{2}e^{i\omega}\right|^{-2}$

$\rho = 1$

Figure 1: Optimal information acquisition and waterfilling



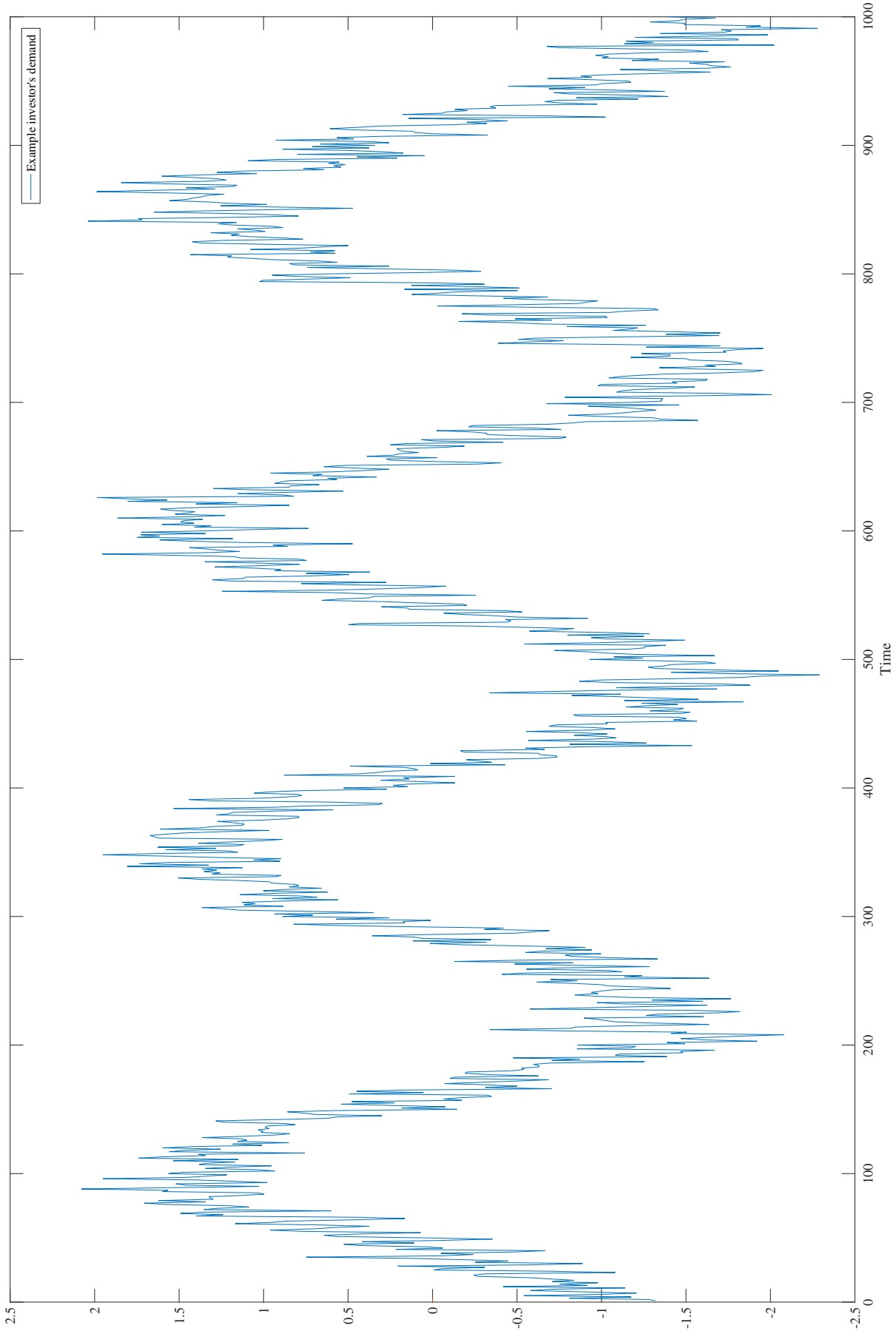


Figure 2: Example investor's demand

Figure 3: Effects of restricting high-frequency trade

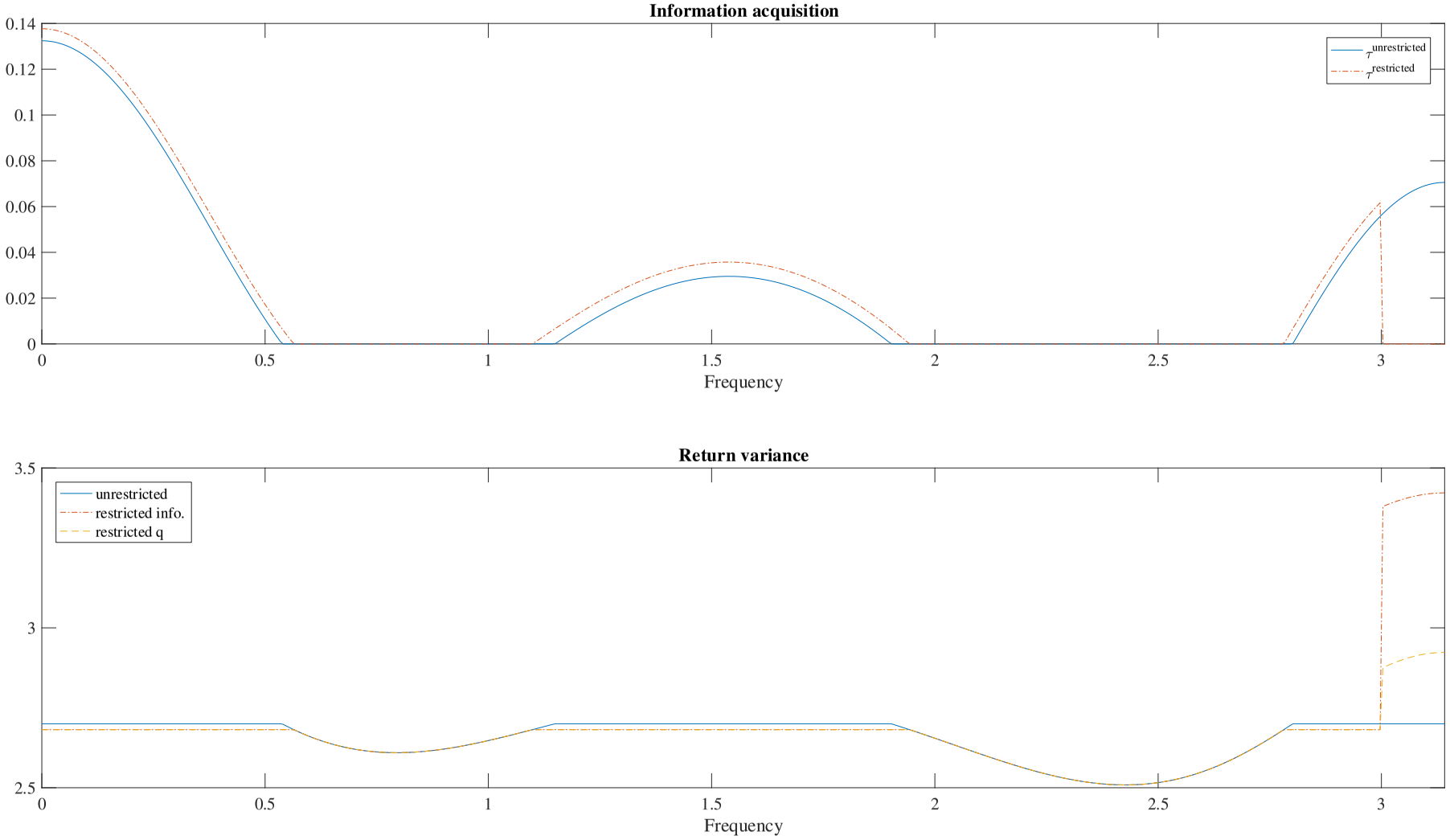


Figure 4: Weights across frequencies and time

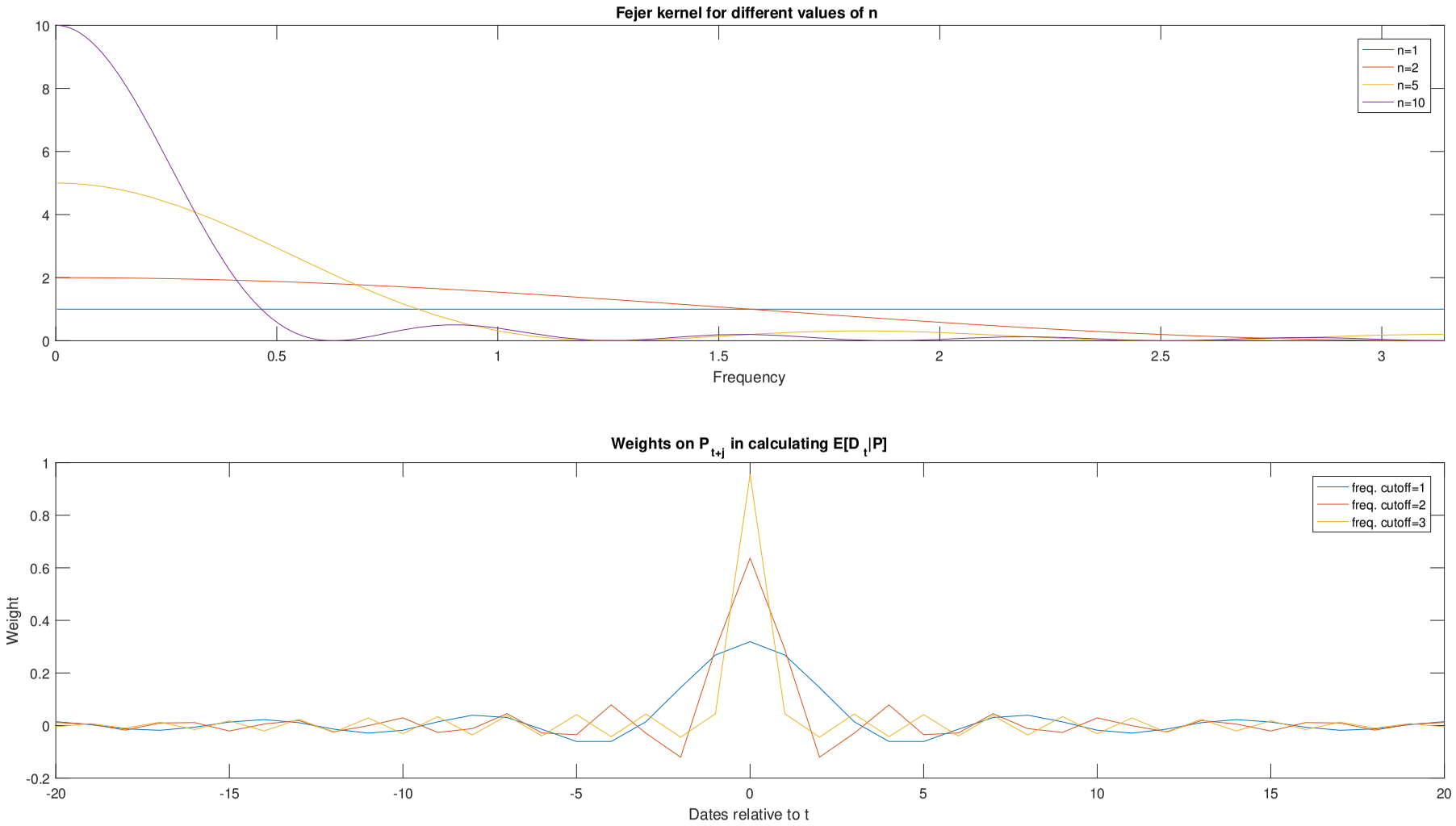


Figure 5: Persistence of the churn rate over time

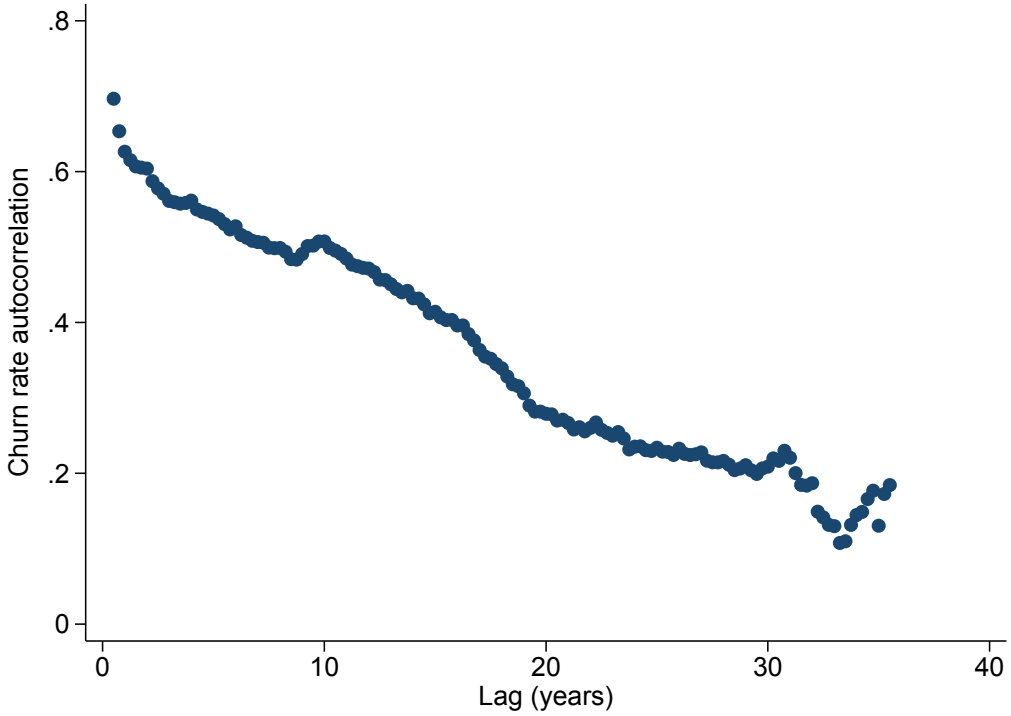
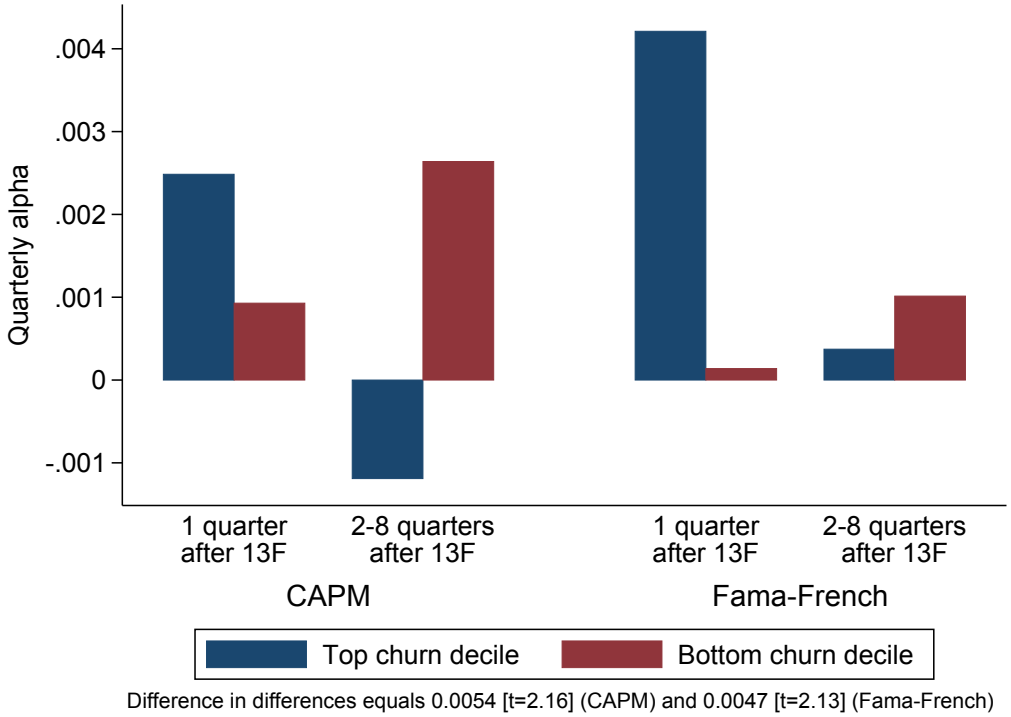


Figure 6: Out-performance of institution holdings at different horizons



Secondary Market Trading and the Cost of New Debt Issuance

Ryan L. Davis, David A. Maslar, and Brian S. Roseman*

February 8, 2017

ABSTRACT

We show that secondary market activity impacts the cost of issuing new debt in the primary market. Specifically, firms with existing illiquid debt have higher costs when issuing new debt. We also find that with the improvement in the price discovery process brought about by the introduction of TRACE reporting, firms that became TRACE listed subsequently had a lower cost of debt. Our results indicate that the secondary market functions of liquidity and price discovery are important to the primary market. Overall, the results presented in this paper provide a greater understanding of the connection between the secondary market and the real economy.

*Ryan L. Davis is at The Collat School of Business, University of Alabama at Birmingham; David A. Maslar is at The Haslam College of Business, University of Tennessee; Brian S. Roseman is at The Mihaylo College of Business and Economics, California State University, Fullerton. We would like to thank Andrew Lynch, Joseph Greco, David Nanigian, Matthew Serfling, and seminar participants at the University of Mississippi and California State University, Fullerton for their helpful suggestions and feedback.

Understanding how financial market activity impacts the real economy is one of the most important topics studied by financial economists. Since firms only raise capital in the primary market it is easy to conclude that trading in the secondary market does not directly affect firm activity, or in turn, the real economy. This potential disconnect leads some to view secondary markets as merely a sideshow to the real economy, an idea that has been debated in the academic literature since at least Bosworth (1975). Recent events have revived and added new dimensions to this debate.[1] The discussion that is now taking place in both the academic literature and the popular press indicates that that this question remains both prevalent and ever changing.

To contribute to this debate, we consider whether two important benefits of secondary markets, liquidity and price discovery, impact the primary market. To this end, we empirically investigate two questions: 1. do firms with illiquid bonds face higher costs when issuing new debt, and 2. does price discovery in the secondary bond market impact a firm's cost of issuing new debt? By answering these questions, we seek to address the broader question: how does secondary market activity affect the real economy?

The view that secondary markets impact the real economy begins with the argument that access to capital is an important determinant of growth. The results in the literature consistently indicate that this relation holds at the country, industry, and firm levels. This question has been examined in numerous studies, including the seminal paper by Rajan and Zingales (1996). The literature has evolved to the point where we now better understand the channels that connect growth and access to capital. Empirical evidence, for example, indicates that access to financing is important for firm investment (Stein (2003); Chava and Roberts (2008); Campello and Graham (2013)). Surveys of corporate decision makers also support this view. For

[1] Some examples include: bailouts given during the financial crisis and the resulting "Main Street" versus "Wall Street" debate arising from the Occupy Wall Street protests (Kuziemko, Norton, Saez, and Stantcheva (2015)), questions regarding the relation between economic growth and equity returns (Ritter (2005); Ritter (2012)), and questions related to the controversial practice of using corporate repurchases to prop up firm growth (Driebusch and Eisen, "Buybacks Pump Up Stock Rally," *The Wall Street Journal*, 7/13/2016, Section C1).

example, after surveying 1,050 Chief Financial Officers (CFOs), Campello, Graham, and Harvey (2010) report that firms facing financial constraints reduce their investment in both technology and fixed capital and also reduce employment.

Based on the theoretical and empirical evidence provided in the literature, we begin with the view that access to capital affects firm activity. From this, we argue that frictions affecting firm access to capital may impact the real economy. The channels we focus on are secondary market liquidity and price discovery. If, for example, an increase in secondary market illiquidity raises a firm's cost of capital or prevents a firm from accessing capital all together, then we can conclude that secondary market illiquidity could hamper a firm's growth.[2]

As Maureen O'Hara discusses in her AFA Presidential Address (O'Hara (2003)), liquidity and price discovery are two of the most important functions of a market. The precise roles that liquidity and price discovery play are still being explored in the literature, with many papers logically focusing on whether secondary market liquidity and price discovery affect trading in the secondary market. For example, when framing the question, O'Hara (2003) focuses on the importance of liquidity and price discovery for asset pricing. These questions are clearly important to the literature, and would likely be important regardless of whether there is a connection between the primary and secondary markets. However, if frictions that arise in the secondary market impact the primary markets as well, then questions of liquidity and price discovery take on an additional level of importance. As Morck, Shleifer, and Vishny (1990) argue, if the secondary market is in fact a sideshow, then any inefficiencies that arise in the secondary market solely represent wealth transfers amongst secondary market participants. While we by no means intend to trivialize the understanding of what could be "wealth transfers" and believe that understanding the trading process is important for its own sake, it is also important

[2] There is some evidence that greater liquidity can actually be detrimental to the real activities of a firm. Fang, Tian, and Tice (2014), for example, find that greater liquidity can actually impede firm innovation. The authors attribute the relation to an increase in liquidity leading to an increase chance of a hostile takeover and a decrease in monitoring by institutional investors. Given the question raised by Fang, et al. (2014), understanding precisely how secondary market liquidity impacts a firm's cost of debt is important.

to note that connecting this process to the primary market may significantly change the scope of inquiry.

We thus examine whether liquidity and price discovery in the corporate bond market impact the primary market for new debt issues. Mauer and Senbet (1992) and Ellul and Pagano (2006) argue that the secondary market affects pricing in the primary market for IPOs. The latter, for example, suggests that greater expected after-market illiquidity results in greater IPO underpricing. While liquidity and price discovery are important elements of all markets, as Green, Li, and Schürhoff (2010) argue, they are especially important in less liquid markets. In the corporate bond market, for example, Chen, Lesmond, and Wei (2007), Bao, Pan, and Wang (2011), Friewald, Jankowitsch, and Subrahmanyam (2012), and Dick-Nielsen, Feldhütter, and Lando (2012) show that bond illiquidity is positivity related to the cross-section of bond returns. As the evidence in the literature indicates that illiquidity impacts expected returns, there is an implied argument that secondary market illiquidity influences a firm's cost of raising new capital (Amihud and Mendelson (1986)). Fundamental to this argument is the view that expected equity returns and bond yields are proxies for a firm's cost of capital.[3] While this view implies that secondary market illiquidity and the cost of raising new capital are linked, we look to test this conjecture directly.

Using the laboratory of publicly traded debt, we examine the effects of secondary market illiquidity and price discovery on the primary market. Using publicly traded debt in our study is advantageous because firms frequently enter, and often revisit, the bond market. While firms can reenter the equity market using SEOs, this activity is comparatively limited: firms tend to enter the bond market with greater frequency. Moreover, firms frequently have multiple bond issues outstanding, and may issue new bonds before the existing bonds mature. Because some firms have

[3] There is a debate in the literature that raises questions as to whether ex post returns are a precise proxy for a firm's cost of capital. As Chen, Chen, and Wei (2011) discuss, ex post returns may reflect other information than a firm's cost of capital, such as grown opportunities and changes in investors' risk preferences (Stulz (1999); Hail and Leuz (2009)), and are also susceptible to questions with respect to the selection of asset pricing model (Fama and French (1997)).

multiple bonds simultaneously trading in the secondary market, we are able to measure the expected illiquidity of a new issue before it begins trading using the illiquidity of the firm's outstanding bonds as a proxy for anticipated illiquidity. By doing so, we can examine the relation between the actual cost of debt and expected market illiquidity, rather than the relation between the expected cost of capital and actual market illiquidity. With varying maturities, coupon structures, and credit risk, the degree of heterogeneity amongst bonds, as well as the cross-sectional differences in bond risks and characteristics, produce cross-sectional variation in bond liquidity.[4] Our empirical approach also allows us to determine if firms with more liquid bonds are disproportionally able to access the debt markets during periods of distress, such as the financial crisis. If secondary market liquidity affects access to capital, then a regulatory objective designed to improve market liquidity will impact a firm's ability to raise new funds. Understanding this channel is generally important, but may be particularly relevant during a liquidity crisis.[5]

Additionally, the staggered implementation of the Trade Reporting and Compliance Engine (TRACE) and the subsequent release of all bond trading data through the Enhanced TRACE files provides us with a unique setting for testing the impact of secondary market price discovery on the primary market. As TRACE now provides two data files, one containing information that was disseminated at the time and one that backfills additional data, we are able to examine the impact of trading when prices are not disseminated to the public – an important component of price discovery. Because TRACE was implemented in 2002, we now have a sufficient time series available to conduct empirical tests. The available data also allows us to

[4] Chen, et al. (2007), Bao, et al. (2011), Friewald, et al. (2012), and Dick-Nielsen, et al. (2012) each not only examine the relation between expected returns and bond illiquidity, but also consider the characteristics that impact this relation.

[5] As many papers have shown (Amihud (2002), for example), both individual security illiquidity and aggregate market illiquidity change over time. Furthermore, both managers and regulators can institute changes that directly influence market liquidity. Managers, for example, can alter secondary market liquidity and price discovery by changing the information environment (disclosure) and changing their exchange listing. The results of this paper also offer important implications for changes in regulation. If channels exist that connect the real economy to the secondary market, then regulations intended to improve secondary market transparency have implications for the real economy.

circumvent many of the objections raised in the literature regarding the estimation of a firm's cost of capital (Fama and French (1997)).

In total, our results suggest a direct relation between the secondary market illiquidity of existing bonds and the cost of new debt issued by the same firm in the primary market. Furthermore, we find evidence that greater illiquidity is a significant predictor of a firm's ability to issue new debt. Thus, not only is issuing new debt costlier for firms with illiquid debt, but firms with illiquid debt may have difficulty accessing credit markets altogether. We also find that TRACE-reported bonds experience lower underwriting costs relative to bonds that were not immediately subject to TRACE-reporting requirements. As the staggered implementation of TRACE provides us a way to capture the benefits of secondary market price discovery for primary market participants, we conclude that a more efficient price discovery process also leads to lower costs in the primary market for new debt issuances. The evidence presented in this paper supports theory suggesting that secondary market activity affects the real economy. Efforts to improve liquidity and price discovery, such as changes in disclosure and the implementation of TRACE, serve to not only improve the secondary market trading environment, but also to provide firms with better access to capital. Better access to capital, in turn, provides firms with better investment options and could potentially improve employment prospects.

In this regard, our analysis contributes to the growing literature that explores connections between secondary market trading and the real economy.[6] In his AFA

[6] As this question is important to the academic literature, it takes on many forms. Aslan and Kumar (2016), for example, show that hedge fund activism in a given firm can impact rival firms' product market performance. Grullon, Michenaud, and Weston (2015) show that short selling constraints impact a firm's ability to access capital and thus impact firm investment. Using the conversion of non-tradable to tradable stocks in China, Campello, Ribas, and Wang (2014) show, how secondary market trading can directly impact corporate activity. And, as McLean and Zhao (2014) discuss, the recent financial crisis not only emphasizes the importance of understanding the connection between financial markets and the real economy, but also provides a laboratory for assessing the extent of the connection. While all of these papers examine different channels, the important underlying commonality is that they all contribute to a better understanding of connections between primary and secondary market activity.

Presidential Address (Zingales (2015)): Does Finance Benefit Society? Luigi Zingales states (p. 1337): "To this day, empirical measures of the benefits of an efficient market are fairly elusive." By directly examining the link between two defining features of the secondary market, liquidity and price discovery, and the real economy, we seek to identify and quantify just such a benefit.

I. Overview of New Corporate Bond Issuances

A. *The underwriting process*

We begin by describing the underwriting process and primary market for new debt issuances. The underwriting process motivates our examination of the link between secondary market activity and the cost of new issues in the primary market.

When a firm decides to raise capital through the issuance of new bonds, it will seek an investment banker to underwrite the new issue and act as an intermediary between the firm and investors. The choice of a lead underwriter(s) is critical to the bond's success. An underwriter's ability, experience, reputation, and strength of relationships with investors are all considered in the selection process (Fang (2005)). Potential underwriters will submit an initial prospectus detailing pricing, strategies, and underwriter compensation. Once chosen, a lead underwriter may form an underwriting syndicate to spread the risk of the new issue and improve the likelihood of selling all of the securities.[7] The underwriter(s) typically has a prearranged group of institutional investors interested in the new debt issue. Underwriters must balance the preferences of these institutional clients with a debt structure (i.e. bond maturity, coupon, and price) that meets the needs of the issuing firm. Satisfying both institutional investors and the issuing firms requires adjusting the bond's yield.

Underwriters make known the firm's intention to issue new debt, help the issuer prepare disclosure documents and prospectuses, and accept indications of interest from investors. Unlike new equity issuances, bond issues typically forego the lengthy roadshow and conference call process. As a result, the time between the

[7] The underwriter may also employ a selling group to assist in selling shares to investors.

announcement and when the bond begins to trade varies from a few hours to days.[8] Even though the timeline for the bookmaking process may vary, many of the details of the issue are not set until the end of the process. Consequently, issuers maintain some flexibility in issue size as well as which orders, if any, to fill. The underwriting process concludes by setting the coupon and initial issue price.

The underwriter not only provides expertise throughout the process, but may also agree to buy a portion or even the entirety of the bond issue until the securities are resold to the public or broker-dealers. The difference between the underwriter's purchase price and the price at which the bonds are sold to investors is known as the underwriting spread or underwriting discount. While the initial bond price may be set at par, or at a premium or discount to par value, the pricing structure itself does not affect the underwriter's compensation. The underwriter's compensation is based on the discount it pays relative to the markup on the initial issuance.[9]

The underwriting spread will depend on a variety of factors including the size and type (public or private) of the issue, as well as demand for the new issue at the initial offering price.[10] In this paper, we examine whether underwriters similarly consider the secondary market illiquidity of existing bonds when pricing new issues by the same firm. We also examine whether the price discovery process aids in the pricing of new debt issuances. We hypothesize that with illiquid securities and barriers to price discovery, underwriter fees, and thus the issuing firm's cost of capital, will increase. While the gross underwriting spread is a function of an underwriter's ability to place a security, it is not immediately clear, however, that secondary market illiquidity or price discovery will influence underwriting costs. If, for example, an issue is purchased entirely by a small number of large institutions,

---

[8] Some participants complain that this condensed process does not allow enough time to reliably evaluate the issue, its structure, or the issuing firm's financial position.

[9] The gross underwriting spread consists of fees paid the lead underwriter, the syndicate and the selling group.

[10] The firm must also choose whether to issue bonds in the public or private market. Public issues will not only appeal to a larger group of investors, but may also help firms gain visibility in the marketplace. A firm that obtains financing through private placements will avoid some of the costs associated with a public offering, including the costs of registering the securities with the SEC and complying with GAAP. Private placements are typically less conventional, marketed to a smaller group of investors, and are inherently riskier.

such investors may intend to hold the bonds until maturity. Accordingly, an active secondary market for the firm's other bond issues may not sway an institution's willingness to buy.

B. Underwriting statistics in our sample

To provide context to our discussion of the issuance process in the above section, we include descriptive data that highlights the frequency and magnitude of new corporate debt issues. As reported in Table I, beginning with the start of TRACE coverage in January 2002, through December 2012, 1,231 firms issued over $4.95 trillion in new debt. Many of these firms frequently revisited the debt market and issue new bonds. Our sample of 1,231 firms initiated 21,247 new debt placements during the sample period, an average of over 17 issues per firm. The subsequent issuances by firms with outstanding debt allows us to measure the costs of new issues resulting from prior illiquidity. The 21,247 issues consist of 10,687 investment grade issues, 1,299 speculative grade issues, and 9,261 unrated issues. From Figure 1, which displays the issuance size characteristics, the average firm raises approximately $200 million with each new debt issue.

< Table I >

< Figure 1 >

In Figure 2, we document the aggregate amount of outstanding debt for each year during the sample period. Approximately $1.80 trillion in total corporate debt was outstanding in 2002, of which $1.15 trillion stemmed from unrated corporate bond issues, $560 billion from investment grade debt, and $92 billion from speculative grade bonds. By the end of our sample period in 2012, the amount of outstanding corporate debt ballooned to $3.54 trillion, comprised of $2.06 trillion in investment grade bonds, $1.50 trillion in unrated debt, and $354 billion in speculative grade bonds.[11]

< Figure 2 >

[11] While firms raised over $4.95 trillion in new debt during the sample period, we report only $3.54 trillion outstanding at the end of the sample. The difference is largely due to bonds that mature during the sample period. The median term to maturity for bonds in our sample is 7 years.

Figure 3 highlights the number and volume of new issues during the sample period. Although time series fluctuations are evident, new issues have increased over time. Even during the financial crisis, firms were able to raise capital through the issuance of investment grade debt. However, the number of unrated bonds decreased, and speculative grade issues were almost nonexistent during this time. From the figures described above, it is apparent that the size and scale of the bond market continues to grow. We believe these results highlight both the importance of our empirical analysis as well as the implications of our study for managers, investors, and regulators alike.

< Figure 3 >

II. Data and Sample

The primary data used in our analysis comes from the Mergent FISD database, which includes information for all debt issuances. The FISD database includes the issue size, initial yield, coupon rate, credit rating at issuance, difference between the yield and the Treasury rate at issuance, underwriting fees paid, as well as many other characteristics of newly issued corporate bonds. We augment the Mergent database with bond trading data from the Trade Reporting and Compliance Engine (TRACE) database. Corresponding with TRACE coverage, our sample contains all new corporate bond issuances from July 2002 through December 2012. Last, for the subset of firms in our sample that are public companies, we collect cash flow, leverage, and firm size measures from Compustat. The final merged database contains information on all new corporate bond issues, including underwriting costs, coupons, and credit ratings, as well as information on subsequent trading that occurs after a bond is issued. The data allows us to determine the costs and characteristics of new issues in the primary market, as well as the capability to calculate secondary market illiquidity measures once the bonds begin trading. In addition to examining the characteristics of new issues, we are also able to account for the features of a firm's previously issued bonds.

< Table II >

We present descriptive statistics of the issuance characteristics in Table II. For the new issues in our sample, the average and median coupon rates are 4.48% and 4.89% respectively. The average (median) years to maturity is 9.89 (7.05) years. Last, 34% of the bonds in our sample are callable, while 1% of the bonds are convertible.

A. Cost of New Debt Variables

In this paper, we use two measures to identify the costs associated with issuing new bonds: the Treasury spread and the gross underwriting spread. The Treasury spread is defined as the difference between the yield to maturity and the yield of a duration-matched Treasury security at the time of issuance. We believe the Treasury spread is a more suitable measure of a firm's cost of debt than the yield to maturity at issuance. Because our sample runs through the financial crisis, the yield on corporate bonds varies significantly over the 11-year period our sample covers. As we also control for credit risk in our analysis, the Treasury spread provides a more stable measure of cost than the yield to maturity at issuance. While the Treasury yield spread will be small on safer bonds issued by large firms, investors typically demand higher returns on smaller, riskier bonds, which results in a higher Treasury spread.

Similar to Butler (2008), we also use the gross underwriting spread as a measure of underwriting costs. While the Treasury spread is intended to account for the costs incurred by secondary market traders, our second cost measure here captures revenues to underwriters. When a corporation issues new debt, the immediate cost that the corporation bares is the gross underwriting spread, which is direct compensation to the underwriter.

We present summary statistics for the above cost measures in Panel A of Table III. As expected, investment grade bonds have lower yields and smaller underwriting spreads than those of speculative grade bonds. During our sample period, newly issued bonds have an average yield to maturity of 4.89%, which is, on average, 1.94% higher than the related Treasury security. New issues pay a gross spread of 11.94%.

Management fees are also higher for more speculative bond issues. In dollar terms, this implies that the average debt issue of $200 million produces approximately $23 million in underwriting fees.

< Table III >

B. Illiquidity Variables

We use the illiquidity of a firm's existing bonds as a proxy for the future expected illiquidity of a new issue. This approach allows us to calculate expected illiquidity measures prior to a bond's initial trading. We compute multiple measures of secondary market illiquidity. The first measure of secondary market illiquidity, $PNT_{i,t}$, is the percentage of days in month t that security i does not trade. It is calculated as:

$$PNT_{i,t} = \frac{Zero\ Volume\ Trading\ Days_{i,t}}{Trading\ Days\ in\ Month\ t} \times 100.$$

$PNT_{i,t}$ measures an investors ability to trade a bond at all, which is especially relevant in the highly illiquid bond market. Higher values of $PNT_{i,t}$ imply greater bond illiquidity.

Our second measure of bond illiquidity is the Kyle and Obizhaeva (*KO*) measure of price impact. This metric is constructed from the illiquidity measure presented in Kyle and Obizhaeva's (2011) model of market microstructure invariance. The measure is calculated using the variance of monthly bond returns, scaled by the dollar volume traded within the month. Dollar volume is calculated as the final trade price of each day multiplied by daily volume, then summed to aggregate the monthly totals. We compute the return variance using all TRACE reported transactions for each month.

$$Kyle\ Obizhaeva\ Illiquidity_{i,t} = \left(\frac{Return\ Variance_{i,t}}{Price_{i,t} * Volume_{i,t}}\right)^{\frac{1}{3}} \cdot 10^6.$$

Because a large return variance for smaller dollar volumes indicates greater illiquidity, larger values of the *KO* measure specify greater bond illiquidity.

Our third measure of bond illiquidity is the Amihud (2002) illiquidity measure, given by:

$$Amihud\ Illiquidity = \frac{1}{D_{i,t}} \sum_{n=1}^{D_{i,t}} \frac{\left|Ret_{i,t,n}\right|}{Price_{i,t,n} * Volume_{i,t,n}} \cdot 10^6 ,$$

where $D_{i,t}$ is the number of observations for security i in month t. We use TRACE reported transactions to identify the return, price, and volume for each bond. Similar to the *KO* measure above, the intuition behind the *Amihud* ratio is that larger returns per dollar of trading volume provides an indication of greater bond illiquidity.

Our last measure of bond illiquidity, $AdjTurnover_{i,t}$, is from Liu (2006). This adjusted turnover measure is similar in construction to one proposed by Lesmond, Ogden, and Trzcinka (1999), and is computed for security i in month t as follows:

$$AdjTurnover_{i,t} = \#\ Zero\ Volume\ Trading\ Days_{i,t} + \frac{\frac{1}{turnover_{i,t}}}{Deflator} \times \frac{21}{\#\ Trading\ Days}.$$

$\#ZeroVolumeTradingDays_{i,t}$ is the number of trading days on which the bond did not trade; $turnover_{i,t}$ is the quotient of the total number of bonds traded per month and the total number of outstanding bonds. Following Liu (2006), we use a deflator of 480,000 that allows $0 < \frac{1/turnover_{i,t}}{Deflator} < 1$. Last, we standardize the number of trading days from one month to the next using $\frac{21}{\#\ Trading\ Days}$. The $AdjTurnover_{i,t}$ illiquidity metric is similar to $PNT_{i,t}$, but distinguishes between two bonds with similar zero volume trading days. This measure is increasing in illiquidity.

One additional benefit of measuring turnover is that it may provide insight into the price discovery process. Although turnover is typically used as a liquidity measure, Barinov (2014) suggests that turnover may more appropriately measure firm-specific uncertainty or investor disagreement surrounding the trading process. In this light, turnover may capture elements of the price discovery process, whereby information is incorporated into prices through the interaction of market participants.

In Panel B of Table III, we present summary statistics for the four measures described above. The average bond in our sample trades 5.40 days per month. Bonds in our sample trade, on average, 148 times per month, which generates over $283

million in trading volume. When partitioning the sample by credit rating, we find that the average speculative grade bond trades more frequently than investment grade bonds as well as bonds that are not rated. Median levels of the *KO* and *Amihud* illiquidity measures are smaller than their respective means.[12]

III. The Economic Effects of Secondary Market Illiquidity in the Primary Market

In this section, we seek to identify an economic link between the primary and secondary debt markets. We conjecture that the two principal functions of the secondary market, liquidity and price discovery, each have a direct impact on the cost of issuing new debt in the primary market. We begin by examining the effects, if any, of secondary market liquidity on the primary market. Then, in the subsequent section, we study the significance of secondary market price discovery in the primary market for new issues.

A. Tests of secondary market illiquidity and the cost of new debt issues

We begin by identifying corporate bonds issued between 2002 and 2012. We then link each newly issued bond with existing bonds issued by the same firm. Since we are interested in whether secondary market illiquidity affects the cost of new issues, we require firms to have outstanding bonds issued after 1975. From this set of prior issues, we eliminate those that mature more than three years prior to the new issues. Bond characteristics may not only change over time, but the market's perception of a new issue may not incorporate the characteristics of bonds that have already matured. We also exclude prior issues that originated within the previous month, since there is insufficient data to measure illiquidity.

For each previously issued bond, we calculate the four illiquidity measures described in Section II, for each month of the sample period. To aid our

[12] Because there is a great deal of skewness in the illiquidity measures, we winsorize our data at the 1% and 99% levels.

understanding of how prior illiquidity affects the cost of new debt, we average the monthly illiquidity variables from all prior issues over the previous year. Should a firm have multiple prior issues, we weight our illiquidity measure by prior issue size.[13]

In our first set of empirical tests, we investigate how the illiquidity of prior-issues affects the cost to issue new debt. We consider the full sample of public and private corporate debt issues from 2002 through 2012. To determine if prior illiquidity influences future underwriting costs, we regress our cost measures, the Treasury spread and gross underwriting spread, on each of the four illiquidity measures. To isolate the effects of prior illiquidity on the underwriting costs of new issues, we control for the heterogeneous characteristics of bonds by including variables for the new issues' term to maturity, duration, size, as well as for the issuers' outstanding debt. We also include indicator variables that identify whether the issue is callable, convertible, senior/junior, privately placed, asset-backed, and if the bond is a 144A bond. To control for credit ratings, we include an indicator variable for each possible rating (AAA, AA+, AA, AA-, etc.), as well as indicator variables that identify if the credit rating of the new issue is higher or lower relative to the most recent issue. For the subset of firms that are publicly traded (for which data is readily available), we include controls for firm characteristics that could influence a firm's cost of debt. These variables include *Cash Flow*, *Leverage*, and *Firm Size* for the year end prior to the new issue. *Cash Flow* is the firm's operating income before depreciation divided by total assets; *Leverage* is the percentage of total financing represented by debt; and *Firm Size* is the log of the sum of debt and stockholder equity. All regressions include firm-year fixed effects.

We report the coefficient estimates of our cross-sectional regression tests in Table IV. In Panel A, we report results using the Treasury spread as the dependent variable. Our findings suggest that the illiquidity of previously issued, outstanding

[13] To address concerns that investors may place more emphasis on recent issues (since these bond characteristics may be similar to the current issue), we repeat all of our analyses using only prior issues that originated within five years of the current issue. The results presented in this paper are robust to this alternative specification.

bonds is directly related to the yield placed on new bond issues by the same firm. Because larger Treasury spreads on new issues are typically associated with greater risk, the positive coefficients on the *KO* and *Amihud* measures indicate that the secondary market illiquidity of existing bonds likely captures the potential illiquidity risks associated with the new issues. This result implies that investors purchasing new issues demand a premium for the expected illiquidity of the bonds. In turn, these costs are directly passed to the issuing firm. Economically, our findings demonstrate that each one percent increase in the *KO* (*Amihud*) measure of illiquidity corresponds to a four (two) basis point increase in the bond yield beyond the maturity-matched Treasury security. The coefficient on $AdjTurnover_{i,t}$ is also positive and significant, specifying that the uncertainty surrounding the trading process, another dimension of illiquidity, affects the cost of new issues as well. The coefficients of *Years to Maturity* and *Issue Size* are positive as well, suggesting that investors require a larger yield for longer maturity bonds as well as for larger debt issuances. Because new debt offerings affect a firm's capital structure, larger bond issues increase the default risk of the issuer.

In columns (5) through (8), we consider the subset of public firms and include additional variables that potentially impact a firm's cost of debt financing. We partition our results to address concerns that more illiquid private debt might be driving our results. The subsample also allows us to control for other firm characteristics that may influence a firm's cost of debt (e.g., *Cash Flow*, *Leverage*, and *Firm Size*). Here, we find a similar outcome as before when considering the full sample: a one percent increase in the *KO* (*Amihud*) illiquidity measure is associated with a five (three) basis point increase in the yield paid at issuance. Our results imply that investors view new publicly-traded debt offered by firms with illiquid outstanding issues as riskier, and thus demand higher yields in return. In sum, our findings are robust to the type of debt (i.e. public or private) issued.

In Panel B of Table IV, we report the coefficients resulting from regressions of underwriter fees on illiquidity. While our approach as presented in Panel A is designed to isolate costs associated with secondary market trading, the results in

Panel B should capture the costs levied by underwriters. While we, again, find a direct relation between the illiquidity of existing bonds and the costs of new issues in the primary market, our results offer added insight into a different dimension of illiquidity charged by underwriters. Our results indicate that underwriters are less concerned about secondary market transactions costs, as seen in the insignificant coefficients of the *KO* and *Amihud* price impact measures, but are more attentive to the ability to trade a bond at all. As seen in the positive and significant coefficients of $PNT_{i,t}$ and $AdjTurnover_{i,t}$, firms incur higher costs on new issues if their previously issued bonds trade on fewer days of the month. From the underwriting process description presented in Section I, underwriters may agree to buy bonds that cannot be sold to investors. Given the consequences of being unable to place new issues, underwriters place a premium on new issues with higher levels of expected illiquidity. Specifically, a one percent increase in the number of days that existing debt does not trade is associated with a 1.65% increase in the underwriting spread paid to the syndicate. Similarly, we find that a one percent increase in $AdjTurnover_{i,t}$ is associated with a 14 basis point increase in the underwriting spread. In dollar terms, these results suggest that for the average issue of $200 million, a one percent increase in illiquidity is associated with an increase in underwriting fees of between $280,000 and $3,300,000.

We also find that *Issue Size* is negatively related to the gross underwriting spread. The sign of this coefficient is in sharp contrast to the same variable presented in Panel A, when considering the Treasury spread. One potential explanation of this result is that because investment banks collect a portion of the total debt issued as compensation, underwriters may be more willing to offer a quantity discount for larger issues. In total, the results in Table IV offer compelling evidence that secondary market illiquidity leads to higher underwriting costs for firms issuing new debt.

<Table IV>

As discussed in Yasuda (2005), underwriters consider first-time issuances more difficult to market, relative to bonds offered by seasoned and frequent issuers.

Because first-time issuers have no historical track record, their new placements exhibit high "informational sensitivity," and consequently, may be charged a premium by underwriters. We consider a subsample of second offerings by first-time debt issuers. By removing seasoned firms with greater debt exposure, and instead study the second debt offering of these first-time issuers, we believe that we are able to isolate the effects of prior illiquidity on the costs of a new issue. The initial bonds issued by the firms in our study have varying degrees of secondary market illiquidity. Accordingly, if both underwriters and investors have limited information regarding the first-time issuers, we expect the illiquidity premium to be even more pronounced for firms with more illiquid debt outstanding.

As reported in Panel A of Table V, we identify 948 firms that first issue debt during our sample window. We examine the relation between the secondary market illiquidity of the initial issue and the costs associated in 597 second issues by the same firm. As reported in Panel B, these firms return to the debt market, on average, 1.83 years later, and typically raise more money in the second issue relative to the first.

< Table V >

We report the results of our multivariate analysis in Table VI. Our approach in this portion of our investigation is similar to that presented in Table IV. However, in Table VI, we consider the marginal change in issuing costs between the first and second issue, and not the costs associated with any other subsequent issues. In addition to years to maturity and the size of the issue, we include controls for whether the bonds are rated, as well as controls for the time between the debt offerings. Given that debt offerings by first-time issuers are more challenging to underwrite than subsequent offerings by frequent and seasoned issuers, including these control variables allows us to isolate the illiquidity effects on the costs of a subsequent issue.

In our test of first time issuers, we find that the cost of a second debt offering is higher than the costs of the first issue for firms whose initial issue is illiquid. When considering the credit spread results in Panel A, we find that $PNT_{i,t}$ and $AdjTurnover_{i,t}$ are priced into the cost of new issues by the same firm. Similarly, in

Panel B, we find that an increase in the *KO* and *Amihud* price impact measure, as well as the $PNT_{i,t}$ and $AdjTurnover_{i,t}$ metrics increase the gross underwriting spread beyond what was paid in the first issue. A one percent increase in price impact and turnover is associated with an incremental increase of 3 basis points beyond what was paid in the first issue. A one percent increase in the number of days a bond doesn't trade is associated with a 28 basis point marginal increase beyond the cost of the first issue. These results confirm that underwriters account for expected secondary market illiquidity when determining their compensation structure. For the average size of a second issue of $536 million, the findings in Table VI suggest that a firm will pay an additional $160,000 to $1,500,000 in underwriting fees for every one percent increase in illiquidity. In total, the results in Table VI suggest that both investors and underwriters demand higher premiums to compensate for the potential illiquidity risks associated with new debt offerings, costs directly incurred by the issuing firms.

< Table VI >

B. Secondary market illiquidity and access to debt

The results in the previous section demonstrate that illiquidity can alter a firm's cost of debt. In turn, secondary market illiquidity influences a firm's access to capital. The results indicate that firms pay a premium for issuing new debt when their previously issued debt is comparatively illiquid. Our tests to this point, however, are predicated on firms being able to access credit markets at all. In our next set of tests, we further explore this relation by determining whether secondary market illiquidity for a firm forecasts the issuance of new credit.

If illiquidity results in a higher cost of debt for firms, as our prior results indicate, then, on the margin, this relation will affect the set of profitable projects available to a firm. Firms with a higher cost of debt may forgo valuable projects that they could have otherwise undertaken. Note, too, that difficulties in raising new capital may be driven by both firm-specific factors as well as market events. Thus, firms may experience changes in their access to capital if either firm-level or market-

level illiquidity changes. Understanding how aggregate market conditions and macroeconomic factors impact a firm's access to capital is also an important question (Erel, Julio, Kim, and Weisbach (2012)).

We begin this portion of our analysis by considering firms with outstanding bonds trading in the secondary market. We compare this total with the number of firms that actually issue new debt in that year. We present descriptive statistics of firms that issue debt, as well as statistics for firms that do not issue debt during the same period, in Table VII. We partition the sample based on credit rating. In 2008, for example, 28 percent of firms with outstanding debt issued new bonds during the year, whereas 21% (30%) of firms with speculative grade debt (debt that is not rated) are able to return to the debt markets during 2008. However, during 2012, 40%, 54%, and 68% of firms with investment-grade debt, speculative-grade debt, and debt that is not rated, respectively, issue new bonds.

< Table VII >

To determine if prior illiquidity poses a hurdle that firms must overcome when issuing new debt securities, we report the results of cross-sectional probit tests in Table VIII. The dependent variable is an indicator variable equal to one if a firm issues debt in the current year (year *t*). The independent variable of interest is the average monthly illiquidity measures for the same firm in year *t-1*. We include the total dollar volume of current debt outstanding in order to control for a firm's need for new debt. Given that the financial crisis provided a market-wide shock, we also include an indicator variable for the years 2008 and 2009, as well as an interaction between illiquidity and the recession-year indicator variables. As a final control, we include indicator variables for the median credit rating of each firm's outstanding bonds.

As seen in Table VIII, the negative and significant coefficients on three out of the four illiquidity measures indicate that prior year illiquidity provides predictive power to identify firms that subsequently issue new debt. We believe these results imply that firm-specific illiquidity represents an impediment to accessing credit.

Firms with comparatively illiquid debt may find it more difficult to fund or expand operations, even after accounting for system wide shocks to liquidity.

Overall, the results in Tables IV, V, and VI suggest that firms with illiquid bonds experience higher costs of new issues. The results in Table VIII indicate that illiquidity also serves as a predictor of a firm's ability to access public credit markets entirely. Our results offer practical implications for managers as they indicate that secondary market trading provides real economic benefits. Collectively, our results indicate that illiquidity improvements are not only associated with the potential to lower a firm's cost of debt, but also indicate that illiquidity improvements might affect the ability of a firm to access debt financing at all.

< Table VIII >

IV. The Economic Effects of Secondary Market Price Discovery in the Primary Market

In the previous section, we provide evidence that the illiquidity of existing bonds has a significant economic impact on the underwriting costs incurred by firms when issuing new bonds. As previously discussed, however, liquidity is only one major function provided by secondary markets. The other important role of secondary markets is to provide the opportunity for price discovery, the process by which new information is assimilated into prices. In this section, we explore whether the price discovery process that occurs in the secondary market also has an economic impact on underwriting costs incurred by firms in the primary market.

One difficulty in determining the effects of liquidity and price discovery is that the two are often indistinguishable in empirical tests. An improvement in one typically produces an improvement in the other. The corporate bond market allows us a novel approach to disentangle the two effects. The staggered implementation of the Trade Reporting and Compliance Engine (TRACE) and the subsequent release of all bond trading data through the Enhanced TRACE files provides us a way to test the effects of secondary market price discovery on the primary market.

TRACE is the vehicle that requires mandatory transaction reporting for corporate bonds. Prior to the implementation of TRACE, investors did not have access to real-time information on transaction sizes and prices. While traders were still able to find liquidity in the pre-TRACE period, investors were forced to transact with an information set that included only stale prices. Consequently, the price discovery process was severely inhibited prior to the implementation of TRACE. Because TRACE allows traders to see prices in real-time, the price discovery process was much more efficient for TRACE-reported bonds than for bonds that were not TRACE reported.

Not all new debt offerings issued in 2002 were immediately TRACE-reported. As presented in Table IX, only 26% of all new debt issuances were TRACE-reported. This percentage increases every year until 2006, the year in which all new issues are TRACE-reported and thereby provide real-time transparency to traders.[14] The staggered implementation of TRACE allows us to examine the impact of trading when prices are not yet disseminated to the public. Specifically, we compare the cost of new bond issues that are TRACE-reported to the costs of new bond issues that were not yet subject to TRACE reporting. Greater price discovery in a firm's outstanding bonds should benefit underwriters when pricing new issues.

< Table IX >

[14] As reported in the TRACE fact book: During Phase I, effective on July 1, 2002, public transaction information was disseminated immediately upon receipt for the larger and generally higher credit quality issues: (1) Investment-Grade debt securities having an initial issue of $1 billion or greater; and (2) 50 Non-Investment-Grade (High-Yield) securities disseminated under FIPS that were transferred to TRACE. Under these criteria, FINRA disseminated information on approximately 520 securities by the end of 2002. Phase II, fully effective on April 14, 2003, expanded public dissemination to include transactions in smaller Investment-Grade issues: (1) all Investment Grade TRACE-eligible securities of at least $100 million par value (original issue size) or greater rated A3/A- or higher; and (2) a group of 120 Investment-Grade TRACE-eligible securities rated Baa/BBB and 50 Non-Investment-Grade bonds. As Phase II was implemented, the number of disseminated bonds increased to approximately 4,650 bonds. In Phase III, fully effective on February 7, 2005, approximately 99 percent of all public transactions and 95 percent of par value in the TRACE-eligible securities market were disseminated immediately upon receipt by the TRACE System. However, transactions over $1 million in certain infrequently traded Non-Investment-Grade securities were subject to dissemination delays, as were certain transactions immediately following the offering of TRACE-eligible securities rated BBB or below.

To determine if firms with TRACE-reported bonds experience lower costs in the primary market, we perform a similar analysis to that presented in Tables IV and VI. In this model, we include an indicator variable specifying whether outstanding bonds issued by the same firm are TRACE reported. We include, but do not report, the same control variables presented in previous tables. To disentangle the effects of price discovery from that of liquidity, we also control for illiquidity using each of the four illiquidity measures reported in our analysis to this point.

The results in Table X indicate that firms with TRACE-reported bonds experience lower underwriting costs in the primary market relative to firms who had bonds that were not TRACE reported. After considering both the Treasury spread in Panel A, as well as the underwriting spread in Panel B, we find that bonds with greater transparency and price discovery in the secondary market experience lower costs in the primary market. Specifically, the negative and significant coefficient of the TRACE-reported indicator variable suggests that bonds with greater transparency and price discovery in the secondary market have lower underwriting costs and yield spreads in the primary market. While numerous studies document an improvement in secondary market liquidity with the implementation of TRACE on July 1, 2002 (see, for example, Bessembinder, et al. (2006) and Goldstein, Hotchkiss, and Sirri (2007)), none of these studies look at the effects of TRACE reporting on the costs of new issues in the primary market. We are the first to show that improved price discovery in the secondary market leads to lower costs of new debt in the primary market.

< Table X >

V. Conclusion

Primary markets, where securities are initially purchased from the issuing firm, serve a clear and necessary purpose. Through the issuance of new securities in the primary market, firms are able raise capital to fund or expand operations. After underwriting fees are subtracted, all proceeds from security issuances go directly to the issuing firm. Issuing firms do not, however, receive a direct capital inflow from

transactions occurring in the secondary market, where investors trade with other investors. While issuing firms are unable to directly collect new investment in the secondary market, firms may still indirectly benefit from trading in the secondary market. Greater secondary market liquidity for equity securities, for example, is shown to lower a firm's cost of capital and lead to significant improvements in firm performance (Butler, et al. (2005); Fang, et al. (2009)).

In this paper, we explore whether secondary market liquidity for corporate bonds provides a positive and significant benefit to the issuing firm. Unlike the primary equity markets of IPOs and SEOs, which are accessed infrequently, the sheer volume of bond issues and reissues, along with the scope of firms and entities issuing debt allow us to address a question posed by Zingales (2015) as to whether finance, in this case secondary markets, benefit society.

Our results indicate that the illiquidity of outstanding bonds is priced into new debt issues by the same firm, where firms with current illiquid debt pay higher prices for subsequent debt issues. We also find that greater illiquidity reduces the likelihood that firms return to the debt market during periods of market turmoil. Additionally, our results suggest that a more efficient price discovery process in the secondary market reduces the cost of new issues in the primary market. The practical inference from our results is that secondary markets are not simply a sideshow, but do in fact provide real economic benefit to issuing firms. Our paper contributes to the growing body of research that sheds light on the societal benefits provided by secondary market. We conclude by suggesting that efforts to improve liquidity and price discovery in secondary markets is warranted, not only because they improve secondary market trading, but also because they provide firms better access to capital to fund growth opportunities.

REFERENCES

Amihud, Yakov, 2002, Illiquidity and stock returns: Cross-section and time-series effects, *Journal of Financial Markets* 5, 31-56.

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

Aslan, Hadiye, and Praveen Kumar, 2016, The product market effects of hedge fund activism, *Journal of Financial Economics* 119, 226-248.

Bao, Jack, Jun Pan, and Jiang Wang, 2011, The illiquidity of corporate bonds, *The Journal of Finance* 66, 911-946.

Barinov, Alexander, 2014, Turnover: Liquidity or Uncertainty?, *Management Science* 60, 2478-2495.

Bessembinder, Hendrik, William Maxwell, and Kumar Venkataraman, 2006, Market transparency, liquidity externalities, and institutional trading costs in corporate bonds, *Journal of Financial Economics* 82, 251-288.

Bosworth, Barry, 1975, The stock market and the economy, *Brookings Papers on Economic Activity* 1975, 257-300.

Butler, Alexander W, 2008, Distance still matters: Evidence from municipal bond underwriting, *Review of Financial Studies* 21, 763-784.

Butler, Alexander W, Gustavo Grullon, and James P Weston, 2005, Stock market liquidity and the cost of issuing equity, *Journal of Financial and Quantitative Analysis* 40, 331-348.

Campello, Murillo, and John R Graham, 2013, Do stock prices influence corporate decisions? Evidence from the technology bubble, *Journal of Financial Economics* 107, 89-110.

Campello, Murillo, John R Graham, and Campbell R Harvey, 2010, The real effects of financial constraints: Evidence from a financial crisis, *Journal of Financial Economics* 97, 470-487.

Campello, Murillo, Rafael P Ribas, and Albert Y Wang, 2014, Is the stock market just a side show? Evidence from a structural reform, *Review of Corporate Finance Studies* 3, 1-38.

Chava, Sudheer, and Michael R Roberts, 2008, How does financing impact investment? The role of debt covenants, *The Journal of Finance* 63, 2085-2121.

Chen, Kevin CW, Zhihong Chen, and KC John Wei, 2011, Agency costs of free cash flow and the effect of shareholder rights on the implied cost of equity capital, *Journal of Financial and Quantitative Analysis* 46, 171-207.

Chen, Long, David A Lesmond, and Jason Wei, 2007, Corporate yield spreads and bond liquidity, *The Journal of Finance* 62, 119-149.

Das, Sanjiv, Madhu Kalimipalli, and Subhankar Nayak, 2014, Did cds trading improve the market for corporate bonds?, *Journal of Financial Economics* 111, 495-525.

Dick-Nielsen, Jens, Peter Feldhütter, and David Lando, 2012, Corporate bond liquidity before and after the onset of the subprime crisis, *Journal of Financial Economics* 103, 471-492.

Dow, James, and Gary Gorton, 1997, Stock market efficiency and economic efficiency: Is there a connection?, *The Journal of Finance* 52, 1087-1129.

Ellul, Andrew, and Marco Pagano, 2006, Ipo underpricing and after-market liquidity, *Review of Financial Studies* 19, 381-421.

Erel, Isil, Brandon Julio, Woojin Kim, and Michael S Weisbach, 2012, Macroeconomic conditions and capital raising, *Review of Financial Studies* 25, 341-376.

Fama, Eugene F, and Kenneth R French, 1997, Industry costs of equity, *Journal of Financial Economics* 43, 153-193.

Fang, Lily Hua, 2005, Investment bank reputation and the price and quality of underwriting services, *The Journal of Finance* 60, 2729-2761.

Fang, Vivian W, Thomas H Noe, and Sheri Tice, 2009, Stock market liquidity and firm value, *Journal of Financial Economics* 94, 150-169.

Fang, Vivian W, Xuan Tian, and Sheri Tice, 2014, Does stock liquidity enhance or impede firm innovation?, *The Journal of Finance* 69, 2085-2125.

Friewald, Nils, Rainer Jankowitsch, and Marti G Subrahmanyam, 2012, Illiquidity or credit deterioration: A study of liquidity in the us corporate bond market during financial crises, *Journal of Financial Economics* 105, 18-36.
Goldstein, Michael A, Edith S Hotchkiss, and Erik R Sirri, 2007, Transparency and liquidity: A controlled experiment on corporate bonds, *Review of Financial Studies* 20, 235-273.
Green, Richard C, Dan Li, and Norman Schürhoff, 2010, Price discovery in illiquid markets: Do financial asset prices rise faster than they fall?, *The Journal of Finance* 65, 1669-1702.
Grullon, Gustavo, Sébastien Michenaud, and James P Weston, 2015, The real effects of short-selling constraints, *Review of Financial Studies* 28, 1737-1767.
Hail, Luzi, and Christian Leuz, 2009, Cost of capital effects and changes in growth expectations around us cross-listings, *Journal of Financial Economics* 93, 428-454.
Hotchkiss, Edith S, and Tavy Ronen, 2002, The informational efficiency of the corporate bond market: An intraday analysis, *Review of Financial Studies* 15, 1325-1354.
Julio, Brandon, Woojin Kim, and Michael Weisbach, 2007, What determines the structure of corporate debt issues?, (National Bureau of Economic Research).
Kuziemko, Ilyana, Michael I Norton, Emmanuel Saez, and Stefanie Stantcheva, 2015, How elastic are preferences for redistribution? Evidence from randomized survey experiments, *The American Economic Review* 105, 1478-1508.
Kyle, Albert Pete, and Anna Obizhaeva, 2011, Market microstructure invariants: Empirical evidence from portfolio transitions, Working Paper, Available at SSRN 1978943.
Kyle, Albert Pete, and Anna Obizhaeva, 2011, Market microstructure invariants: Theory and implications of calibration, Working Paper, Available at SSRN 1978932.
Lesmond, David A, Joseph P Ogden, and Charles A. Trzcinka, 1999, A New Estimate of Transaction Costs, *Review of Financial Studies* 12, 5, 1113-1141.
Liu, Weimin, 2006, A liquidity-augmented capital asset pricing model, *Journal of Financial Economics* 82, 631-671.
Mauer, David C, and Lemma W Senbet, 1992, The effect of the secondary market on the pricing of initial public offerings: Theory and evidence, *Journal of Financial and Quantitative Analysis* 27, 55-79.
McLean, R David, and Mengxin Zhao, 2014, The business cycle, investor sentiment, and costly external finance, *The Journal of Finance* 69, 1377-1409.
Morck, Randall, Andrei Shleifer, and Robert W Vishny, 1990, The stock market and investment: Is the market a sideshow?, *Brookings Papers on Economic Activity* 1990, 157-215.
O'Hara, Maureen, 2003, Presidential address: Liquidity and price discovery, *The Journal of Finance* 58, 1335-1354.
Rajan, Raghuram G, and Luigi Zingales, 1996, Financial dependence and growth, (National Bureau of Economic Research).
Ritter, Jay R, 2005, Economic growth and equity returns, *Pacific-Basin Finance Journal* 13, 489-503.
Ritter, Jay R, 2012, Is economic growth good for investors? 1, *Journal of Applied Corporate Finance* 24, 8-18.
Roll, Richard, 1984, A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market, *Journal of Finance* 39, 4, 1127-1139.
Ronen, Tavy, and Xing Zhou, 2013, Trade and information in the corporate bond market, *Journal of Financial Markets* 16, 61-103.
Stein, Jeremy C, 2003, Agency, information and corporate investment, *Handbook of the Economics of Finance* 1, 111-165.
Stulz, René M, 1999, Golbalization, corporate finance, and the cost of capital, *Journal of applied corporate finance* 12, 8-25.
Yasuda, Ayako, 2005, Do bank relationships affect the firm's underwriter choice in the corporate-bond underwriting market?, *The Journal of Finance* 60, 1259-1292.
Zingales, Luigi, 2015, Presidential address: Does finance benefit society?, *The Journal of Finance* 70, 1327-1363.

Table I
Corporate Bond Issues (2002-2012)

This table reports summary statistics for new issues of corporate bonds during the sample period covering 2002 through 2012. Panel A reports the statistics for the entire sample period, while Panel B reports the statistics averaged by year. Statistics are partitioned by investment rating at the time of issue. Number of issuers is the number of unique corporations that issue bonds during the sample period, and number of issues is the total number of unique issues from the issuers in the sample. Total volume is the sum of the issue amount, and average issue size is the average amount issued.

	Investment Grade	Speculative Grade	Not Rated	Full Sample
Panel A: Full Sample				
Number of Issuers	440	99	692	1,231
Number of Issues	10,687	1,299	9,261	21,247
Volume Issued (Millions)	2,645,152	343,210	1,966,952	4,955,313
Avg. issue size	247.51	264.21	212.39	233.22
Panel B: Average per year				
Number of Issuers	40.00	9.00	62.91	111.91
Number of Issues	971.55	118.09	841.91	1,931.55
Volume Issued (Millions)	240,468	31,201	178,814	450,483
Avg. issue size	360.56	542.65	292.53	335.91

Table II
Corporate Bond Characteristics at Issuance

In this table, we present the characteristics of newly issued corporate bonds. Major characteristics include the coupon paid to investors, the time in years until the bond matures as a percent of par. We also include the proportion of new issues that are callable and convertible. We report means and medians. Characteristics of bonds are partitioned according to investment rating at the time of issue.

	Investment Grade	Speculative Grade	Not Rated	All
Mean				
Coupon	4.40	5.54	4.42	4.48
Years to Maturity	9.97	11.28	9.61	9.89
Offer Yield	4.86	5.64	4.83	4.89
Offering Price of Par	99.87	99.86	99.85	99.86
Proportion Callable	0.32	0.42	0.34	0.34
Proportion Convertible	0.01	0.01	0.02	0.01
Median				
Coupon	4.88	5.68	4.75	4.89
Years to Maturity	7.02	9.99	7.54	7.05
Offer Yield	5.00	5.65	5.00	5.03
Offering Price of Par	100.00	100.00	100.00	100.00
Total New Issues (2002-2012)	10,687	1,299	9,261	21,247

Table III
Liquidity and Cost of Newly Issued Corporate Bonds

In this table, we report the main variables used to identify illiquidity and the cost of issuing bonds. In Panel A, we report the principal costs of issuing bonds, which includes the yield to maturity at issue, the gross spread paid to the underwriting syndicate, the management fee, the reallowance fee, and the difference between the Treasury yield and the bond's yield to maturity at issuance. In Panel B, we report the average issue illiquidity variables, which include the number of days in a month that a bond is traded, the dollar volume traded in a month, the number of trades in a month, the Kyle-Obizhaeva (2011) illiquidity measure, and the Amihud (2002) illiquidity measure.

	Investment Grade	Speculative Grade	Not Rated	All
Panel A: Cost of Issuing Descriptive Statistics				
Mean				
YTM at issuance	4.86	5.64	4.83	4.89
Credit spread	1.72	2.97	2.07	1.94
Bond issue gross spread	11.09	12.76	12.79	11.94
Management Fee	4.96	6.95	7.91	6.64
Reallowance Fee	2.19	2.18	2.47	2.34
Median				
YTM at issuance	5.00	5.65	5.00	5.03
Credit Spread	1.45	2.63	1.52	1.50
Bond issue gross spread	8.75	10.00	10.00	9.75
Management fee	4.00	4.00	5.00	4.00
Reallowance fee	2.50	2.50	2.00	2.00
Panel B: Illiquidity Statistics				
Mean				
Monthly Trading Days	5.35	6.00	5.37	5.40
Monthly $ Volume per issue	266,134,012	667,601,291	249,713,397	283,521,628
Monthly trades per issue	144.44	182.11	147.72	148.17
Kyle-Obizhaeva illiquidity	3.36	3.62	3.35	3.37
Amihud Bond illiquidity	3.27	3.04	3.65	3.45
Adjusted Turnover illiquidity	15.42	15.03	15.87	15.61
Median				
Monthly Trading Days	4.00	4.00	4.00	4.00
Monthly $ Volume per issue	3,671,000	6,291,000	4,150,060	4,000,000
Monthly trades per issue	18.00	32.00	23.00	21.00
Kyle-Obizhaeva Liquidity	1.40	1.35	1.73	1.52
Amihud Bond Liquidity	1.06	1.35	1.42	1.24
Adjusted Turnover	17.00	16.80	17.35	17.18
Total New Issues (2002-2012)	10,687	1,299	9,261	21,247

Table IV
The Cost of Issuing Illiquid Bonds

In this table, we report cross-sectional regression tests of the costs of new issues on prior illiquidity of outstanding bonds. The sample includes new public and private corporate bond issues during the period from 2002 to 2012. To measure secondary market illiquidity, each bond is required to have at least one other debt issuance prior to the current new issue. When computing the liquidity of existing debt, we use the average monthly liquidity of all outstanding bonds for the year prior to the new issue, weighted by issue size. The illiquidity variables include the percentage of days in a month that a bond does not trade, the Kyle-Obizhaeva (2011) measure of price impact, the Amihud (2002) measure of price impact, and Liu's (2006) adjusted turnover measure. The dependent variable in all specifications is a form of issuing costs, including the difference between the yield to maturity and the Treasury yield at issuance (Panel A), and the gross spread paid to the underwriter (Panel B). The independent variables include the years to maturity, log of the new issue size, log of prior outstanding issues, and the duration of the issue. For public firms where the data is available, we include the leverage ratio, cash flow, and log of firm size. Other control variables include indicators for convertible, callable, senior, junior, 144A eligible, privately placed, and asset-backed issues, as well as indicators for credit upgrades and downgrades since the last issue. All regressions include firm and year fixed effects. Robust test-statistics are reported in parentheses, where ***, **, and * indicate significance at the 1%, 5%, and 10% levels.

Panel A: Credit Spread

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PNT	KO	Amihud	Adj. TO	PNT	KO	Amihud	Adj. TO
	All Corporate Debt Issues				Corporate Debt Issues of Public Firms			
Illiquidity	0.04	0.04***	0.02**	0.01**	0.05	0.05***	0.03**	0.01
	(0.42)	(2.59)	(2.53)	(1.99)	(0.31)	(3.06)	(2.44)	(1.52)
Years to Maturity	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***	0.03***
	(2.80)	(2.76)	(2.70)	(2.79)	(2.73)	(2.66)	(2.61)	(2.73)
Log Issue Size	0.08***	0.08***	0.08***	0.08***	0.07***	0.07***	0.07***	0.07***
	(6.86)	(6.84)	(6.76)	(6.82)	(4.55)	(4.53)	(4.44)	(4.49)
Log Outstanding Debt	0.06	0.07	0.07	0.06	-0.03	-0.02	-0.03	-0.03
	(0.89)	(0.98)	(0.88)	(0.89)	(-0.26)	(-0.21)	(-0.29)	(-0.25)
Duration	-0.02	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03	-0.03
	(-1.09)	(-1.08)	(-0.99)	(-1.09)	(-1.24)	(-1.19)	(-1.10)	(-1.24)
Cash Flow					-3.22**	-3.30**	-3.37**	-3.24**
					(-1.97)	(-2.03)	(-2.03)	(-1.98)
Leverage					-0.02	-0.02	-0.04	-0.01
					(-0.03)	(-0.03)	(-0.07)	(-0.01)
Log Firm Size					-0.00	-0.00	-0.01	-0.00
					(-0.12)	(-0.07)	(-0.17)	(-0.11)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm and Year Fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R^2	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39
N	919	919	918	919	731	731	730	731
Observations	11,364	11,364	11,281	11,364	7,793	7,793	7,768	7,793

Panel B: Underwriting Spread								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PNT	KO	Amihud	Adj. TO	PNT	KO	Amihud	Adj. TO
	All Corporate Debt Issues				Corporate Debt Issues of Public Firms			
Illiquidity	1.65***	0.18	0.00	0.14***	1.21**	0.30*	0.05	0.11**
	(2.80)	(1.55)	(0.04)	(3.27)	(2.11)	(1.91)	(1.19)	(1.98)
Years to Maturity	0.20**	0.19**	0.20**	0.20**	0.27***	0.26***	0.27***	0.27***
	(2.43)	(2.40)	(2.42)	(2.42)	(3.92)	(3.89)	(3.91)	(3.92)
Log Issue Size	-0.85***	-0.87***	-0.87***	-0.87***	-1.10***	-1.11***	-1.11***	-1.11***
	(-5.18)	(-5.18)	(-5.15)	(-5.16)	(-5.88)	(-5.83)	(-5.77)	(-5.81)
Log Outstanding Debt	-0.47*	-0.40	-0.45*	-0.43*	-0.48	-0.46	-0.50	-0.48
	(-1.87)	(-1.52)	(-1.65)	(-1.68)	(-1.27)	(-1.24)	(-1.31)	(-1.25)
Duration	0.77***	0.77***	0.76***	0.77***	0.53***	0.54***	0.53***	0.53***
	(4.19)	(4.17)	(4.13)	(4.16)	(3.20)	(3.22)	(3.18)	(3.20)
Cash Flow					-7.12	-8.20	-7.92	-7.54
					(-1.47)	(-1.54)	(-1.50)	(-1.49)
Leverage					-0.67	-0.84	-0.86	-0.74
					(-0.26)	(-0.32)	(-0.33)	(-0.29)
Log Firm Size					0.22*	0.24*	0.26**	0.22*
					(1.84)	(1.88)	(2.35)	(1.85)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm and Year Fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R^2	0.34	0.34	0.34	0.34	0.54	0.54	0.54	0.54
N	943	943	942	943	753	753	752	753
Observations	15,257	15,257	15,153	15,257	10,160	10,160	10,122	10,160

Table V
First time issuers

This table presents summary statistics for first-time bond issuers during the period from 2002 through 2012. This subsample of bond issuers includes firms that potentially have little information regarding the expected risks of the issue. Panel A reports the frequency of first time issuers, including the total number of initial issues, the total number of second issues, as well as the total number of subsequent issues for the remainder of the sample period. In instances where a firm issues two different bonds with differing maturities on the same day, both count as a second issue. Panel B summary statistics of issues by first time issuers.

Panel A: Frequency of New Issuers	
First Time Issuers: 2002-2012	948
First time issuers with a Second Issue: 2002-2012	597
Total subsequent issues: 2002-2012	6,331

Panel B: Summary Statistics of Secondary Issues				
	Mean	Median	Min	Max
Total Subsequent issues per issuer	2.18	1.00	1.00	85.00
Size of Initial Issue	508.83	362.50	0.37	5,000.00
Size of Second Issue	536.72	400.00	0.15	4,625.00
Years between Initial and Second issue	1.83	1.24	0.01	10.70

Table VI
Change in issuing costs following first time issues

This table presents the cross sectional regression tests of marginal underwriting costs on a firm's second bond issue. To be included in the sample, a firm must issue its second bond, where the only other bond issued by the corporation is the initial issue that occurred previously. The dependent variable includes the change in underwriting costs of the second bond issue beyond the first issue by a firm, namely the difference between the initial yield to maturity and the Treasury yield, as well as the underwriting spread paid to the syndicate. The principal independent variable is the average monthly illiquidity of the previously issued bond. Control variables include an indicator variable identifying whether the bond is speculative grade or not rated, the time in years between initial and second issue, years to maturity of the current issue, the log of the issue size, and indicator variables indicating 144A, senior, junior, callable, and convertible issues. Indicator variables also indicate whether the new issue receives a higher or lower grade relative to its previous issue. All specifications include year fixed effects. Robust t-statistics are reported in parentheses, with ***,**, and * indicating significance at the 1%, 5%, and 10% levels respectively.

Panel A: Credit Spread	(1)	(2)	(3)	(4)
	PNT	KO	Amihud	Adj. TO
Illiquidity	0.53**	0.02	0.02	0.03**
	(2.51)	(0.44)	(1.07)	(1.99)
Years to Maturity	0.01	0.01	0.01	0.01
	(0.49)	(0.44)	(0.47)	(0.46)
Log Size of Issue	0.13**	0.13**	0.12**	0.13**
	(2.41)	(2.43)	(2.29)	(2.35)
Time Between Issues	0.00	0.00	0.00	0.00
	(1.32)	(1.26)	(1.25)	(1.31)
Controls	Yes	Yes	Yes	Yes
Firm & Year	Yes	Yes	Yes	Yes
Adj-R^2	0.18	0.18	0.18	0.18
n	386	386	386	386
Observations	2,830	2,830	2,769	2,830
Panel B: Underwriting Spread				
Illiquidity	0.28***	0.03***	0.04***	0.03***
	(3.26)	(2.60)	(4.49)	(3.85)
Years to Maturity	0.06***	0.06***	0.06***	0.06***
	(35.48)	(35.28)	(34.77)	(35.34)
Log Size of Issue	-0.06***	-0.06***	-0.06***	-0.05***
	(-5.96)	(-6.03)	(-5.99)	(-5.85)
Time Between Issues	0.00	-0.00	-0.00	0.00
	(0.24)	(-0.02)	(-0.11)	(0.22)
Controls	Yes	Yes	Yes	Yes
Firm & Year	Yes	Yes	Yes	Yes
Adj-R^2	0.36	0.36	0.36	0.36
n	386	386	386	386
Observations	3,985	3,985	3,904	3,985

Table VII
Proportion of Firms Issuing Bonds by Year

This table reports the number of firms that issue bonds each year of the sample period. For each year, the number of firms eligible to issue debt (Potential Repeat Issuers) is estimated by summing the number of unique firms that have outstanding debt trading in the secondary market. The credit rating for firms that do not issue new debt is estimated using the existing issues. In the instances where current issues have multiple credit ratings, or if multiple ratings differ among agencies, the median credit rating across all issues is used. The trading data comes from TRACE, while the issuing data comes from Mergent FISD.

	Investment Grade			Speculative Grade			Not Rated		
Year	Potential Repeat Issuers	Issuers	%	Potential Repeat Issuers	Issuers	%	Potential Repeat Issuers	Issuers	%
2002	822	213	26	226	39	17	642	299	47
2003	778	236	30	141	43	30	771	353	46
2004	687	158	23	134	42	31	891	257	29
2005	689	172	25	136	32	24	895	246	27
2006	712	228	32	140	61	44	809	253	31
2007	748	269	36	146	70	48	766	277	36
2008	722	202	28	136	29	21	712	215	30
2009	801	297	37	157	61	39	617	341	55
2010	855	259	30	186	78	42	571	270	47
2011	905	298	33	202	60	30	519	232	45
2012	1,020	408	40	218	118	54	435	297	68

Table VIII
Does Illiquidity Impede Access to Capital?

This table reports coefficient results from a cross sectional probit analysis of the determinants of a firm's ability to issue new debt. To understand the firm's choice to issue new debt we regress the following equation:

$$Pr(Issued_{k,t} = 1) = \alpha_0 + \beta_1 Avg\ Illiquidity_{k,t-1} + \beta_2 OutstandingDebt_{k,t-1} + \beta_3 Recession_t + \beta_4 Recession_t * AvgIlliquidity_{k,t} + \varepsilon_{k,t}.$$

The dependent variable is an indicator variable equal to one if a firm k issues new debt in year t, zero otherwise. To be included in the sample the firm must have existing debt that currently trades in the secondary market. The principal independent variable is the average monthly illiquidity of existing bonds issued by the same firm in the year prior. We include as control variables the median credit rating of the existing bonds issued, the log of the outstanding debt issued by the firm at the end of the prior year, an indicator variable that marks the year 2008 and 2009 as crisis years, and an interaction of the firm's illiquidity variable and the recession variable. Robust test statistics are reported in parentheses, with ***,**, and * indicating significance at the 1%, 5%, and 10% levels respectively.

	Probit (Issuer = 1)			
	(1)	(2)	(3)	(5)
	PNT	KO	Amihud	Adj. TO
Intercept	-0.19	-2.24***	-2.69***	-0.58**
	(-0.69)	(-7.22)	(-7.63)	(-2.08)
Prior Year Illiquidity	-0.14**	-0.12***	-0.06***	0.04**
	(-2.02)	(-9.38)	(-8.27)	(2.39)
Prior Year Outstanding Debt	0.00	0.04***	0.06***	0.03**
	(0.12)	(4.28)	(5.83)	(2.51)
Recession	0.21***	-0.20	0.51	-0.03
	(2.97)	(-0.64)	(1.30)	(-0.36)
Recession * Prior Year Illiquidity	-0.36***	-0.02	0.02	0.00
	(-3.57)	(-0.69)	(1.30)	(0.02)
Firm and Credit Fixed Effects	Yes	Yes	Yes	Yes
N	2,473	2,460	2,455	2,473
Observations	17,870	16,179	16,123	17,842

Table IX
TRACE Reporting of New Debt Issuances

This table reports the number and volume of new issues during the years 2002-2006. During this sub-period, FINRA reported trades of bonds in waves depending on issue size and credit rating. We report the average number of new issues that are reported on TRACE at issuance.

	All New Issues		TRACE reported		Not Reported		Percent Trace
	#	Volume	#	Volume	#	Volume	
2002	229	$86,522,700	59	$58,475,000	170	$28,047,700	26%
2003	777	374,147,728	399	266,239,800	378	$107,907,928	51%
2004	546	313,509,987	398	280,635,237	148	$32,874,750	73%
2005	500	294,787,014	478	262,742,014	22	$32,045,000	96%
2006	566	386,532,825	566	386,532,825	0	0	100%
Total	2618	1,455,500,254	1900	1,254,624,876	718	$200,875,378	

Table X
The Real Effect of Price Impact on Issuing Costs

This table presents cross sectional regression results of the impact of TRACE reporting on underwriting costs. The independent variable includes the costs of underwriting, either the difference between the yield to maturity and the Treasury yield at the time of issuance or the gross spread paid to the underwriting syndicate. The principal independent variable is an indicator variable equal to one if the firm's prior issues are TRACE reported, zero otherwise. Control variables include the years to maturity, log of the issue size, the bond's duration, and indicator variables marking whether the bond is callable, convertible, 144A, senior, or a junior issue. We include crediting rating dummy variables, as well as firm fixed effects. Robust test statistics are reported in parentheses, with ***,**, and * indicating significance at the 1%, 5%, and 10% levels respectively.

Panel A: Credit Spread								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	PNT	KO	Amihud	Adj. TO	PNT	KO	Amihud	Adj. TO
	2002-2003				2002-2005			
Illiquidity	-0.21*	0.18***	0.03***	0.02***	0.35	0.31***	0.17***	-0.08
	(-1.91)	(3.52)	(2.68)	(2.69)	(0.72)	(3.04)	(5.01)	(-0.96)
Prior Bonds Trace Reported	-0.12	-0.17**	-0.15*	-0.12	-0.22***	-0.21***	-0.23***	-0.23***
	(-1.33)	(-1.98)	(-1.73)	(-1.26)	(-3.66)	(-3.78)	(-3.78)	(-3.67)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm and Year Fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R^2	0.09	0.14	0.09	0.09	0.14	0.18	0.15	0.13
N	280	282	263	308	416	411	414	416
Observations	576	571	571	576	1,085	1,078	1,077	1,085
Panel B: Underwriting Spread								
Illiquidity	2.04***	0.75***	0.37***	-0.26**	1.85***	0.24	0.04	0.19***
	(3.76)	(7.10)	(9.83)	(-1.99)	(2.61)	(1.31)	(0.89)	(2.65)
Prior Bonds Trace Reported	-0.49	-0.34	-0.92***	-0.32	-1.16***	-1.13***	-1.17***	-1.15
	(-1.52)	(-1.11)	(-3.00)	(-0.91)	(-3.97)	(-3.95)	(-4.10)	(-1.57)
Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm and Year Fixed	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj-R^2	0.47	0.51	0.52	0.47	0.22	0.23	0.23	0.22
N	304	306	285	389	459	453	458	459
Observations	1,119	1,096	1,075	1,119	2,285	2,229	2,212	2,285

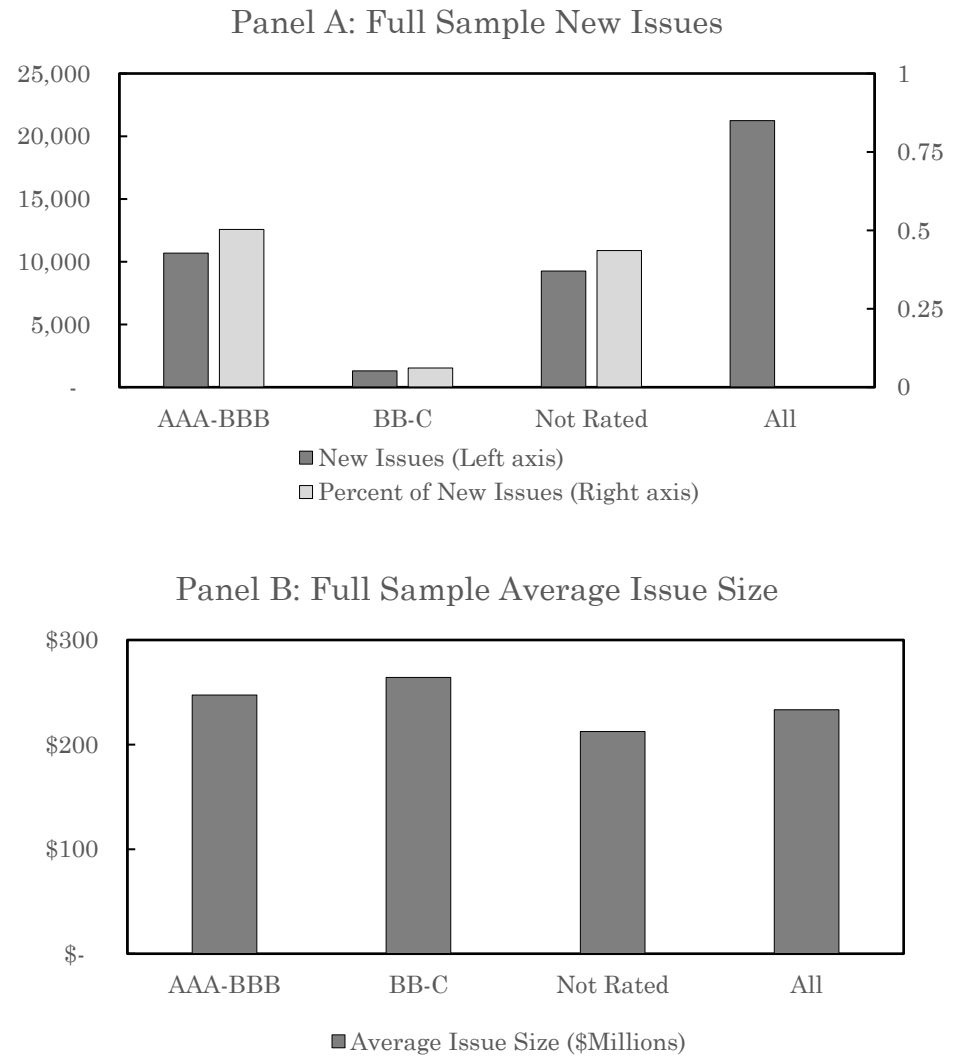


Figure 1. Corporate Bond Issuances (2002-2012)

The figures display primary market activity for corporate bonds issued from 2002 through 2012, partitioned by credit rating at the time of the issue. Panel A reports the total amount of capital raised through corporate bonds, whereas Panel B reports the average issue size. Both figures provide aggregate totals of the full sample of firms issuing bonds.

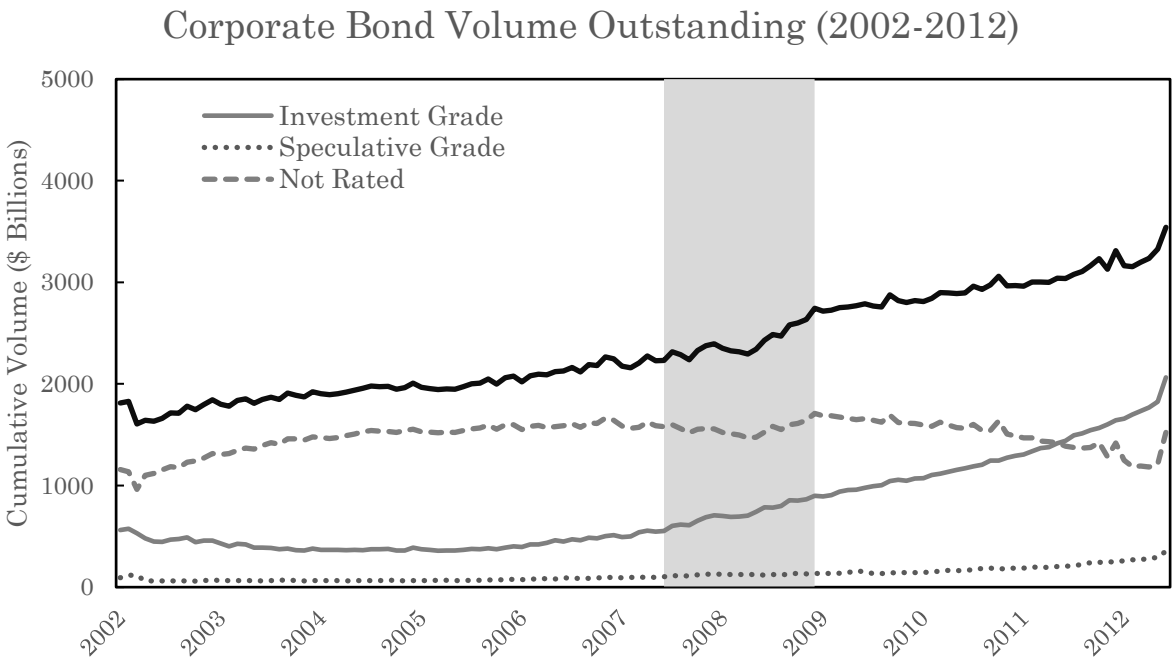


Figure 2. Monthly Corporate Debt Outstanding (2002-2012)
The figure displays the aggregate amount of corporate debt outstanding during the sample period from January 2002 through December 2012.

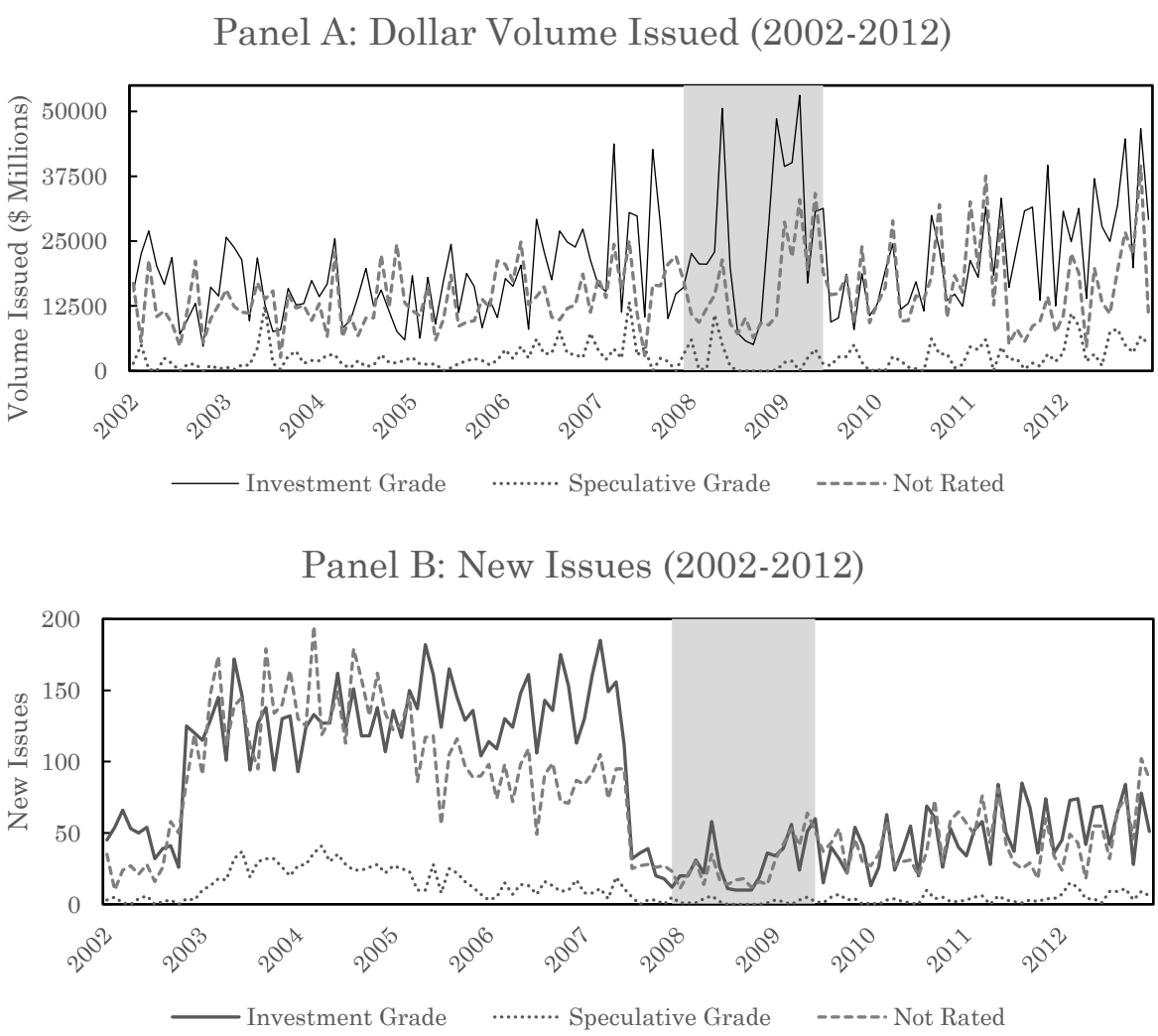


Figure 3. Monthly Corporate Bond Issues (2002-2012)

The figure displays the monthly amount of capital issued through U.S. corporate bonds during the sample period from January 2002 through December 2012. Panel A reports the monthly volume issued. Panel B reports the number of monthly issues.

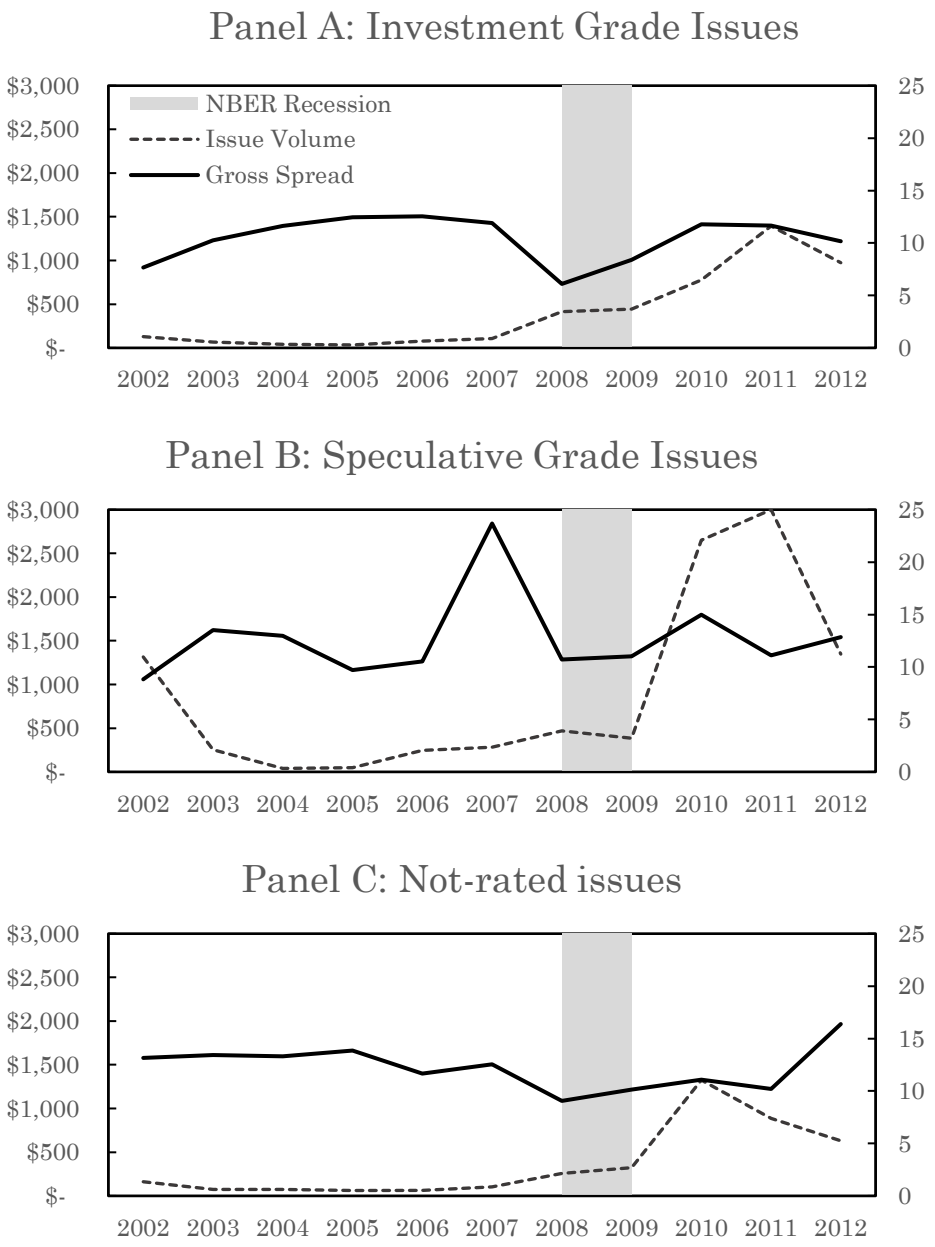


Figure 4. **Bond Trading Volume and fees from 2002-2012**

This figure displays the yearly average trading volume alongside the gross spread, the percentage of the issue amount paid to the underwriting syndicate. Panels A, B, and C report issues for investment grade, speculative grade, and non-rated grade issues respectively. Issue volume and gross spread are averaged by firm and issue.

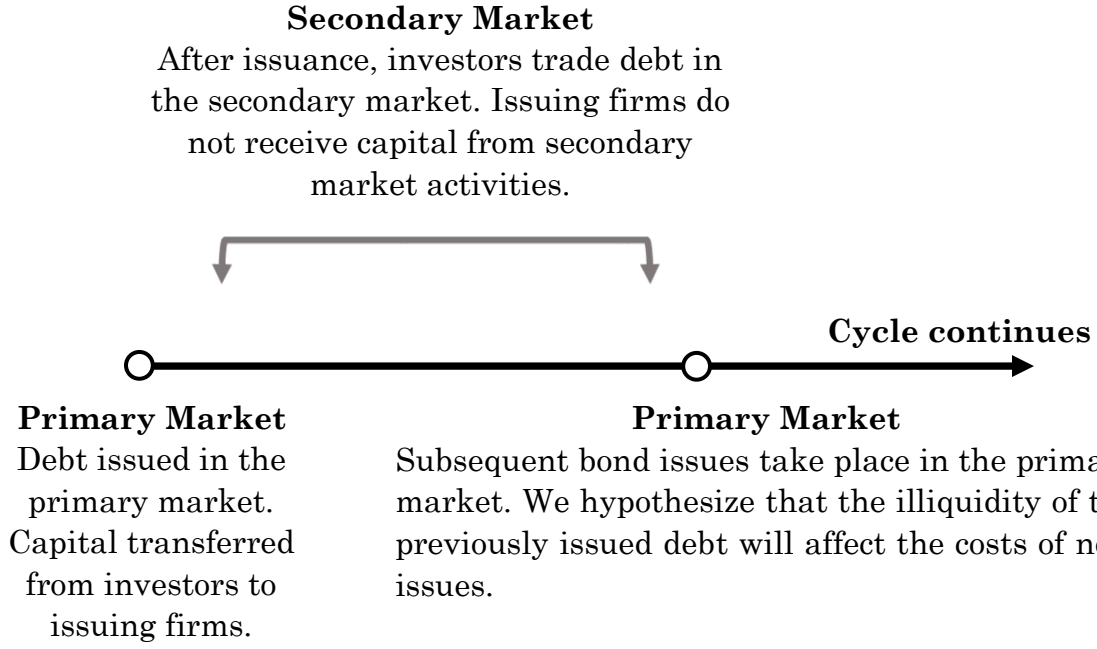


Figure 5. **How Secondary Market Liquidity Affects Underwriting Costs**
This study links secondary market activity with the primary market for new issues. When considering liquidity, this paper postulates that the characteristics of previously issued bonds will influence the fees associated with new issues. Both underwriters and investors estimate the potential risks of new bond issues by examining the past performance of outstanding bonds by the issuing firm.