Toward a Fully Continuous Exchange*

Albert S. Kyle† Jeongmin Lee‡

July 4, 2017

Abstract

We propose a new market design for a securities exchange that matches "continuous scaled limit orders." This new order type differs from standard limit orders in two ways. First, orders to buy and sell represent flows of shares over time rather than stocks of shares available for immediate purchase or sale. Second, orders are expressed as continuous piecewise linear functions relating price to quantity rather than step functions defined on a discrete grid of prices and quantities. Continuous scaled limit orders implement Fischer Black's vision of traders limiting temporary price impact by trading gradually over time. They dramatically lessen the rents high frequency traders earn from the current market design. The proposal is compatible with frequent batch auctions and random time delays.

Keywords: Market microstructure, auction design, market design, continuous scaled limit orders, smooth trading, frequent batch auction, VWAP, TWAP, high frequency trading, liquidity, bid-ask spread, price impact.

*We are grateful to Eric Budish, Peter Cramton, Vince Crawford, Larry Glosten, P.K. Jain, Charles Jones, Bruce Lehmann, Bill Longbrake, Anna Obizhaeva, Bart Taub, Yajun Wang, and Haoxiang Zhu for helpful comments.

†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.

Almost half a century ago, Fischer Black (1971*a*,*b*) made bold predictions about how stock market trading would change if the market design for trading stocks moved from the human-dominated specialist system to an electronic 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 temporary price impact makes it expensive to trade large quantities over short horizons, he conjectured that customers would spread large trades out over time to reduce 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.

Fischer Black was remarkably prescient. Large institutional traders around the world nowadays spread their trading out over time exactly like he said they would. Algorithmic trades are often executed by breaking large intended meta-orders into many small pieces and trading the many small pieces over time. Even as computers have become much cheaper, the technology gap among traders remains economically significant in a manner Fischer Black may not have foreseen. High frequency traders earn substantial profits from being a few microseconds faster than other traders. This has prevented the bid-ask spread on small orders from disappearing.

The purpose of this paper is to introduce a fully continuous exchange, a new type of exchange designed to implement Fischer Black's vision of an efficient market design. The continuous exchange matches buyers and seller who use a new order type, which we call continuous scaled limit orders.

A continuous scaled limit order is different from a standard limit order in two ways. First, it allows traders to choose the maximum rate at which the order executes over time. This allows all traders to submit one order to trade gradually. In effect, this converts supply and demand for securities into flows over time, just like textbook supply and demand schedules for goods and services. Second, it allows the actual execution rate to vary continuously between zero and the maximum rate as the price varies between upper and lower limit prices chosen by the trader. There is a uniquely determined market clearing price at which all supply and demand are satisfied. These two differences allow our proposal for continuous exchanges to address problems related to

discreteness in current stock market design.

The trading of equities in the United States and Europe has in recent decades become dominated by continuous limit order books which handle millions of buy and sell orders each day. Continuous limit order books have elements of discreteness in price, quantity, and time. 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 (one cent per share). A limit order book processes discrete orders sequentially in the order of their arrival. Because a discrete price grid makes demand and supply schedules step functions, there is typically excess supply or demand at every price. To allocate quantities when there is excess supply or demand, exchanges often use time priority, which executes incoming orders against the existing orders that arrived first in the limit order book. Because sending, receiving, and processing messages take time, latencies make it impossible for a trader to trade continuously in time. The ability to spread trades out over time depends on traders' message costs. Traders with inferior technology submit a smaller number of larger orders. In this way, standard limit order books promote provision of instantaneous liquidity by having larger-than-optimal orders available for execution at any given time. With discreteness in prices, quantities, and time, so-called "continuous" limit order books are hardly continuous.

The discreteness of today's market design creates rents for traders who have invested in costly superior technology. High frequency traders use their speed to take advantage of time priority by placing orders quickly to be the first in the queue. Time priority is profitable for traders at the front of the queue because large tick size limits price competition between ticks. High frequency traders use their speed to "pick off" slower traders' orders by hitting or lifting stale bids or offers before the slower traders can cancel them. They also use their speed to cancel their own stale bids and offers before other traders can hit or lift them. High frequency traders further benefit from a reluctance by slow traders to compete for time priority due to slow traders seeking to avoid being picked off.

By making trading continuous in price, quantity, and time, a fully continuous ex-

change dramatically lowers the rents that high frequency traders earn. Because traders with inferior technology can spread their orders over time easily, the orders picked off by high frequency traders are small. Since continuous scaled limit orders make flow demand schedules piecewise linear functions of the price, there is a unique market clearing price with no excess supply or demand. Therefore, an allocation rule such as time priority is unnecessary. Exchanges do not need to send numerous messages to traders to update partial order execution because traders can infer exact quantities traded from a public feed of the market clearing price. Since traders compete based on the price and not their relative speed, winner-take-all rents for being the fastest trader almost vanish.

If continuous scaled limit orders were widely used, there would be winners and losers. The current market design, promoting provision of instantaneous liquidity, favors the fastest traders and traders with short-term information. High frequency traders would lose because the gains from being the first to act would virtually disappear. Trading desks at investment banks might lose because they trade on very short-term information. In contrast, retail investors with inferior technology and little information would gain because having high message costs and slow speed would not prevent efficient trading algorithms. Large institutional investors would gain because their slightly slower speed would not be a relative disadvantage and because the opportunity cost of trading more slowly on long-term information would be low.

There are multiple sources of net welfare gains. First, there are gains from cost savings of reduced investment in the arms race to be the fastest trader in the market (Harris, 2013; Li, 2014; Biais, Foucault and Moinas, 2015; Budish, Cramton and Shim, 2015). Second, there are further gains from reducing the number of messages that traders need to send to spread out large trades over time and to receive updates on partial order execution. Third, the ability of institutional traders to trade on long-term information at lower cost results in increased production of long-term information while discouraging production of short-term information. Long-term information is potentially socially valuable because a more accurate stock price can induce more efficient investment decisions. Continuous scaled limit orders are consistent with the regulatory objectives the U.S. Securities and Exchange Commission (SEC), the U.S. Commodity Futures Trading Commission (CFTC), and the United Kingdom's Financial Conduct Authority (FCA),

whose various objectives include investor protection; fair, open, effective, transparent, and competitive financial markets; and efficient capital formation.

Continuous scaled limit orders make financial market design consistent with the theoretical models of Vayanos (1999), Du and Zhu (2017), and Kyle, Obizhaeva and Wang (2017). These models prove that spreading large trades out over time is equilibrium behavior when traders have market impact. Traders limit the urgency of their trading because dumping large quantities on the market quickly creates immediate, negative, temporary price impact. Smooth trading is an optimal response to smooth trading by others even when the rules of trading allow front-running, bluffing, and arbitrarily aggressive trading.

Budish, Cramton and Shim (2015) propose a market design in which frequent batch auctions are held at intervals of 100 milliseconds or one second. Frequent batch auctions reduce the profits high frequency traders earn by picking off resting limit orders because orders arriving between auctions are allocated quantities and prices that do not depend on which orders arrive first and which arrive last. Harris (2013) proposes random time delays, which similarly limit high frequency trader profits by shuffling the order in which messages are processed. Unlike continuous scaled limit orders, frequent batch auctions and random time delays do not reduce the tendency for traders with high message costs to submit orders which are too large and do not eliminate the problem that there is typically excess supply or demand at the prices at which trade takes place.

Black (1995) proposed "average-price market orders" and "indexed limit orders" as modifications of standard order types which allow gradual trading to match the Volume Weighted Average Price (VWAP) or Time Weighted Average Price (TWAP). Our more radical proposal represents a new market design which allows traders to achieve TWAP perfectly by placing one order.

The plan of this paper is as follows. Section 1 describes the continuous market design by contrasting continuous scaled limit orders with standard limit orders. Section 2 examines the implications of traders having different message costs, the algorithmic simplicity of implementing continuous scaled limit orders in a matching engine, and the implications of continuous scaled limit orders for high frequency trading. Section 3 discusses how continuous scaled limit orders are consistent with recent theoretical lit-

erature. Section 4 shows how continuous scaled limit orders are consistent with random time delays, frequent batch auctions, price speed bumps, and quantity speed bumps. Section 5 concludes.

1 Order Types and Market Clearing

In this section, we contrast standard limit orders with continuous scaled limit orders. Standard limit orders offer instantaneous execution of a given quantity; thus, cumulative trading volume is a discontinuous function of time. The discrete price grid defined by minimum tick size makes market supply and demand schedules step functions for which is typically necessary to ration supply or demand when trade takes place at any allowed price. In this sense, the market is discontinuous in quantities, prices, and time. Continuous scaled limit orders define flow-supply and flow-demand schedules for shares. Since supply and demand schedules are continuous, monotonic, piecewise linear functions, there is a unique market clearing price at which flow supply equals flow demand for shares.

1.1 Standard Limit Orders

The standard limit orders used in current exchanges are messages with three parameters: a buy-sell indicator, a quantity $Q_{\max}$, and a limit price P. A standard buy limit order conveys the message "Buy up to $Q_{\max}$ (shares) at a price of P (dollars per share) or better." For an order placed at time t_0, let $Q(t_0,\, t)$ denote the cumulative quantity executed during the time interval $[t_0,\, t]$ for $t > t_0$. Let $p(t)$ denote the most recent transaction price, and define $p_{\min}(t_0,\, t)$ as the minimum price $p(t)$ during the interval $t \in [t_0,\, t]$. Then $Q(t_0,\, t)$ satisfies

$$Q(t_0,\, t) = \begin{cases} Q_{\max} & \text{if } p_{\min}(t_0,\, t) < P, \\ \alpha(t_0,\, t) Q_{\max} & \text{if } p_{\min}(t_0,\, t) = P, \\ 0 & \text{if } p_{\min}(t_0,\, t) > P. \end{cases} \qquad \text{where } \alpha(t_0,\, t) \in [0,\, 1], \tag{1}$$

If the minimum market price $p_{\min}(t_0, t)$ is above the limit price P, nothing is bought; if it is below the limit price, the order is fully executed ($Q(t_0, t) = Q_{\max}$); if it exactly equals the limit price, then the quantity executed $Q(t_0, t)$ depends on the rule for allocating traded quantities when it may not be possible to satisfy all demands the limit price P. This allocation rule determines $\alpha(t_0, t)$, which is a monotonically nondecreasing step function of time measuring the fraction of the order executed up to time t. Depending on how the order interacts with other orders, it might receive a full execution ($\alpha(t_0, t) = 1$), a partial execution ($0 < \alpha(t_0, t) < 1$), or no executed quantity at all ($\alpha(t_0, t) = 0$).[1]

With standard limit orders, the limit price P is an integer multiple of the minimum tick size,[2] and the quantity $Q_{\max}$ is an integer multiple of the minimum lot size.[3] The discrete price grid makes an allocation rule to determine $\alpha(t_0, t)$ necessary because the market supply and demand schedules calculated by aggregating all sell and buy orders define quantities as discontinuous step functions of price, and 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 chooses the price that maximizes the quantity traded. Since there is typically excess supply or demand at this price, some rule is needed to allocate prices and quantities. Different allocation rules determine the fractional allocation $\alpha(t_0, t)$ in different ways. For example, time priority specifies that older orders must receive full execution ($\alpha(t_0, t) = 1$) before newer orders receive any execution ($\alpha(t_0, t) > 0$). Instead of time priority, some markets use a "pro rata" or proportional allocation rule according

[1] The notation in equation (1) is meant to convey intuition; it is not meant to be mathematically precise. With more formal notation, the quantities $Q_{\max}$, P, $\alpha(t_0, t)$, and $Q(t_0, t)$ would have superscripts indicating the identity of the specific message, which could be mapped to a specific trader. Instead of equation (1), we could write $Q(t_0, t) = \alpha(t_0, t)Q_{\max}$ with $\alpha(t_0, t) = 1$ when $p_{\min}(t_0, t) < P$ and $\alpha(t_0, t) = 0$ when $p_{\min}(t_0, t) > P$.

[2] In the U.S. market, the stated minimum tick size 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 $0.01 (one cent per share) in 2001. The issue of tick size has been a surprisingly controversial political issue over the years. In 2015, the SEC approved a "Tick Size Pilot Program," an experiment in which stock exchanges temporarily increased the tick size on some stocks to five cents while leaving the tick size on other stocks unchanged at one cent.

[3] Historically, "round lots" (integer multiples of 100 shares) have been subject to different order execution and price reporting practices than "odd lots" (integer multiples of one share which are not integer multiples of 100 shares).

to which all orders receive a partial execution proportional to the unexecuted quantity in the limit order book when executed against an incoming order.

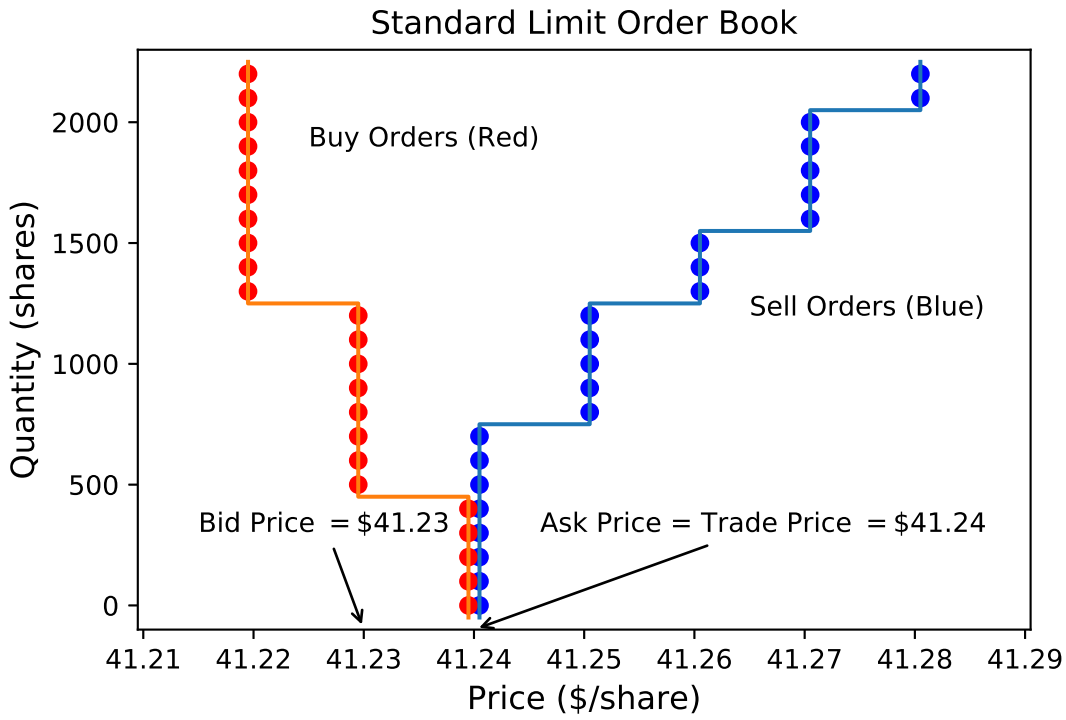

Figure 1: Market Clearing with Standard Limit Orders

Figure 1 depicts a transaction for 500 shares in a standard limit order book. Each solid dot represents demand or supply for a minimum-round-lot quantity of 100 shares. An executable limit order to buy 500 shares at $41.24 has just arrived into the limit order book. Since there are limit orders to sell 800 shares at the same price, all 500 shares in the buy order are immediately executed, but 300 shares of the sell orders at $41.24 remain unexecuted. Both immediately before and immediately after this order execution, the best bid price is $41.23 and the best offer price is $41.24. In addition to unexecuted buy orders at $41.23 and unexecuted sell orders at $41.24, the limit order book also has unexecuted buy orders at lower prices and unexecuted sell orders at higher prices. The

graphs of the supply and demand schedules look like step functions.[4]

Sending, receiving, and processing order messages take time, and the orders are sequentially processed based on the arrival of orders. Even though messages are sent and received in continuous time, economically significant time lags of a few milliseconds effectively prevent traders from trading continuously with standard limit orders. Continuous scaled limit orders, discussed next, allow trading to take place essentially continuously in time by incorporating a dynamic execution strategy into the order itself.

1.2 Continuous Scaled Limit Orders

Define a "continuous scaled limit order" as a message with five parameters: a buy-sell indicator, a quantity $Q_{\max}$, two price levels P_L and P_H (with $P_L < P_H$), and a maximum trading speed $U_{\max}$.[5] Such an order conveys the message, "Buy up to a cumulative total of $Q_{\max}$ (shares) at maximum rate $U_{\max}$ (shares per hour) at prices between P_L and P_H (dollars per share)."

A trader could mimic a continuous scaled limit order by sending a sequence of very small standard limit orders one at a time, with a new small order sent to replace the previous order after it is executed. Since submitting orders takes time—even at the speed of light—a sequence of small orders can mimic a continuous scaled limit order only imperfectly. By moving numerous small orders into the matching engine, one continuous scaled limit order replaces thousands of standard limit orders and does it more efficiently because performing operations within the matching engine is much faster and cheaper than performing operations by sending and processing messages. A continuous scaled limit order defines a demand function for a flow of assets just like the demand schedules in economics textbooks represent flow demands for goods and ser-

[4] Mathematically, supply and demand schedules are step functions which define the price as a function of the quantity. This can be seen by rotating the vertical and horizontal axes. Discontinuities in the step functions occur at quantity levels which require a jump in price to accommodate a changing quantity supplied or demanded.

[5] An order may also contain parameters defining the time when the order begins execution and the time when execution stops. We assume for simplicity that orders are for immediate execution and are good until canceled. We conjecture that it would be possible to develop more complex order types that would allow $U_{\max}$ to be a function of other market characteristics such as trading volume, price volatility, or some measure of market liquidity.

vices over time. The trading speed or flow demand $U\left(p\left(t\right)\right)$ is a function of the the market clearing price $p\left(t\right)$ given by

$$U\left(p\left(t\right)\right) \coloneqq \begin{cases} U_{\max} & \text{if } p\left(t\right) < P_L, \\ \left(\frac{P_H - p(t)}{P_H - P_L}\right) U_{\max} & \text{if } P_L \leq p\left(t\right) \leq P_H, \\ 0 & \text{if } p\left(t\right) > P_H. \end{cases} \tag{2}$$

If the price is strictly below P_L, the trader buys at rate $U_{\max}$. If the price is strictly above P_H, the trader does not trade. If the price is between P_L and P_H, the demand schedule is interpolated linearly, making its slope $U_{\max}/(P_H - P_L)$.

For an order placed at time t_0, the cumulative quantity executed by time $t > t_0$ is given by the integral

$$Q\left(t_0, t\right) \coloneqq \int_{t_0}^{t_0+t} U\left(p\left(\tau\right)\right) \mathrm{d}\tau. \tag{3}$$

If the order is canceled at time t without being filled, then $Q\left(t\right) < Q_{\max}$. If the order has been filled at time t, then $Q\left(t\right) = Q_{\max}$.[6] If the price remains low enough so that that order is executed at maximum rate $U_{\max}$, the order will be fully executed exactly after time $t = Q_{\max}/U_{\max}$ (hours). If the price fluctuates above and below P_L and P_H, the full execution will take longer than time $Q_{\max}/U_{\max}$. If the price stays above P_H, the order will not be executed at all.

Continuous scaled limit orders are different from the standard limit orders used on current exchanges in two important ways. First, trading is made continuous in time and quantity by converting stocks $\alpha(t_0, t)Q_{\max}$ (shares) to flows $U(t)$ (shares per hour). This limits the rate at which orders can be executed. Second, the quantity is made a continuous function of a uniquely defined market clearing price by converting step functions (for shares) into piecewise linear functions (for shares per hour). Continuous scaled limit orders make an allocation rule unnecessary, even when the quantities $Q_{\max}$ (shares) and $U_{\max}$ (shares per hour) are multiples of a minimum lot size and prices P_H and P_L are multiples of a minimum tick size.

A set of continuous scaled limit buy orders defines an aggregate flow-demand sched-

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

ule, denoted $D(p)$, as the sum of the flow demand $U(p)$ 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 $S(p)$, is defined analogously to a demand schedule as the graph of a continuous, weakly monotonically increasing, piecewise linear function.

Suppose that the aggregate demand and supply schedules intersect at a point where either of the two is not flat. Then the excess demand schedule $D(p) - S(p)$ is strictly decreasing in the neighborhood of the intersection. There exist a "best bid" price P_B and "best ask" price P_A, both on the tick grid, where P_A is one tick size larger than P_B, and there is excess demand at the best bid and excess supply at the best ask price given by

$$D(P_B) - S(P_B) \geq 0 \qquad \text{and} \qquad S(P_A) - D(P_A) > 0. \tag{4}$$

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

$$\omega := \frac{D(P_B) - S(P_B)}{D(P_B) - S(P_B) + S(P_A) - D(P_A)}. \tag{5}$$

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

$$p(t) = P_B + \omega (P_A - P_B). \tag{6}$$

Intuitively, the price is a weighted average of the two prices P_B and P_A, with weights $1-\omega$ and ω proportional to the excess demand and supply at these prices.[7]

Figure 2 depicts market clearing with continuous scaled limit orders. Both the downward sloping demand schedule and upward sloping supply schedule are piecewise lin-

[7] 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.

ear functions of price with knot points at which prices are integer multiples of the minimum tick size of $0.01 and quantities are integer multiples of one share per hour. The downward sloping piecewise linear flow-demand schedule and the upward sloping piecewise linear flow-supply schedule intersect at a unique price and quantity. The unique price, $41.246 per share, is not an integer multiple of minimum tick size. The unique quantity, 20 800 shares per hour, happens to be an integer number of shares per hour, but this is not generally the case.

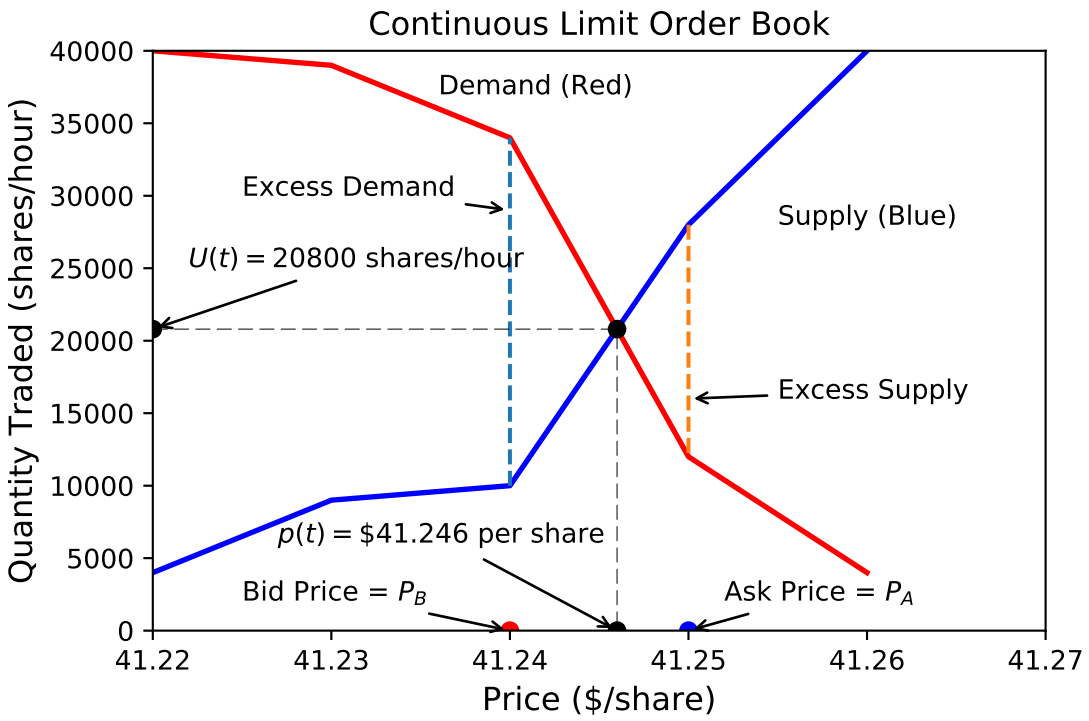

Figure 2: Market Clearing with Continuous Scaled Limit Orders

In the United States, bid and ask prices are published in integer multiples of one cent. Although there is technically no bid-ask spread with continuous scaled limit orders, the exchange can define as its "best bid" and "best ask" for public quotation purposes the prices obtained by rounding the market clearing price down and up to the nearest one-cent tick increments. These prices correspond to P_B and P_A in equations (4), (5), and (6).

In Figure 2, there is net excess demand of 24 000 shares per hour at the best bid price of $P_B = \$41.24$ per share and net excess supply of 16 000 shares per hour at the best offer price of $P_A = \$41.25$ per share.

Unlike with standard limit orders, all continuous scaled limit orders are treated symmetrically and executed simultaneously. With standard limit orders, order matching respects price and time priority. With continuous scaled limit orders, all flow demand and supply is satisfied at the market clearing price.

2 Economics and Technology

Trading involves real resource costs which are borne both by the exchange and by traders. There are huge economies of scale in processing messages, operating a matching engine, and disseminating and processing trading information. These economies of scale imply that exchanges can process orders more efficiently than traders, and high-volume traders can process orders more efficiently than low-volume traders. Continuous scaled limit orders have the economic effect of replacing messages which traders send to exchanges with instructions that are carried out inside the exchange's matching engine. This makes trading more efficient in a manner that benefits small traders relatively more than large traders. In this way, continuous scaled limit orders reduce the deadweight costs of the high-frequency-trading arms race in a direct and more efficient manner than other indirect mechanisms, including frequent batch auctions and random order-processing delays (discussed in section 4).

2.1 Message Costs

As discussed in Section 3 below, theory suggests that traders would use continuous scaled limit orders. Today's matching engines are based on standard limit order books, not continuous scaled limit orders. Using standard limit orders to mimic the effect of continuous scaled limit orders requires sending thousands of standard limit orders to trade tiny quantities. An infrastructure of costly computers, communication lines, software, and professional skills is needed to handle huge numbers of messages.

Message costs vary across traders. Technologically unsophisticated retail traders do not have access to the best technologies; as a result of high message costs, they place a small number of standard limit orders. Technologically sophisticated large institutions do have access to such technologies; as a result of low message costs, it is economically feasible for them to mimic continuous scaled limit orders with thousands of standard limit orders. High frequency traders have the fastest technologies; they use their speed advantage to pick off other traders' orders, gain time priority in standard limit order books, and cancel their own limit orders before other traders can pick them off.

Large institutional investors often attempt to match VWAP or TWAP over some time interval by placing many small orders to participate in a constant fraction of trading volume and to match the prices at which other traders are trading. TWAP is easier to implement than VWAP because VWAP implementation requires forecasting random total volume minute by minute over the horizon of execution. VWAP and TWAP are, in practice, quite similar. A TWAP order corresponds precisely to a fully executable continuous scaled limit order. It would be possible to define a modified form of continuously scaled VWAP orders to restrict execution to a maximum fraction of contemporaneous volume rather than a maximum number of shares. In principle, a VWAP version of continuous scaled limit orders would reduce message costs if traders otherwise would submit messages to change the execution rate of their orders as volume fluctuated in the market.

By moving the costs from traders to the matching engine, continuous scaled limit orders can level the playing field for large traders and small traders by allowing all traders to slice their orders into small pieces automatically and gradually trade toward their target inventories without being equipped with large bandwidth and processing power. Order-shredding strategies like VWAP and TWAP suggest that institutional investors try to mimic with many messages what continuously scaled limit orders are designed to do more efficiently with one message.

The tradeoff traders face between incurring lower message costs for large orders and incurring higher message costs for order shredding can be seen in data. Since the introduction of a one-cent tick size in 2001, O'Hara, Yao and Ye (2014) and Kyle, Obizhaeva and Tuzun (2016) show that average trade size has declined dramatically. Improved computer technology during this period suggests that the decline in trade size is the

result of declining message costs making it easier to mimic the outcome of continuous scaled limit orders.

While falling message costs make it cheaper for all traders to mimic continuous scaled limit orders, the inequality in message costs across traders maintains an unleveled playing field in which small differences in message costs can have large differences in market outcomes.

Since retail traders have higher message costs than institutional investors, they may adopt execution strategies which require fewer messages but result in worse prices obtained on executed orders. These strategies include placing executable orders to hit bids or lift offers immediately, giving up a large price concession than could be obtained by placing limit orders close to the market and adjusting limit prices over time. They also include leaving resting limit orders away from the market, where they suffer a risk of being picked off if market prices change quickly over a short period of time.

Since high frequency trading requires placing and canceling many orders, high frequency traders must have very low message costs to be competitive. It is reasonable to conjecture that high frequency traders achieve low message costs by exploiting economies of scale and thus submit smaller orders than other traders. Consistent with this conjecture, Kirilenko et al. (2017) show that high-frequency traders have trades that are half as large (five versus ten contracts) as other traders.

High frequency traders also have faster speeds (lower latency) than other traders. Low message costs and fast speeds allow high frequency traders to make small profits on trades with other traders. Their profits per share or per contract are higher when trading against traders with relatively higher message costs or relatively lower speed. In a study of high frequency trading in the S&P E-mini futures market, Baron, Brogaard and Kirilenko (2012) find that high frequency traders earn profits when taking the other side of trades with both retail (low volume) and other (higher volume) traders, but they earn larger profits per contract trading against retail traders than others. These results are likely due to relatively higher message costs and lower speeds making it difficult for retail traders to time their trades strategically.

Even if technological progress reduces absolute message costs for institutional investors so much that the message cost savings of using continuous scaled limit orders

becomes small relative to sending many messages to a standard limit order book, the relative differences in speed would allow high frequency traders to continue to trade profitably in standard limit order books based only on using their *relative* speed advantage.

2.2 Matching Engine

To make it easy for both sophisticated and unsophisticated users to submit demand schedules, the user interface might use intuitive units. We require the price limits P_H and P_L to be integer multiples of the minimum tick size of one cent, the order size $Q_{\max}$ to be an integer number of shares, and the maximum rate of trading $U_{\max}$ to be an integer number of shares per hour. These assumptions make the market supply and demand schedules piecewise linear functions with knot points at prices that are integer multiples of minimum tick size and quantities that are integer numbers of shares per hour.

For the matching engine to calculate the market clearing price and the quantities each trader is allocated, some degree of discretization is necessary. For illustrative purposes, we assume that quantities are measured as nanoshares (10^{-9} shares), prices are measured in microdollars (10^{-6} dollars or ten-thousandths of a cent per share), time is measured in milliseconds (10^{-3} seconds), and maximum trade rates are measured in nanoshares per millisecond. These units are chosen so that quantities, prices, time, and trade rates can be rounded off to integer numbers of nanoshares, microdollars, milliseconds, and nanoshares per millisecond with little loss of economic significance. Importantly, such discretization for internal calculations is economically different from the discreteness in the current market design because gains from gaming become negligible. As we discuss next, these units and rounding conventions allow the matching engine to perform fast integer arithmetic calculations.

Suppose a customer submits a continuous scaled limit order to buy $Q_{\max} = 8000$ shares at maximum rate $U_{\max} = 4000$ shares per hour between limit prices $P_L = \$41.25$ and $P_H = \$41.50$ per share. Let K denote the minimum tick size, say \$0.01 per share, and write prices as $P = nK$ where n is a positive integer. If the stock price is \$41.35 per

share, then $n = 4135$. Let $U = (U_{MIN}, \ldots, U_{MAX})$ denote a vector defining the order's trade rate $U_n = U(p)$ when $p = nK$. For example, if $MIN = 0$ and $MAX = 10\,000$, then U defines trade rates at all increments of one cent from $p = \$0.00$ to $p = \$100.00$ per share. Then the limit order to buy at maximum rate $U_{\max}$ between $P_L = n_L K$ and $P_H = n_H K$, consistently with equation (2), can be written

$$U_n := \begin{cases} U_{\max} & \text{if } n < n_L, \\ \left(\frac{n_H - n}{n_H - n_L}\right) U_{\max} & \text{if } n_L \le n \le n_H, \\ 0 & \text{if } n > n_H. \end{cases} \tag{7}$$

Within the matching engine, the trader's demand schedule in equation (7) can be expressed as a vector $\hat{U} = (\hat{U}_{MIN}, \ldots, \hat{U}_{MAX})$, where $\hat{U}_n$ rounds the quantity U_n to an integer number of nanoshares per millisecond in a computationally convenient manner discussed next. The hat-notation indicates quantities rounded to integers in the matching engine. When the user places a continuously scaled limit buy order at a rate of 4000 shares per hour between \$41.25 and \$41.50 per share, each one-tick price increment of one cent between these prices theoretically changes the rate of buying by $\Delta U = U_{n+1} - U_n = 44\,444.44444\ldots$ nanoshares per millisecond.[8]. For internal calculations, this is rounded to the integer rate of $\Delta\hat{U} = 44\,444$ nanoshares per millisecond per one-cent tick. Using this approximation, the order in (7) can be written

$$\hat{U}_n := \begin{cases} (n_H - n_L)\,\Delta\hat{U} & \text{if } n < n_L, \\ (n_H - n)\,\Delta\hat{U} & \text{if } n_L \le n \le n_H, \\ 0 & \text{if } n > n_H, \end{cases} \tag{8}$$

where the trade rates $\hat{U}_n$ are integer numbers of nanoshares per millisecond. Note that $(n_H - n_L)\,\Delta\hat{U}$, equivalent to 3999.96 shares per hour, is approximately equal to $U_{\max}$ (4000 shares per hour), but these numbers may differ by more than one nanoshare per second due to rounding conventions. There is little economic significance associated with such rounding. At the fastest possible execution rate, it would take a fraction of a second longer than one hour to trade 4000 shares.

[8] This amount is calculated as $4000 \cdot 10^{9-3}/(60^2 \cdot 25)$.

Given an integer vector of demand schedules $\hat{U}$ for all traders, the aggregate demand schedule can be calculated by adding the vectors representing all demand schedules element-by-element. The computational complexity of keeping track of piecewise linear demand schedules would be much greater if the knot points occurred at arbitrary prices rather than at integer multiples of a fixed tick size. Since computers are optimized to perform integer vector operations efficiently, such calculations are performed at very low cost. For optimized implementation, the calculations may by further simplified by exploiting the sparse, simple structure of the vectors $\hat{U}$.

With supply schedules defined analogously to demand schedules, it is computationally easy to calculate the best bid and ask prices because demand schedules are downward sloping and supply schedules are upward sloping. Bisection is an efficient algorithm for this purpose: Given a high price with excess supply and a low price with excess demand as a starting guess, choose the price half-way in between as the next guess, observe whether there is excess demand or supply at this price, then update the initial guess and continue guessing until the best bid P_B and best offer P_A are calculated in a manner consistent with (4). These calculations have computationally trivial cost.

As an order executes, the theory implies that executed quantities change continuously as functions of time. As a practical computational matter, executed quantities will be calculated based on the minimum increment in time stamps, which we are assuming to be one millisecond. As long as the exchange receives no new messages which add, cancel, or change existing orders, and as long as all existing orders are not fully executed, each executable order in the limit order book will execute in small constant increments as the matching engine's clock ticks. When new messages arrive or existing order finish executing, market clearing prices and trading rates are recalculated.

In calculating market clearing prices and quantities, it remains necessary for the matching engine to deal with some rounding of fractional prices and quantities so that they can be expressed as integers. Write the bid price as $P_B = n_B K$ and the ask price as $P_A = n_A K$ (with $n_A = n_B + 1$). Suppose the bid and ask prices are within the order's upper and lower limits ($n_L \leq n_B < n_A \leq n_H$), and let $\hat{U}_B := (n_H - n_B)\Delta\hat{U}$ denote an order's demand rate at the bid price. If the interval between price and quantity recalculations is Δt (such as seven milliseconds), then the theoretical number of shares

$\Delta Q = Q(t, t + \Delta t)$ traded by a limit buy order over the time interval $[t, t + \Delta t]$ is given by

$$\Delta Q = \left(\hat{U}_B - \omega \cdot \Delta\hat{U}\right) \Delta t. \tag{9}$$

Since the fraction ω in (5) is the ratio of two arbitrary non-negative integers, it is an arbitrary rational number. We require that ω is rounded to 4 decimal places and let $\hat{\omega}$ denote the rounded result, so that the market clearing price is approximated as an integer multiple of 10^{-6} dollars, or, equivalently, 10^{-4} cents. With ω rounded to $\hat{\omega}$, the price is given by

$$\hat{p}\,(t) = P_B + \hat{\omega} \cdot (P_A - P_B) = P_B + \hat{\omega} \cdot K. \tag{10}$$

This rounded price $\hat{p}(t)$ is an integer multiple of 10^{-6} dollars because P_B and P_A are measured in cents (integer multiples of 10^{-2} dollars) and the factor $\hat{\omega}$ is an integer multiple of 10^{-4}. The trade rate $\hat{U}(\hat{p}(t))$ is calculated by rounding the implied rate $\hat{U}_B - \hat{\omega} \cdot \Delta\hat{U}$ to an integer number of nanoshares per millisecond.[9] The number of shares bought over the time interval Δt (an integer multiple of milliseconds) is

$$\Delta\hat{Q} = \hat{U}(\hat{p}(t)) \cdot \Delta t \approx \left(\hat{U}_B - \hat{\omega} \cdot \Delta\hat{U}\right) \cdot \Delta t. \tag{11}$$

Because of rounding, the aggregate quantity bought by all buyers will not be exactly equal to the aggregate quantity sold by all sellers. To deal with this issue, the exchange itself can take the difference—a few nanoshares—into its own inventory each millisecond. To move the imbalances toward zero, the exchange may calculate the rounded quantity $\hat{\omega}$ by rounding the fraction ω up or down in the direction which reduces its cumulative inventory. Obviously, these inventories are likely to be economically inconsequential.

At the end of each day, the cumulative quantity traded by each customer is an integer number of nanoshares that is unlikely to be an exact integer number of shares for partially filled orders. One way to deal with the fractional shares is to clear and settle fractional shares. If, alternatively, shares must be cleared and settled as integer num-

[9] While $\hat{U}_B$ is an integer number of nanoshares per millisecond, rounding is necessary because $\hat{\omega} \cdot \Delta\hat{U}$ may have a fractional number of nanoshares per millisecond.

bers of shares, 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 and fractional part ϵ by writing $X = \text{int}(X) + \epsilon$. To clear the fractional part of X, we propose cash-settling the fraction ϵ by randomly 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.[10]

Since the price $\hat{p}(t)$ and thus the trading rates $\hat{U}\left(\hat{p}\left(t\right)\right)$ are uniquely defined, traders can accurately infer the quantities they trade from a public feed of prices, or equivalently from a time-stamped data feed providing the history of P_B, P_A, and $\hat{\omega}$ in (5). Exchanges need not send constant updates of prices and quantities for each fractional share bought by 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.

For standard limit orders, the situation is entirely different. Of course, when a trader sees a price lower than the price on a limit buy order, the trader can infer that his order was fully executed and expect a prompt confirmation of full execution. But when a trader observes the asset trading at the exact limit price on his order, the trader does not know whether any portion of his own order was executed. Furthermore, even if the trader also observes the quantity traded at the reported price, the trader cannot infer from a public feed whether any of this quantity represents a portion of his own order unless the trader knows his exact position in the queue of time priority. As a consequence of time priority, it becomes necessary for an exchange offering standard limit

[10] One approach is the following simple mechanism: Pick two traders randomly from the pool of traders with fractional shares. Suppose one trader has a fraction ϵ_1 shares and the other trader has a fraction ϵ_2 shares with $0 \leq \epsilon_1 < 1$ and $0 \leq \epsilon_2 < 1$. There are two cases. If (case 1) $\epsilon_1 + \epsilon_2 < 1$, then trader 1 buys ϵ_2 shares from trader 2 with probability $\epsilon_1/(\epsilon_1+\epsilon_2)$, and trader 2 buys ϵ_1 shares from trader 1 with probability $\epsilon_2/(\epsilon_1+\epsilon_2)$. Then the expected quantity traded by each trader is zero, and there can be no gaming. If (case 2) $\epsilon_1 + \epsilon_2 > 1$, then trader 1 sells $1-\epsilon_2$ shares to trader 2 with probability $(1-\epsilon_1)/(1-\epsilon_1+1-\epsilon_2)$ and trader 2 sells $1-\epsilon_1$ shares to trader 1 with probability $(1-\epsilon_2)/(1-\epsilon_1+1-\epsilon_2)$. As a result of this process, one trader has a fractional share and the other has an integer number of shares. The trader with fractional shares is added back to the pool and the process repeated until all traders have disposed of all fractional shares.

orders to send numerous messages to traders so that the traders can know the cumulative quantities traded on partially executed orders. The cheapest way to provide this information is to send each trader a message reporting partial execution of any order. This makes it necessary for traders in the back of the queue to infer from non-receipt of a message that nothing has happened to their order. If these confirmation messages are sent more slowly than market-wide prices and quantities are executed, the trader must deal with ambiguity concerning partial execution while waiting for a confirmation to arrive or not to arrive. If confirmation messages are sent more quickly than market-wide prices and volumes are updated, then traders receiving confirmation of partially executed orders have potentially valuable information about market conditions before other traders and might use this information to their own advantage at the expense of other traders. For these reasons, continuously scaled limit orders are not only computationally more economical than standard limit orders, but they also create a more level playing field concerning timely access to information about current market conditions.

The way in which bids, offers, traded quantities, and trade prices are reported publicly is highly regulated. In U.S. markets, bids and offers have historically been reported in integer multiples of one cent, and quantities are reported in multiples of 100 shares. An exchange offering continuous scaled limit orders might report P_B and P_A as bid and ask prices and report the exact price $\hat{p}(t)$ every time it changed (time-stamped to milliseconds). This would allow traders to infer from public feeds the exact cumulative quantities executed on their orders. Since there is no instantaneous market depth in a continuous limit order book, reporting the rate at which depth is supplied over time in a continuous limit order book does not fit within the current price reporting structure. An exchange offering continuous scaled limit orders could offer its own private feed providing—at a minimum—the price $\hat{p}(t)$, the rate of trade $D(\hat{p}(t)) = S(\hat{p}(t))$, and excess flow demand and excess flow supply at best bid and offer prices $P_B(t)$ and $P_A(t)$, respectively. The exchanges' private feed might also supply information about the shape of market demand and supply schedules away from the best bid and offer prices.

Operating a matching engine requires sending and receiving, encrypting and decrypting, confirming and reconfirming, sorting and queuing, and matching numerous messages related to placing, modifying, canceling, and confirming full or partial execu-

tion of orders. The computational costs of such message processing, incurred by both exchange and its customers, are much greater than the computational costs of running the matching engine itself. By automatically spreading order execution out over time, continuous scaled limit orders replace messages that shred orders with internal calculations within the matching engine. Since making internal calculations is dramatically cheaper than processing messages, exchanges offering continuous scaled limit orders will reduce both their own costs and the costs of customers if the customers demand the gradual execution of orders that continuous scaled limit orders make possible.

2.3 High Frequency Trading

Continuous scaled limit orders level the playing field by neutralizing the speed advantage relatively faster traders enjoy in markets where matching engines match standard limit orders. In a limit order book with standard limit orders, speed benefits faster traders in three ways: (1) It allows faster traders to pick off resting limit orders when prices change by letting faster traders hit bids and lift offers faster than slower traders can cancel resting bids and offers. (2) It allows faster traders to avoid being picked off by canceling their own stale resting limit orders faster than slower traders can hit them. (3) It allows faster traders to gain the best position in the time-priority queue at the best bid and offer prices by allowing them to add orders to the limit order book more quickly than slower traders. These speed advantages are a function of relative speed, not absolute speed; speeding up all traders by a factor of ten does not change the economic magnitude of these advantages.

Example of Standard Limit Order Book. To illustrate these advantages, consider the following typical hypothetical scenario in a standard limit order book. The best bid for a stock is $39.99 and the best offer price is $40.01. Both of these bids and offers represent orders previously placed by the fastest high frequency trader in the market. An institutional investor would like to purchase 5000 shares by making 50 purchases of 100 shares each at a gradual rate over time. The midpoint is an attractive price at which to buy, so the institutional investor places a standard limit order to buy 100 shares at the

prevailing midpoint of $40.00. The high frequency trader bidding at $39.99 also thinks $40.00 is an attractive price at which to buy, so the high frequency trader submits a limit order to buy 100 shares at the same price at approximately the same time. Because his technology is faster, the high frequency trader's order reaches the limit order book before the order of the institutional investor. After both orders have arrived into the limit order book, there are 200 shares at the new best bid of $40.00 and 100 shares at the best offer of $40.01.

Next, a different institutional seller submits a limit sell order to sell 100 shares at $40.00 per share. Since the high frequency trader's buy order arrived into the limit order book first, it has time priority and therefore is executed against the incoming order. The original institutional buyer, frustrated by his inability to buy at $40.00, places an executable limit order to buy at $40.01. This order executes against the fastest high frequency trader's sell order that was in the limit order book to begin with. The fastest high frequency trader has now bought 100 shares at $40.00 and sold 100 shares at $40.01, making a profit of $1.00. The fastest high frequency trader was able to make this $1.00 profit because his limit orders reached the limit order book faster than other traders and gave him time priority on both orders. Due to his speed disadvantage, the institutional investor missed out on the trade at $40.00 and paid $40.01 instead, incurring $1.00 in additional transactions costs.

Next, new information which reduces the value of the asset suddenly hits the market in such a way that all traders observe this information approximately simultaneously. The new information reduces the value of the asset by about $0.20, from about $40.00 to about $39.80. At approximately the same time, the fastest high frequency trader sends a message to cancel his limit buy order at $39.99 and submits a limit sell order to hit the institutional investor's bid at $40.00; the institutional investor sends a message to cancel his limit buy order at $40.00; and other slower high frequency traders submit limit sell orders at prices of $40.00 or lower. Because his technology is slower, the institutional investor's order is hit by the fastest high frequency trader before he can cancel it. The fastest high frequency trader is successful in canceling his limit buy order at $39.99 because he is faster than the other high frequency traders. As a result of his speed, the fastest high frequency trader makes $0.20 on 100 shares, or $20.00, by hitting the institu-

tional investor's bid. The institutional investor loses the same amount as an opportunity cost.

This example illustrates how high frequency traders can use their faster speed to (1) pick off other traders' orders, (2) cancel their own orders, and (3) gain time priority in a standard limit order book. These three advantages are complimentary. Slower high frequency traders will be hesitant to compete for time priority with the fastest high frequency traders because they know their orders are more likely to be picked off if market conditions change. Consistent with this analysis, Yao and Ye (2015) find that small stocks with relatively large tick size attract high frequency traders more than large stocks with relatively small tick size.

Some exchanges use "pro rata" allocation rather than time priority to allocate trades when the number of buyers does not match the number of sellers at a given price. If pro rata allocation had been used in the above example, then both the fastest high frequency trader and the institutional investor would have purchased 50 shares rather than the fastest high frequency trader having purchased 100 shares under a time priority rule. Pro rata allocation gives traders incentives to place larger orders. Thus, the fastest high frequency trader might have placed an order to buy 900 shares rather than 100 shares, in which case he would have been allocated 90 shares and the institutional investor only 10 shares. Putting 900 shares instead of 100 shares into the limit order book multiplies by a factor of 9 the losses suffered from being picked off. Taking this into account, the institutional investor might choose not to place an order for 900 shares himself because he knows his order has a higher probability of being picked off due to his slower speed. Like time priority, a pro rate allocation rule also rewards the fastest traders for their speed and deters slower traders from competing with them.

Example of Continuous Limit Order Book. An exchange offering continuous scaled limit orders would dramatically lower the rents that high frequency traders can earn due to their speed. To illustrate how this works, consider another hypothetical example. An institutional buyer and an institutional seller both 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 $U_{\max}^{BUY} = 3600$ shares per hour (one share per

second). The seller places an order to sell $Q_{\max}^{SELL} = 10\,000$ shares between $P_L^{SELL} =$ \$40.00 and $P_H^{SELL} =$ \$40.01 per share at maximum rate $U_{\max}^{SELL} = 3600$ shares per hour (also one share per second). As long as the buyer and the seller are the only traders in the market, the equilibrium price is the midpoint \$40.0050, and the buyer and seller trade with each other at a rate of 1800 shares per hour (1/2 share per second).

Next, 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 $U_{\max}^{HFT} = 7200$ shares per hour (2 shares per second). Since

$$D(P_B) = 10\,800, \qquad D(P_A) = 0, \qquad S(P_B) = 0, \qquad S(P_A) = 3600, \tag{12}$$

we obtain

$$\omega = \frac{D(P_B) - S(P_B)}{D(P_B) - S(P_B) + S(P_A) - D(P_A)} = \frac{3}{4}, \qquad p(t) = (1-\omega)P_B + \omega P_A = 40.0075. \tag{13}$$

The price increases by one-quarter-cent from \$40.0050 to \$40.0075. 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 at \$40.0075. As a result of his participation, the high frequency trader drives the price above the midpoint, but does not change the sum of the buyer's rate of buying and the seller's rate of selling, which is 1 share per second.[11]

Continuous scaled limit orders dramatically limit the gains from picking off and canceling orders. As before, suppose market conditions change and all traders suddenly receive information that the value of the asset has fallen by 20 cents. The high frequency trader cancels his buy order and places an order to sell at a very high rate between \$40.00 and \$40.01. This drives the price down arbitrarily close to \$40.00 but does not let the high frequency trader sell at a rate faster than one share per second. The institutional investor also cancels his order, but the high frequency trader's sell order arrives faster than the in-

[11] In this hypothetical example, the sum of the buyer's rate of buying and seller's rate of selling is unchanged because the buyer and seller both have the same value of their trading rates $U_{\max}^{BUY} = 1 = U_{\max}^{SELL} =$ 1 share per second. With different assumed values, the combined rate would be different.

stitutional investor's cancelation. The high frequency trader earns a profit proportional to the difference in message processing speeds. If the institutional investor is one second slower than the high frequency trader, then the high frequency trader makes a profit of $0.20 per share on one share, or $0.20; these profits are 100 times less than the profits on the corresponding example with standard limit orders. More likely, the institutional investor can cancel the order only a fraction of a second more slowly than the high frequency trader's order arrives. If the institutional investor is slower by 50 milliseconds, his total losses are $0.01, or 2000 times less than with standard limit orders.

Economic Efficiency and Winner-Take-All. Standard limit order books create a winner-take-all race in which gains go to the fastest trader. As improved message processing technology increases the speed of all traders, the winner-take-all payoffs do not change if traders do not otherwise change how they trade. The winner-take-all system encourages technological investment in speed but does not improve the efficiency of the economy (Harris, 2013; Li, 2014; Biais, Foucault and Moinas, 2015; Budish, Cramton and Shim, 2015).

As technology improves, traders may shred orders into tinier and tinier pieces until the limits of minimum tradable lot size are reached; they may also revise their orders more and more frequently, subject to constraints imposed by minimum tick size. If minimum lot size and minimum tick size are decreased to accommodate the demand for smaller orders and smaller price increments, the number of messages will explode as message costs fall. The result will be that falling message costs do not create economic efficiency by reducing the aggregate economic costs of a given number of messages. Instead, the result will be continued large dollar investments in message technology to increase speed and accommodate an ever-increasing number of messages.

With continuous scaled limit orders, the winner-take-all rewards to being the fastest trader are almost eliminated. If the speed of all traders doubles, the gains from picking off and canceling continuous scaled limit orders are halved. As a result of improvements in technology since the beginning of the century, the potential rents earned from using faster speed to compete with continuous scaled limit orders have already become negligible, and they will become more negligible in the future. The negligibility of these

gains is due to the fact that continuous scaled limit orders allow traders to trade gradually without incurring high message costs.

With continuous scaled limit orders, high frequency traders are unlikely to disappear. Cross-market arbitrage opportunities will still exist, and high frequency traders will likely compete to make such arbitrage opportunities disappear quickly, holding inventories for short periods of time. Continuous scaled limit orders improve the efficiency with which these socially useful arbitrage services are performed by reducing message costs and 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.

3 Theoretical Equilibrium Trading Strategies

Recent developments in theoretical modeling of financial markets support continuous scaled limit orders. Before the 1970s, finance and economics researchers generally believed that the stock market was the best example of perfect competition operating in the economy. The concept of perfect competition requires that each trader takes the price as given, acting as if large quantities could be bought at the same price as large quantities could be sold. If all traders were perfectly competitive, there would be no demand for the use of continuous scaled limit orders. Instead, all traders would want to trade immediately the quantity which would move their inventories to optimal levels.

From a theoretical modeling perspective, it is tricky to explain how trading can move prices in models of perfect competition. The easiest modeling shortcut is to assume that trading does not move prices at all; instead, new information exogenously moves supply and demand schedules up and down, but buying and selling large quantities has no effect on prices. We think that the belief that trading in the stock market reflects the forces of perfect competition—and the companion belief that information and not trading moves prices—is an empty ideology.

Taking the view that the market is perfectly competitive, Grossman and Miller (1988) argue that market liquidity is determined by the supply and demand for immediacy. Customers exogenously demand immediacy: they are willing to pay whatever price the

market makers charge for immediate execution of their desired quantity. This price reflects a discount or premium to compensate market makers for the risk of bearing positive or negative inventories. If the market were to reopen for another round of trading, there would be no trade. In this sense, a competitive model like this generates no demand for continuous scaled limit orders, which spread trading out over multiple rounds of trading.

As Black (1995) discussed, many theoretical models such as Grossman and Miller (1988) impose unrealistic restrictions on trading strategies. Grossman and Stiglitz (1980) assume one round of trading. Kyle (1985) prevents noise traders from timing their trades, prevents market makers from trading on information, and prevents informed traders from submitting limit orders. Glosten and Milgrom (1985) and Grossman and Miller (1988) require customers to trade through dealers without posting their own limit orders. Admati and Pfleiderer (1991) only allow uninformed traders to submit sunshine orders. Glosten (1994) does not allow customers to switch from market orders to limit orders.

Traders themselves have always known that trading costs are economically significant. Historically, high trading costs were the automatic result of monopolistically fixed commissions—which were not deregulated until the 1970s—and privileges of exchange memberships, which gave specialists and other members access to better trading opportunities than nonmembers. More importantly, large traders have long realized that execution of very large trades moves prices and results in market impact costs much larger than commissions and bid-ask spread costs. Sophisticated asset managers invest significant resources measuring market impact costs and calculating the effects of transactions costs on optimal trading strategies.

In models of imperfect competition, traders are allowed to trade strategically, taking into account their price impact. When executable (market) orders or non-executable (limit) orders are allowed with one round of trading, Kyle (1989) shows that imperfect competition induces traders to restrict the quantity they trade to a fraction of the competitive demand, just like a product supplier with market power restricts output. Kyle and Lee (2017) show that the market may stay imperfectly competitive and traders continue to restrict their quantity even as infinitely many traders compete with one another.

This shading of quantities in one round of trading suggests that if trade were to take place over multiple rounds, traders would choose to spread their trading out over time.

In the dynamic model of Kyle (1985), an informed trader is restricted to submitting market orders and not allowed to submit limit orders. He exercises market power by trading gradually over time; if he were a perfect competitor, he would trade huge quantities immediately.

Our proposal for continuous scaled limit orders is theoretically strongly grounded in recent dynamic trading models. Kyle, Obizhaeva and Wang (2017) show that traders use limit orders to trade gradually to limit price impact in a continuous-time model. Since inventories are shown to be differentiable functions of time, the results map directly into the trading speed $U(t)$ associated with continuous scaled limit orders. The similar models of Vayanos (1999) and Du and Zhu (2017) have a discrete-time setting which allows traders to trade at evenly spaced time intervals. This discrete approach makes it possible to study how the frequency with which auctions are held affects the equilibrium. As the time between rounds of trading goes to zero, the discrete-time model converges to the corresponding continuous-time model.

The models place essentially no restrictions on how traders are allowed to trade. There is no exogenous demand for immediacy. Traders are not required to trade through dealers. Since trading is anonymous, traders cannot reveal their private information or commit to being uninformed. Instead, they can attempt to bluff, front-run, or otherwise strategically throw their weight around in the market to affect the expectations and the trading of others. All traders trade rationally given their beliefs. They apply Bayes Law correctly, take into account their price impact correctly, and trade optimally to limit price impact costs.

In these theoretical models, there exists an equilibrium in which traders trade gradually using limit-order strategies very similar to our proposed continuous scaled limit orders. Traders face temporary and permanent price impact costs. Each trader correctly conjectures that the price is a function of his own inventory (permanent impact) and the rate at which he is buying (temporary impact). Because traders have private information, the price moves against the trader due to adverse selection. This means that buying pushes prices up and selling pushes prices down. Just like textbook supply and demand

schedules, which are flows over time, the extent to which the price moves against the trader increases in the rate at which a trader buys or sells; buying a given quantity over a shorter period of time requires paying a higher price because more urgency signals stronger private information. To reduce price impact due to adverse selection, traders smooth their trading over time with optimal trading strategies that map closely into continuous scaled limit orders.

Because of linearity, optimal trading strategies can be implemented with continuous scaled limit orders that are exactly linear, not piecewise linear. For each trader, there is a particular price such that the trader wants to buy at lower prices and sell at higher prices. Therefore, each trader can implement an optimal trading strategy by submitting one buy order and one sell order. The upper limit on the buy order and lower limit on the sell order are the prices at which the trader is content neither to be a buyer nor a seller. The lower limit on the buy order and upper limit on the sell order are arbitrary very low and very high prices, respectively, that are unlikely to be reached during a trading day. The slopes of traders' demand and supply schedules are theoretically constant over time; this implies that the slope $U_{\max}/(P_H - P_L)$ does not have to be updated. The limit price at which each trader chooses not to trade changes over time as traders' inventories change and new private information is received. As a practical matter, we believe that the theory suggests that traders might implement nearly optimal trading strategies by updating the price limits on their continuous scaled limit buy and sell orders at frequencies like once per hour or once per day, not once per second or subsecond.

To summarize, in contrast to Grossman and Miller (1988), models of imperfect competition show that traders who take into account their temporary price impact would never choose to demand immediacy even though delayed order execution is costly. This is because the equilibrium cost of very fast execution of orders is very high because traders providing immediacy to others rationally infer from a desire for immediacy that the trader demanding immediacy has valuable private information. This makes immediacy so expensive that, given a choice between paying a high price to trade quickly and a low price to trade slowly, traders choose to trade somewhat slowly.

If an exchange were to offer both continuous scaled limit orders, these recent theoretical models imply that, exactly consistent with Fischer Black's conjectures, the in-

stantaneous liquidity in standard limit order books would dry up, and almost all traders would use continuous scaled limit orders to access continuous liquidity over time.

4 Related Proposals

Frequent batch auctions, random time delays, and participating orders are different alternative proposals for limiting the gains associated with being the fastest high frequency traders in a market with standard limit orders. Our proposal is compatible with these proposals and also with quantity and price speed bumps.

Frequent Batch Auctions. Budish, Cramton and Shim (2015) propose frequent batch auctions as a mechanism for limiting the rewards from being the fastest trader in a standard limit order book. The idea is for exchanges to process messages at fixed time intervals, such as every second or every 100 milliseconds. The batching interval is large relative to the speed differences between traders, which may be only a few milliseconds. When messages are processed as a batch, the order of arrival of messages processed in the same batch does not matter. As a result, the first orders received do not obtain any advantages of time priority over orders received later in the same batch.

Our proposal for continuous scaled limit orders is fully compatible with frequent batch auctions. Suppose, as we have assumed, that the clock for calculating quantities executed with continuous scaled limit orders ticks once every millisecond. Suppose two continuous scaled limit orders are received during a sub-millisecond time interval during which the clock does not tick. Since the clock does not tick, no trading occurs during this time interval at any rate. The order in which the messages are processed therefore does not matter because there is no time priority in the limit order book. Thus, our proposal for continuous scaled limit orders will treat orders arriving during the same millisecond time interval in the same way regardless of the order in which they arrive. In this respect, our proposal for continuous scaled limit orders is compatible with frequent batch auctions with a batching interval of one millisecond. This batching interval is a parameter which could be changed to be 100 milliseconds or one second to match the interval proposed by Budish, Cramton and Shim (2015). This is the sense in which

continuous scaled limit orders are compatible with frequent batch auctions.

Since frequent batch auctions contemplate a standard limit order book, frequent batch auctions do not eliminate the need for an allocation rule like time priority or pro rata order matching. Since the market clears in increments of one tick, there will typically be excess supply or demand at the market clearing price. Frequent batch auctions require an allocation rule for determining which executable orders do not get executed at the market clearing price. One possibility is to give time priority to orders received in earlier batches while using pro rata matching for orders received in the same batch. Another possibility is to use a pro rata rule for all orders. Regardless of whether orders are processed in batches or sequentially, a pro rata allocation rule for all orders encourages traders to submit orders for larger quantities than they expect to execute. A pro rata rule favors faster traders because frequent batch auctions only lessen, but do not eliminate, the gains from speed.

Budish, Cramton and Shim (2015) examine the rewards for speed using the simplistic finance theory which assumes that prices are moved by information, not trading. Suppose that the batching interval is 100 milliseconds and the fastest traders are 5 milliseconds faster than the second fastest traders. Now suppose new information arrives at some random time, and all traders try to act on it by submitting orders with revised prices as fast as possible. Then the fastest orders will arrive in an earlier batch than the second fastest in 5 percent of the batch auctions. With a speed difference of 5 milliseconds and a batching interval of 100 milliseconds, the gains from being the fastest trader are reduced by 95 percent.

This logic implies that the gains from speed could be reduced further by making the batching interval longer. Suppose the batching interval is increased from 100 milliseconds to 1 second. Then this logic implies that the gains from being faster fall by 99.5 percent, not 95 percent.

The logic changes when the underlying economic model to justify it is changed from the simplistic model that quantities do not move prices to a dynamic equilibrium model in the spirit of Vayanos (1999) or Du and Zhu (2017). These more realistic models imply that sophisticated traders will spread their participation out over time so that they purchase desired quantities more or less continuously over the course of a day. If the

batching interval increases by a factor of ten from 100 milliseconds to 1 second, then sophisticated traders will increase the size of their orders by a factor of ten. Even though, with the longer batching interval, attempts to pick off orders are successful only 0.5 percent of the time rather than 5.0 percent of the time, the quantity picked off will increase by a factor of ten. Thus, the gains are unaffected by making the batching interval longer because the size of sophisticated orders is proportional to the batching interval.

Our proposal for continuous scaled limit orders is based on the idea that message costs induce traders using a standard limit order book to submit orders that are suboptimally large. If this is the case, then increasing the batching interval to make optimal order size larger helps the traders submitting suboptimally large orders. Consistent with our view, Lee et al. (2004) and Barber et al. (2009) find that individual traders in the Taiwanese stock market lose more than two percent of Taiwan's GDP trading stock. From 1995 to 1999, the Taiwan Stock Exchange operated with batch auctions held every 90 seconds. They show that while large institutions smooth out their trading by participating in numerous auctions, individual traders place less frequent orders and suffer economically significant losses. This evidence suggests it would require a batching interval of more than 90 seconds to reduce the losses of retail investors with high message costs.

With continuous scaled limit orders, a trader with high message costs can protect his orders against being picked off much more effectively than with frequent batch auctions in a standard limit order book. Continuous scaled limit orders protect traders with high message costs by building order shredding into the order execution automatically.

Random Time Delays. Harris (2013) proposes that random time delays be added to the processing of messages to place, cancel, and modify limit orders. For example, the exchange might add a time lag with a uniform distribution over some fraction of a second which is long relative to the differences in speeds of the fastest traders and other fast traders. This has the effect of shuffling the order in which messages are received so that the message which was sent first has about a fifty percent probability of being processed before a message sent a small fraction of a second later. In effect, this levels almost completely the playing field on which the fastest traders and slightly slower

traders compete. While aimed at standard limit order books, random time delays could also be used with continuous scaled limit orders, in which case the already tiny gains from being faster than other traders would become even tinier.

Participating Orders. Black (1995) proposed "average-price market orders" and "indexed limit orders" as modifications to standard limit orders that implement a similar approach to market design. Both of these order types are similar to VWAP and TWAP strategies, which traders use to trade at the same average price as other traders over time in the context of a standard limit order book. Black's proposal is similar to ours in that it attempts to promote gradual trading by incorporating automatic adjustments to orders into the order itself rather than require traders to cancel and replace orders frequently. Our proposal is more radical than Black's because it changes in a fundamental way the manner in which exchanges clear markets. It represents a new continuous market design for securities exchanges, not just a new order type.

Price Speed Bumps. While continuous exchanges allow traders to trade patiently without incurring large message costs, some traders may nevertheless choose to trade with extreme urgency, either intentionally or by mistake. To prevent such orders from creating price disruptions like "flash crashes," continuous exchanges can adopt straightforward rules such as the price and quantity speed bumps.

A price speed bump begins when the price changes quickly over a short period of time, for example, by more than $m+5$ cents over m seconds. 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. If the market clears at a price of \$39.80 per share, it will take 15 seconds of no trading before trading takes place at \$39.80 per share. 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 no longer generates excess supply, the new market clearing price will be the perhaps-slightly-higher price that clears the market for the entire duration of the speed bump. Thus, a trader whose sell order at $39.80 was resting in the market for all 15 seconds will have 15 seconds of cumulative volume executed at $39.80 when trading restarts after a delay of 15 seconds. This particular structure for a speed bump not only protects naive traders but also 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 order cancelation itself is likely to end the speed bump and execute all of his disruptive trades at the worst possible price for him. This mechanism for implementing price speed bumps discourages intentionally disruptive trading.

Quantity Speed Bumps. Bilateral bargaining in a dealer market allows dealers to offer different terms to different customers and to exclude customers from trades they would like to participate in. For example, a scrupulous dealer might charge a lower bid-ask spread to a customer perceived to have little private information, and an unscrupulous dealer might take advantage of a poorly informed customer by offering bad prices that other customers would be willing to improve upon.

Organized exchanges have a natural structure which facilitates all traders having access to the same trading opportunities, with the trades going to the traders willing to offer the best prices. On organized exchanges, the spirit of this principle is violated when a buyer and a seller negotiate a gigantic block in a dealer setting, place prenegotiated matching buy and sell orders at almost exactly the same time, and cross the block so fast that other traders do not have enough time to improve one side of the market or the other by participating in one side of the trade or the other.

A quantity speed bump is a mechanism for slowing down trade on an organized exchange so that all traders have enough time to participate in large blocks. A quantity speed bump implies that orders with large urgency, for example, executing one day's trading volume in less than one minute, should be exposed to the entire market for a sufficiently long time so that traders with moderately slow technology can participate in almost all of the transaction. This, in effect, prevents traders from supplying instan-

taneous liquidity in a targeted manner to specific counterparties by excluding other traders; it allows all traders to participate most of the time that the order is actively being executed in the market.[12] For example, exchanges might have rules which slow down fast trading so that executing one day's "normal" volume takes at least five minutes.

5 Conclusion.

Offering continuous scaled limit orders on securities exchanges would make it possible to implement Fischer Black's vision of continuous electronic markets without requiring traders to place numerous limit orders. Continuous scaled limit orders dramatically reduce the profits that high frequency traders make by using their technology to exploit the discreteness in today's markets. They further protect the rest of the market from being picked off by encouraging high frequency traders to compete among themselves. Continuous scaled limit orders address the inefficient arms race among high frequency traders by almost eliminating the profitability of being slightly faster than other traders. Our proposal is different from other policy proposals such as frequent batch auctions proposed by Budish, Cramton and Shim (2015) and random message processing delays proposed by Harris (2013) in that we directly address the source of the underlying problem, the perverse incentives created by limit order discreteness in price, quantity, and time.

In today's markets, various exchanges operate simultaneously and compete for trading volume. Much of this fragmentation results from standard limit orders books with large minimum tick size (Chao, Yao and Ye, 2017). Different exchanges use different fee structures, such as maker-taker or taker-maker pricing, to mimic trading with fractional tick size. Fragmentation also allows traders to get around time priority by routing orders to different exchanges. We believe that if exchanges offered continuous scaled limit orders, they would attract trading volume and these incentives for fragmentation would disappear. If a continuous exchange operates along with a standard exchange, traders

[12] 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.

with low and moderate technology would trade in a continuous exchange to protect themselves from being picked off without incurring large message costs. This would allow trading to consolidate on a smaller number of exchanges. We speculate that there might be one dominant exchange offering continuous scaled limit orders, facing price competition from a handful of smaller competing exchanges. Studying whether an exchange offering continuous scaled limit orders would dominate various other market designs is an interesting topic for future study.

References

Admati, Anat R., and Paul Pfleiderer. 1991. "Sunshine Trading and Financial Market Equilibrium." *Review of Financial Studies*, 4(3): 443–481.

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. 2012. "The Trading Profits of High Frequency Traders." Available at `http://www.bankofcanada.ca/wp-content/uploads/2012/11/Brogaard-Jonathan.pdf`.

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.

Black, Fischer. 1995. "Equilibrium Exchanges." *Financial Analysts Journal*, 51(3): 23–29.

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.

Chao, Yong, Chen Yao, and Mao Ye. 2017. "Discrete Pricing and Market Fragmentation: A Tale of Two-Sided Markets." *American Economic Review*, 107(5): 196–199.

Du, Songzi, and Haoxiang Zhu. 2017. "What Is the Optimal Trading Frequency in Financial Markets?" *Review of Economic Studies*, forthcoming: Available at `https://doi.org/10.1093/restud/rdx006`.

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

Glosten, Lawrence R., and Paul R. Milgrom. 1985. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics*, 14: 71–100.

Grossman, Sanford J., and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review*, 70(3): 393–408.

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. 2017. "The Flash Crash: High-Frequency Trading in an Electronic Market." *Journal of Finance*, 72(3): 967–998.

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., Anna A. Obizhaeva, and Tugkan Tuzun. 2016. "Microstructure Invariance in US Stock Market Trades." FEDS Working Paper No. 2016-034. Available at SSRN: https://ssrn.com/abstract=2774039 or http://dx.doi.org/10.17016/FEDS.2016.034.

Kyle, Albert S., Anna A. Obizhaeva, and Yajun Wang. 2017. "Smooth Trading with Overconfidence and Market Power." *Review of Economic Studies,* forthcoming: Available at https://doi.org/10.1093/restud/rdx017.

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.

O'Hara, Maureen, Chen Yao, and Mao Ye. 2014. "What's Not There: Odd Lots and Market Data." *Journal of Finance,* 69(5): 2199–2236.

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

Yao, Chen, and Mao Ye. 2015. "Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls." available at https://ssrn.com/abstract=2478216.