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Since the late 1990s, human beings have been replaced by computerized limit order books.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.