Toward a Fully Continuous Exchange: Stock Market Design Based on Flow Trading

Albert S. “Pete”Kyle
University of Maryland

Seminar
Swedish House of Finance
Stockholm, Sweden

September 19, 2019

Trends in Equity Trading

Trends in Equity Trading

- Technology has made placing many small orders cheap.
- Competing exchanges have replaced 1990s upstairs market.

Implications of Trends

- Order shredding replaces block trades.
- High frequency trading algorithms replace human market makers.

Proposal in this Paper: Clear markets with flow orders

- Implements Fischer Black's vision of continuous trading.
- Eliminates high-frequency arms race.

Trends in Trading Driven by Regulation

- Block trading in upstairs market of 1990s.
- SEC–DOJ order handling rules change Nasdaq
- Tick size reduced from 1/8 to 1/16 to one cent.
- Regulation NMS (and MiFiD in Europe) fragment markets.
- Result is competing exchanges with electronic order books.

Trends Driven by Technology

Implications of faster computers and rapid communications technology with high bandwidth:

- Costs of placing and canceling orders is very low: implies many orders and cancelations.
- Small trades can be cleared at very low cost: implies small trade size.
- Rapid arbitrage across exchanges: Forces exchanges to compete in fees (also maker-taker pricing).
- Quantitative strategies need smart electronic systems for order handling to improve order execution quality.

Finance Theory Implies Smooth Trading

Albert S. Kyle, Anna A. Obizhaeva and Yajun Wang, "Smooth Trading with Overconfidence and Market Power," *Review of Economic Studies, Vol. 85*, 2018, pp. 611–652.

Apply game theory to continuous double-auction for "flows" of assets with

- Imperfect competition
- Overconfidence ("agreement to disagree")
- Symmetry

Continuous new Gaussian private information

Results

- For each trader, price depends linearly on other traders' information, own inventory (permanent price impact), time derivative of own inventory (temporary price impact).
- Trader's rate of buying (time derivative of inventory) is linear function of public information ("dividends"), inventory, private information, and market price (which reveals other traders' information).
- Trader smooths trading out gradually over time, trading off decay of information against permanent and temporary market impact.

Equilibrium Model with Optimized Trading

- Traders rationally take into account price impact.
- Traders understand how price impact now affects trading opportunities in the future.
- Traders are allowed to bluff, front run, spoof, etc.—but choose not to do so in equilibrium.
- Suboptimal fast selling leads to a “flash crash”—should not occur in equilibrium but might occur as an out-of-equilibrium mistake.

Stock Market Trading as a Game

The game-theoretic model solution captures the way institutional investors think and trade. Investors ...

- Collect random information about fundamentals continuously.
- Process the raw information statistically, turning it into signals.
- Use the signals to predict fundamental value and future returns rationally (except for overconfidence).
- Calculate a constantly changing optimal portfolio based on the changing signals.
- Trade gradually toward the optimal portfolio (target inventory), optimally taking into account market impact costs and signal decay.

Our Proposal: Trading Stocks as Flows

Albert S. Kyle and Jeongmin Lee, "Toward a Fully Continuous Exchange," *Oxford Review of Economic Policy, Vol. 33*, No. 4, 2017, pp. 650–675.

- Implement a market design consistent with equilibrium theory of speculative trading.
- Make equity trading continuous in price, quantity, and time with flow demand and supply curves.
- Virtually eliminate incentives for "arms race" among high frequency traders.
- Compatible with frequent batch auctions (Budish, Cramton, Shim, 2015)

High Frequency Trading and Market Design

Potential benefits and costs of HFT

- Benefits: Provision of liquidity to the other traders
- Costs: Expenditures on inefficient arms race to transfer wealth by
  - Picking off slow traders' stale limit orders
  - Obtaining time priority in limit order book

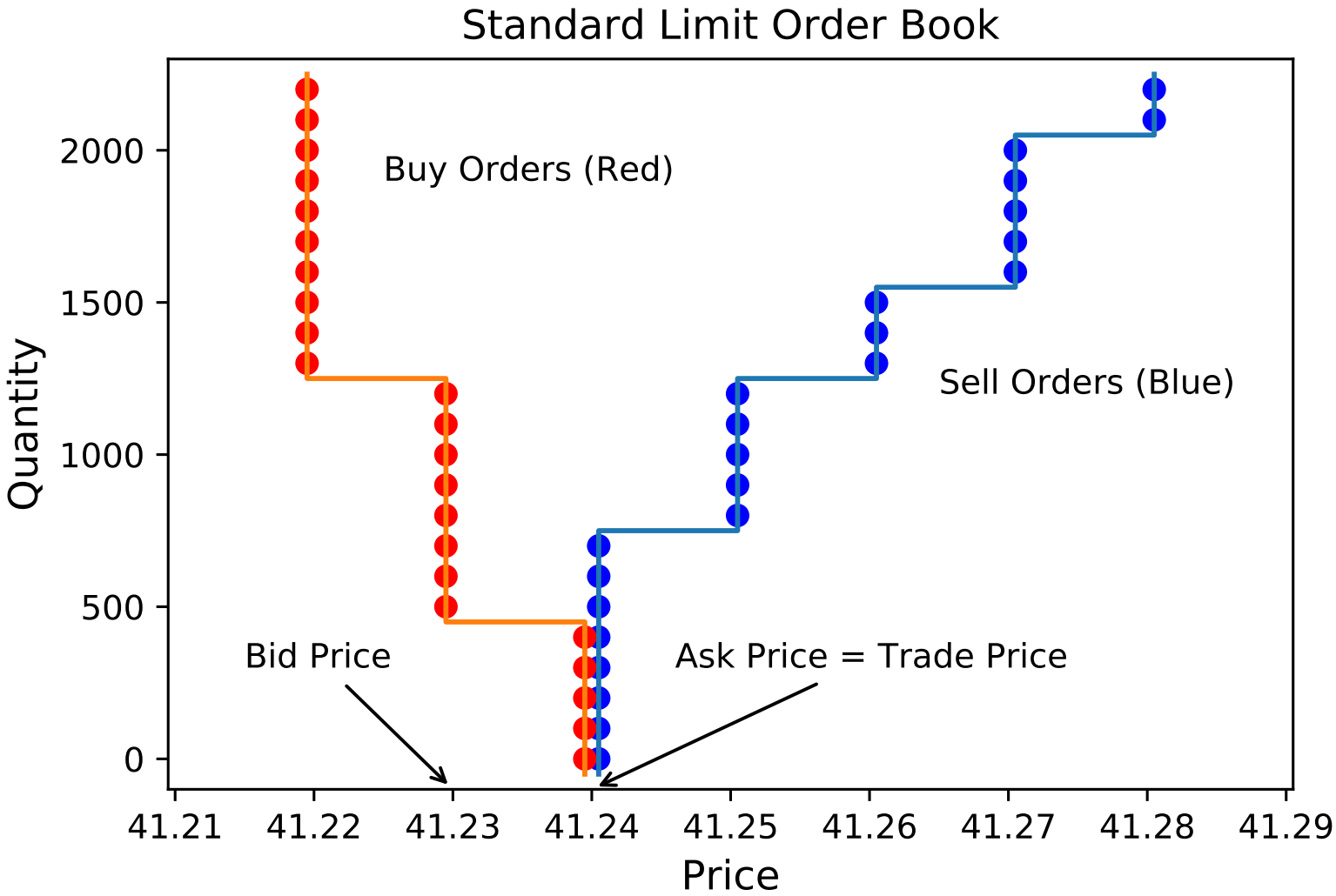
We propose a new market design called a "**fully continuous exchange**" to level the playing field for all traders

Discreteness in Today's Markets

"Continuous limit order books," which dominate equities trading in the U.S. and Europe, have elements of discreteness in price, quantity, and time

- Price is an integer multiple of a minimum tick size ($0.01)
- Quantity is an integer multiple of minimum lot size (one share or one hundred shares)
- Orders are processed sequentially; latencies prevents anyone from trading continuously in time

Market Clearing (?) in a CLOB



HFTs in Today's Markets

Daily rents that HFTs can earn at the expense of slow traders:

$$\Pi = Q \times F \times M \tag{1}$$

- Q: the size of the trade (in shares) at each instant
- F: the frequency of the opportunity to (1) pick off and run over slow traders and (2) buy at the bid or sell at the offer in one day
- M: the dollar trading profit per share; related to the tick size ($0.01) and time priority

Frequent Batch Auctions

Budish, Cramton, Shim (2015) propose a new market design in which auctions are held at discrete intervals

- All orders arrived within the batching intervals are treated equally: no time priority within the interval
- This lowers F, the trading frequency
- If Q, the size of the trade, and M, the dollar trading profit per share at each instant, remain the same, lowering F reduces the daily rents Π

- Would lowering trade frequency F affect trade size Q or profit margin M? If so, then how?

Frequent Batch Auctions

- The dollar per-share trading profit M is likely unaffected by HFT trade frequency F because
  - Tick size, which limits price competition, is not changed
  - Price change greater than the tick when news arrives
  - Profits from time priority are based on tick size
- But the size of the trade Q likely increases when trade frequency decreases because
  - Traders have fewer auctions per day at which they implement target trading volumes
    - May also depend on message costs and serial correlations of trading motives

Dynamic Models

Dynamic models of Vayanos (1999), Du and Zhu (2017), Kyle, Obizhaeva, Wang (2017) all show traders choose to trade gradually to reduce their price impacts

- Consistent with large institutional traders spreading their large trades into many small pieces
- Du-Zhu show each order becomes larger as trading becomes less frequent

$$F \downarrow \quad \Rightarrow \quad Q \uparrow \tag{2}$$

The effect of FBAs would be (partly) offset by traders submitting larger orders at each batching interval

What if the Tick Size Goes Down?

Eliminating the tick in today's markets is practically infeasible and inefficient

- Traders would try to beat one another by offering price improvements $\epsilon \to 0$
- Flashing quotes and numerous messages

Changing tick size has ambiguous effects

- Lowering tick size makes prices go up or down more often when information changes, so $M \downarrow \quad \Rightarrow \quad F \uparrow$

Raising tick size

- Further limits price competition and makes gaining time priority more valuable

Our Proposal (Step I): Make Make Quantities a Continuous Function of Price

Let traders choose two limit prices, P_L and P_H, which respect the minimum tick size

- Standard limit buy order:

$$Q = \begin{cases} Q_{\max} & \text{if } p \leq P_L \\ 0 & \text{if } p > P_L \end{cases} \tag{3}$$

- **Scaled** limit buy order:

$$Q = \begin{cases} Q_{\max} & \text{if } p \leq P_L \\ \left(\frac{P_H - p}{P_H - P_L}\right) Q_{\max} & \text{if } P_L < p \leq P_H \\ 0 & \text{if } p > P_H \end{cases} \tag{4}$$

Our Proposal (Step I): Make Quantities a Continuous Function of Price

With standard limit orders, aggregate demand and supply schedules are decreasing and increasing step functions

- Market does not clear: excess supply or demand

With scaled limit orders, aggregate supply and demand functions are decreasing and increasing piecewise-linear functions

- Unique intersection that clears the market
- No time priority
- HFTs must compete on the price

What about Liquidity Provision?

Potential benefits and costs of HFT

- Benefits: Provide liquidity to the other traders
- Costs: Pick off slow traders' stale limit orders
  - Inefficient arms race wastes resources

Why must liquidity be provided by HFTs? Why not other traders?

- Submitting limit orders implies all traders provide some liquidity to the others
- HFTs' technology allows them to participate in the market more continuously than slow traders in today's markets
- Faster HFTs may deter liquidity provision by others.

Message Costs

Since trading gradually is an optimal strategy

- Institutional investors use order-shredding strategies like VWAP and TWAP

But in today's markets

- The extent to which traders can shred orders is limited by the minimum lot size
- Implementing such strategies require sending numerous order messages
- HFTs technology lowers their **message costs**
- More costly for slow traders to shred their orders

Our Proposal (Step II): Make Quantities a Continuous Function of Time

Let traders submit a dynamic schedule of limit orders at once

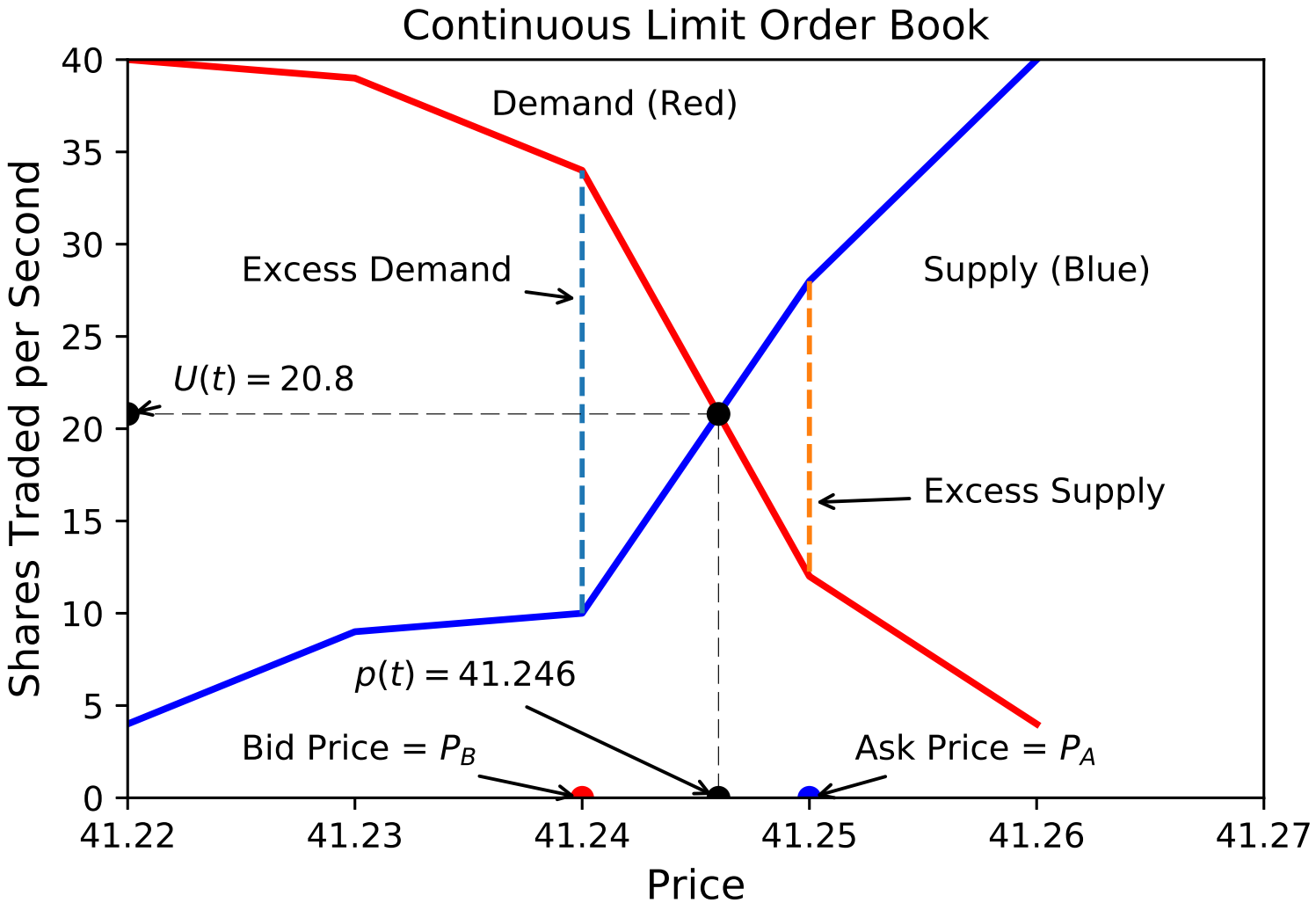
- **Continuous** scaled limit buy order: "Buy up to $Q_{\max}$ shares at maximum rate $U_{\max}$ shares per second at prices between P_L and P_H"

$$U(p) = \frac{dQ}{dt} = \begin{cases} U_{\max} & \text{if } p \leq P_L \\ \left(\frac{P_H - p}{P_H - P_L}\right) U_{\max} & \text{if } P_L < p \leq P_H \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

  where the number of shares bought between t_0 and t until canceled

$$Q(t) = \int_{t_0}^{t} U(p(\tau))\, d\tau \qquad \text{for } t \leq Q^{-1}(Q_{\max}) \tag{6}$$

Market Clearing with CoSLOs



Effects of CoSLOs

- Continuous scaling in price prevents HFTs from being rewarded for racing to achieve time priority
- Continuous scaling in time prevents HFTs from being rewarded for picking off stale limit orders.
  - Slower traders may cancel orders in a few milliseconds, resulting in only tiny fractions of shares being picked off
  - Slower HFTs will move prices in a direction favorable to resting limit orders, so the orders not canceled trade only fractions of shares at unfavorable prices.
- A slow trader can guarantee order execution at TWAP exactly by placing an executable CoSLO
  - Our approach can probably be adapted to guarantee execution at VWAP as well

History: Fischer Black's (1971) Predictions

If trading and market making moved from a human specialist system to an electronic system,

- Bid-ask spreads on small trades would be reduced to a vanishingly small level
- Liquidity would not be supplied cheaply, especially over short periods of time
- Customers would spread large trades out over time to reduce trading costs

He was prescient:

- Large institutional traders and algorithmic traders nowadays spread their trading out over time by breaking large trades into many small pieces

Technology Gap

But, not all of his predictions were correct:

- Bid-ask spreads on small trades did not disappear
- Retail traders still pay large trading cost

Perhaps Fischer Black did not foresee . . .

- The “technology gap” would remain economically significant even with improved technology and competition
- High frequency traders would earn profits by being a few microseconds faster than their competitors even though absolute speeds approached the speed of light

Our Proposal

A “fully continuous exchange” implements Fischer Black (1971)’s vision of an efficient market design:

- Customers would spread large trades out over time to reduce trading costs
- Bid-ask spreads on small trades would vanish
- Liquidity would not be supplied cheaply, especially over short periods of time

The new market design allows traders to choose **two limit prices** and **trade gradually** to level the playing field

Demand for Immediacy (?)

Grossman and Miller (1988) view that market liquidity is determined by the supply and demand for immediacy

- Customers demand and market makers supply immediacy
- Customers are willing to pay whatever price the market makers charge to achieve their desired quantity
- This view has its origin to a competitive REE model like that of Grossman and Stiglitz (1980)

In a fully continuous exchange, liquidity is supplied and demanded over time

Discreteness in the Matching Engine

Internal calculations would require discretizing time, quantity, and price (millisecond, nanoshares, and microdollars)

- Calculating the price follows simple integer vector algebra
- Allocating quantities is straightforward
- Far fewer messages from the exchanges as well as from traders

Discretization in the matching engine is economically different from discreteness in the current market design because the **gains from gaming it** would be negligible

Details on Price Speed Bumps

Price speed bump prevents execution of orders if price would move a large amount in a short period of time.

- Our proposal lets price move, say, 5 cents plus 1 cent per second, with numbers scaled for typical volume in stock
- When sell imbalance occurs, orders accumulate over time without being executed, as price falls at maximum rate
- Traders can place new orders or cancel old orders, back-tracked to time when trading delay began
- Trading delay stops and markets clear as soon as maximum falling price clears market.

This proposal creates good incentives to provide liquidity, punishes bad incentives to sell aggressively. Details of speed bump implementation proposal are still work in progress.

Details on Quantity Speed Bumps

Quantity speed bumps attempt to allow all traders to participate equally in price formation by defeating incentives of traders to use “dealer market” to exclude other traders from trading. Idea: If large block trade is negotiated between two parties, it cannot be “crossed” instantaneously. Instead, order must be executed continuously at a rate slower than a maximum allowed rate

- Maximum rate is function of past volume, say one day’s volume in five minutes
- Trading gradually over time allows all traders in market to participate in price formation
- Proposal prevents targeting better prices at more informed customers

Front-Running

Do CoSLOs make slow traders more vulnerable to front-running?

- Suppose a HFT learns about a slow trader's intended buy CoSLO
- The HFT will have to buy faster than the slow trader and sell back to the slow trader
- CoSLOs make trading quickly more expensive and trading slowly less expensive since all traders can easily trade slowly unless they have special reasons not to
- Front-running would be less profitable

Random Delays

Harris (2013) proposed random delays

- Shuffling the queue of the limit order book
- Creates a perverse incentive for HFTs to submit so many orders to increase the probability of getting time priority

Future Directions for Smooth Trading Research

- Combine smooth trading with market microstructure invariance: Requires thinking of financial markets as "infinitely non-competitive."
- Do continuous scaled limit orders dominate other market structures (with minimum tick size, minimum lot size, numerous messages in limit order book)?
- Smooth trading for multiple assets simultaneously?
- Combine volume-weighted-average price (VWAP) into smooth order type?
- Can "square root puzzle" be explained by slower execution of large orders?

Conclusion

We propose a **fully continuous exchange** as a new market design for organized stock exchanges

- Making quantities continuous in price eliminates race for time priority
- Making quantities differentiable in time dramatically reduces reward from picking off stale limit orders

By converting expensive messages into cheap internal calculations, CoSLOs allow **all traders** to trade (optimally) gradually without costly technology or fear of being picked off