The Changing Nature of Trading Markets: Past, Present, and Future

Albert S. “Pete” Kyle

University of Maryland

University of Maryland–UBS Conference
Asset Management in 2017: Pioneers and New Frontiers
New York, NY
Wednesday, May 24, 2017

Outline

- The Past: Buttonwood agreement (fixed commissions and exclusive dealing), commission deregulation, upstairs markets, Nasdaq odd-eighths scandal, one-cent tick size, electronic order handling, algorithmic trading, regulation NMS.
- The Present: Market fragmentation, order shredding, high frequency trading, trend toward agency trading.
- The Future: Continuous trading as envisioned by Fischer Black (1971)?

Organized Exchanges with Fixed Commissions

Buttonwood Agreement establish NYSE inn late 1700s) with a textbook cartel agreement:

- Charged fixed commissions (price fixing)
- Conducted all trading with member on exchange floor (monitoring).
- Avoided trading with non-members (limited entry)

Fixed Commission Structure: Enforcement

NYSE needed a mechanism to enforce fixed commissions.

How to prevent dealers from offering kickbacks to customers to undermine fixed commission structure?

- All trades required be agency trades, conducted on exchange floor in public (open outcry).
- Required separation of floor brokers from market makers or specialist.
- Exchanges audited brokerage accounts, especially "error" accounts. Requires good data.

Centralized trading protected small traders from bad prices.

How can securities industry collect rents without fixed commissions?

Traditional Organized Exchange

Example: NYSE with specialists in 1980s and 1990s, CME and CBOT with floor traders in 1980s and 1990s

Network externalities and regulatory barriers to entry kept trading on exchange floor.

- Specialist charged high fees to match buyers and sellers.
- Supply of market makers (floor traders) limited by costly exchange memberships.
- Tick size, physical layout of trading floor, rules and practices benefited floor traders at expense of customers.
- Customers imperfectly protected by time and price priority, ticker with bids, offers, or trades.
- Large customers matched trades in “upstairs” dealer market.

Dealer Market

Example: Nasdaq in 1980s and early 1990s. Treasury notes and bonds until recently. Corporate bonds even now.

Decentralized trading between dealers and customers.

- Customers had to trade with dealers, not with one another.
- Dealer resisted price reporting; have informational advantage over customers. Operated "inside" market among themselves.
- Quoted prices sometimes wider than prices for favored customers.

Result: Bid-ask spreads wider for small trades by less sophisticated customers.

Today: De-mutualized Exchanges: CME

"De-mutualization": Exchanges makes profit from transaction fees and other fees. No longer owned by members (brokers, market makers, specialists) but rather by someone else: private equity of public company which.

- Efficient central limit order book allows electronic interfaces, strict enforcement of time and price priority.
- Market data is valuable, especially time value. Co-location.
- If exchange is a monopolist, then has incentive to charge monopoly prices.
- Monopoly power limited by ability of customers to match trades outside the market or on a competing exchange.

Today: Competition in Fragmented Markets

Example: Regulation NMS and MiFiD: US and EU in 2015.

- Computer technolodgy realizes network externalities so cheaply that barriers to entry are low.
- Fragmentation rewards speedy arbitrage by high-frequency traders.
- Fragmentation multiplies amount of market data.

Fragmentation undermines rents for exchanges.

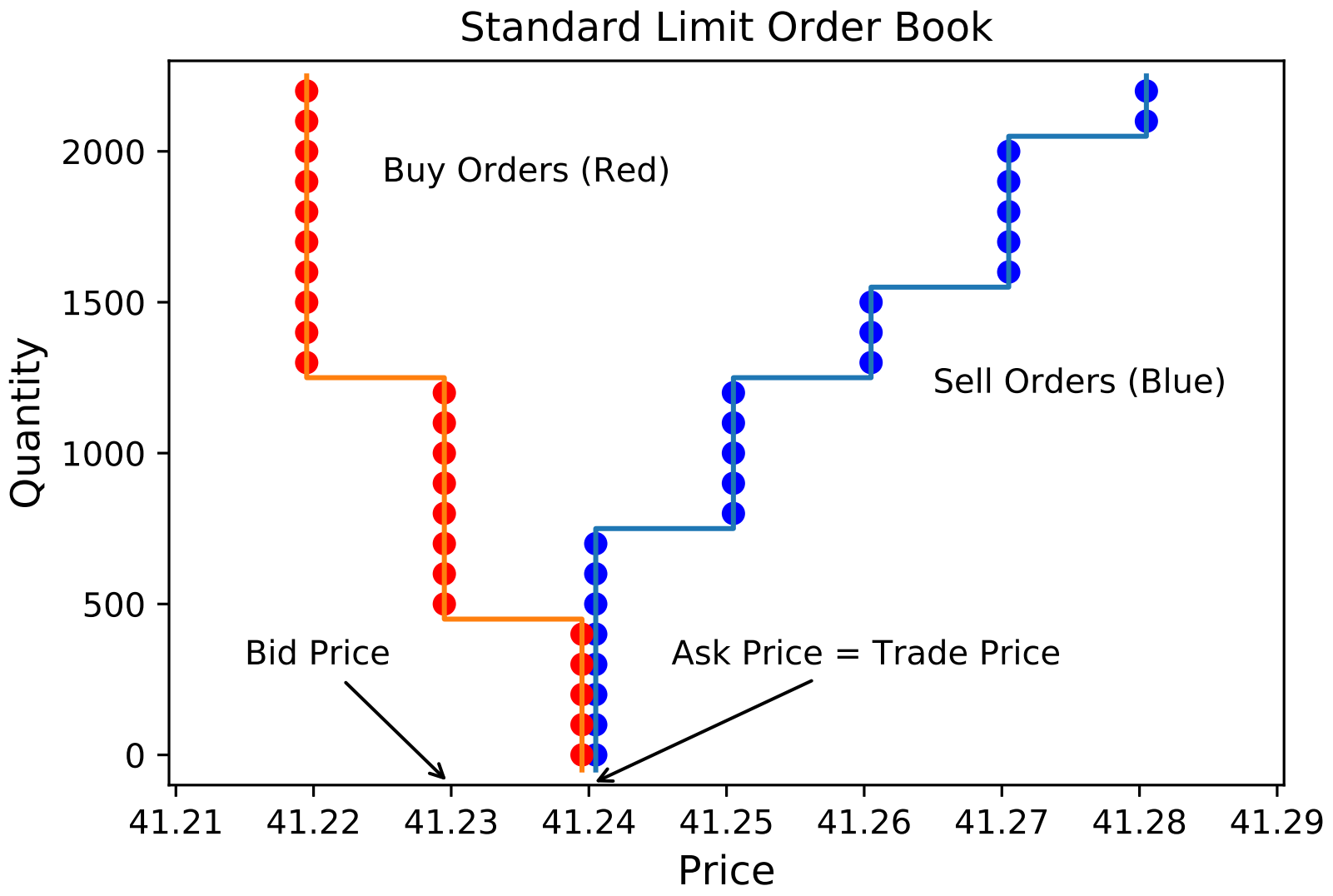
Current Equity Market Design

"Continuous limit order book" is hardly continuous:

- Price is discrete: minimum tick size of one cent.
- Quantity is discrete: minimum lot size of one share or 100 shares.
- Time has lags: Sending, receiving, and processing messages take time, even at the speed of light.

Discreteness amplifies the rents that faster traders can earn:

- Minimum tick size allows high frequency traders to obtain time priority.
- High frequency traders can pick off size in limit order book.


Standard Limit Order Book
Buy Orders (Red)
Sell Orders (Blue)
Bid Price
Ask Price = Trade Price
Quantity
2000
1500
1000
500
0
41.21
41.22
41.23
41.24
41.25
41.26
41.27
41.28
41.29
Price

Our Proposal: Kyle and Lee (2017)

- Implement Fischer Black's vision of an efficient market design by introducing a "continuous scaled limit order".
- Black (1971) predicted that in a "fully automated" exchange:
  - The bid-ask spread for small quantities would be small.
  - There would be ittle depth for immediate execution of large trades.
  - Liquidity would be provided over time.
  - Traders would attempt to reduce trading costs by trading gradually over time.

Continuous Scaled Limit Order

Trader sends one message to trade gradually over time.

- Quantity is a flow instead of stock
- Price is a range of two prices instead of one limit price.

Result is piecewise linear flow supply and demand schedules.

A continuous scaled limit buy order message:
"Buy up to $Q_{\max} = 10\,000$ total shares at prices between $P_L =$ \$41.20 and $P_H =$ \$41.25 at maximum rate $U_{\max} = 3600$ shares per hour (one share per second)."

- Quantities ($Q_{\max}$ and $U_{\max}$) and prices (P_L and P_H) respect minimum lot size and minimum tick size.
- Traders can modify or cancel their order after some minimum resting time.

Algebra: Quantities

- The cumulative quantity executed after time τ

$$Q(\tau) := \int_{t_0}^{t_0+\tau} U(p(t))\, dt, \tag{1}$$

  where $p(t)$ is the market clearing price.

- The trading speed is

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

Algebra: Aggregate Demand and Supply

- Aggregate flow demand schedule $D(p)$ is the sum of the trading speed $U(p)$ of all buy orders.
- Aggregate flow supply schedule $S(p)$ is the sum of the trading speed $U(p)$ of all sell orders.
- $D(p)$ is a downward slopping piecewise linear function and $S(p)$ is a upward slopping piecewise linear function.

Algebra: Best Bid and Best Ask

- P_B is the maximum price with positive excess flow demand

$$X_D = D(P_B) - S(P_B) \geq 0. \tag{3}$$

- P_A is the minimum price with positive excess flow supply

$$X_S = S(P_A) - D(P_A) \geq 0. \tag{4}$$

- If $P_B = P_A$, the market clearing price is

$$p(t) = P_B = P_A. \tag{5}$$

Algebra: Market Clearing

If $P_B \neq P_A$, P_B and P_A are one tick apart and

$$X_D + X_S > 0, \tag{6}$$

- The relative order imbalance is

$$\omega := \frac{X_D}{X_D + X_S} \in [0, 1]\,. \tag{7}$$

- The market clearing price is

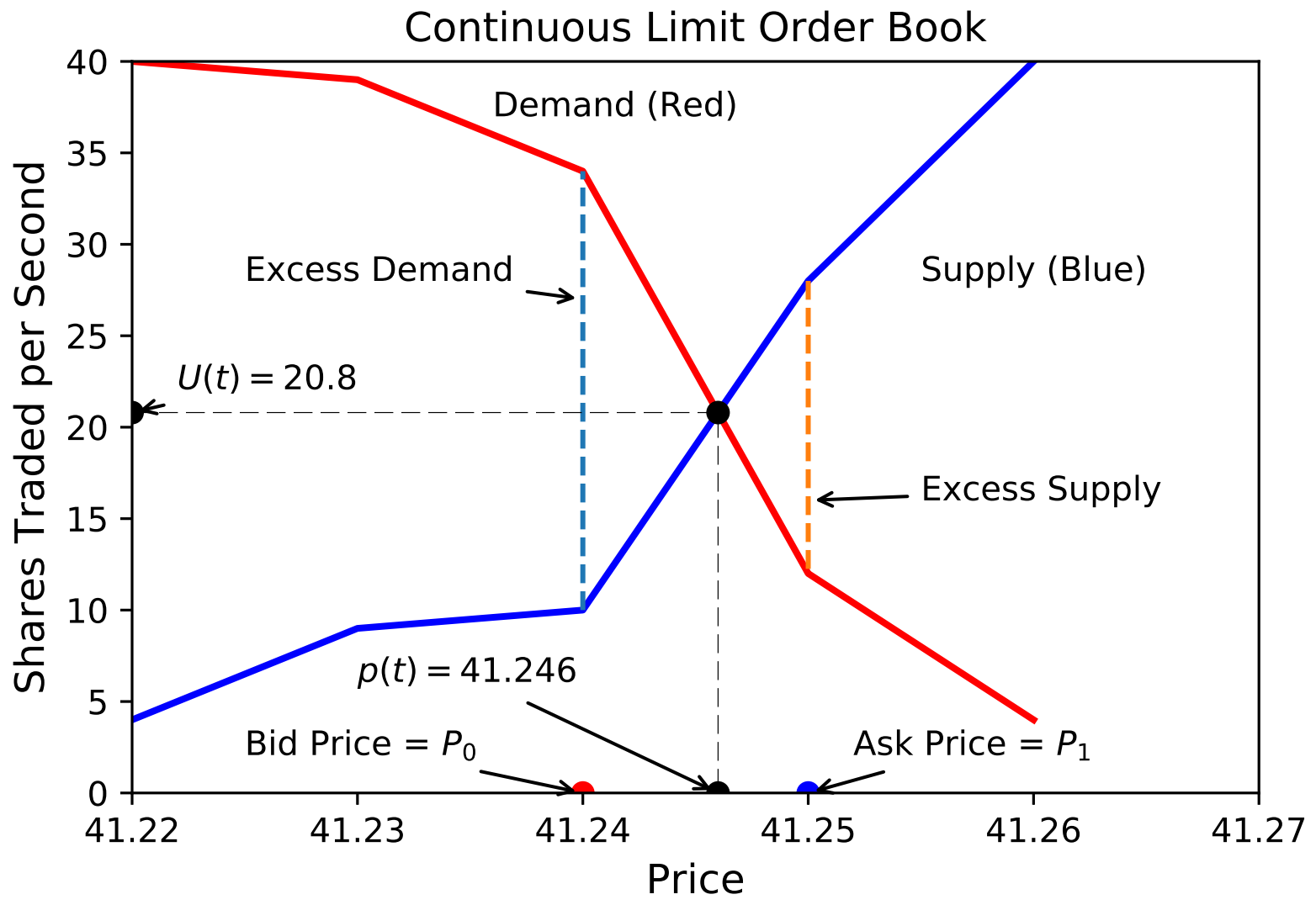
$$p\,(t) = (1 - \omega) P_B + \omega P_A, \tag{8}$$

  which is not an integer multiple of the tick size.

Implementation

The exchange publicly announces $\{P_B, P_A, X_D X_S\}$.

- Everyone can infer $p(t)$.
  - There is neither excess supply nor excess demand.
  - No need for time priority or an allocation rule.
- Traders can infer the execution of their order $U(p(t))$ and $Q(t)$.
  - No need for messages providing continuous updates of cumulative quantity executed.


Continuous Limit Order Book
Demand (Red)
Supply (Blue)
Excess Demand
Excess Supply
$U(t) = 20.8$
$p(t) = 41.246$
Bid Price = P_0
Ask Price = P_1
Shares Traded per Second
Price
0
5
10
15
20
25
30
35
40
41.22
41.23
41.24
41.25
41.26
41.27

Continuous Scaled Limit Orders: Implications

- All traders can trade gradually without incurring costs for placing numerous messages.
- No time priority for faster traders to exploit.
- No size in limit order book for faster traders to pick off.

No instantaneous liquidity.

- Depth offered gradually over time.
- Infinite cost of trading any quantity immediately.

Conserving Message Costs

- Trader submits continuous scaled limit order as one message rather than many small messages
- No need for messages providing continuous updates of cumulative quantity executed
  - Exchanging messages between computers is thousands of times more expensive than performing calculations on the same CPU due to need for encryption, handshake, verification, reconciliation, bandwidth

Matching Engine Efficiency

- Since “kink points” on continuous scaled limit order are multiples of minimum tick size, each order can be represented as a fixed-size vector of quantities
- Requiring minimum continuous lot size (e.g., one share per hour), makes limit order a vector of integers
- Aggregate demand and supply schedules are fixed-length integer vectors
  - easy to calculate by integer vector addition
  - easy to calculate “best bid” and “best ask” prices

Comparison: Frequent Batch Auctions

Budish, Cramton, Shim (2015) propose frequent batch auctions

- Compatibility: A discrete approximation to continuous scaled limit orders can be implemented with frequent batching.
- Difference in motivation: We assume traders want to trade slowly and continuously by chopping orders into into small pieces; this defends against being picked off by high frequency traders.

Empirical Evidence

- Kirilenko, Kyle, Samadi, Tuzun: HFTs had smaller trader size in S&P 500 E-min futures market.
- Lee, Liu, Roll, Subrahmanyam: Taiwan had frequent batch auctions in which
  - Small traders incur high cost participating in few auctions (lose more than two percent of GDP)
  - Large foreign institutions earn significant profits and participate in auctions more continuously

Theoretical Evidence

Albert S. Kyle, Anna A. Obizhaeva, and Yajun Wang, "Smooth Trading with Overconfidence and Market Power," 2017.
Theoretical model: Optimal trading with continuous limit order book. Everyone trades slowly if everyone else trades slowly.

- Symmetric trade continuously to profit from continuous flow of information.
- Traders believe their signals are more precise than other traders believe them to be.
- Each trader submits an "flow-order" according to which the derivative of his inventory is a linear function.
- Traders reduce price impact by trading more slowly.
- Decay of signals places a limit on how slowly traders can afford to trade.

Is CoSLO vulnerable to “Front-Running”?

- In theoretical model, traders believe price is linear function of own inventory (permanent price impact) and time derivative of own inventory (temporary price impact).
- Theoretical model proves that front-running is not profitable.
- Intuition: Price already reveals so much of a trader’s information that other traders want to trade against it based on disagreement, not front run.

Competing Exchanges

If a continuous exchange competes with other exchanges using standard limit orders, we conjecture that the continuous exchange will dominate

- Work in progress related to Glosten (1994): “Is the Electronic Limit Order Book Inevitable?”
- Conjecture: Winning exchange looks like Fischer Black’s electronic market

After Continuous Scaled Limit Orders

Move execution of trading strategies into matching engines with different order types:

- Orders to buy multiple assets as functions of multiple prices: automatic legging of spreads, automated index arbitrage, automated basket trades:
  - All are technologically easier with continuous scaled limit orders.
- Integrate repo market, spot market, and futures market.
- Integrate “real time” clearing, settlement, and custody.
- Complete audit trail, including customer account number of legal entity identifier on all trades.

Summary

- Continuous scaled limit orders level the playing field for traders with high message costs
- Regular traders can lower transaction costs by trading gradually
- Eliminating time priority makes HFTs compete with one another. Speed therefore benefits regular traders instead of harming them.
- When an HFT picks off a stale order, rapid price adjustment resulting from competition benefits regular traders.