The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response

Eric Budish, Peter Cramton and John Shim

Forthcoming, *Quarterly Journal of Economics*

Seminar Slides, Sept 2015

The HFT Arms Race: Example

CHICAGO
NEW YORK

- In 2010, Spread Networks invests $300mm to dig a high-speed fiber optic cable from NYC to Chicago.
- Shaves round-trip data transmission time . . . from 16ms to 13ms.
- Industry observers: 3ms is an "eternity". "Anybody pinging both markets has to be on this line, or they're dead."
- Joke at the time: next innovation will be to dig a tunnel, "avoiding the planet's pesky curvature".
- Joke isn't that funny . . . Spread's cable is already obsolete!
- Not tunnels, but microwaves (first 10ms, then 9ms, now 8ms).
- Analogous races occurring throughout the financial system, sometimes measured as finely as microseconds or nanoseconds

The HFT Arms Race: Market Design Perspective

- We examine the HFT arms race from the perspective of market design.
  - We assume that HFT's are optimizing with respect to market rules as they're presently given
  - But, ask whether these are the right rules
    - Avoids much of the "is HFT good or evil?" that seems to dominate the discussion of HFT
    - Instead, ask at a deeper level what is it about market design that incentivizes arms race behavior, and is this design optimal
- Central point: HFT arms race is a *symptom* of a simple flaw in modern financial market design: continuous-time trading.
- Proposal: discrete-time trading.
  - Replace continuous-time limit order books with *discrete-time frequent batch auctions*: uniform-price double auctions conducted at frequent but discrete time intervals, e.g., every tenth of a second.

Frequent Batch Auctions

A simple idea: discrete-time trading.

1. Direct-feed millisecond level data from exchanges: continuous market violates basic asset pricing principles at HFT time horizons.
   - Market correlations completely break down.
   - Frequent mechanical arbitrage opportunities.
   - Mechanical arbs –> arms race. Arms race does not compete away the arbs, looks like a "constant".
2. Theory model: critique of the continuous limit order book
   - Mechanical arbs are "built in" to the market design. Sniping.
   - Harms liquidity (spreads, depth).
   - Induces a never-ending, socially wasteful, arms race for speed.
3. Frequent Batch Auctions as a market design response
   - Eliminates mechanical arbs and sniping.
   - Competition on speed –> competition on price.
   - Enhances liquidity and stops the arms race.

Frequent Batch Auctions

A simple idea: discrete-time trading.

1. **Direct-feed millisecond level data from exchanges: continuous market violates basic asset pricing principles at HFT time horizons.**
   - **Market correlations completely break down.**
   - **Frequent mechanical arbitrage opportunities.**
   - **Mechanical arbs –> arms race. Arms race does not compete away the arbs, looks like a "constant".**
2. Theory model: critique of the continuous limit order book.
   - Mechanical arbs are "built in" to the market design. Sniping.
   - Harms liquidity (spreads, depth).
   - Induces a never-ending, socially wasteful, arms race for speed.
3. Frequent Batch Auctions as a market design response
   - Eliminates mechanical arbs and sniping.
   - Competition on speed –> competition on price.
   - Enhances liquidity and stops the arms race.

Brief Description of the Continuous Limit Order Book

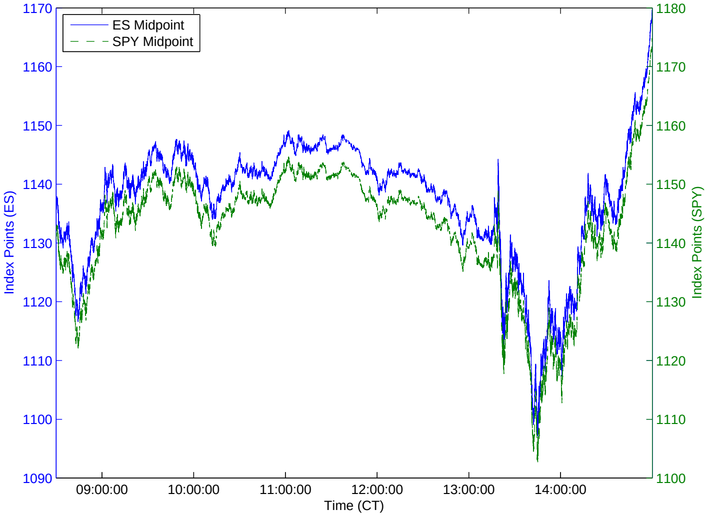
- Basic building block: limit order
  - Specifies a price, quantity, and buy/sell (bid/ask)
  - "Buy 100 shares of XYZ at $100.00"
- Traders may submit limit orders to the market at any time during the trading day
  - Also may cancel or modify outstanding limit orders at any time
  - Orders and cancelations are processed by the exchange one-at-a-time in order of receipt (serial process)
- Set of outstanding orders is known as the limit order book
- Trade occurs whenever a new limit order is submitted that is either (i) bid $\geq$ lowest ask; (ii) ask $\leq$ highest bid
  - New limit order is interpreted as accepting (fully or partially) one or more outstanding orders

Data

- "Direct feed" data from Chicago Mercantile Exchange (CME) and New York Stock Exchange (NYSE)
  - Gives "play by play" of limit order book
  - Millisecond resolution time stamps
  - These are the data HFT firms subscribe to and parse in real time
- Focus primarily on a pair of instruments that track the S&P 500 index
  - ES: E-Mini S&P 500 Future, traded on CME
  - SPY: SPDR S&P 500 Exchange Traded Fund, traded on NYSE (and other equities exchanges)
- Time period: 2005-2011

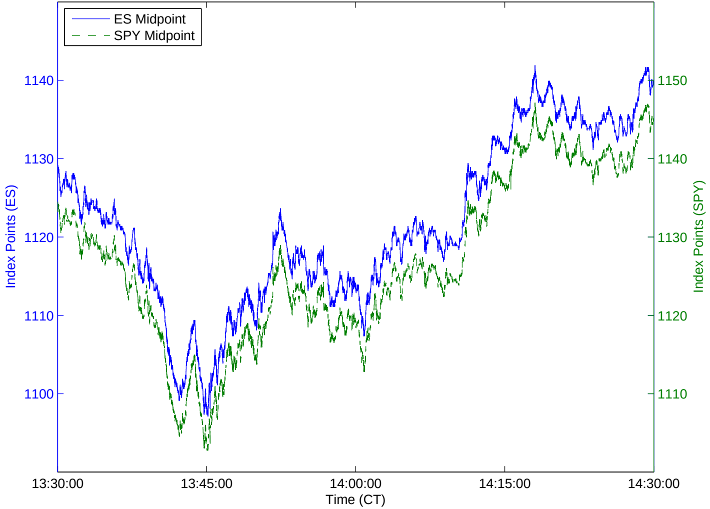
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 Day



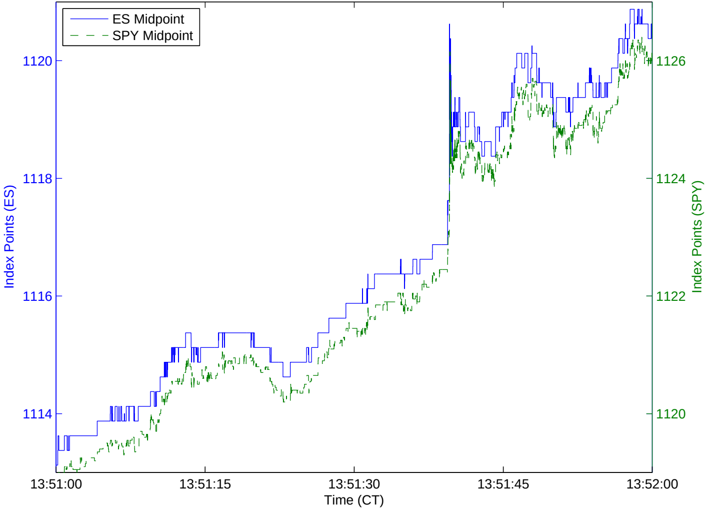
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 hour



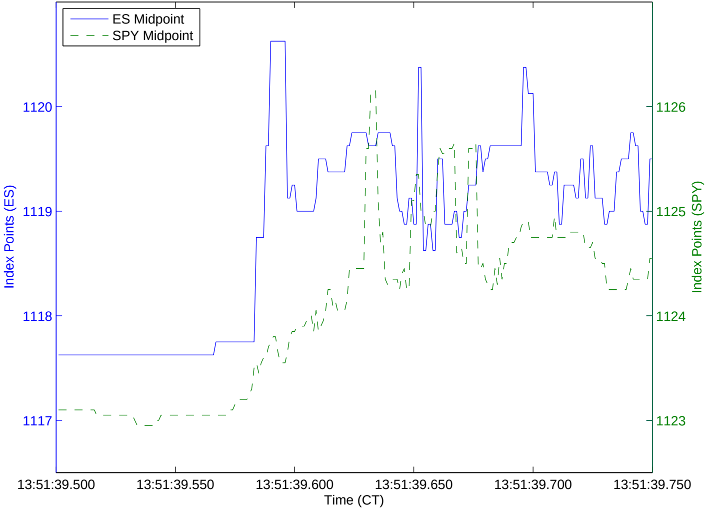
Market Correlations Break Down at High Frequency

ES vs. SPY: 1 minute



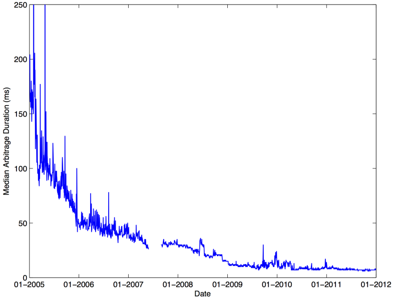
Market Correlations Break Down at High Frequency

ES vs. SPY: 250 milliseconds

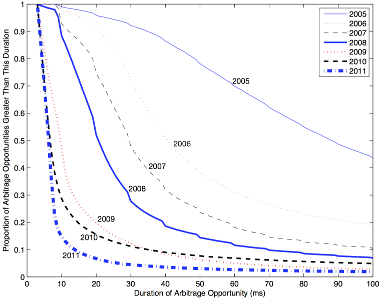


Arb Durations over Time: 2005-2011

Median over time

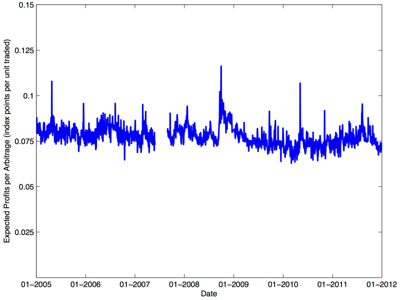


Distribution by year

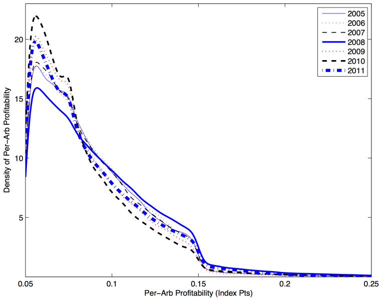


Arb Per-Unit Profits over Time: 2005-2011

Median over time

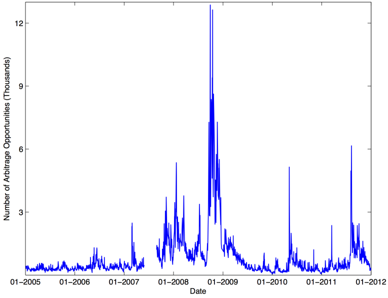


Distribution by year

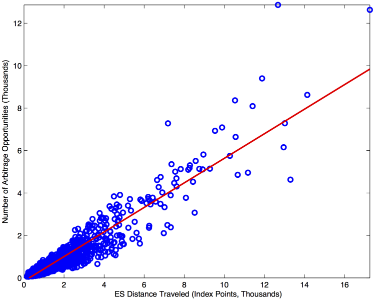


Arb Frequency over Time: 2005-2011

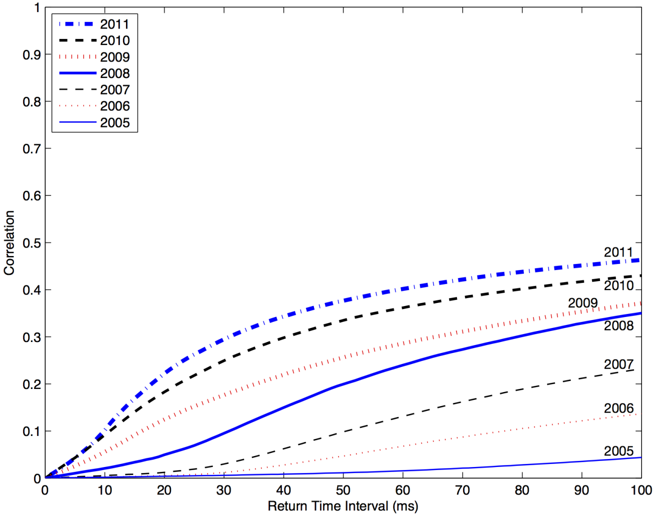
Frequency over time



Frequency vs. Volatility



Correlation Breakdown Over Time: 2005-2011



Arms Race is a "Constant" of the Market Design

- ES-SPY arbitrage and resulting arms race looks like a "constant" of the continuous limit order book.
  - Rather than a profit opportunity that is competed away over time
- Correlation Breakdown
  - Competition does increase the speed with which information is incorporated from one security price into another security price
  - Competition does not eliminate correlation breakdown
- Mechanical arbitrage
  - Competition does increase the speed requirements for capturing arbs ("raises the bar")
  - Competition does not reduce the size or frequency of arb opportunities
- These facts both inform and are explained by our model

Total Size of the Arms Race Prize

- Estimate annual value of ES-SPY arbitrage is $75mm (we suspect underestimate, details in paper)
- And ES-SPY is just the tip of the iceberg in the race for speed:

1. Hundreds of trades very similar to ES-SPY: highly correlated, highly liquid
2. Fragmented equity markets: can arbitrage SPY on NYSE against SPY on NASDAQ! Even simpler than ES-SPY.
3. Correlations that are high but far from one can also be exploited in a statistical sense.
4. Race to top of book (artifact of minimum price tick)
5. Race to respond to public news (eg Business Wire, Fed)

We don't attempt to put a precise estimate on the total prize at stake in the arms race, but common sense extrapolation from our ES-SPY estimates suggest that the sums are substantial

Frequent Batch Auctions

A simple idea: discrete-time trading.

1. Direct-feed millisecond level data from exchanges: continuous market violates basic asset pricing principles at HFT time horizons.
   - Market correlations completely break down.
   - Frequent mechanical arbitrage opportunities.
   - Mechanical arbs –> arms race. Arms race does not compete away the arbs, looks like a "constant".
2. **Theory model: critique of the continuous limit order book.**
   - **Mechanical arbs are "built in" to the market design. Sniping.**
   - **Harms liquidity (spreads, depth).**
   - **Induces a never-ending, socially wasteful, arms race for speed.**
3. Frequent Batch Auctions as a market design response
   - Eliminates mechanical arbs and sniping.
   - Competition on speed –> competition on price.
   - Enhances liquidity and stops the arms race.

Model: Goal

Simple new model which is motivated by, and helps to explain, these empirical facts. The model serves two related purposes

1. Critique of the continuous limit order book market design
2. Identifies the economic implications of the HFT arms race

Model: Preliminaries

- There is a security, x, that trades on a continuous limit-order book market
  - (Real, continuous-time limit order book. Not discrete-time sequential-move modeling abstraction like in Glosten-Milgrom)
- There is a publicly observable signal, y, of the value of security x
- Purposefully strong assumption:
  - Fundamental value of x is *perfectly* correlated to the public signal y
  - x can always be costlessly liquidated at this fundamental value
- "Best case" scenario for price discovery and liquidity provision in a continuous limit order book
  - No asymmetric info, inventory costs, etc.
- We think of x and y as a metaphor for pairs or sets of securities that are highly correlated
  - Ex: x is SPY, y is ES
  - Ex: x is SPY on NYSE (NASDAQ, dark pools, etc.), y is SPY on BATS

Evolution of y

- The signal y evolves as a compound Poisson jump process
- Arrival rate λ_{jump}
- Jump distribution F_{jump}
    - Finite support
    - Symmetric with mean zero
- Let J denote the random variable formed by drawing randomly according to F_{jump}, and then taking the absolute value.
    - The "jump size" distribution

Players: Investors and Trading Firms

Investors

- Represent end users of financial markets: mutual funds, pension funds, hedge funds, etc.
- Since there is no asymmetric information about fundamentals, could be called “liquidity traders” or “noise traders”
- Arrive stochastically to the market with an inelastic need to either buy or sell 1 unit of x
- Poisson arrival rate is λ_{invest}. Equal probability of need to buy vs. need to sell
- Mechanical strategy: trade at market immediately upon arrival
    - This is microfounded in the paper (all else equal prefer to transact sooner rather than later; assume that investors act only as “takers” of liquidity, not “makers”; investors not fast enough to snipe)

Players: Investors and Trading Firms

Trading Firms

- Equivalently: HFTs, market makers, algorithmic traders
- No intrinsic demand to buy or sell x
- Their goal in trading is simply to buy x at prices lower than y and sell at prices higher than y. Payoffs:
  - Buy x at price p at time t: earn $y_t - p$
  - Sell x at price p at time t: earn $p - y_t$
- Objective is to maximize profits per unit time
- Entry
  - Initially: # of trading firms is exogenous, $N \geq 2$
  - Below, we will endogenize entry

Latency

Exogenous entry case

- No latency in observing y
  - Trading firms observe innovations in the signal y with *zero time delay, for free.*
- No latency in submitting orders to the exchange
  - If multiple orders reach the market at the same time, the order in which they are processed is random (serial processing)
  - Alternatively, orders are transmitted with small random latency, and processed in order of receipt (eg, colocation)
- Again, best case scenario for the continuous market

Endogenous entry case

- Will add latency in observing y

“Sniping”

- Given the model setup – no asymmetric information, no inventory costs, everyone risk neutral – one might conjecture that (Bertrand) competition among trading firms leads to effectively infinite liquidity for investors
    - That is, trading firms should offer to buy or sell x at price y in unlimited quantity at zero bid-ask spread
- But that is not what happens in the continuous limit order book market, due to a phenomenon we call “sniping”

“Sniping”

- Suppose y jumps, e.g., from $\underline{y}$ to $\bar{y}$
  - This is the moment at which the correlation between y and x temporarily breaks down
- Trading firms providing liquidity in the market for x send a message to the continuous limit order book
  - Withdraw old quotes based on $\underline{y}$
  - Replace with new quotes based on $\bar{y}$

“Sniping”

- However, at the exact same time, *other* trading firms send a message to the continuous market attempting to “snipe” the stale quotes before they are adjusted
  - Buy at the old quotes based on $\underline{y}$, before these quotes are withdrawn
- Since the continuous market processes messages in *serial* – that is, one at a time – it is possible that a message to snipe a stale quote will get processed before the message to adjust the stale quote
- In fact, not only possible but *probable*
  - For every 1 liquidity provider trying to get out of the way
  - $N-1$ other trading firms trying to snipe him
  - Hence, when there is a big jump, liquidity provider gets sniped with probability $\frac{N-1}{N}$

"Sniping"

- Hence, in a continuous limit order book, *symmetrically observed public information creates arbitrage rents*.
  - Obvious mechanical arbitrages are not supposed to exist in an efficient market (Fama, 1970)
  - Closely associated with correlation breakdown phenomenon
  - Mechanically very similar to Glosten-Milgrom (1985) adverse selection, but caused by the market design not asymmetric information
    - Interpretation: Glosten-Milgrom adverse selection is "built in" to the market design
    - Symmetrically observed information is processed by the market as if it were asymmetric
    - (i.e. no such thing as symmetric information, at least during mkt hours)
  - In equilibrium, gets passed on to investors

Equilibrium, Exogenous Entry

The unique static Nash equilibrium is described as follows:

- Investors: trade immediately when their demand arises, buying or selling at the best available ask or bid, respectively.
- Trading Firms: of the N trading firms, 1 plays a role we call "liquidity provider" and $N-1$ play a role we call "stale-quote sniper".
- Liquidity provider
  - Maintain a bid and ask for 1 unit of x at spread of $s > 0$, derived below (stationary)
  - If y_t jumps, send a message to cancel old quotes and replace w new quotes
- Snipers
  - if y_t jumps such that $y_t > y_{t^-} + \frac{s}{2}$ or $y_t < y_{t^-} - \frac{s}{2}$, attempt to trade at the stale quote ("immediate or cancel")
- Trading firms are indifferent between these two roles in equilibrium. (Eqm is unique up to sorting into roles)

Equilibrium Bid-Ask Spread

In equilibrium, the bid-ask spread is such that trading firms are indifferent between liquidity provision and sniping.

- Return to liquidity provision
  - Benefits: $\lambda_{invest} \cdot \frac{s}{2}$
  - Costs: $\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N}$
- Return to sniping
  - Benefits: $\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N}$
- Indifference condition:

$$\lambda_{invest} \cdot \frac{s^*}{2} = \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2}) \qquad (1)$$

- Uniquely pins down s. Interpretation:
  - LHS: revenue from investors due to non-zero bid-ask spread
  - RHS: rents to trading firms from mechanical arbitrages

Remark: Thin Markets

- What happens if investors sometimes need to trade 1 unit but sometimes need to trade 2 units?
- If the liquidity provider provides a quote with depth 2 at the same bid-ask spread as above
  - Benefits scale less than linearly with quote size: sometimes investors only want 1
  - Costs scale linearly with quote size: if get sniped, get sniped for the whole amount!
- Hence, equilibrium bid-ask spread is wider for second unit than first
- Not only is there a positive bid-ask spread even without asymmetric information about fundamentals, but markets are thin too

Equilibrium, Endogenous Entry

- Now, endogenize entry.
  - Trading firms observe the signal y with a small time delay, $\delta_{slow} > 0$, for free
  - Can pay a cost c_{speed} to reduce latency from δ_{slow} to δ_{fast}, with $0 \leq \delta_{fast} < \delta_{slow}$. Let $\delta = \delta_{slow} - \delta_{fast}$
- Equilibrium
  - Very similar structure to above: 1 liquidity provider, $N - 1$ stale-quote snipers
  - N now endogenous: the number of fast traders (for simplicity, allow N real not integer. mild assumption on c_{speed} ensures $N \geq 2$ in eqm)
  - Fast traders earn zero profits
    - Market design creates rents; speed competition dissipates rents
    - Could generalize to give inframarginal fast traders positive profits
  - No role for slow traders in equilibrium

Equilibrium, Endogenous Entry

- Zero-profit condition for liquidity provider

$$\lambda_{invest} \cdot \frac{s}{2} - \lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{N-1}{N} = c_{speed} \tag{2}$$

- Zero-profit condition for stale-quote snipers

$$\lambda_{jump} \cdot \Pr(J > \frac{s}{2}) \cdot \mathbb{E}(J - \frac{s}{2} | J > \frac{s}{2}) \cdot \frac{1}{N} = c_{speed} \tag{3}$$

- Together, equations (2) and (3) describe equilibrium, by uniquely pinning down the bid-ask spread s^*, the total entry of trading firms N^*, and the indifference of trading firms between the two roles they might play.

Equilibrium, Endogenous Entry

- Adding (2) and $N-1$ times (3) yields

$$\lambda_{invest} \cdot \frac{s^*}{2} = N^* \cdot c_{speed} \tag{4}$$

- Economic interpretation: all of the expenditure by trading firms on speed technology ultimately is borne by investors, via the bid-ask spread.
  - Arms-race prize = expenditures on speed = cost to investors
  - Remember: arms-race profits have to come from *somewhere*

What's the Market Failure?

Chicago question: isn't the arms race just healthy competition? what's the market failure?

What’s the Market Failure?

Market Failure 1: Sniping

- Mechanical arb opportunities are “built in” to the market design
- These arb opportunities violate weak-form EMH (Fama, 1970)
    - Market looks highly efficient in time space, but it isn’t efficient in volume space
- Arbs create rents from symmetrically observed public information

Market Failure 2: Arms Race

- The arb rents then induce an arms race for speed
- Mathematically, a prisoners’ dilemma

Remarks on the Equilibrium

Arms Race is a "constant"

- Arms race prize = expenditures on speed = cost to investors $= \lambda_{jump} \cdot \Pr(J > \frac{s^*}{2}) \cdot \mathbb{E}(J - \frac{s^*}{2} | J > \frac{s^*}{2})$
- Comparative static: the negative effects of the arms race on liquidity and welfare do not depend on either
  - the cost of speed (if speed is cheap, there will be more entry)
  - the magnitude of speed improvements (seconds, milliseconds, microseconds, nanoseconds, ...)
- The problem we identify is an equilibrium feature of continuous limit order books
  - not competed away as HFTs get faster and faster
  - ties in nicely with empirical results

Remarks on the Equilibrium

Role of HFTs

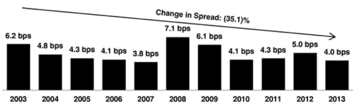
- In our model HFTs endogenously perform two functions
  - Useful: liquidity provision / price discovery
  - Rent-seeking: sniping stale quotes
- The rent-seeking seems like zero-sum activity among HFTs
  - but we show that it ultimately harms real investors
- Clarification
  - Our results do *not* imply that on net HFT has been bad for liquidity or social welfare
  - Our results *do* say that sniping is bad for liquidity and the speed race is socially wasteful
  - Frequent batch auctions preserve (in fact, enhance) the useful function that HFTs perform while eliminating sniping and the speed race

Remark: Empirical Evidence of Effect of HFT on Liquidity

Consistent with “IT Good, Speed Race Bad”

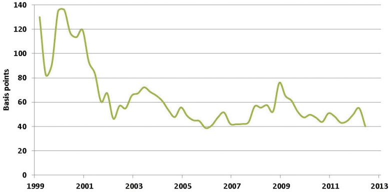
Virtu IPO Filing (Spreads)

Narrowing Bid/Ask Spreads (S&P 500)



Angel, Harris and Spatt (Cost to Trade Large Blocks)

Average Transaction Cost Estimate
for 1M Shares in a $30 Stock



Source: Authors' analysis of Ancerno trade data.

Frequent Batch Auctions

A simple idea: discrete-time trading.

1. Direct-feed millisecond level data from exchanges: continuous market violates basic asset pricing principles at HFT time horizons.
    - Market correlations completely break down.
    - Frequent mechanical arbitrage opportunities.
    - Mechanical arbs –> arms race. Arms race does not compete away the arbs, looks like a "constant".
2. Theory model: critique of the continuous limit order book.
    - Mechanical arbs are "built in" to the market design. Sniping.
    - Harms liquidity (spreads, depth).
    - Induces a never-ending, socially wasteful, arms race for speed.
3. **Frequent Batch Auctions as a market design response**
    - **Eliminates mechanical arbs and sniping.**
    - **Competition on speed –> competition on price.**
    - **Enhances liquidity and stops the arms race.**

Frequent Batch Auctions: Overview

- High level: analogous to the current market design but for two key differences
  - Time is treated as discrete, not continuous
  - Orders are processed in batch, not serial

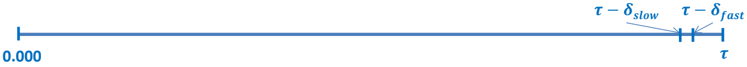
Frequent Batch Auctions: Definition

- The trading day is divided into equal-length discrete batch intervals, each of length $\tau > 0$.
- During the batch interval traders submits bids and asks
  - Can be freely modified, withdrawn, etc.
  - If an order is not executed in the batch at time t, it automatically carries over for $t+1, t+2, \ldots,$
- At the end of each interval, the exchange "batches" all of the outstanding orders, and computes market-level supply and demand curves
- If supply and demand intersect, then all transactions occur at the same market-clearing price ("uniform price")
  - Priority: still price-time, but treat time as discrete. Orders submitted in the same batch interval have the same priority. Rationing is pro-rata.
- Information policy: orders are not visible during the batch interval. Aggregate demand and supply are announced at the end.
  - Discrete time analogue of current practice in the continuous limit order book market

Why Batching Solves the Problems with Continuous-time

Reason 1: discrete time reduces the economic relevance of tiny speed advantages

- Consider a slow trader who attempts to provide liquidity to investors
- There is 1 fast trader present in the market
- Continuous market: liquidity provider is vulnerable to being sniped by the fast trader for *all* jumps in y.
- Discrete market: liquidity provider is vulnerable to being sniped by the fast trader for only $\frac{\delta}{\tau}$ proportion of jumps in y:



Why Batching Solves the Problems with Continuous-time

Reason 2: the auction changes the nature of competition: from competition on speed to competition on price

- As above, suppose a slow trader attempts to provide liquidity to investors
- There are $N \geq 2$ fast traders present in the market (exogenously)
- Suppose y jumps in the interval $[\tau - \delta_{slow}, \tau - \delta_{fast}]$ where the liquidity provider is vulnerable.
- All of the fast traders wish to exploit the stale quote ... but the auction means trade goes not to who is first but to who offers the best *price*
- Eqm price competition drives the price of x to its new, correct level. Slow liquidity provider trades at the auction price, not the stale price.
  - Continuous market: competition on speed. *Someone is always first.*
  - Batch auction market: competition on price. Lots of orders reach the exchange by the end of the batch interval.

Why Batching Solves the Problems with Continuous-time

- Another way to think about these two points:
  - Discrete-time dramatically reduces the likelihood that tiny speed advantages lead to (economically relevant) asymmetric information
  - Auction eliminates rents from symmetrically observed information
- N.B.: with batch intervals of e.g. 100 ms, there is still plenty of scope for market participants to develop genuinely asymmetric information about security values, for which they would earn a rent.
  - Frequent batch auctions just eliminates rents from information that many observe at basically the same time and understand equally well.

Equilibrium of Frequent Batch Auctions, Exogenous Entry

- $N \geq 2$ trading firms, exogenously in the market, any $\tau > 0$
- Description of equilibrium:
  - Bertrand competition drives bid-ask spread to zero, effectively infinite depth
  - No sniping
- Highlights the central differences between frequent batch auctions and the continuous limit order book
  - No rents from symmetrically observed public information
  - No mechanical arbitrage opportunities
  - Bertrand competition on price drives spread to zero, as expected given model setup
- Note: eqm obtains for any $\tau > 0$; discontinuity as go from continuous+serial -> discrete+batch

Equilibrium of Frequent Batch Auctions, Endogenous Entry

- Description of equilibrium
  - If τ sufficiently long relative to δ, then in equilibrium no trading firms pay c_{speed} to be δ faster
  - Slow trading firms provide $\bar{Q}$ units of liquidity at zero-bid ask spread
- Key condition: not worth it for a fast trader to enter to pick off the slow traders:

$$\frac{\delta \lambda_{jump}}{\tau} E(J) \cdot \bar{Q} < c_{speed} \tag{5}$$

- The fraction $\frac{\delta \lambda_{jump}}{\tau}$ is the proportion of time during the trading day during which the fast trader has a profitable sniping opportunity.
- For any finite $\bar{Q}$, the condition is satisfied for long enough τ.
  - Hence, any desired market depth can be provided by slow traders at zero cost if the batch interval is sufficiently long.

How Long is Long Enough to Stop the Speed Race

$$\frac{\delta \lambda_{jump}}{\tau} E(J) \cdot \bar{Q} < c_{speed}$$

- In appendix we use combination of ES-SPY data and info from HFT public documents to calibrate the lower bound on τ
- Extremely rough. Intended just for sense of magnitudes.
- δ open to multiple interpretations
  1. Year-on-year speed improvement for state-of-the-art HFT (about 100 microseconds for NYC-Chicago in 2014)
  2. Difference between state-of-the-art HFT and sophisticated non-HFT algo traders (a few milliseconds)
- Interpretation 1 –> lower bound on order of 10 or 100ms.
- Interpretation 2 –> lower bound on order of 100ms or 1s.
- Again, extremely rough.

Computational Benefits of Discrete Time

- Overall
  - Continuous-time markets implicitly assume that computers and communications technology are infinitely fast.
  - Discrete time respects the limits of computers and communications. Computers are fast but not infinitely so.
- Exchanges
  - Continuous: Computational task is mathematically impossible; latencies and backlog unavoidable; clock sync hard
  - Discrete: Computation is easy; clock sync easy
- Regulator
  - Continuous: Audit trail is difficult to parse; need to adjust for relativity, clock sync issues
  - Discrete: Simple audit trail; state at t, $t+1$,...
- Algorithmic traders
  - Continuous: Always uncertain about current state; temptation to trade off robustness for speed (MacKenzie, 2014)
  - Discrete: Everyone knows state at time t before decision at time $t+1$

Policy Debates Cleaned Up By Discrete Time

- Clock Synchorinization across exchanges
  - Continuous-time: challenging.
  - Discrete-time: trivial.
- Exchange Message Priority Rules
  - Continuous-time: details of message priority matter. Book updates vs. trade confirmation messages. CME controversy.
  - Discrete-time: issue goes away. plenty of time to disseminate all of the relevant info.
- "Level Playing Field" in access to info
  - Continuous-time: even if in principle info is released to all simultaneously, someone receives / acts on it first. arbitrage rents even from symmetrically observed public information.
  - Discrete-time: restores possibility of meaningfully symmetric information.
- Payment for order flow, Dark Pool debates
  - Continuous time: paper trail makes it hard for investors to know whether they got a fair price, versus a stale price
  - Discrete time: paper trail clean. Easier to discover if exploited.

Alternative Responses to the HFT Arms Race

- Tobin Tax
  - Does partially mitigate sniping
  - But: blunt instrument. Needs to be large to effect the arms race, and the cost gets passed on to investors.
- "Bans" on HFT: Message Ratios, Minimum Resting Times
  - Misunderstand cause and effect
  - Resting times likely to exacerbate sniping
- Random delay
  - Does *not* mitigate sniping
  - Encourages redundant messages: each message to snipe is like a lottery ticket
  - Explosion in message traffic
- IEX speed bump + price sliding to NBBO midpoint
  - Ingenious, eliminates sniping
  - But, only works while IEX is small relative to the rest of the continuous market (free-rides off price discovery elsewhere)
  - Still continuous-time serial-process, so does not fully eliminate incentive to be tiny bit faster than competition

Open Questions

- Another Chicago question: if this is such a good idea, why hasn't an exchange already tried it? Potential reasons:
    - Relatively new problem
    - Coordination challenge
    - Regulatory ambiguities
    - Vested interests in the current market structure
- Issues due to fragmentation of US equities markets
    - What is equilibrium if there are both batch and continuous exchanges operating in parallel?
    - Mechanics if multiple exchanges each run batch (how to ensure law of one price)
    - Interaction with Reg NMS

Open Questions

- Market stability
  - Common claim among policy makers is that stopping the HFT arms race would enhance market stability (meaning vulnerability to flash crashes, exchange outages, programming glitches, etc.)
  - This is another potential welfare benefit of frequent batching, but not yet modeled
- Open questions re frequent batch auction design
  - Would be desirable to consider a richer model with asymmetric information, inventory management, investors needing to trade large quantities
  - Optimal batch interval, and how this varies by security
  - Tick sizes?

New York Attorney General Speech, March 18th, 2014

We have to review, and it's something I want to raise, and I'm sure it will be discussed at the panel, and **carefully consider a proposal that I like very much. It was put forward by economists at the University of Chicago School of Business – not an enemy of free markets, the University of Chicago School of Business, by any means.**

In December, they issued a detailed and thoughtful proposal for reforms that would **fundamentally reorient the markets in a very simple way that would help restore confidence in them. Their proposals would reaffirm the basic concept that the best price – not the highest speed – should win.**

The University of Chicago proposal – which I endorse – would, in effect, put a speed bump in place. Orders would be processed in batches after short intervals – potentially a second or less than a second in length – but that would ensure that the price would be the deciding factor in who obtains a trade, not who has the fastest supercomputer and early access to market-moving information.

SEC Chair White's Speech, June 5th, 2014

We must consider, for example, whether the increasingly expensive search for speed has passed the point of diminishing returns. I am personally wary of prescriptive regulation that attempts to identify an optimal trading speed, but I am receptive to more flexible, competitive solutions that could be adopted by trading venues. These could include frequent batch auctions or other mechanisms designed to minimize speed advantages. They could also include affirmative or negative trading obligations for high-frequency trading firms that employ the fastest, most sophisticated trading tools.

...

A key question is whether trading venues have sufficient opportunity and flexibility to innovate successfully with initiatives that seek to deemphasize speed as a key to trading success in order to further serve the interests of investors.[14] **If not, we must reconsider the SEC rules and market practices that stand in the way.**

Bloomberg Editorial, June 18th, 2014

Today's stock market is falling short. A wasteful arms race among high-frequency traders, the growth of dark pools (private trading venues) and assorted conflicts of interest have undermined its performance. If investors don't trust the market, that hurts capital formation, not to mention retirement and college savings.

...

Fixing the problems will require more than a tweak here and there. One idea that's winning converts would replace the 24-hour, continuous trading of stocks with frequent auctions at regular intervals.

Why would that help? Because it would lessen the emphasis on speed and direct more attention to the price that investors are willing to pay for stocks, given the prospects of the companies concerned, their industries and the broader economy. **The high-speed arms race would subside, because shaving another millisecond off the time it takes to trade would confer no benefit.**

...

Mary Jo White, the Securities and Exchange Commission chair, indicated in a June 5 speech her interest in batch auctions. She should make it a priority to conduct a test program. It's a promising idea.

Summary

- We take a market design perspective to the HFT arms race.
- Root problem isn't "evil HFTs", it's continuous-time / serial-process trading.
- Alternative: discrete-time / batch-process trading

1. Direct-feed data: continuous-time markets don't actually work in continuous time: correlations completely break down; frequent mechanical arbs; never-ending arms race
2. Theory: root cause is the current market design
   - Arms race is a never-ending, equilibrium feature
   - Arms race harms liquidity and is socially wasteful
3. Frequent Batch Auctions as a market design response
   - Benefits: eliminates sniping, stops arms race, enhances liquidity, computational advantages
   - Costs: investors must wait a small amount of time to trade, unintended consequences