Liquidity Cycles and Make/Take Fees in Electronic Markets

Thierry Foucault (HEC, Paris) Ohad Kadan (Washington U)
Eugene Kandel (Hebrew U)

April 2011

Algorithmic Trading

- Algorithmic trading is the use of powerful computer algorithms to make high-frequency trading decisions.
- In recent years 60%-70% of all trades are believed to be initiated by computer algorithms.
- Result: Dramatic reduction in the time it takes traders to respond to new trading opportunities.
- The effects of this evolution are not yet fully understood.
- Algorithmic trading is the subject of heated debates: maker/taker pricing, “flash orders,” and the “flash crash” of May 2010.

Latency and Attention

- Traders do not instantaneously react to a change in the state of the market because obtaining, processing and acting upon new information takes time.
- Algorithmic traders refer to this delay as "latency."
- For human traders, reducing latency is costly as it requires attention, and traders must allocate their limited attention among multiple tasks.
- Algorithmic trading considerably relaxes the cost of attention but it does not eliminate this cost completely.
- Even computers have limited processing capacity that needs to be allocated among multiple tasks (e.g., parallel trading in hundreds or thousands of securities within fractions of seconds).
- Traders face a trade-off between the benefit and the cost of attention in the face of limited monitoring capacity.

In This Paper

- We study the trade-off between the costs and benefits of reducing latencies in a particular asset by monitoring it more frequently.
- We endogenize latencies, and analyze the effects of drastic reductions in the cost of attention (algorithmic trading).
- We shed light on several issues regarding current market structures:
  - The widespread adoption of the maker/taker pricing model.
  - The consequences of algorithmic trading on liquidity, volume, and welfare.

Liquidity Cycles in Limit Order Markets

- Trade in limit order markets is characterized by liquidity cycles with two phases (Biais, Hillion, and Spatt (1995)):
  - A "make liquidity" phase - traders post prices (limit orders). Liquidity is plentiful.
  - A "take liquidity" phase - traders consume liquidity by posting market orders.
- The submission of a market order creates profit opportunities for limit order submitters.
- The submission of a limit order creates profit opportunities for market order submitters.
- This creates cross-side externalities - multiplicity of equilibria.

Make/Take Fees

Trading platforms charge different fees on "market makers" (limit orders) and "market takers" (market orders). These fees add-up to millions of dollars per trading day.

	Tape A - NYSE Stocks		Tape B - Other Stocks		Tape C - NASDAQ Stocks	
	Make Fee	Take Fee	Make Fee	Take Fee	Make Fee	Take Fee
NYSE Arca	-23	30	-22	30	-23	30
Nasdaq	-20	30	-20	30	-20	30
BATS	-24	25	-24	25	-24	25
EDGX	-25	30	-30	30	-25	30
LavaFlow	-24	27	-24	27	-24	27

Fees per 100 shares. Source: Traders' Magazine, August 2009

Make/Take Fees (cont.)

- Maker/taker pricing results in significant monetary transfers between market participants.
- Example:
  - Average monthly volume on NYSE-Arca during 2009 was about 32 billion shares.
  - A net fee of 7 cents per 100 shares generates an approximate annual revenue of $270 million to the exchange.
  - Compared to an equal breakdown, the maker/taker model results in an approximate annual wealth transfer of $1.0 billion from market-takers to high-frequency market-makers on NYSE-Arca alone (14% of 2009 US volume).

Make/Take Fees - Controversy

- Some high-frequency market-making firms follow rebate-capture strategies, and strongly support maker/taker pricing.
- Other market participants have voiced concerns that this model could result in excessive fees for takers.
- Both sides are needed for a transaction. Why compensate one side and not the other?
- Angel, Harris, and Spatt (2010) argue that make/take fees are irrelevant since traders can undo their effect through transaction prices.
- The SEC decided to cap take fees in equity markets at three cents per 100 shares.
- No formal economic analysis of make/take fees and their link to high-frequency trading.

Summary of Questions

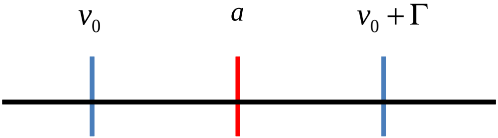
- What is the economic rationale of maker/taker pricing? What determines the fee structure?
- Do make/take fees have an economic effect and when?
- What is the relation between maker/taker pricing and algorithmic trading?
- What is the effect of algorithmic trading on liquidity, volume, and welfare?
- What can be said about the new form of cross-side liquidity externality. Can it be identified empirically.

Market Participants

- Three types of market participants:
  - M market makers: post quotes (e.g., high frequency electronic market makers such as GETCO, ATD, Tradebot Systems, Inc., Optiver).
  - N market takers: hit quotes (e.g., brokers slicing orders).
  - One "matchmaker": the trading platform.

Gains from trade

- Market makers value the security at v_0.
- Market takers value the security at $v_0 + \Gamma$.
- $\implies$ Market takers are buyers and market makers are sellers.
- Gains from trade per transaction $= \Gamma$.
- Model of the "upper half" of the market - limit sell and market buy only.
- Market-makers and market-takers meet on a trading platform where they trade at a price denoted by a.



Make/Take Fees

- The trading platform charges make/take fees each time a trade occurs.
- The fee (per share) paid by a market-maker is denoted by c_m.
- The fee paid by a market-taker is denoted by c_t.
- Negative fees reflect rebates.
- The platform earns a profit of

$$\bar{c} \equiv c_m + c_t.$$

Gains from Trade per Transaction

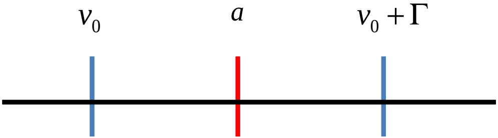
- Order size is normalized to 1.
- Total gains from trade = Γ.
- The market-taker obtains

$$\pi_t = v_0 + \Gamma - a - c_t.$$

- The market-maker obtains

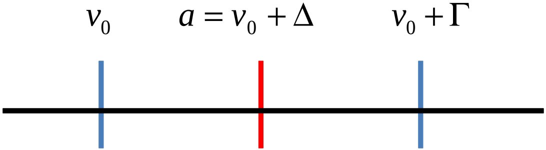
$$\pi_m = a - v_0 - c_m.$$

- The platform obtains $\bar{c} = c_m + c_t$.



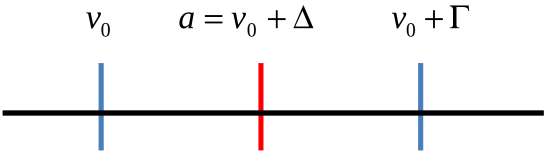
Price Grid

- The price at which market-makers and market-takers trade, a, must be on a grid with a tick size (minimum price variation) equal to $\Delta(\ell) = \frac{\Gamma}{2\ell}$, where ℓ is a positive integer.
- In the baseline case, we set $\ell = 1$ and we denote the tick size by $\Delta = \Delta(1)$.
- Assume that make and take fees are smaller in absolute value than Δ.
- When $\ell = 1$, the only price on the grid at which market-makers and market-takers agree to trade is $a = v_0 + \Delta$.



Timing and Cycles

- An infinite horizon model with a continuos time line.
- At each point in time the market can be in one of two states:
  - State E – liquidity is low: no offer is posted at a.
  - State F – Liquidity is high: an offer for one share is posted at a.
- A cycle is the flow of events from the moment the market gets into state E until it returns into this state.



Monitoring

- Market-makers and market-takers have an incentive to monitor the market to be the first to seize profit opportunities.
- Each market-maker $i = 1, ..., M$ inspects the market according to a Poisson process with parameter μ_i.
- Each market-taker $j = 1, ..., N$ inspects the market according to a Poisson process with parameter τ_j.
- Total inspection frequencies:

$$\bar{\mu} \equiv \mu_1 + ... + \mu_M \quad \text{and} \quad \tau \equiv \tau_1 + ... + \tau_N.$$

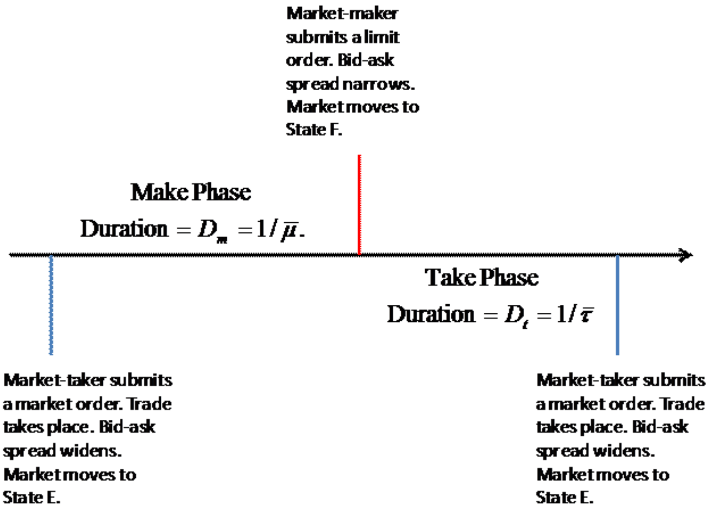
- The duration of a cycle is

$$\mathcal{D}\left(\bar{\mu}, \bar{\tau}\right) = \frac{1}{\mathcal{R}\left(\bar{\mu}, \bar{\tau}\right)} = \frac{1}{\bar{\mu}} + \frac{1}{\bar{\tau}}.$$

- The transaction rate is

$$\mathcal{R}\left(\bar{\mu}, \bar{\tau}\right) \equiv \frac{1}{\frac{1}{\bar{\mu}} + \frac{1}{\bar{\tau}}} = \frac{\bar{\mu}\bar{\tau}}{\bar{\mu} + \bar{\tau}}.$$

Make/Take Cycles



Monitoring Costs

- Monitoring costs over a time interval of length T quadratic in monitoring intensities μ_i and τ_j:

$$
\begin{aligned}
C_m(\mu_i) &= \frac{1}{2}\beta\mu_i^2 T \\
C_t(\tau_j) &= \frac{1}{2}\gamma\tau_j^2 T
\end{aligned}
$$

- γ and β measure the unit costs of monitoring.
- The trade-off:
  - More monitoring of a security increases the chance of being "first" to capture a profit opportunity.
  - But - more monitoring of one security implies lost profit opportunities elsewhere.
- Algorithmic trading - reduction in γ or β.

Timing and Equilibrium

- Timing:
  1. The Trading Platform chooses its fee structure: $\{c_m, c_t\}$,
  2. Market Participants choose simultaneously their monitoring intensities: $\{\mu_i\}_{i=1}^{M}$, $\{\tau_j\}_{j=1}^{N}$, taking fees as given.
  3. Rest of the game: the game unfolds as described previously on a continuous time line, forever.
- Nash equilibrium in monitoring intensities.
- Solve backwards.

Objective Functions

- The expected payoff to market-maker i per transaction (cycle) is

$$\frac{\mu_i}{\mu_1 + ... + \mu_i + ... + \mu_M}\pi_m = \frac{\mu_i}{\bar{\mu}}\pi_m.$$

- Let $\tilde{n}_T$ be the (random) number of completed transactions (cycles) until time T.
- The expected payoff to market-maker i until time T is

$$\Pi_i(T) = E_{\tilde{n}_T}(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}}\pi_m) - \frac{1}{2}\beta\mu_i^2 T.$$

- This expected payoff is undefined as $T \to \infty$.

Objective Functions (cont.)

- Each player maximizes the long-term payoff per unit of time (best utilization of computer equipment).
- Thus, market-maker i chooses monitoring intensity to maximize

$$\Pi_{im} \equiv \lim_{T\to\infty} \frac{\Pi_i(T)}{T} = \lim_{T\to\infty} \frac{E_{\tilde{n}_T}\left(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}} \pi_m\right)}{T} - \frac{1}{2}\beta\mu_i^2.$$

- By the "Renewal Reward Theorem":

$$\lim_{T\to\infty} \frac{E_{\tilde{n}_T}\left(\sum_{k=1}^{\tilde{n}_T} \frac{\mu_i}{\bar{\mu}} \pi_m\right)}{T} = \frac{\frac{\mu_i}{\bar{\mu}} \cdot \pi_m}{\mathcal{D}\left(\bar{\mu}, \bar{\tau}\right)} = \mathcal{R}\left(\bar{\mu}, \bar{\tau}\right) \cdot \frac{\mu_i}{\bar{\mu}} \cdot \pi_m.$$

Objective Functions (cont.)

1. **Expected profit for Market Makers** = transaction rate times probability of winning times share in profits less monitoring costs.

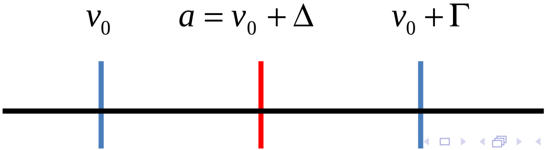
$$\Pi_{im}(\mu_i, \bar{\tau}, \bar{\mu}) = \mathcal{R}\left(\bar{\mu}, \bar{\tau}\right) \cdot \frac{\mu_i}{\bar{\mu}} \cdot \left(\Delta - c_m\right) - \frac{1}{2}\beta\mu_i^2.$$

2. **Expected profit for Market Takers** = transaction rate times probability of winning times share in profits less monitoring costs.

$$\Pi_{jt}(\tau_j, \bar{\tau}, \bar{\mu}) = \mathcal{R}\left(\bar{\mu}, \bar{\tau}\right) \cdot \frac{\tau_i}{\bar{\tau}} \cdot \left(\Gamma - \Delta - c_t\right) - \frac{1}{2}\gamma\tau_j^2.$$

3. **Expected profit for Platform** = transaction rate times total fees.

$$\mathcal{R}\left(\bar{\mu}, \bar{\tau}\right) \cdot \left(c_m + c_t\right)$$



Objective Functions - Properties

- **Same-Side Competition:** The expected payoff of each participant decreases with aggregate monitoring of remaining participants on the same side.
- **Cross-Side Complementarity**:
  - The marginal benefit from monitoring on one side increases with aggregate monitoring of participants on the other side.
  - Market-makers (resp., market-takers) monitor the state of the market more frequently when they expect market-takers (resp. market-makers) to monitor the state of the market more frequently.
- **Implications:**
  1. Liquidity begets liquidity.
  2. Both sides are equally important for the formation of liquidity!
  3. Which side should be subsidized (if at all) is not obvious...

Equilibrium for Fixed Fees

- Cross-side complementarities $\Longrightarrow$ Coordination Problem between market-makers and market-takers.
- Multiple equilibria. In our case - exactly two.
- **Finding 1:** *There exists an equilibrium with no monitoring and no trade.*

Equilibrium with Monitoring and Trade

Denote $z \equiv \frac{\gamma\pi_m}{\beta\pi_t}$.

Finding 2: *There exists a unique equilibrium with trade. In this equilibrium, traders' monitoring intensities are given by*

$$\mu_i^* = \left(\frac{M + (M-1)\,\mathcal{V}^*}{(1+\mathcal{V}^*)^2}\right)\left(\frac{\pi_m}{M\beta}\right) \quad i = 1, \ldots, M$$

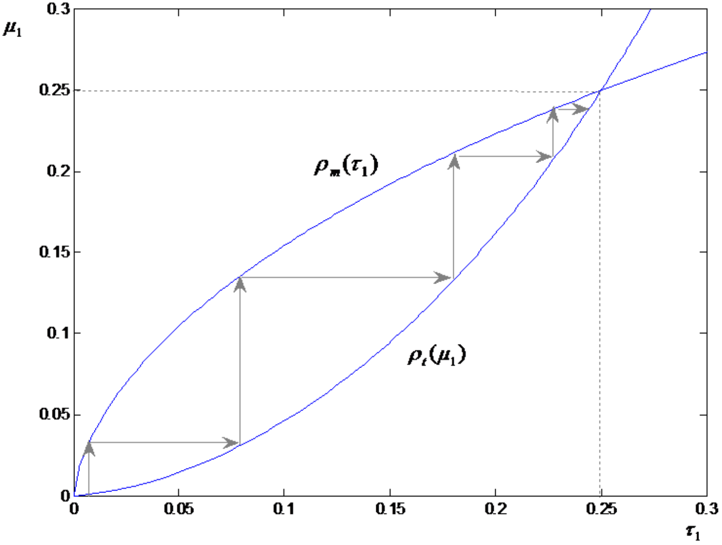
$$\tau_j^* = \left(\frac{\mathcal{V}^*\left((1+\mathcal{V}^*)\,N - 1\right)}{(1+\mathcal{V}^*)^2}\right)\left(\frac{\pi_t}{N\gamma}\right) \quad j = 1, \ldots, N$$

where $\mathcal{V}^* = \frac{\bar{\mu}^*}{\bar{\tau}^*}$ *is the unique positive solution to the cubic equation*

$$\mathcal{V}^3 N + (N-1)\mathcal{V}^2 - (M-1)\,z\mathcal{V} - Mz = 0.$$

$\mathcal{V}^*$ is termed - the velocity ratio.

Stability of No-Trade



Trading Activity for Fixed Fees

- **Finding 3:** *In the unique equilibrium with trade:*

1. The aggregate monitoring level of both sides increases in the number of participants on either side, and decreases in monitoring costs and in the fee per trade charged on either side.
2. The trading rate decreases in the monitoring costs and in the trading fees, and increases in the number of participants on either side. *Decrease with traders' monitoring costs (γ and β).*

- **Finding 4:** Monitoring is typically not balanced. *If* $\frac{\gamma\pi_m(2M-1)}{\beta\pi_t(2N-1)} > 1$ *(resp.* < 1*) then the aggregate monitoring of the market-making side is higher (resp. lower) than the aggregate attention of the market-taking side, i.e.,* $\bar{\mu} > \bar{\tau}$ $(\bar{\mu} < \bar{\tau})$.
- Incentive for the trading platform to choose fees to equilibrate the speed of reaction of both sides.

Breakdown of Fees

- Trading platform's problem:

$$\max_{c_m, c_t} \underbrace{(c_m + c_t)}_{\text{fees per trade}} \cdot \underbrace{\mathcal{R}(\bar{\mu}, \bar{\tau})}_{\text{transaction rate}},$$

$$s.t. \quad c_m + c_t = \bar{c}.$$

- **Finding 5:** *The optimal breakdown of fees* equalizes the volume effects of a small change in fees on both sides:

$$\frac{\partial \mathcal{R}}{\partial c_m} = \frac{\partial \mathcal{R}}{\partial c_t}.$$

- Fee structure:
  - Closed form solutions and full comparative statics for two cases: $M = N = 1$ (dual monopoly) and M, N "very large" (thick market).
  - Numerical simulations for intermediate cases.
  - In general, a flat fee ($c_m = c_t$) is not optimal $\Longrightarrow$ There is a take/make spread: $c_t - c_m \neq 0$.

Determinants of the Make/Take Fees

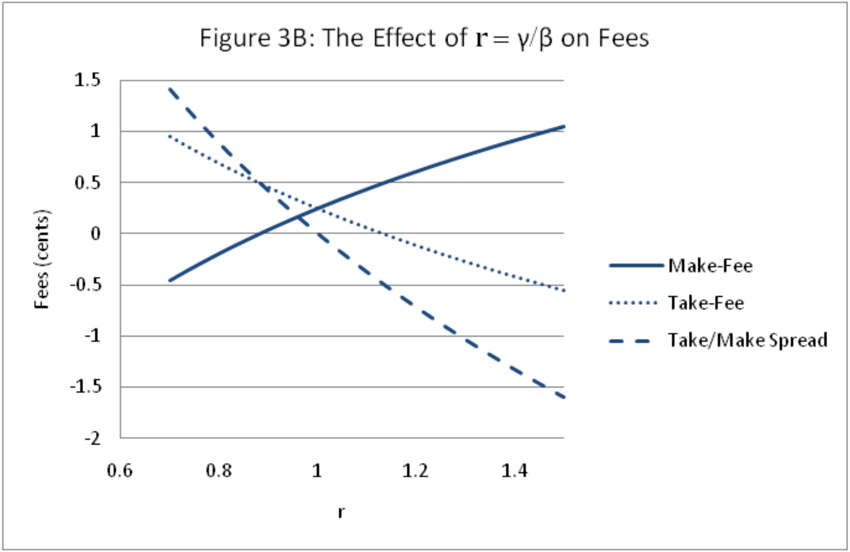
- **General Principle:** The trading platform chooses fees to reduce excessive attention of one side relative to the other.
- **Finding 6:** The take fee decreases and the make fee increases with the relative size of the market-making side, q, and the relative monitoring cost for the market-taking side, r. Thus, the take/make spread $(c_t - c_m)$ decreases with these parameters.
- $\Longrightarrow$Optimality of liquidity rebates is not a foregone conclusion.

Example

- Suppose $M = 10$, $N = 20$; $\beta = 0.2$, $\gamma = 0.1$, $\Gamma = 25$, $\Delta = 12.5$, $v_0 = 300$, and $\bar{c} = 1/10$ (all monetary amounts in cents). Time unit: seconds.

	Uniform Pricing	Optimal Pricing
Make-Fee (cents/share)	0.05	-2.72
Take-Fee (cents/share)	0.05	2.82
Total Fee (cents/share)	0.1	0.1
Trading Rate (shares/second)	267.4	277.3
Platform's Fee Revenue ($/second)	0.2674	0.2773
Estimated Annual Platform Fee Revenue ($ in millions)	3130	3244

Example: Monitoring Costs and Make/Take Fees



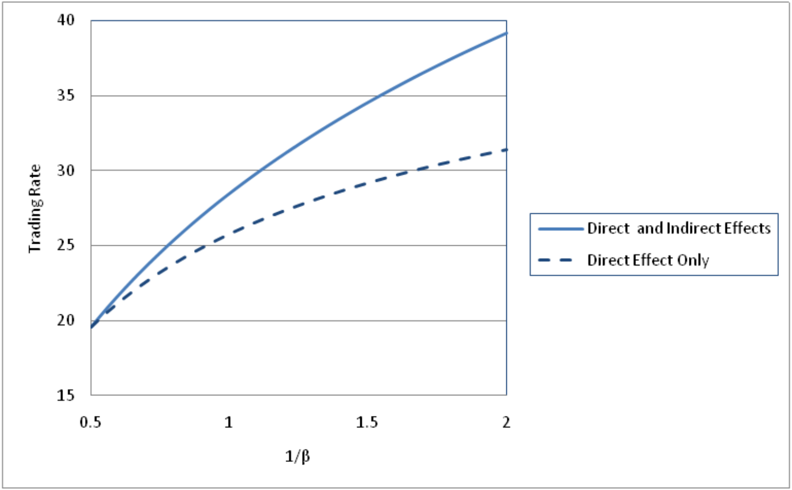
Algo Trading and Trading Activity

- Burst in trading volume in recent years.

$$\text{volume} = \text{Transaction rate} \times \text{transaction size.}$$

- The model suggests two explanations:
  1. Algorithmic trading $\Longrightarrow$ Lower monitoring costs $\Longrightarrow$ Trading rate increases. Cross-side complementarities amplify the effect of small reductions in monitoring costs.
  2. Trading platforms' pricing policies: the make-take spread (adequately chosen) contributes to a larger trading rate.

Algo Trading and Trading Activity



Empirical Implication: Bid-Ask Spread and Algo Trading

- The average bid-ask spread in equilibrium is:

$$ES = \theta^* \cdot a + (1 - \theta^*) \cdot (a + \Delta),$$

where

$$\theta^* = \left(\frac{\mathcal{D}_t}{\mathcal{D}_m + \mathcal{D}_t}\right) = \frac{\mathcal{V}^*}{1 + \mathcal{V}^*}.$$

- Average bid-ask spread is small when market makers monitor more, and large when market takers monitor more.
- Development of algo trading has an ambiguous impact on the bid ask spread
  - Automation of the market-making side $\Longrightarrow$ The bid-ask spread declines (Hendershott, Jones and Menkveld (2011)).
  - Automation of the market-taking side $\Longrightarrow$ The bid-ask spread increases (Hendershott and Moulton (2009)).

Identifying Liquidity Externalities

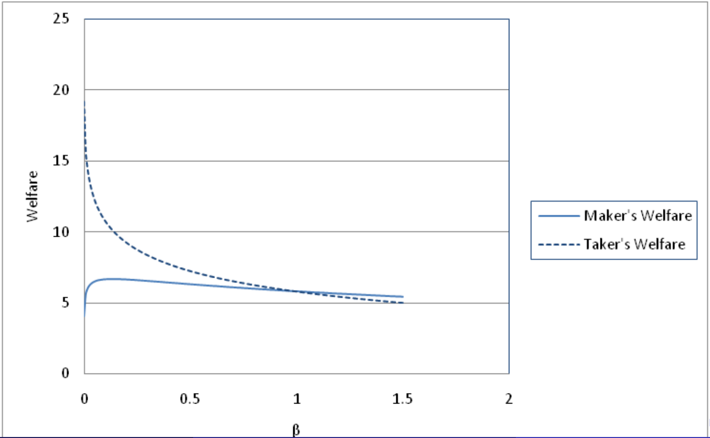
- A new type of liquidity externality: the monitoring decisions of the market-making side and the market-taking side reinforce each other.
- Barclay and Hendershott (2004): The empirical identification of the magnitude of liquidity externalities is challenging.
- The problem is to show a causal effect.
- Our model suggests several variables that can serve as instruments.
- Example: Exogenous shocks in the number of market-makers, their monitoring costs, and make-fees have a direct effect on trading activity of market-makers, but only an indirect effect on the trading activity of market takers.

Empirical Implication: Duration Clustering

- Cross-side complementarities imply clustering in time of orders and trades.
- Duration clustering is a major issue in financial econometrics (Engle and Russell (1998)).
- Example: In equilibrium, an increase in the number of market participants on ONE side triggers:
  1. An increase in aggregate monitoring intensities of BOTH sides.
  2. A decrease in average response time of BOTH sides.
- Result: Positive correlation between durations.
- Different from explanations of clustering in trading activity based on asymmetric information (e.g. Easley and O'Hara (1992)).

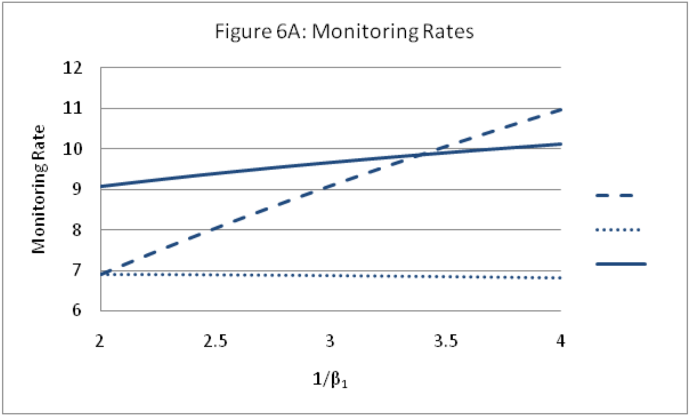
Algorithmic Trading and Welfare

- Trivial effect: Algo trading increases the rate at which gains from trade are being realized.
- Less trivial: In the face of fees, algo trading can be a mixed blessing to those who invest in technology.



Fast and Slow Traders

- Latencies often differ across traders because some have a technological edge over others.
- Co-location - Reduces latencies by positioning a trader's computers in close physical proximity to a trading platform's computer.



Make/Take Fees and the Tick-Size

- When the tick size is $\Delta(\ell) = \frac{\Gamma}{2\ell}$, the grid of possible prices is

$$P_\ell \equiv \{v_0 + n\Delta(\ell) : n \text{ is an integer}\}.$$

- Price on the grid is determined by the market power of makers vs. takers.
- **Finding 7**: When the tick size is zero ($\ell = \infty$), the make/take fee breakdown has no effect on monitoring decisions and on the trading rate.
- **Finding 8:** When the tick size is positive, the results from the baseline model hold. Make/take fees matter.

Make/Take Fees and the Tick Size - Example

	Tick Size = \$1/8		Tick Size = \$1/16		Tick Size = \$1/100	
	Uniform Pricing	Optimal Pricing	Uniform Pricing	Optimal Pricing	Uniform Pricing	Optimal Pricing
Trading Rate (shares/second)	330.6	355.7	330.6	346.4	330.6	333.4
Platform's Profits (\$/second)	0.3306	0.3557	0.3306	0.3464	0.3306	0.3334
Makers' Welfare (per second)	42.3	57.2	42.3	49.9	42.3	43.5
Takers' Welfare (per second)	42.8	34.4	42.8	39.2	42.8	42.3
Total Welfare (per second)	85.4	92.0	85.4	89.5	85.4	86.1
Estimated Annual Platform Fee Revenue (\$ in millions)	3868	4162	3868	4053	3868	3901

Summary of Main Findings

- Cross-side Complementarities - Multiplicity of equilibria. Unstable no-trade equilibrium as opposed to stable equilibrium with trade.
- Make-take fees matter as long as the tick size is not zero: they affect the incentives of traders to monitor and participate in the market. They indirectly affect the trading rate, market liquidity, and traders' welfare.
- Make-take fees are used to balance liquidity demand and supply.
- Algorithmic trading: (i) Burst in the trading rate; (ii) Higher total welfare but ambiguous effect on individual traders; (iii) Ambiguous effect on the bid-ask spread.
- Current fee structure is consistent with a relatively small number of electronic market-makers compared to the number of market-takers.
- New source of liquidity externalities and duration clustering.

Future Research

- Informed trading. Exposure to informed trading reduces market-makers' expected profits and should therefore lead to smaller make fees, other things being equal.
- Endogenize the number of makers and takers by allowing traders to choose their side.
- Competition between exchanges. The economic forces analyzed in our paper should still hold in a multi-market environment as long as monitoring is costly. Inter-market competition may add other considerations.