Market Fragmentation and the Latency Arms Race

Chris Rockwell, Erik Brinkman,
Elaine Wah, and Professor
Michael P. Wellman

High Frequency Trading (HFT) & Latency Arbitrage (LA)

HFT:

- Rapid trading of securities by computers
- Controls over 50% of trading volume today[1]

Latency Arbitrage:

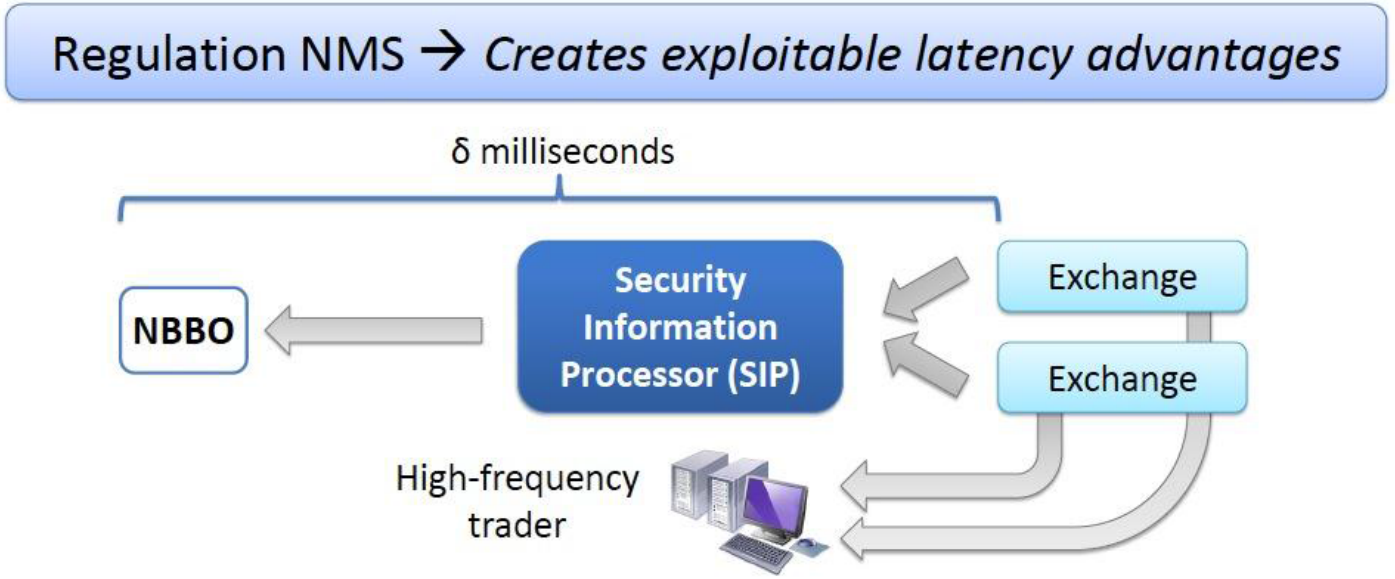
- Type of HFT taking advantage of speed advantages & delay across markets to generate nearly risk-free profit
- Currently accounting for an estimated $21 billion in profit per year[2]

Market Fragmentation

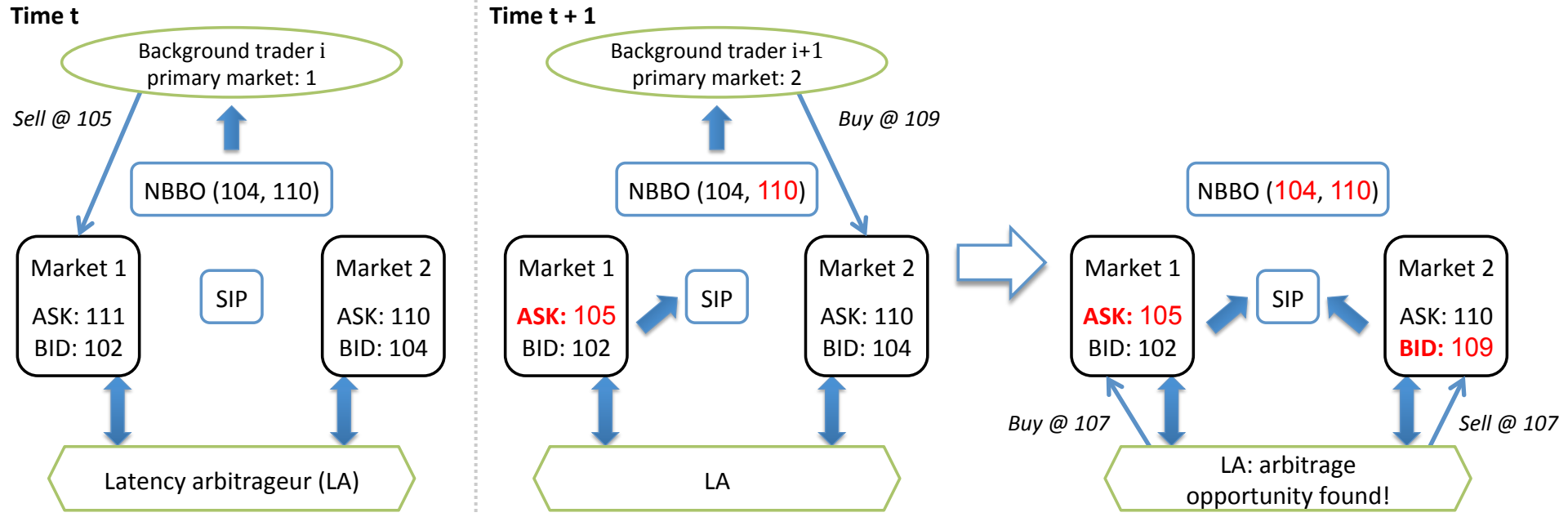
- Over 40 trading venues in U.S. for stocks, each with its own highest bid & lowest sell price
- Implies possibility of price disparity across markets
- Mitigation? **Regulation NMS**: routes orders for best execution based on an aggregated best bid and sell price, the **National Best Bid and Offer (NBBO)**
- Routed via **Security Information Processor (SIP)**

Latency Arbitrage

- Traders face latency in their own market and in getting the NBBO
- Latency arbitrageurs face lower latency and can take advantage of price disparities between NBBO and single markets



Example


Time t
Background trader i
primary market: 1
Sell @ 105
NBBO (104, 110)
Market 1
ASK: 111
BID: 102
SIP
Market 2
ASK: 110
BID: 104
Latency arbitrageur (LA)
Time t + 1
Background trader i+1
primary market: 2
Buy @ 109
NBBO (104, 110)
Market 1
ASK: 105
BID: 102
SIP
Market 2
ASK: 110
BID: 104
LA
NBBO (104, 110)
Market 1
ASK: 105
BID: 102
SIP
Market 2
ASK: 110
BID: 109
Buy @ 107
Sell @ 107
LA: arbitrage
opportunity found!

Previous Work & Our Model

- Elaine Wah's Simulation[3]
  - Two market model
  - Focused on LA effects on surplus and liquidity
- Our Simulation: similar
  - Include delay for latency arbitrageurs
  - Include delay for agents to local markets
  - Multiple latency arbitrageurs
  - Focus on reactions of competing HFTs

Competition: Latency Arms Race

- **Latency Arms Race:** a scenario where HFTs try to keep reducing their latencies until they approach 0
- We are interested in modeling this by looking at revenues in our model to predict how arbitrageurs value reducing latency
- In reality, HFT firms spend millions of dollars a year investing in technology to reduce latency

Methodology

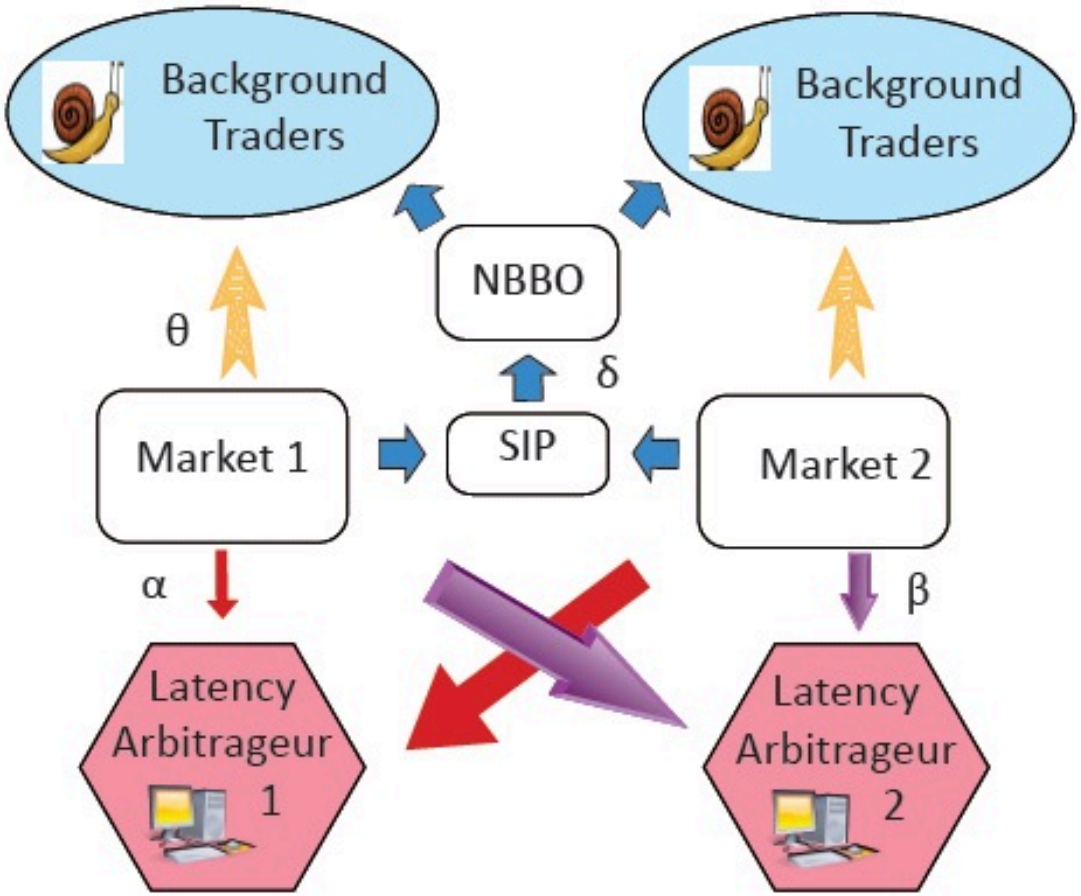
Agent-Based Modeling

- Allows us to specify agent behavior individually → overall market behavior can change over time
- Particularly conductive for modeling interactions between traders, exchanges, and the SIP

Discrete-Event Simulation

- Facilitate isolation of relationship between fragmentation, clearing rules, and latencies
- Allow variable latency of information access for different agents to different markets

Model



Note: Orders are routed to the market offering the best execution, based on a comparison of the NBBO and prices in the alternate market.

Background Traders

- Access to NBBO with latency δ
- Access to their market with latency $\theta < \delta$
- Trade based on private valuation of stocks, using simple strategies

Latency Arbitrageurs

- Access to all markets at a single latency α or $\beta < \theta$
- Arbitrage if market 1's highest buy order (BID) > market 2's lowest sell order (ASK) or vice versa

--Buy in market 1 & sell in market 2

Single-Equity Model

Captures:

- Communication latencies (between exchanges, information processors and traders)
- Current U.S. regulatory environment (order routing, Regulation NMS)
- Relationship between market fragmentation & latency arbitrage
- Competition between latency arbitrageurs

Experiments

- Simulate two-market model with two latency arbitrageurs
- Measure revenues at various latencies

Scenarios:

1. Both agents at equal latency
2. Agents at unequal latencies

- Focus on what reaction each HFT will take to different situations and how these decisions affect shared revenue of agents
- Use **Empirical Game Theoretic Analysis (EGTA)** to ensure fundamental agents respond appropriately to varying HFT strategies

Hypothesis

- Predict HFTs face a prisoners dilemma in the form of a latency arms race
  - HFTs will invest in faster technology regardless of other firm's decision
  - Would be best off not doing this

Hypothesis: HFTs Will Invest

- If the other HFT **does not invest** in faster technology:
  - This HFT has incentive to invest to receive all of revenue from latency arbitrage, rather than half
- If the other HFT **does invest** in faster technology:
  - This HFT has incentive to capture half of revenue from latency arbitrage, rather than none.
  - If possible, it will speed up more as it now finds itself in the situation above

Hypothesis: Prisoner's Dilemma

- Assume cost of investing in technology is nonzero and revenue for arbitrageurs is not significantly correlated with latency when greater than zero
- Arbitragers are best off mutually agreeing not to invest in technology

	No Investment	Small Investment	Large Investment
No Investment	(½, ½)	(0,1)	(0,1)
Small Investment	(1,0)	(½, ½)	(0,1)
Large Investment	(1,0)	(1,0)	(½, ½)

A simplistic model estimating revenues for the two LAs. Note if we were looking at profit, having an investment would involve a higher cost and thus lower profit. Payoffs are a fraction of total revenue earned using latency arbitrage

Future Work

- Changing minimum bid increment to see effects
- More complex latency arbitrage strategies
- Look at different model structures
  - Call market vs. two CDA market[4]

Citations

- [1]According to Financial Times, 29 July 2009, 73% in “SEC runs eye over high-speed trading.”
- [2]D. Schneider. The microsecond market. IEEE Spectrum, pages 66 - 81, June 2012.
- [3]Elaine Wah and Michael P. Wellman. Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model, February 2013
- [4]Eric Budish, Peter Cramton, and John Shim. The High-Frequency Trading Arms Race: Frequenct Batch Auctions as a Market Design Response, July ‘13

Acknowledgments

- This work was supported by NSF Award 1156469, for REU Site: Incentive-Centered Design. PI Erin Lea Krupka.