Latency Arbitrage, Market Fragmentation, and Efficiency: A Two-Market Model

Laina Chin
Terminology Review

- Latency arbitrage
- High Frequency Trading (HFT)
- Market fragmentation
- NBBO
- Continuous market vs. call market
Why is this happening?

- How bad is latency arbitrage?
- How would/could call markets make life better?
Why does this matter?

- One of the first *good* models of latency arbitrage
- Expose the harm of latency arbitrage
- Propose call market as a solution
THE EXPERIMENT

Dependent variables

- Allocative efficiency
  - Overall surplus/welfare
  - Main focus!

- Liquidity:
  - Availability of trade opportunities
  - Often measured by difference between bid and ask

- Volatility

- Price discovery
THE EXPERIMENT

Independent variables

- Presence of latency arbitrage
  - Latency arbitrageur vs. no latency arbitrageur

- Market fragmentation
  - Two-market model vs. centralized (one-market) model

- Market clearing rules
  - Continuous CDA vs. discrete call market

- Amount of latency
  - NBBO update time / length of clearing interval
RESULTS

Latency arbitrage

- Less surplus/lower welfare 😞

- Mixed effect on liquidity: slightly better execution times, larger difference between bid and ask
  - Fastest trade execution (most liquidity): 2-market model with LA
  - Slowest: 2-market model without LA

- Minimal effect on volatility

- Improves price discovery
RESULTS

Market fragmentation

- Single market:
  - More surplus than two-market model with LA
  - Less surplus than two-market model without LA

- Minimal effect on volatility
RESULTS

Market clearing rules

- Call markets ALWAYS way more efficient than continuous
In conclusion…

Call markets rule, latency arbitrage drools

- How could this experiment be improved? Were there any clear limitations?
- What are some different extensions could be done?
- Is it possible to have a 3+ market model? How might that change things, if at all?