Limit order book markets: a queueing systems perspective

Costis Maglaras

(Acknowledgments: Ciamac Moallemi, Hua Zheng, Arseniy Kukanov, Praveen Sharma, John Yao)

Columbia Business School

May 2015 | CFM – Imperial Distinguished Lecture Series

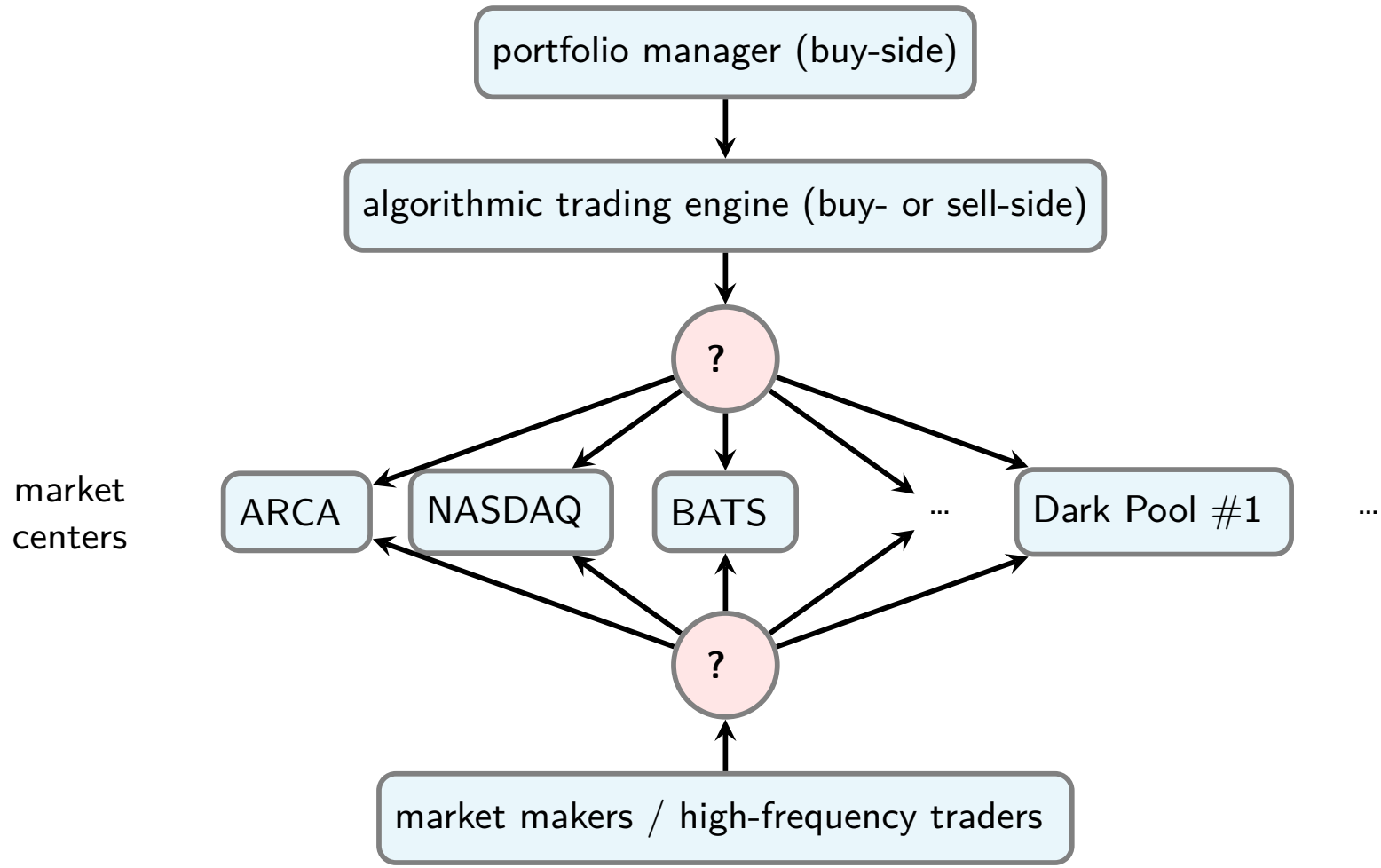
Outline

- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

Overview of algorithmic trading and limit order book markets

1. Overview of algorithmic trading
   - high level view of equities execution ecosystem
   - algorithmic trading systems
   - trade scheduling and the role of market impact models
   - tactical execution in a LOB
   - fragmentation, internalization, incentives, ...

Simplified view of trading



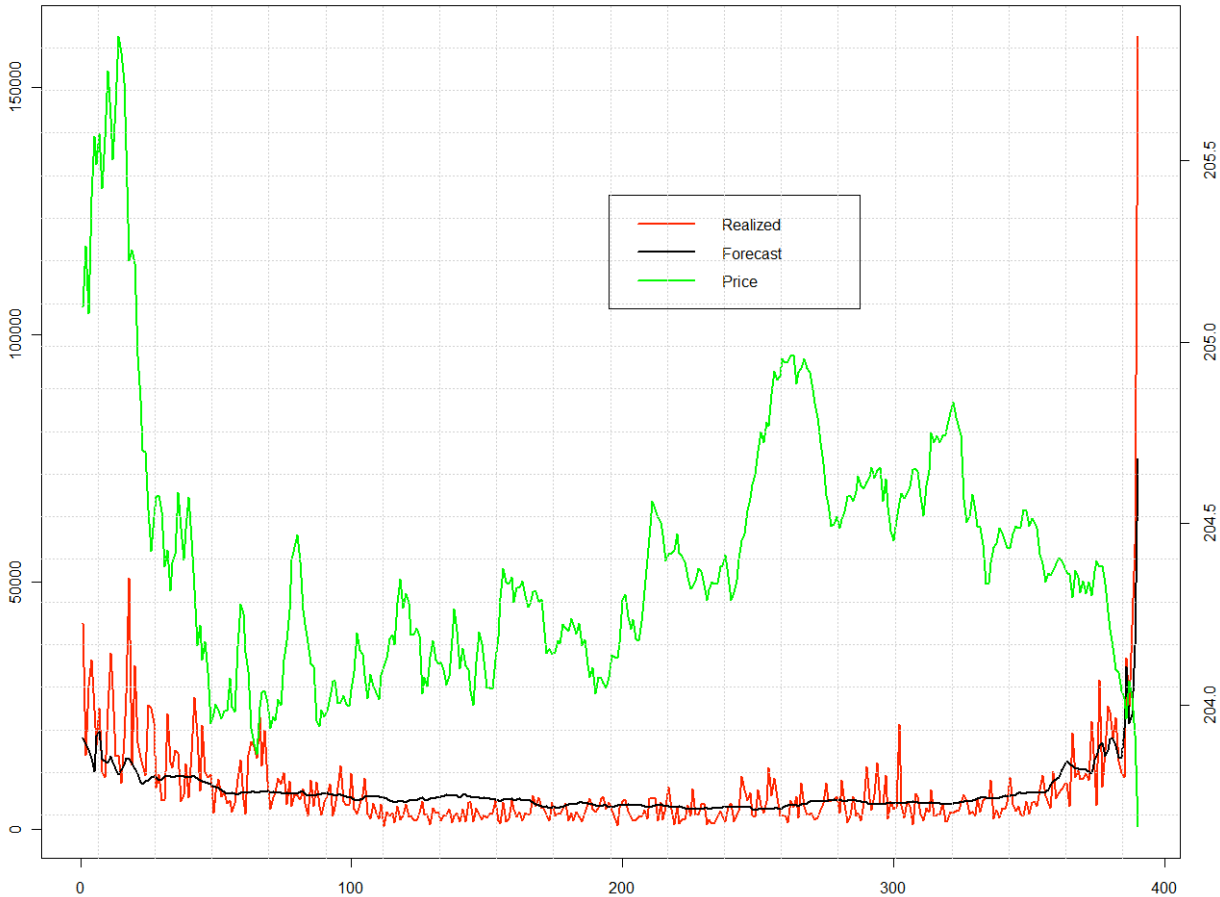
Modern U.S. Equity Markets

- Electronic
- Decentralized/Fragmented
  NYSE, NASDAQ, ARCA, BATS, Direct Edge, ...
- Exchanges ($\sim$ 70%)
  electronic limit order books (LOBs)
- Alternative venues ($\sim$ 30%)
  ECNs, dark pools, internalization, OTC market makers, etc.
- Participants increasingly automated
  - institutional investors: "algorithmic trading"
  - market makers: "high-frequency trading" ($\sim$ 60% ADV)
  - opportunistic/active (price sensitive) investors: "aggressive/electronic"
  - retail: "manual" ($\sim$ 5% ADV; small order sizes)

An example

- How should you buy 250,000 shares of IBM stock between 12:30pm and 4:00pm?
  - Is this order "large"?
  - How fast should you trade? When?
  - How much will it cost you?
  - Who are you trading against?
- How is it done in practice?

Example cont.



– Forecasted Volume 12:30-4pm = 1,525,000 shares; Avg spread = $.04 (1.95bps);
– Expected Market Impact (1230-4pm) ≈ 20bps ≈ 40 pennies/share
– Expected MI (1230-130pm) ≈ 28bps ≈ 56 pennies/share

Institutional traders (broad strokes)

- investment decisions & trade execution are often separate processes
- institutional order flow typically has "mandate" to execute
- trader selects broker, algorithms, block venue, ...
  (algorithm $\approx$ trading constraints)
- main considerations:
  – "best execution"
  – access to liquidity (larger orders)
  – short-term alpha (discretionary investors)
  – information leakage (large orders the spread over hrs, days, weeks)
  – commissions (soft dollar agreements)
  – incentives (portfolio manager & trading desk; buy side & sell side)
- execution costs feedback into portfolio selection decisions & fund perf
- S&P500:
  – ADV $\approx$ <1% MktCap (.1% – 2%)
  – Depth (displayed, top of book) $\approx$ .1% ADV
  – Depth (displayed, top of book) $\approx 10^{-6} - 10^{-5}$ of MktCap
  $\Rightarrow$ orders need to be spread out over time

Market Makers & HFT participants (broad strokes)

- supply short-term liquidity and capture bid-ask spread capture
  mostly intraday flow; limited overnight exposure
- small order sizes $\sim$ depth; short trade horizons / holding periods
- profit $\approx$ (captured spread) - (adverse selection) - (TC)
  - critical to model **adverse selection**:
    short term price change conditional on a trade
- important to model short term future prices ("alpha"):
  - microstructure signals (limit order book)
  - time series modeling of prices (momentum; reversion)
  - cross-asset signals (statistical arbitrage, ETF against underlying, ...)
  - news (NLP)
  - detailed microstructure of market mechanisms
    ...
- risks: adverse price movements; flow toxicity; accumulation of inventory
  & aggregate market exposure

My focus is on limit order book dynamics

Limit order book behavior affects:

- algorithmic trade execution systems & performance
- trading signals & execution for MMs
- key element of modern market microstructure over short time horizons
- regulatory implications

Queueing behavior plays an important role in short-term market dynamics

. . . the specific lens of these lectures

Overview of algorithmic trading and limit order book markets

1. Overview of algorithmic trading
   - high level view of equities execution ecosystem
   - algorithmic trading systems
   - trade scheduling and the role of market impact models
   - tactical execution in a LOB
   - fragmentation, internalization, incentives, ...

Algorithmic Trading Strategies (90+% of institutional flow)

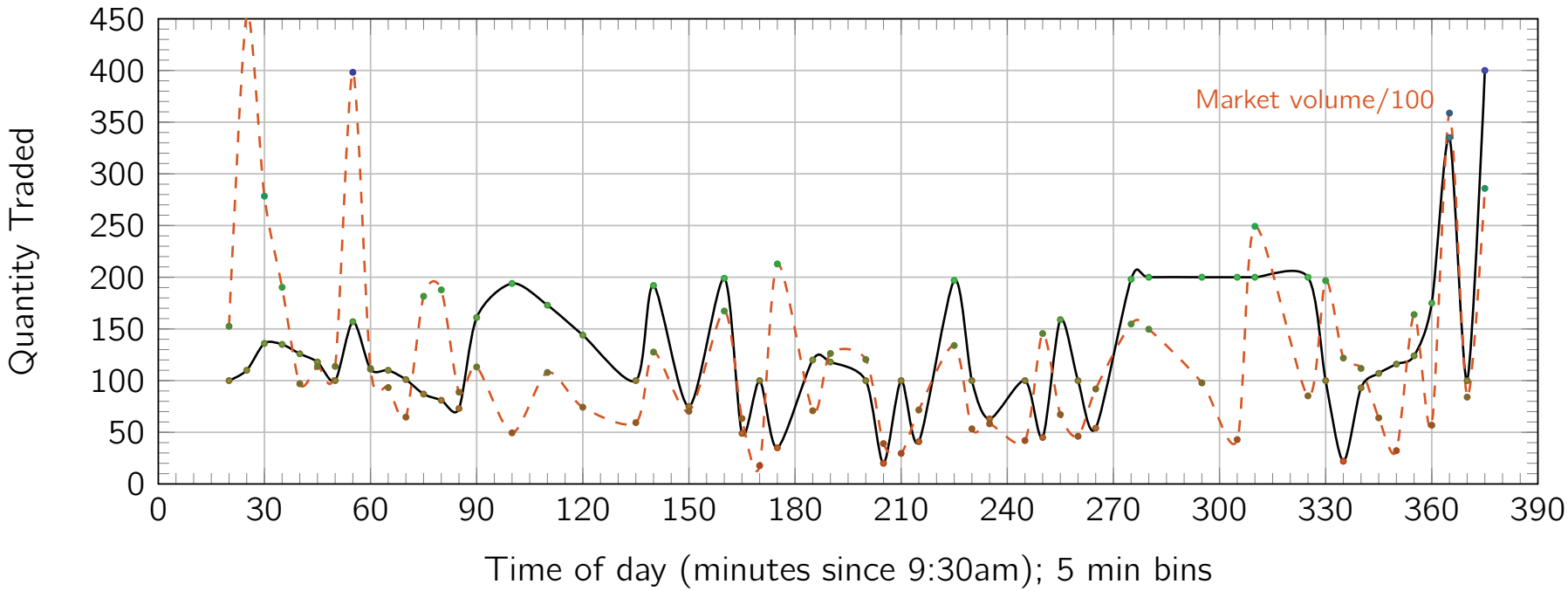
- **VWAP (Volume-Weighted-Average-Price):** trades according to forecasted volume profile to achieve (or beat) the market's volume weighted average price
  - Passive strategy; subject to significant market risk

- **TWAP (Time-Weighted-Average-Price):** trades uniformly over time to achieve (or beat) TWAP benchmark
  - Passive strategy; market risk; not very popular in practice

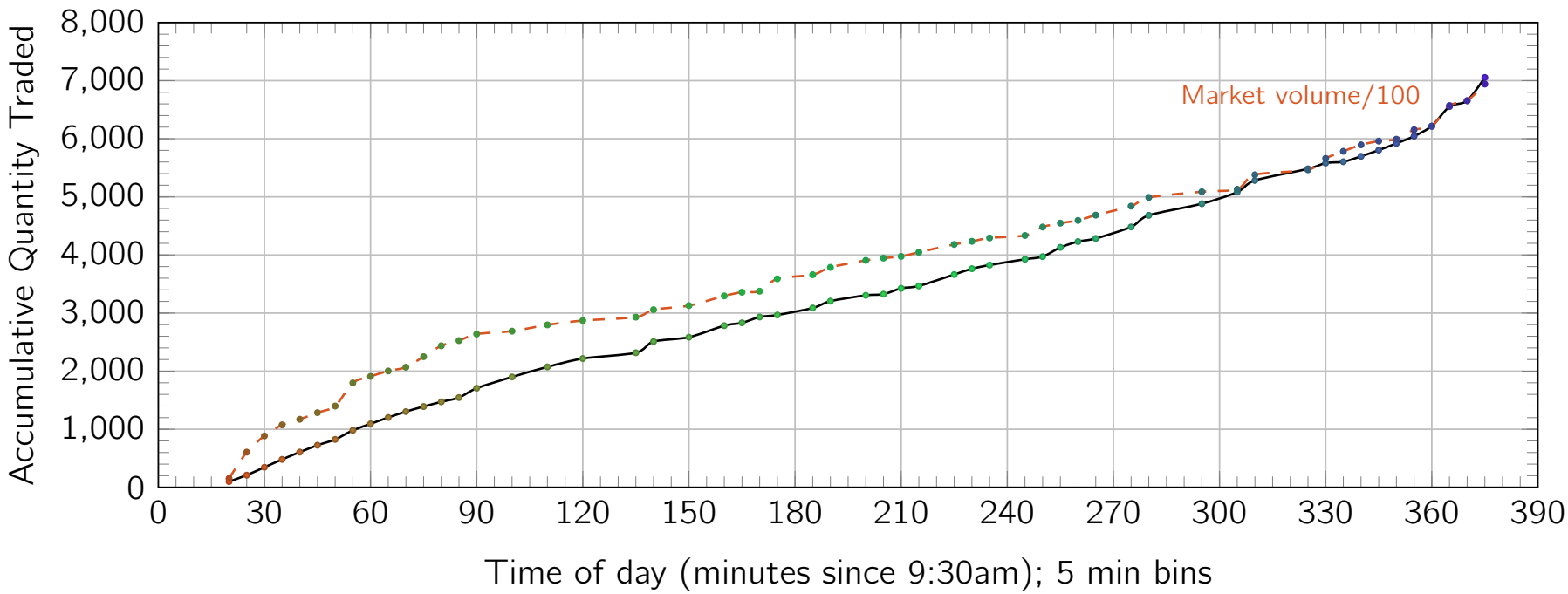
- **POV (Percent-of-Volume):** Executes while tracking the realized volume profile at a target participation rate, e.g., buy IBM at 15% part. rate
  - Controls behavior during volume spikes to avoid excessive cost
  - Popular in practice $\sim$ 5%-30% part.rates; (part.rate $\sim$ cost)

- **IS (Implementation Shortfall):** schedules trade so as to optimally tradeoff expected shortfall (cost) against execution risk
  - variable execution speed; adapts wrt changes in mkt conditions
  - Popular, especially with portfolios where cost/risk tradeoff is intricate

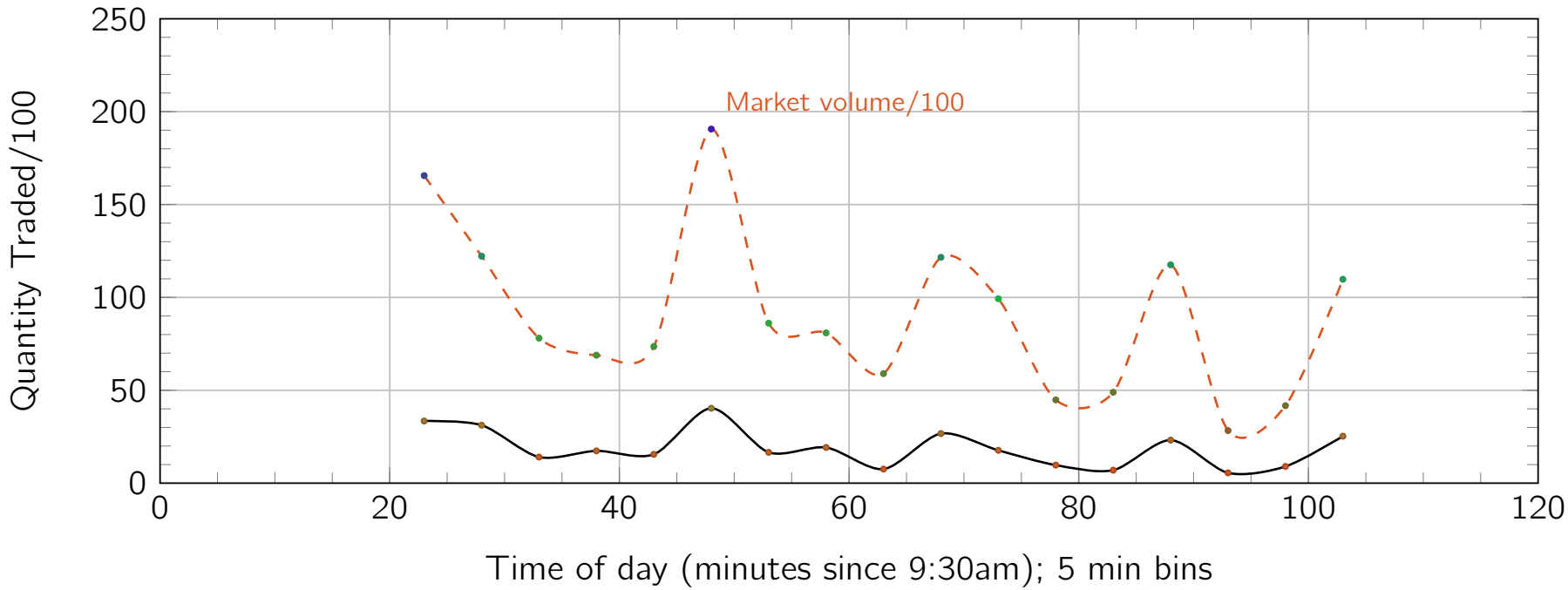
VWAP, XLY, 07/22/2013 ($\approx .15\%ADV$)



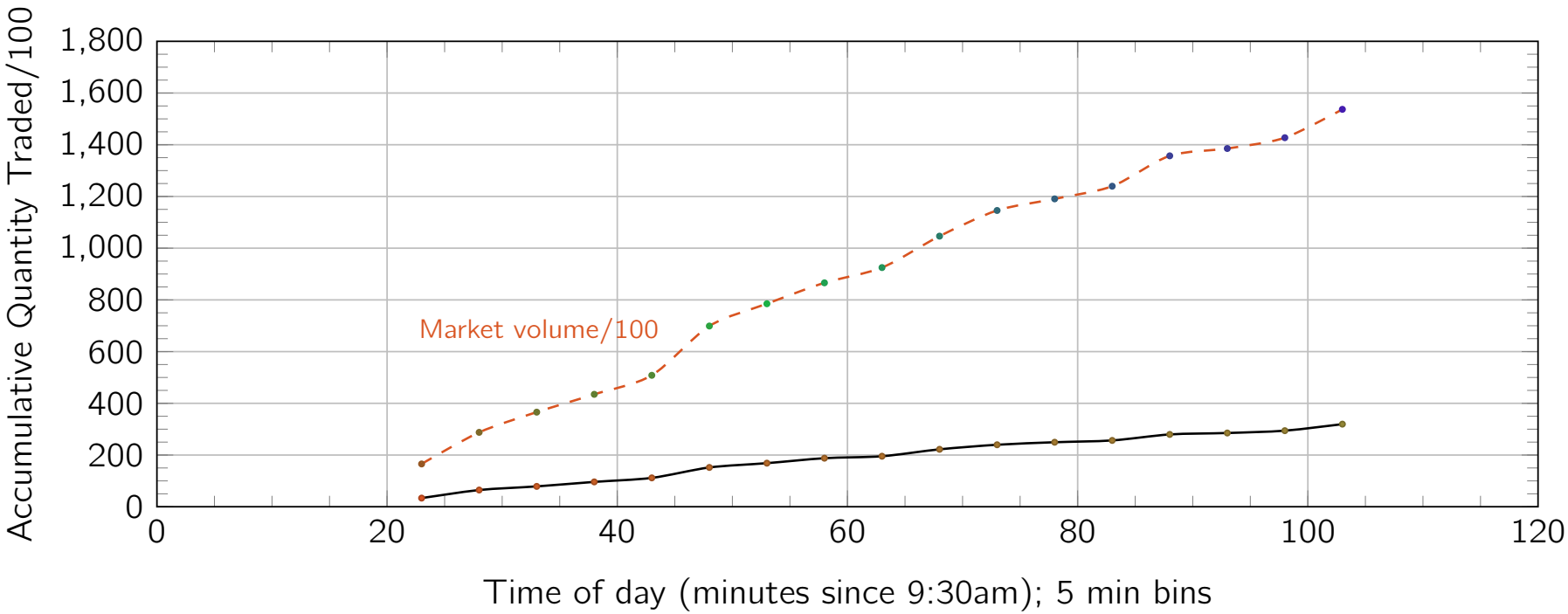
VWAP, XLY, 07/22/2013 (cumulative quantity)



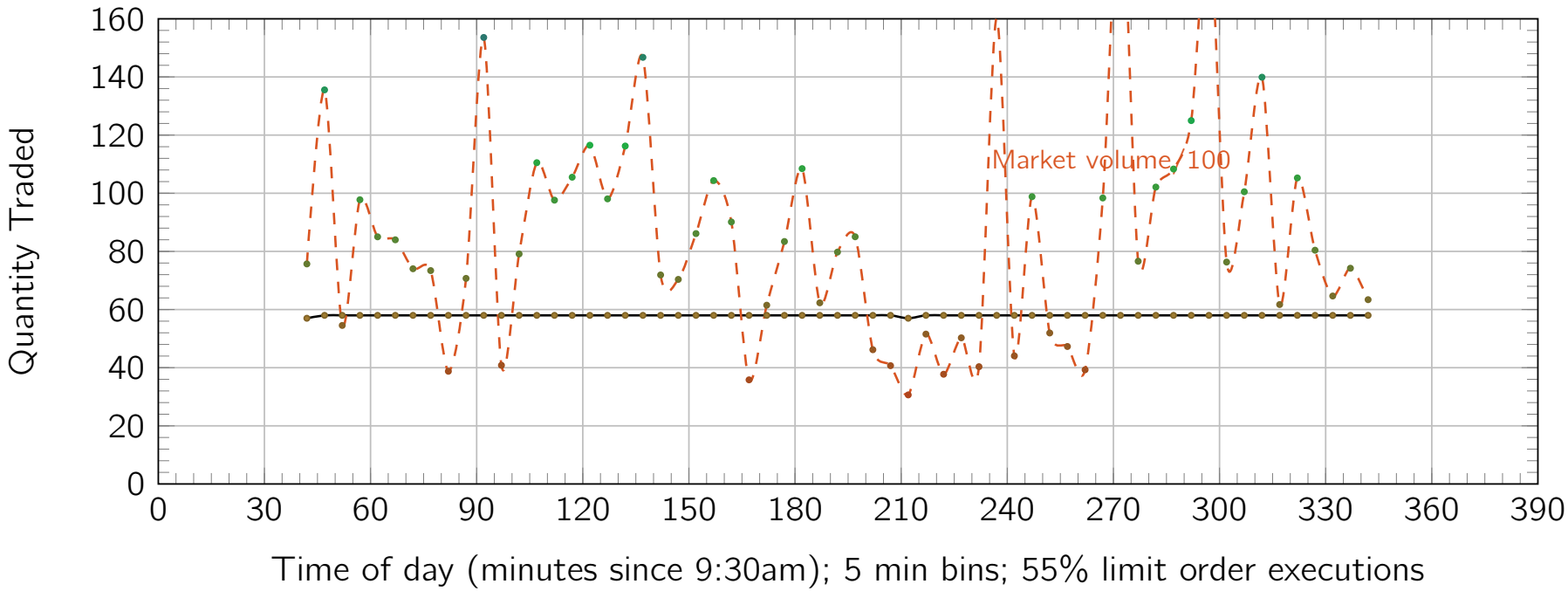
POV 20% ACT, 07/08/2013



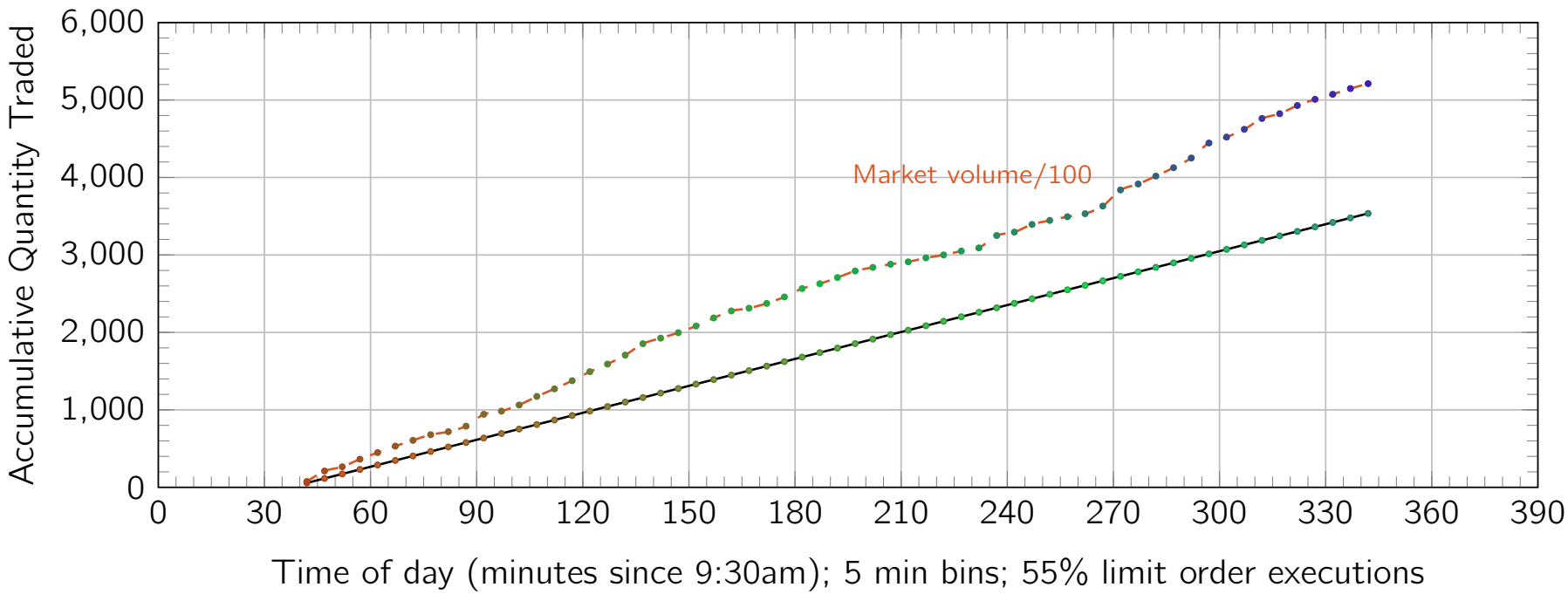
POV, ACT, 07/08/2013 (cumulative quantity)



Schematic of execution profiles: TWAP, XLY, 07/02/2013



TWAP, XLY, 07/02/2013 (cumulative quantity)



What is the high-level architecture of such a system?

Algorithmic Trading Systems: typically decomposed into three steps

- **Trade scheduling:** splits parent order into $\sim$ 5 min "slices"
  – relevant time-scale: minutes-hours
  – schedule follows user selected "strategy" (VWAP, POV, IS, . . .)
  – reflects urgency, "alpha," risk/return tradeoff
  – schedule updated during execution to reflect price/liquidity/. . .

- **Optimal execution of a slice ("micro-trader"):** further divides slice into child orders
  – relevant time-scale: seconds–minutes
  – strategy optimizes pricing and placing of orders in the limit order book
  – execution adjusts to speed of LOB dynamics, price momentum, ...

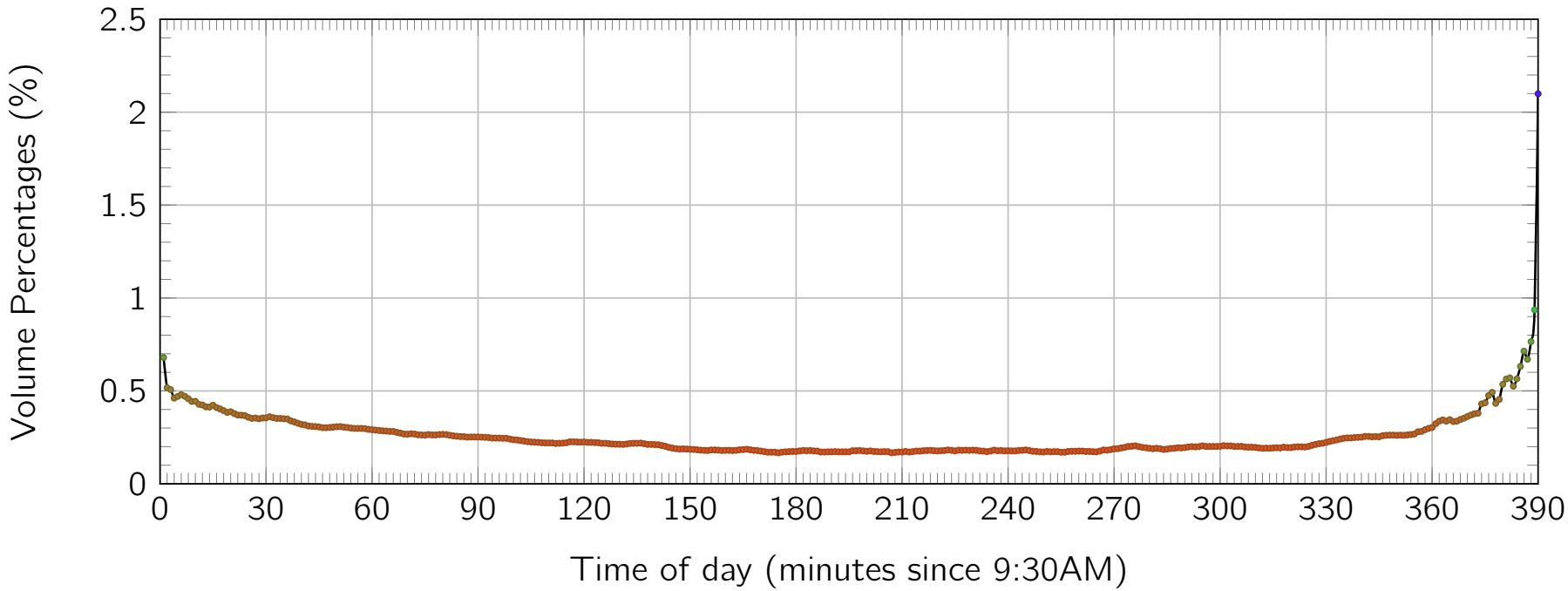
- **Order routing:** decides where to send each child order
  – relevant time-scale: $\sim$ 1–50 ms
  – optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of 2nd and 3rd steps mostly technological/historical artifact
(should not be treated separately)

Algorithmic Trading Systems: basic building blocks

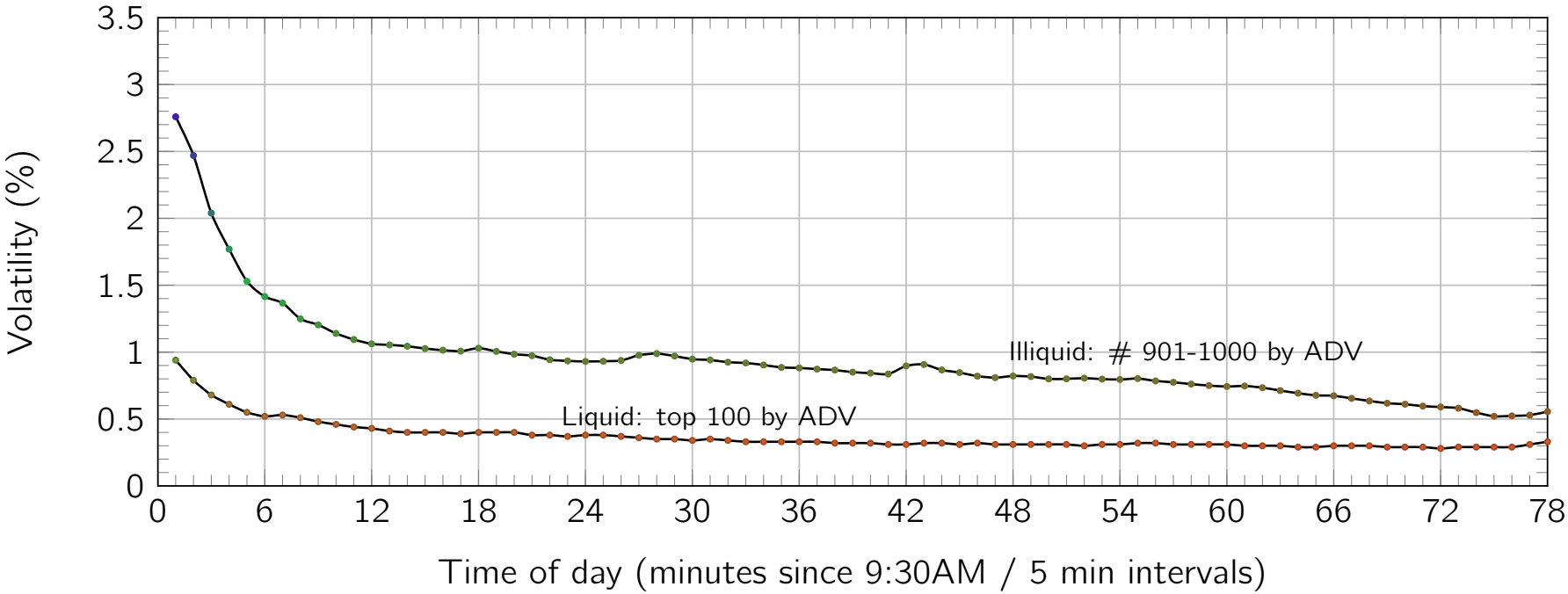
- forecasts for intraday trading patterns
  - volume
  - volatility
  - bid-ask spread
  - market depth
  - ...
- real-time market data analytics
- market impact model (more on this tomorrow)
- risk model
  - "of the shelf" risk models calibrated using EOD closing price data do not incorporate intraday correlation structure
  - intraday data? (tractable for liquid securities, e.g., S&P500 universe)
  - cross-asset liquidity model & market impact model

Intraday volume profile: cross-sectional average of S&P500



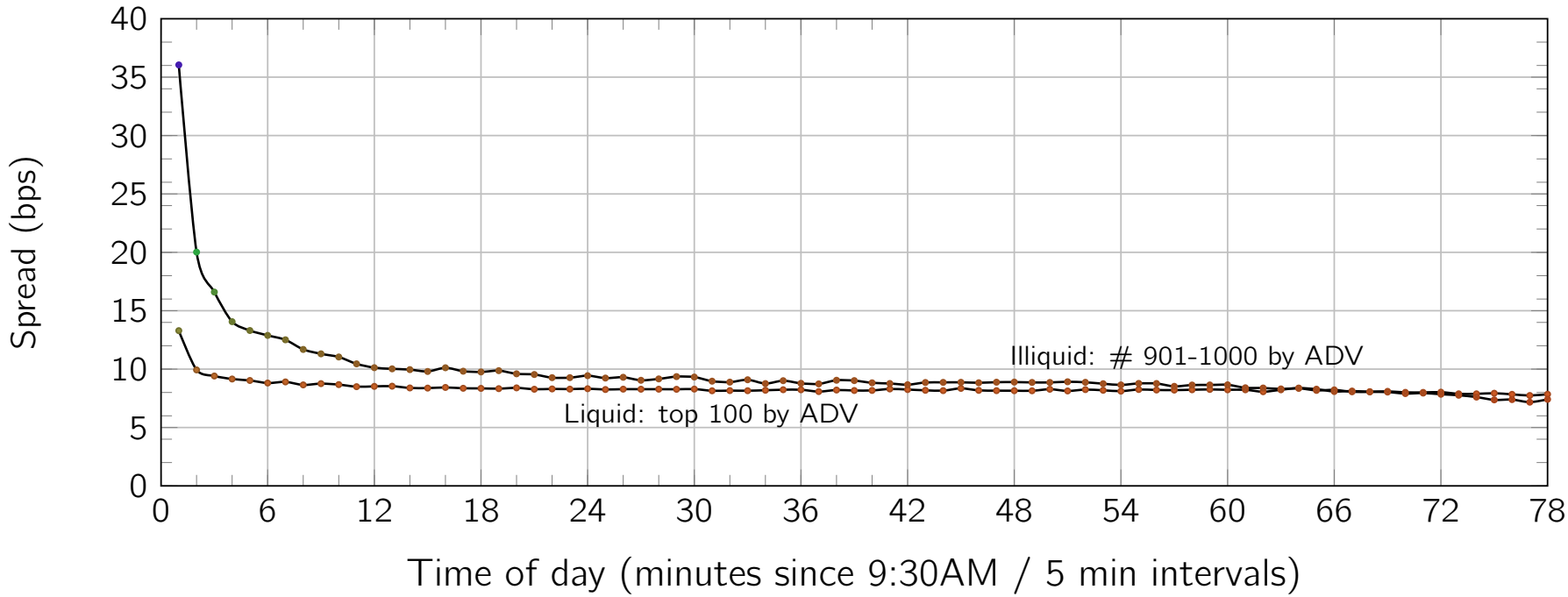
S&P500 cross-sectional, smoothed intraday trading volume profile (min-by-min).

Intraday volatility profile: cross-sectional averages



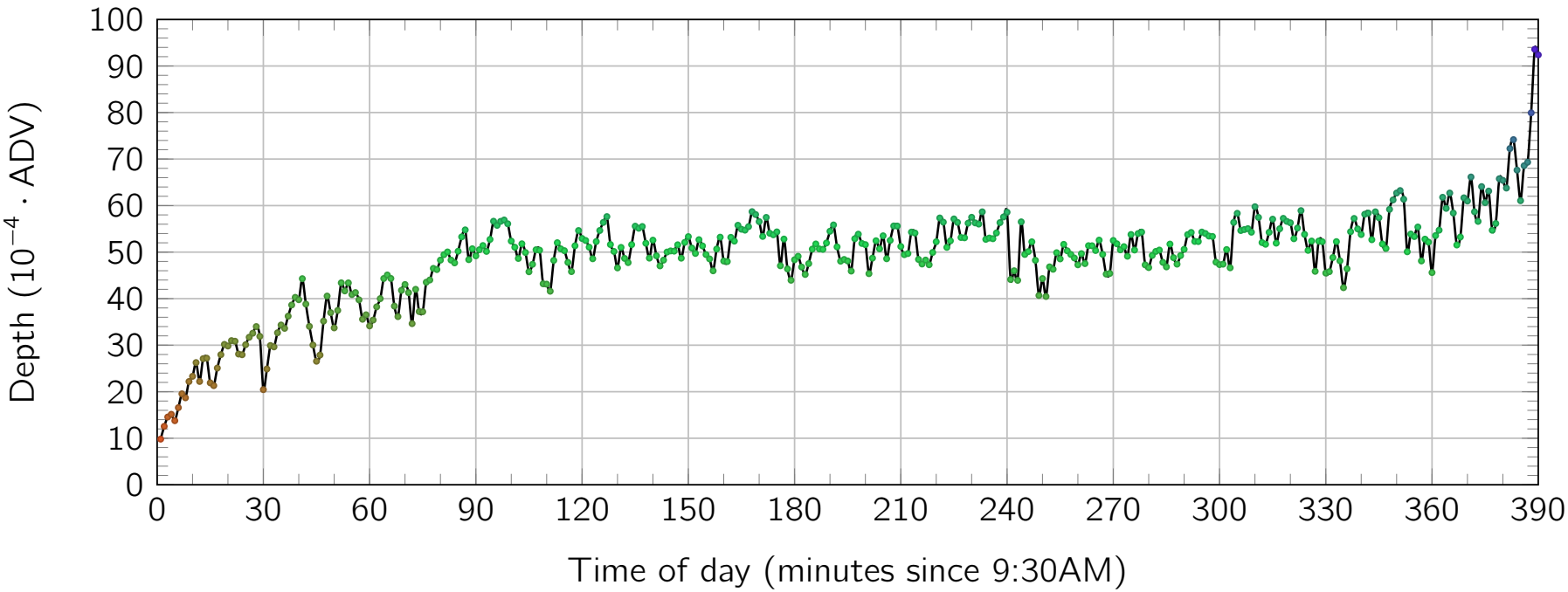
Cross-sectional averaged intraday volatility profiles for US equities.

Intraday spread profile: cross-sectional averages



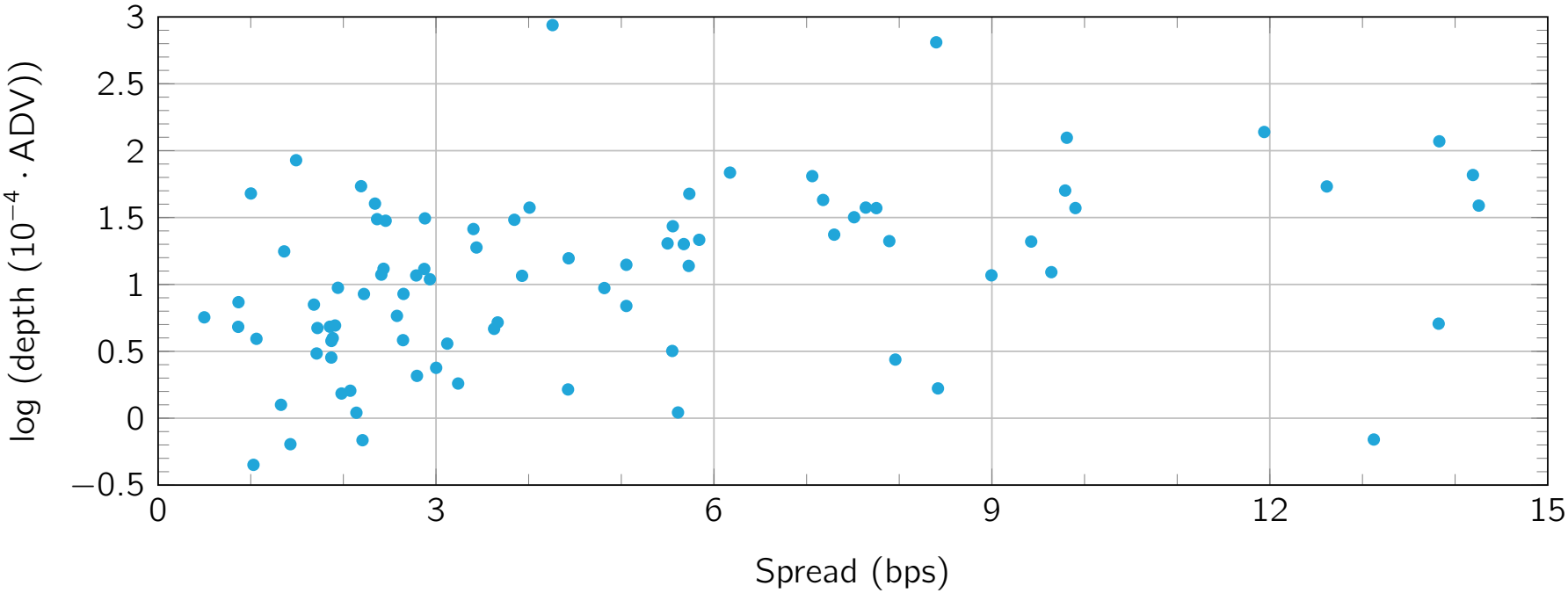
Cross-sectional averaged intraday spread profiles for US equities.

Intraday depth profile



US equities, top 100 securities wrt ADV, cross-sectional, intraday depth profile, in units of $10^{-4} \cdot$ ADV.

Log-depth as a function of spread (top 100 names ranked by ADV)



"Large tick" stocks:

- liquid & low priced stocks, spread $\approx$ \$0.01, but 1 spread = 5 – 15 bps
- depth $\nearrow$ as spread (in bps) $\uparrow$
  . . . capturing spread yields significant return

Algorithmic Trading Systems: trade scheduling

- **Trade scheduling:** splits parent order into $\sim$ 5 min "slices"
  – relevant time-scale: minutes-hours
  – schedule follows user selected "strategy" (VWAP, POV, IS, . . .)
  – reflects urgency, "alpha," risk/return tradeoff
  – schedule updated during execution to reflect price/liquidity/. . .

- **Optimal execution of a slice ("micro-trader"):** further divides slice into child orders
  – relevant time-scale: seconds–minutes
  – strategy optimizes pricing and placing of orders in the limit order book
  – execution adjusts to speed of LOB dynamics, price momentum, ...

- **Order routing:** decides where to send each child order
  – relevant time-scale: $\sim$ 1–50 ms
  – optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of 2nd and 3rd steps mostly technological/historical artifact
(should not be treated separately)

Trade Scheduling: VWAP

- $X(t)$= # shares traded in $[0, t]$ $\quad (X(0) = 0, \;\; X(T) = X$, RCLL, $\nearrow)$

$$\bar{p} = \frac{1}{X} \int_0^T p(t) dX(t)$$

- Benchmark: $L(t)$ cumulative traded volume in market in $[0, t]$

$$\bar{v} = \frac{1}{L(T)} \int_0^T p(t) dL(t)$$

- Control problem:

$$\text{choose } X(t) \text{ to} \qquad \min \;\; \bar{p} - \bar{v} \qquad \text{(for buy order)}$$

- Typical solution:
  - $L_f(t)$ = forecast $L(t)$ using k days of HF trading data
  - schedule trade according to $L_f(t)$
  - model $L(t)$; filter aggressive trades
  - adapt forecast to real time conditions & deviate opportunistically
  - incorporate tactical short-term alpha signals (sec to minutes)

Trade Scheduling: Implementation Shortfall

- Shortfall $S := \bar{p} \; - \; p_{\text{arrival}}$

- Fundamental tradeoff:
  - quick execution $\Rightarrow$ adverse price movement (market impact)
  - slow execution $\Rightarrow$ subject to price risk due to market movement

- Control problem (one possible variation): choose $X(t)$ to

$$\min \quad \mathbf{E}[S] \; + \lambda \mathbf{Var}[S]$$

  where $\lambda > 0$ is a risk aversion parameter

- Typical solution:
  - use a rolling horizon (MPC) control: at t, compute control for $[t, T]$
  - refine price impact estimates to real time conditions
  - adapt trading speed & order placing logic in real-time
  - incorporate tactical short-term alpha signals (sec to minutes)
- typical example: principle trading desks; transitions

Essential building block: market impact (price impact) model

- "Macro" model are variants of the following:

  Price impact = Temporary (:= $f(x(t))$) + Permanent (:= $h(x(t))$)

  $$\tilde{p}(t) = p(t) + f(x(t)) \quad \text{and} \quad p(t+1) = p(t) + h(x(t)) + \sigma(t)N(0,1)$$

  (above expression assumes Temp = Instantaneous; o/w we need convolution eqn . . .)

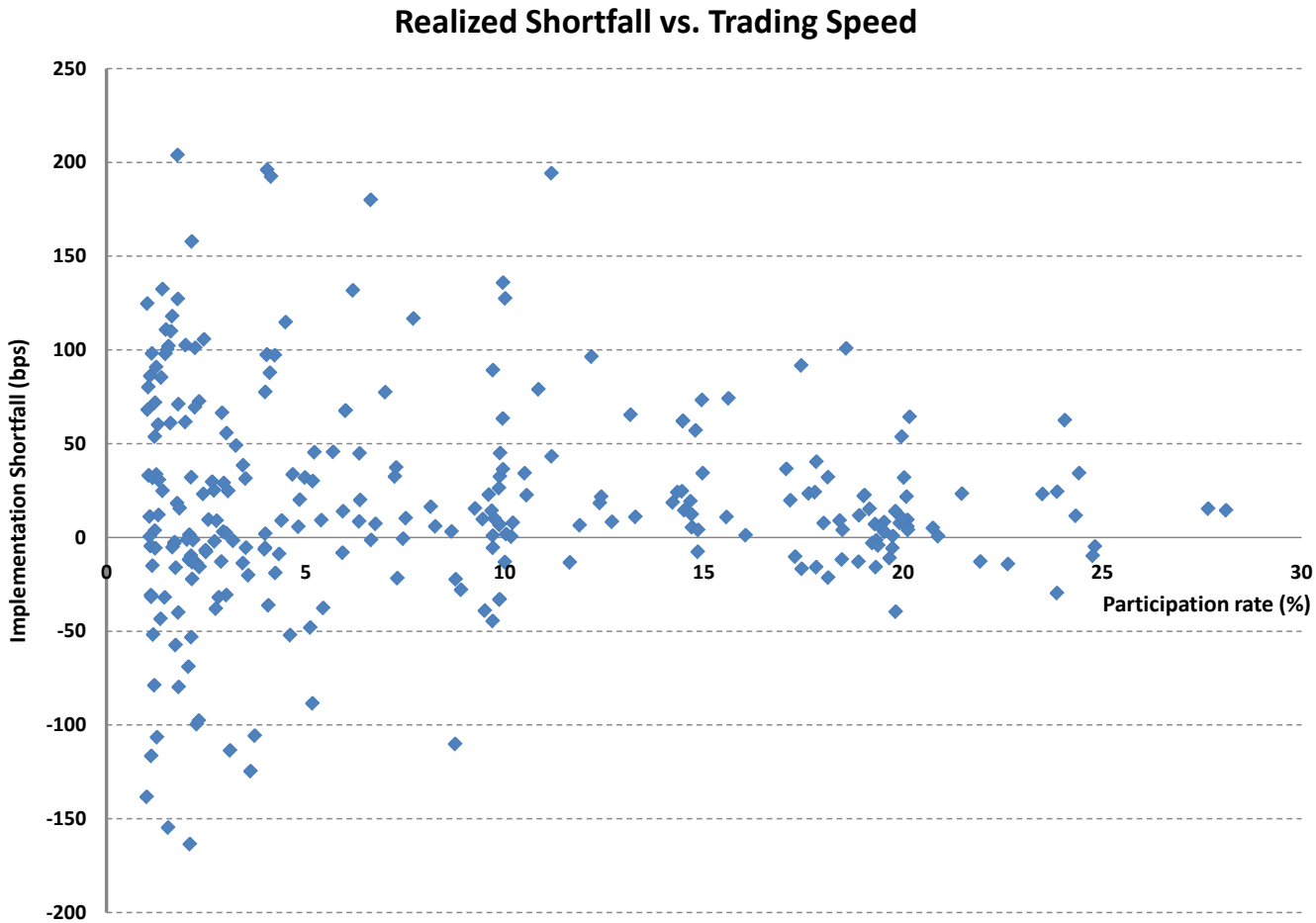
- No-arbitrage argument supports use of linear permanent price impact

- Estimation of MI coefficients via non-linear regression based on realized transaction costs of actual trades. Typical findings

  $$f(x(t)) = \alpha_{0,t} + \alpha_{1,t} \cdot s_t + \alpha_{2,t} x(t)^p, \quad p = 1 \text{ or } p = 1/2, 2/3, \ldots$$

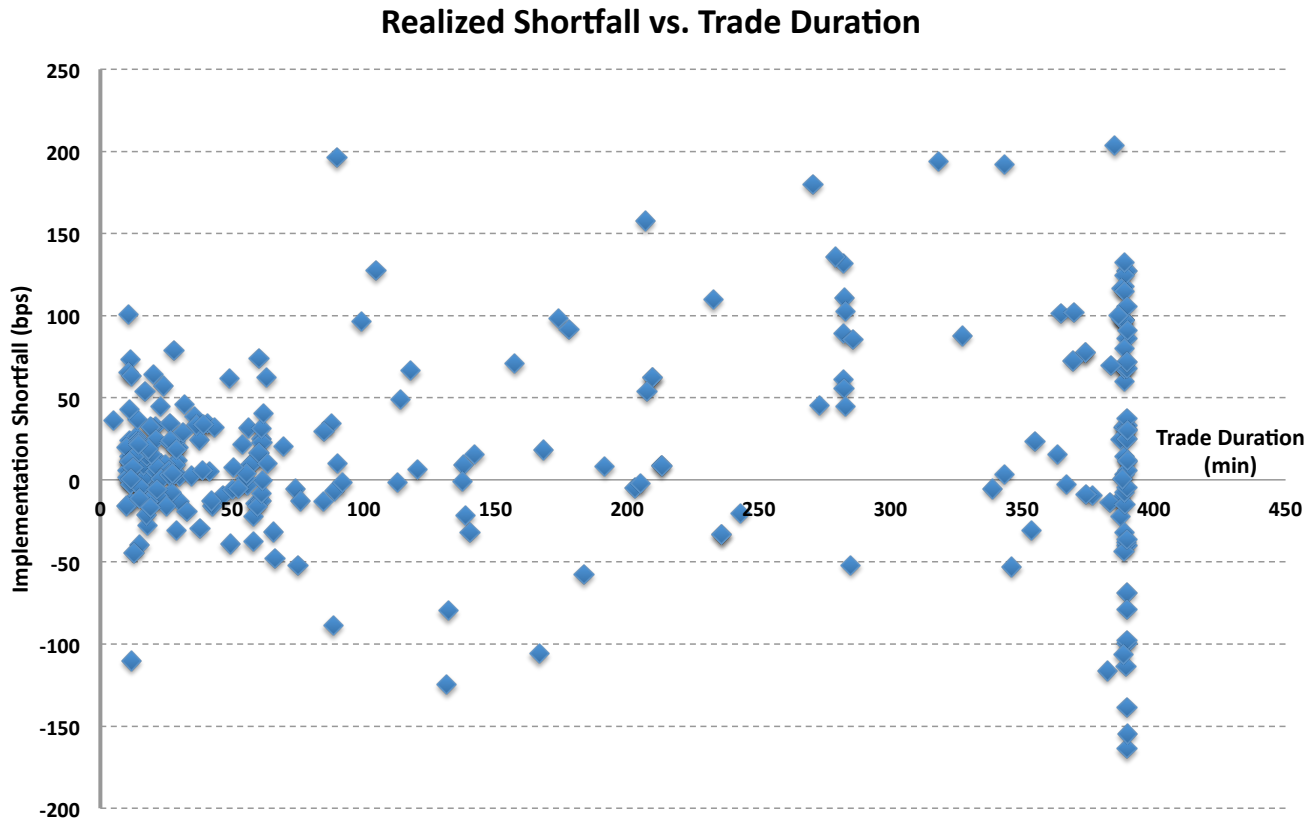
  – $\alpha_{i,t}$ depend on spread, tick size, volume, volatility, . . .
  – solving IS: QP for linear $f(\cdot)$ and SOCP for fractional p model
  – regression fits of market impact are "noisy" (more tomorrow)
  – alternate: impact function decays with time (also noisy to estimate)

Realized Shortfall for a sample of POV and VWAP orders



- model is more estimable for aggressive executions ($\geq$ 10% part rate)

Realized Shortfall for POV and VWAP orders (cont.)



- model is more estimable for slow duration orders ($\leq$ 30min)

Trade scheduling: key modeling and trading decisions

- accurate forecasts for intraday trading patterns
- market impact model (more on this tomorrow)
- mathematical formulation of trade scheduling problem could yield

  a) essentially open-loop (static) or
  b) feedback (adaptive) trade schedules
  (preferred given noisy reference input data)
- how much to trade over a period of k minutes, bounds on permissible deviations from plan, limit prices to control price impact, ...
- adapt to real-time conditions
- optimized control of portfolio executions

1. Overview of algorithmic trading
   - high level view of equities execution ecosystem
   - algorithmic trading systems
   - trade scheduling and the role of market impact models
   - tactical execution in a LOB
   - fragmentation, internalization, incentives, ...

Algorithmic Trading Systems: short horizon execution in LOB

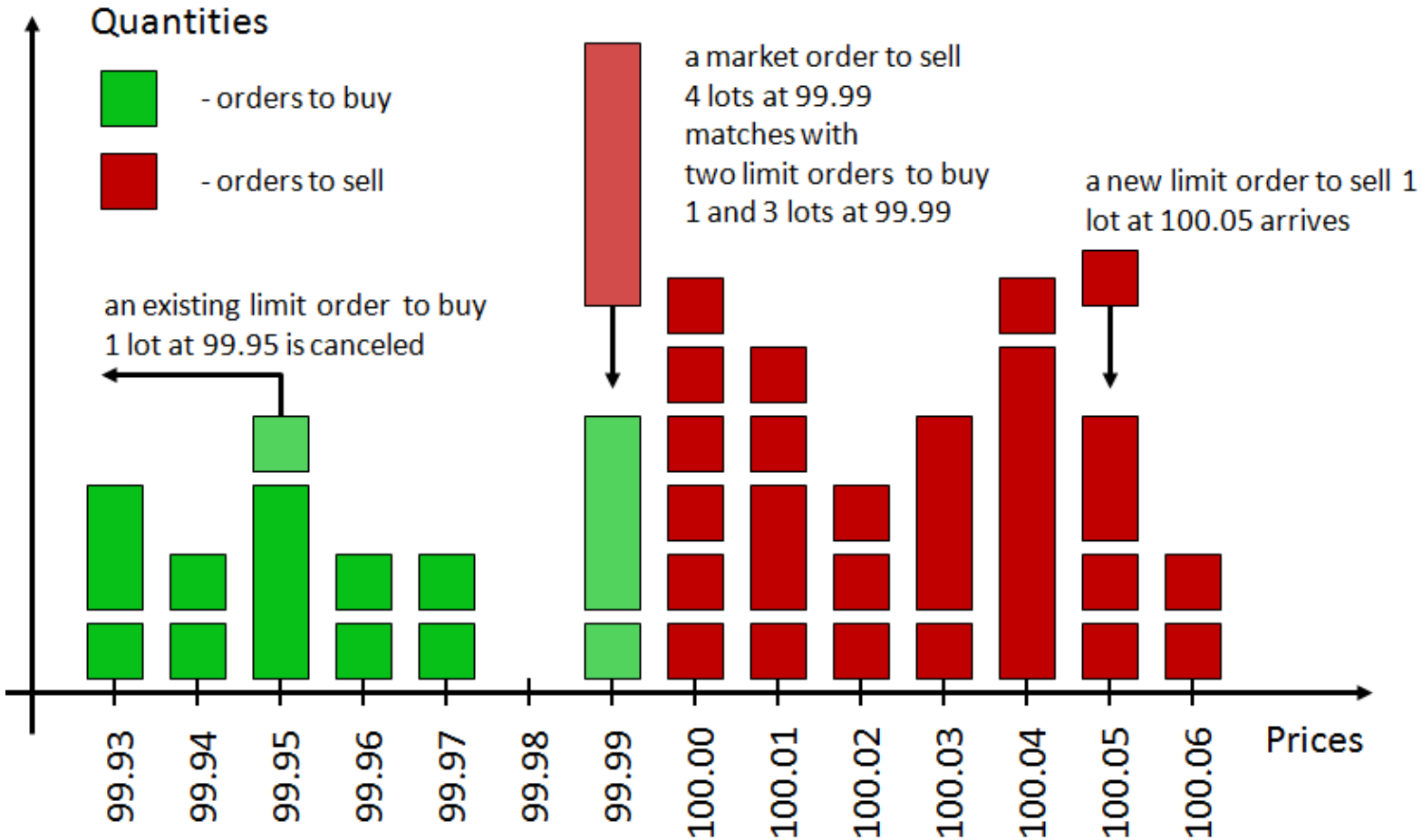
- **Trade scheduling:** splits parent order into $\sim$ 5 min "slices"
  – relevant time-scale: minutes-hours
  – schedule follows user selected "strategy" (VWAP, POV, IS, ...)
  – reflects urgency, "alpha," risk/return tradeoff
  – schedule updated during execution to reflect price, liquidity/...

- **Optimal execution of a slice ("micro-trader"):** further divides slice into child orders
  – relevant time-scale: seconds–minutes
  – strategy optimizes pricing and placing of orders in the limit order book
  – execution adjusts to speed of LOB dynamics, price momentum, ...

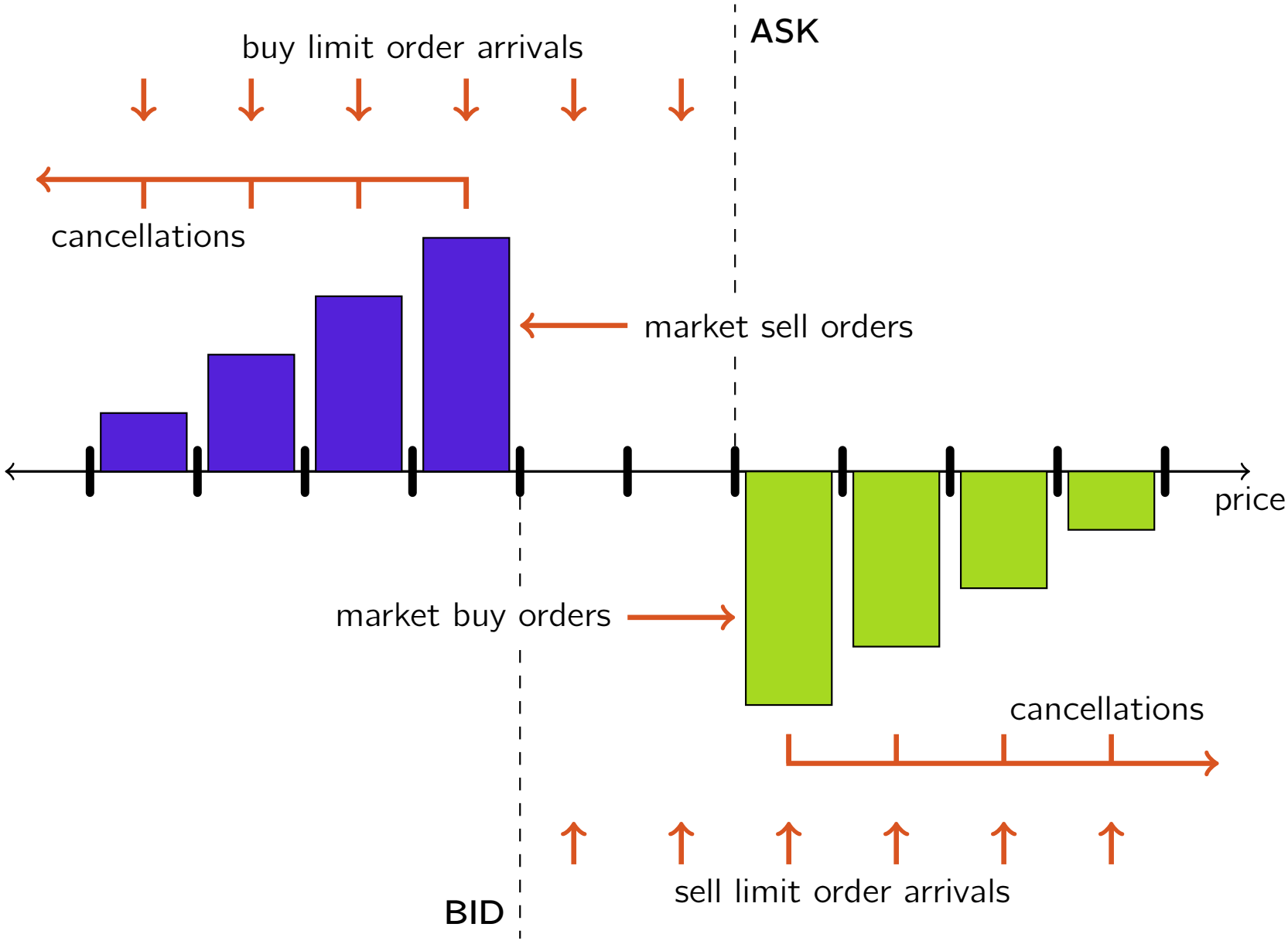
- **Order routing:** decides where to send each child order
  – relevant time-scale: $\sim$ 1–50 ms
  – optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of 2nd and 3rd steps mostly technological/historical artifact (should not be treated separately)

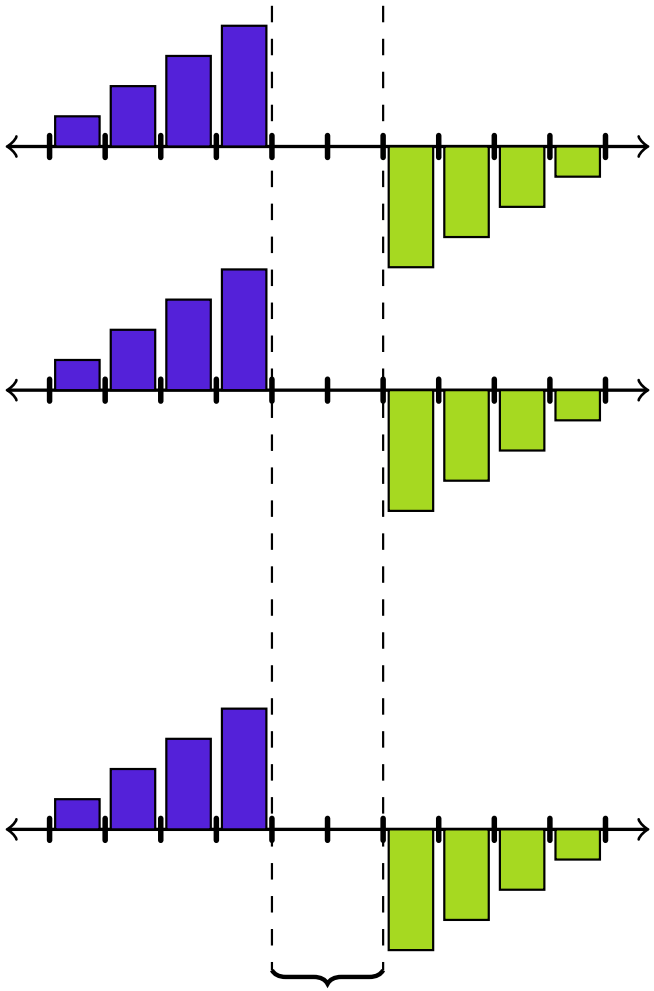
LOB schematic



The Limit Order Book (LOB)



Multiple Limit Order Books



Price levels are coupled through protection mechanisms (Reg NMS)

Execution in a LOB

What are the key considerations & decisions?

Execution in LOB: key modeling and trading decisions

- real-time measurements and forecasts for event rates
  (arrivals, trades, cancellations on each side of the LOB)
- heterogenous flows wrt arrivals, executions, cancellations (tomorrow)
- time/price queue priority:
  – estimate queueing delay & **P**(fill in T time units)
  – limit order placement . . . depends on queueing effects at each exchange
  – maintain / estimate queue position (& residual queueing delay)
  – adverse selection as a function of exchange, depth, queue position, . . .
  – transaction cost models
- microstructure, short-term alpha signals
- optimize execution price by tactically controlling

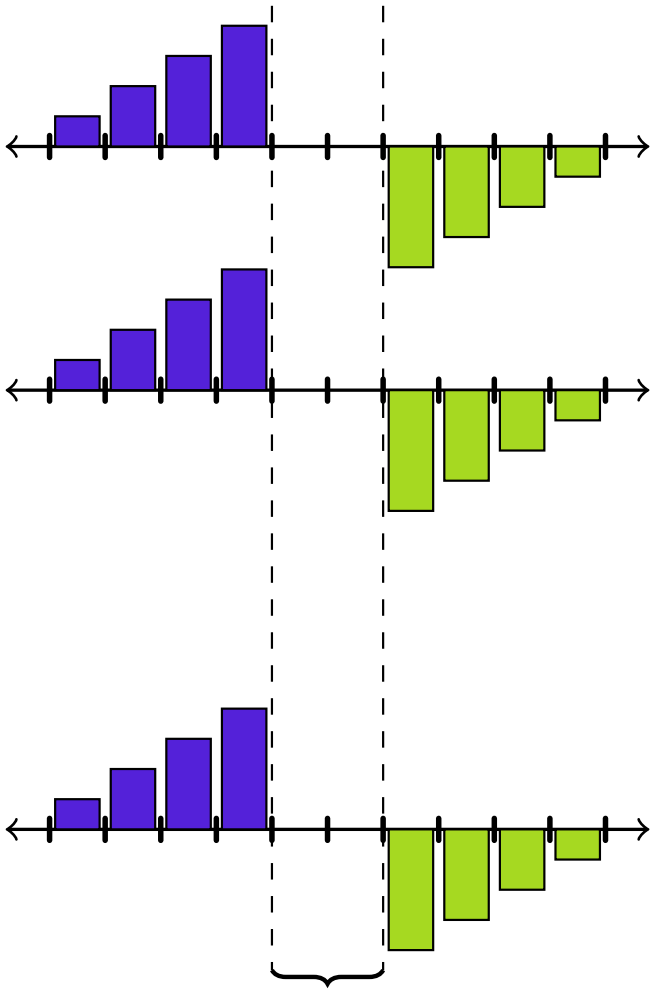
  – when to post limit orders, and to which exchanges
  – when to cancel orders
  – when & how to execute using market orders

We will discuss LOB dynamics and associated control decisions in days 2-3.

Overview of algorithmic trading and limit order book markets

1. Overview of algorithmic trading
   - high level view of equities execution ecosystem
   - algorithmic trading systems
   - trade scheduling and the role of market impact models
   - tactical execution in a LOB
   - fragmentation, internalization, incentives, ...

Fragmentation (more in day 3)



exchanges differ in fee/rebates

traders heterogenous wrt urgency

fragmentation impacts order routing decisions

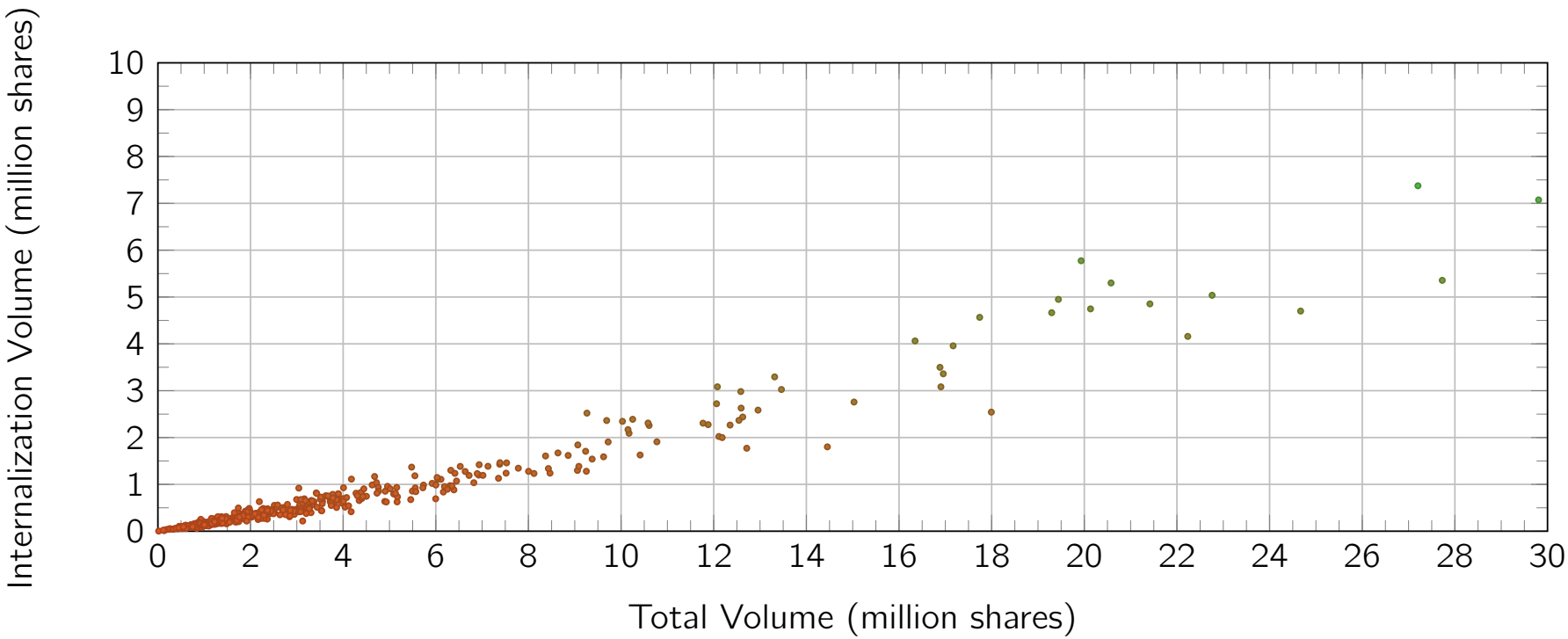
optimized routing $\Rightarrow$ LOB dynamics couple

Internalization & incentives between broker & buy side client

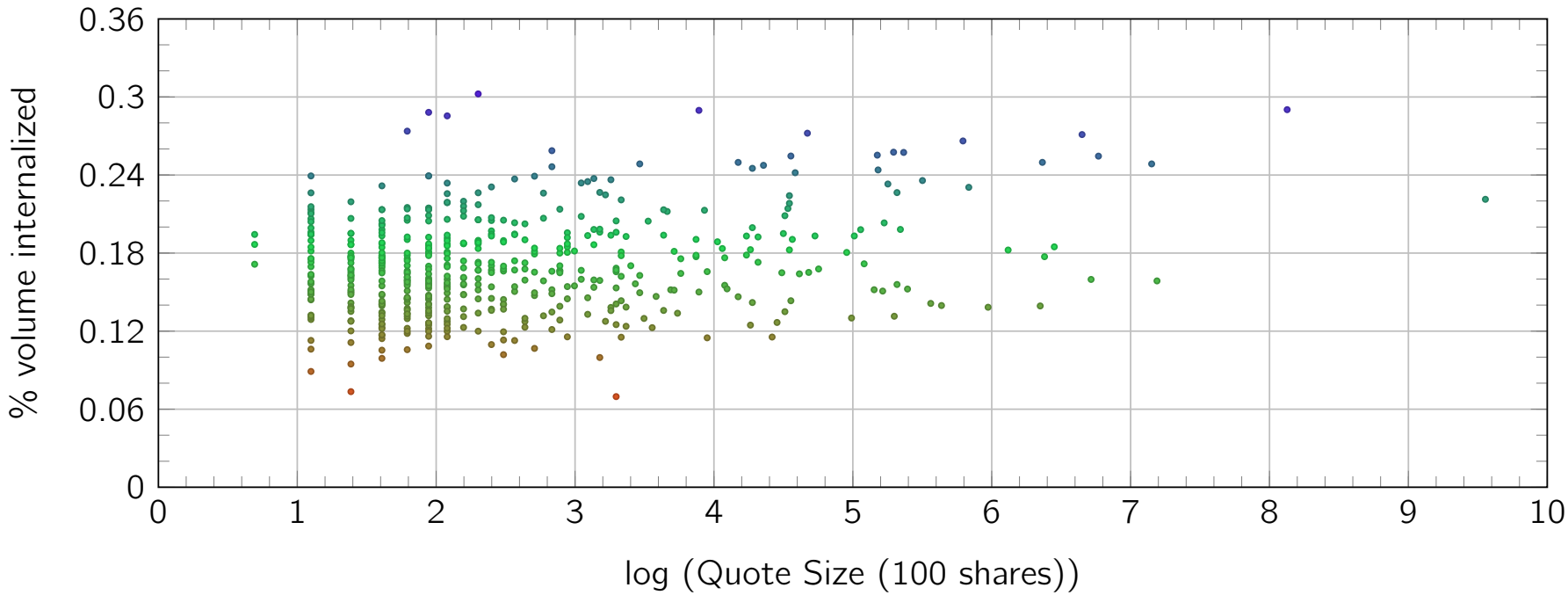
- Typical scenario:
  - Algo orders often “shown" to internalizers prior to going to Exch.
  - internalizer decides whether to fill all-or-part of order
  - algo order avoids exchange fee
  - unfilled portion is subsequently routed to exchange
  - HFT internalizer knows about orders directed to exchange (informational advantage over many participants
  - ($\sim$ all) retail flow is routed through HFT internalizers

- Why do brokers that execute algo flow trade with internalizers?
  - buy-side client typically pays broker an “all-in" rate (incl. exch.fees)
  - all-in rate $\sim$ .5 ¢/share
  - fee/rebate $\sim$ .25 ¢/share $\Rightarrow$ broker net rev $\in$ (.25, .75) ¢/share
    . . . broker wants to avoid paying the exchange fee
  - client indifferent as long as execution quality is good (is it?)

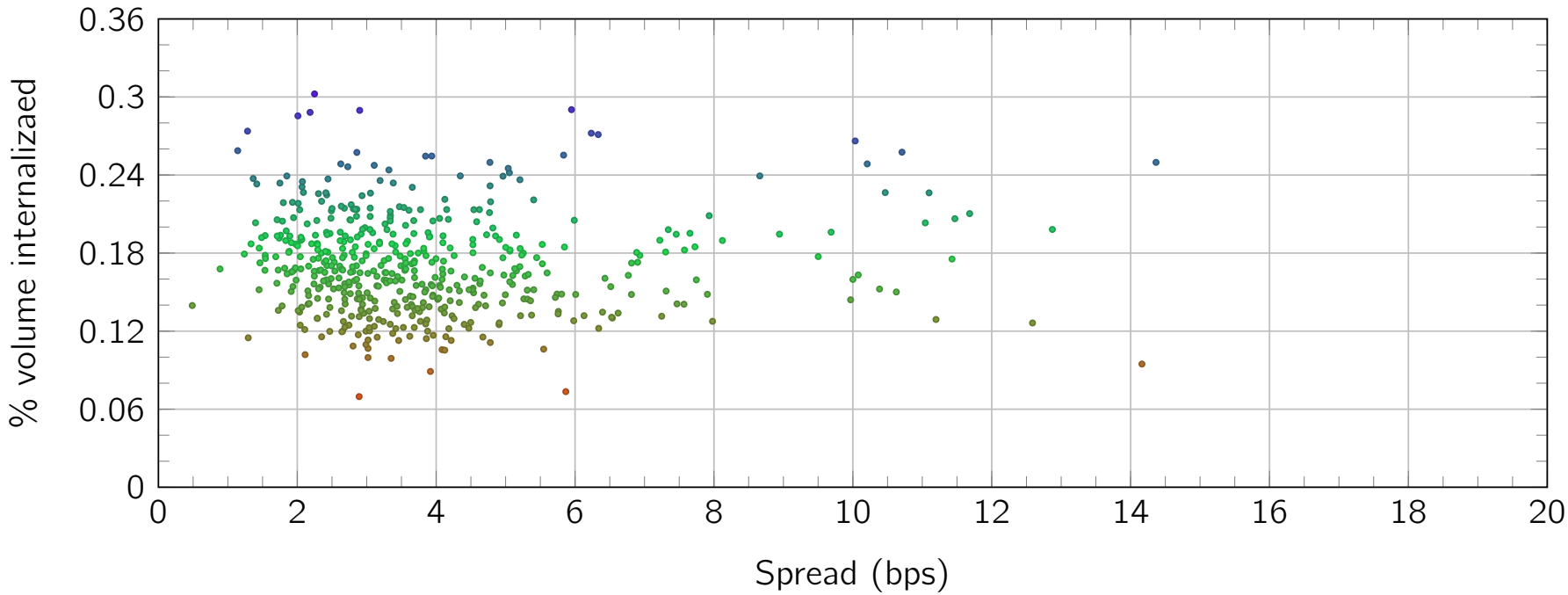
How much volume is internalized? A lot ...



% volume internalized vs quote size



% volume internalized vs spread



Outline

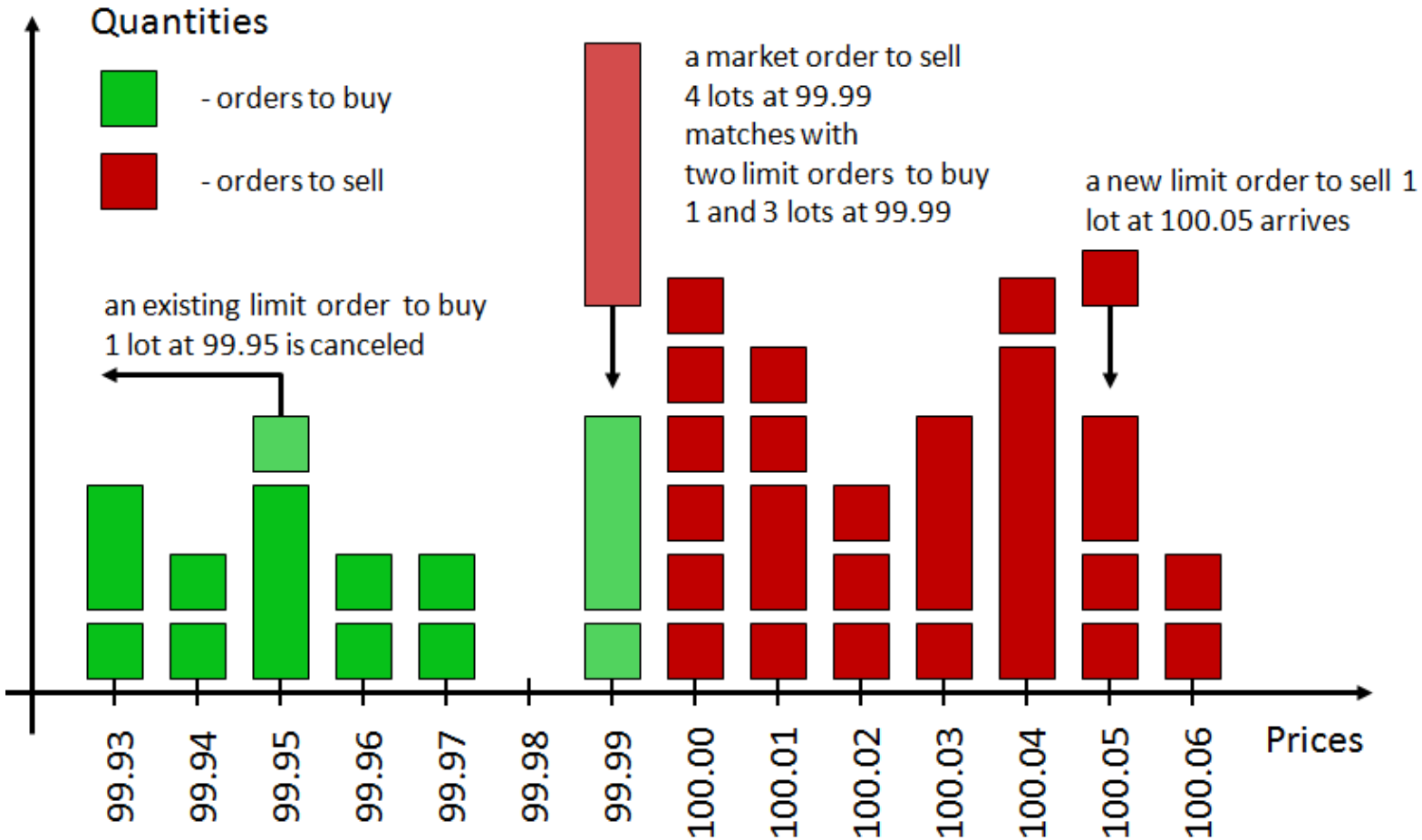
- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

2. Limit order book (LOB) as a queueing system

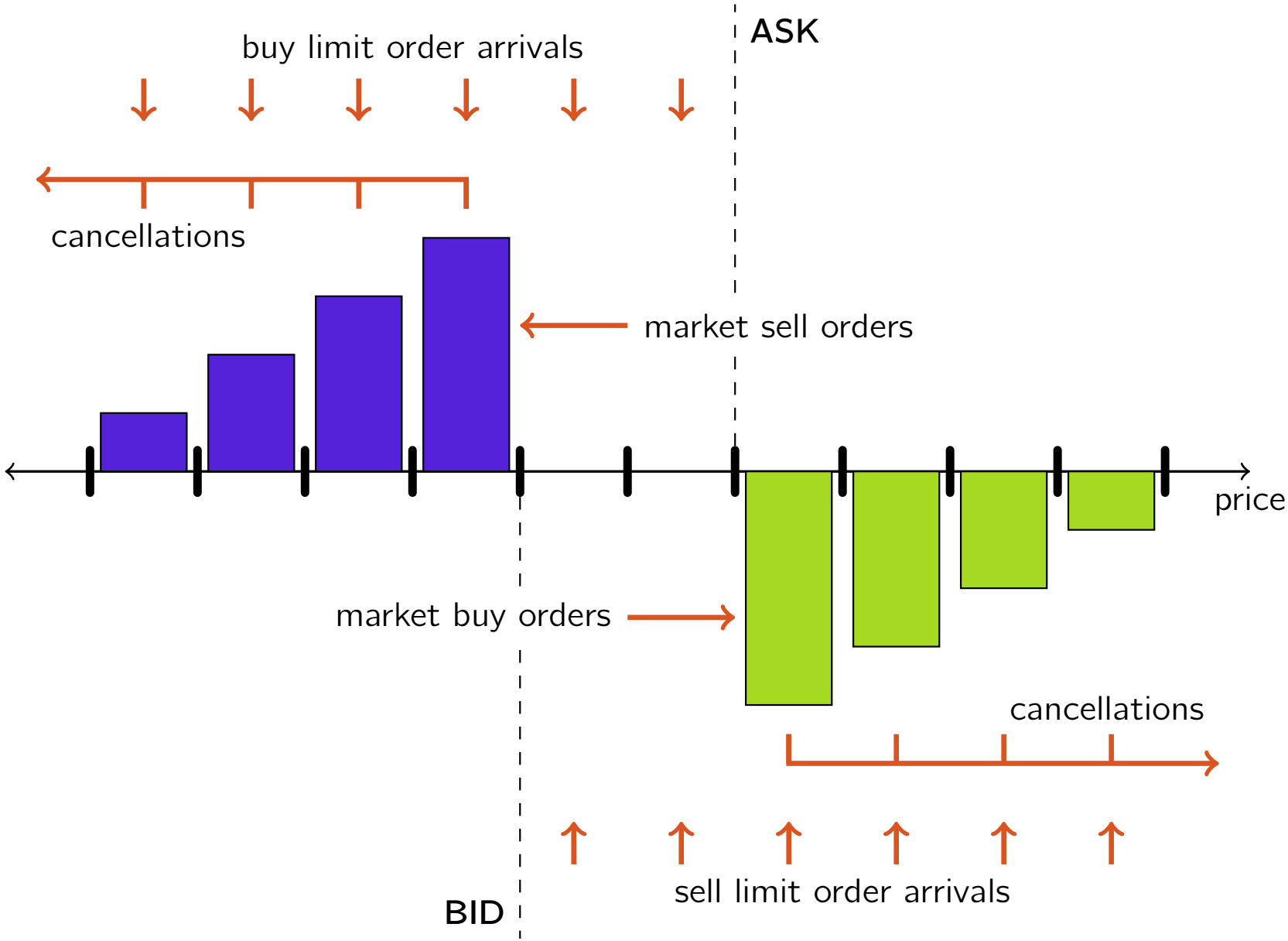
   - time/price priority & LOB as a multi-class queueing system
   - events:
     - limit order arrivals
     - trade executions (service completions)
     - cancellations
   - motivating questions:
     - delay estimation & heterogeneous order cancellation behavior
     - short-horizon optimal execution in the LOB & microstructure cost model
     - adverse selection
     - optimal order routing in a fragmented market structure
   - background on simple queueing models & their asymptotic behavior
     a quick view on time-scales

   (our focus today will be on "top-of-book")

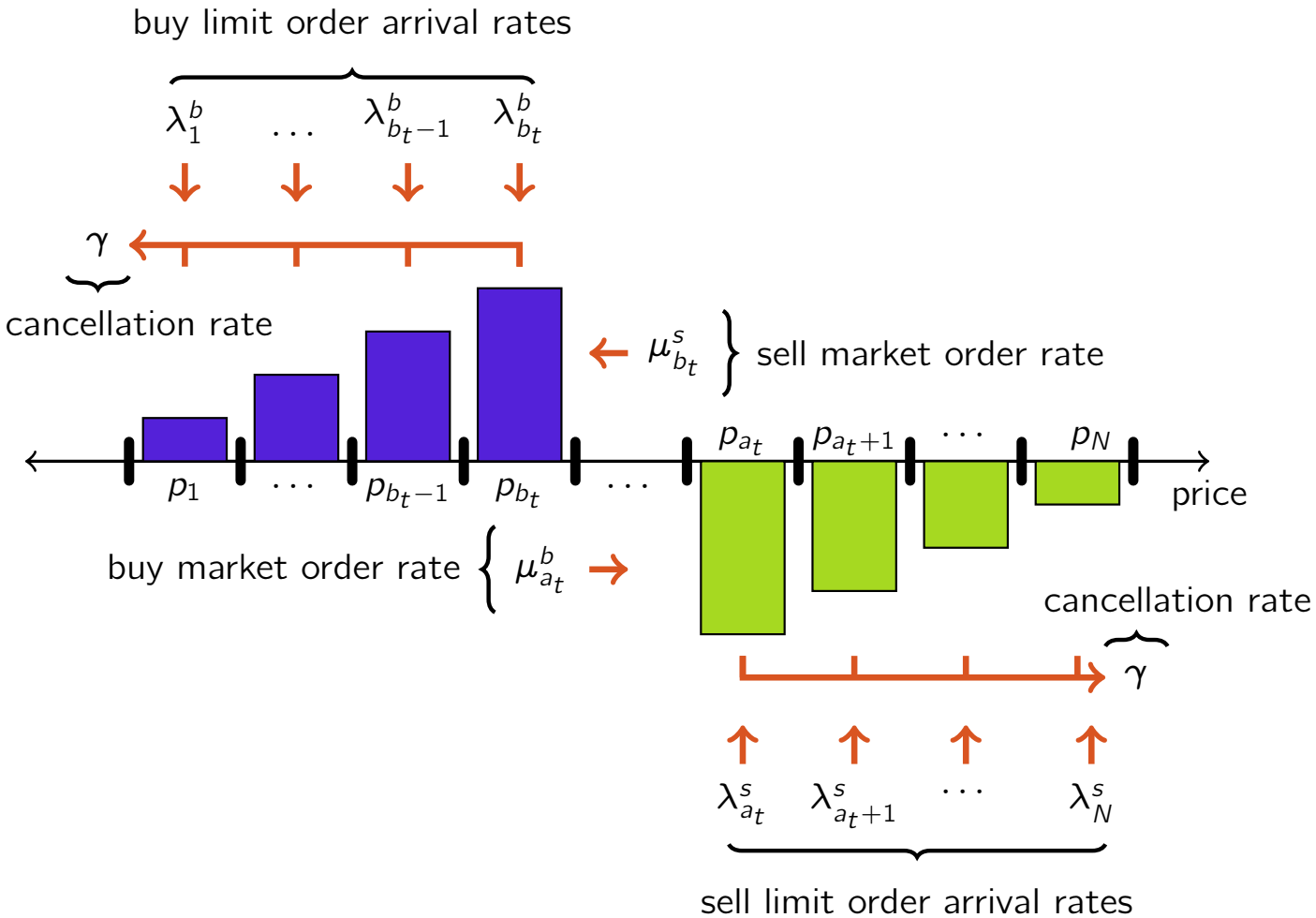
LOB schematic



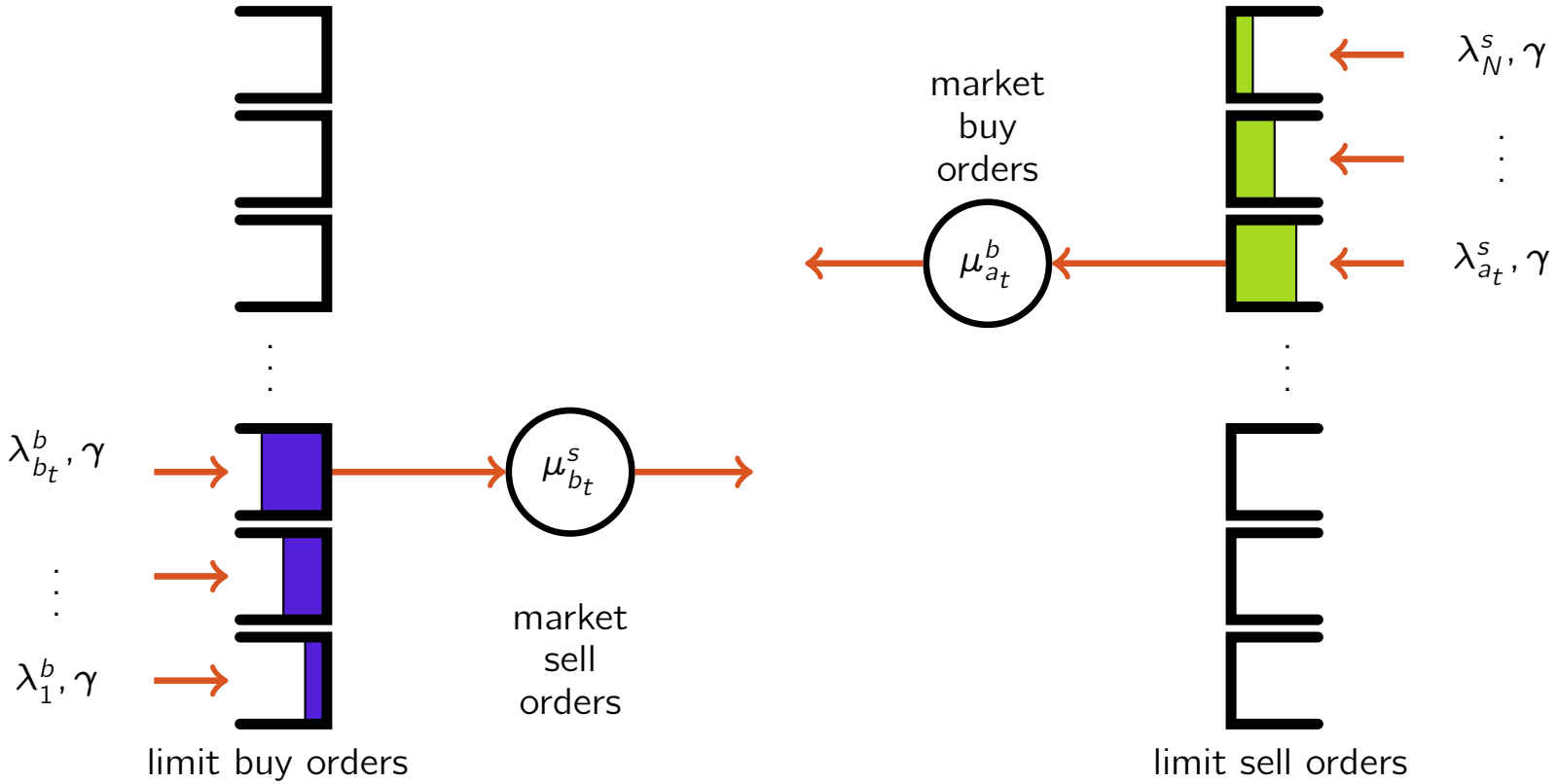
The Limit Order Book (LOB)



LOB: event driven (short-term) view



LOB re-drawn as a multi-class queueing network



Overview of algorithmic trading and limit order book markets

2. Limit order book (LOB) as a queueing system

   - time/price priority & LOB as a multi-class queueing system
   - events:
     - limit order arrivals
     - trade executions (service completions)
     - cancellations
   - motivating questions:
     - delay estimation & heterogeneous order cancellation behavior
     - short-horizon optimal execution in the LOB & microstructure cost model
     - adverse selection
     - optimal order routing in a fragmented market structure
   - background on simple queueing models & their asymptotic behavior
     a quick view on time-scales

     (our focus today will be on "top-of-book")

Limit order arrivals

- Poisson?
- rate fcn's λ (limit order submissions), μ (trades = service completions)

  – time-of-day
  – price level, distance from best bid / best ask, spread
  – depth, certainly at top of book
  – effective tick size
  – rates of other flows; large blocks; ...

  other possible considerations:

  – model "strategies" that generate flow, e.g.,
    - POV responds to (filtered) volume
    - HFT participants respond "quickly" to queue depletion events

      ...

    structurally estimate state-dependent rate fcn
    (complex / over fitting? / depends on intended use)
- jumps or bursts?

Order sizes

- distinguish trades that happen on exchanges (as opposed to dark pools)
- most trades in increments of round lots: 100, 200, ...

	top 500 names (ADV)	top 1000 names (ADV)
Q1 (# shares)	87	84
Q2 (# shares)	101	101
Q3 (# shares)	151	139

- odd lots (mostly < 100 share trades – non-negligible)
- roll up trades over δt to account for "simultaneous" prints triggered by same parent
- think in \$ or in shares (or in depth multiples)?
- trade sizes are heavy-tailed (lognormal gives reasonable fit)

Cancellations

1. disregard cancellations

2. timer-based cancellations:

   – each limit order has associated with it a patience ξ
   – $\xi \sim \exp(\gamma) \;\Rightarrow$ cancellation outflow $\approx -\gamma Q(t)\delta t$
   – some general patience distributions also tractable (asymptotically)
   – state-dependent cancellation flow "stabilizes" queues
   – pretty reasonable model for child orders generated by algorithmic strategies

3. constant cancellation outflow $\approx -\eta\delta t$

   – state independent (not good)
   – no feedback stabilization, i.e., as $Q(t) \uparrow$ cancellation flow constant
   – but, more tractable

– appropriate model to use depends on the context (more in days 2/3)

Heterogenous trading behaviors

- different market participants exhibit significantly different behavior wrt

  – limit order submission
  – cancellations
  – trade sizes & trade submission triggers

- should we model flow through one order generating process? (single type model)

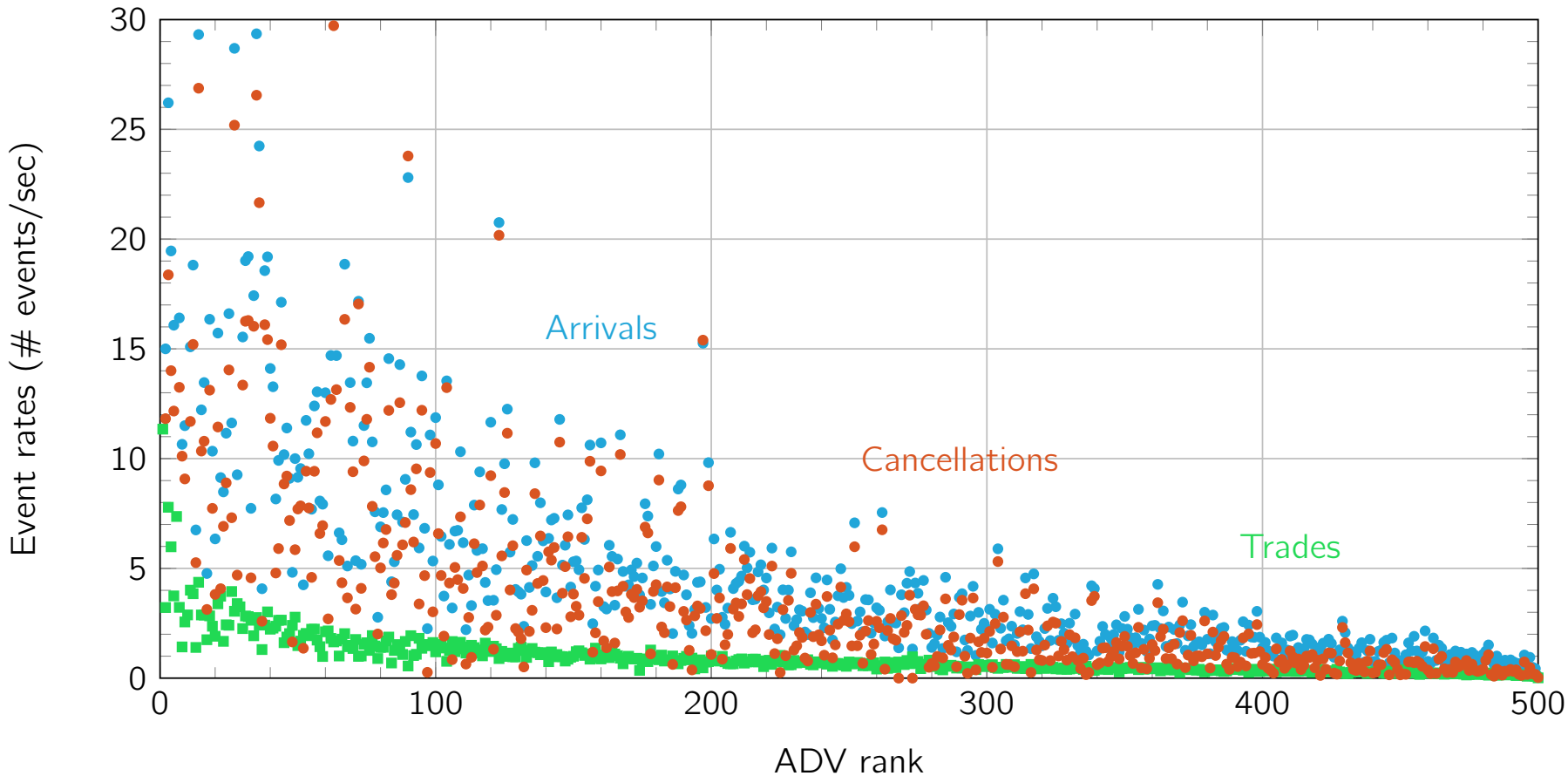
  – e.g., Poisson ($\lambda(t, \text{state vars})$), sizes $\sim G$, patience $\sim F$

- or model heterogenous behavior and use a mixture model, e.g.,

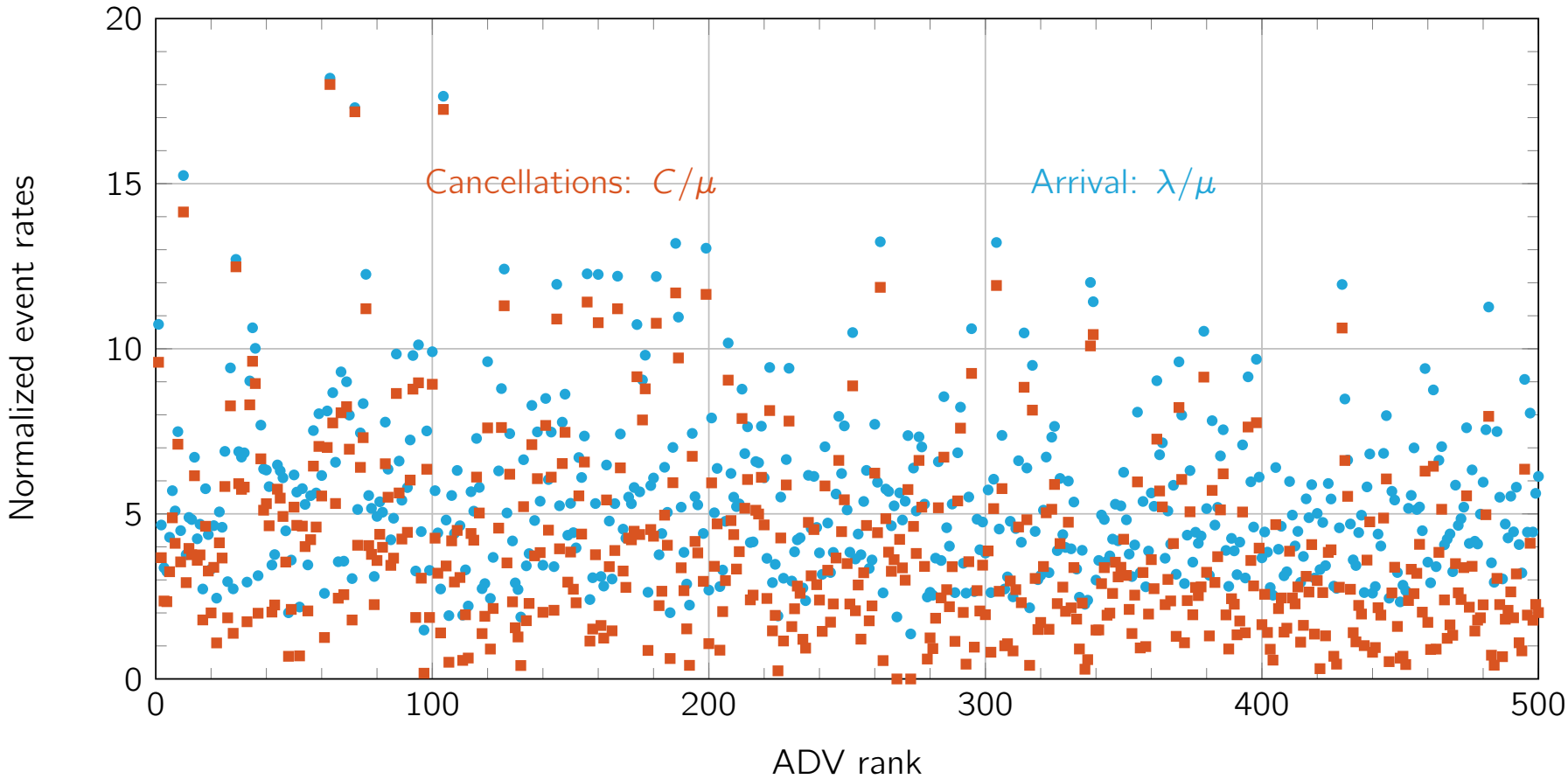
  – algo: Poisson ($\lambda(t, \text{state vars})$), sizes $\sim$ Geo$(1/s)$, patience $\sim \exp(\theta)$
  – MM: event driven arrivals, cancellations, trades (typically as a fcn of state and signals)
  – blocks: Poisson($\eta(t, \text{state vars})$), sizes $\sim$ lognormal

  we will see both styles of models

Event rates (top of book)

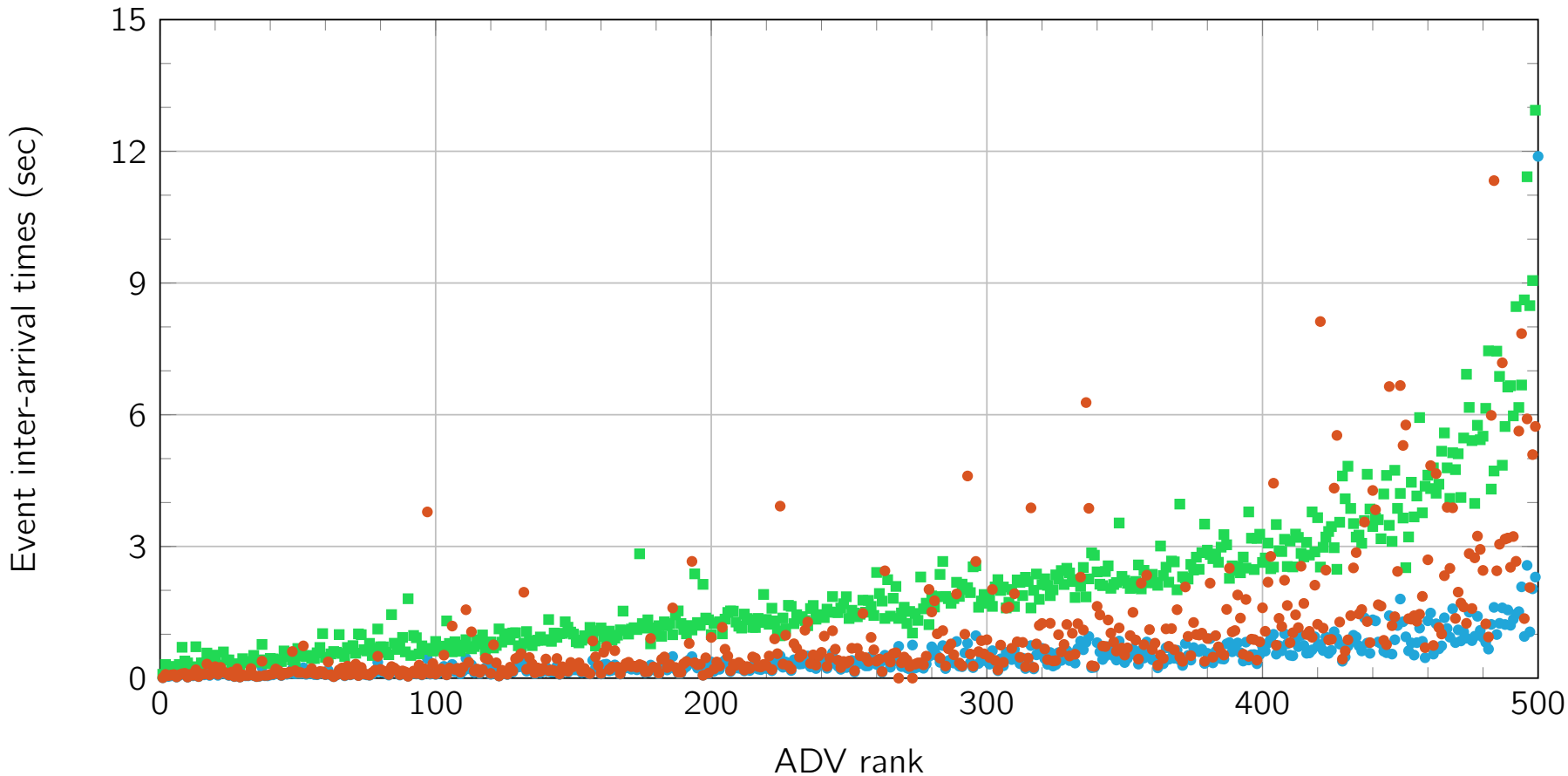


Normalized event rates (top of book)

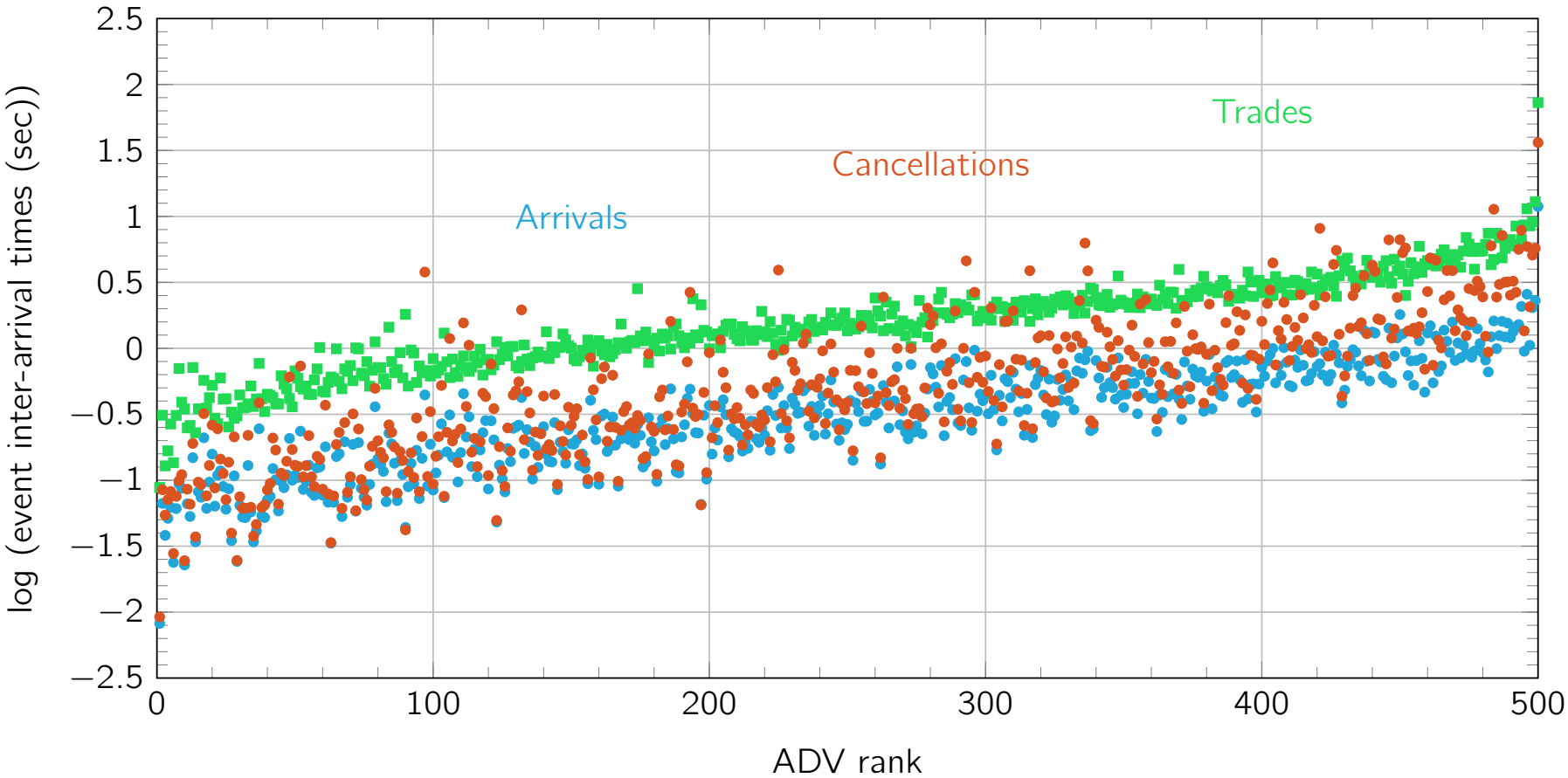


- cancellation volume (at top of book) $\gg$ trade volume
- arrival volume (limit orders at top of book) $\gg$ traded volume

Interarrival times (top of book)

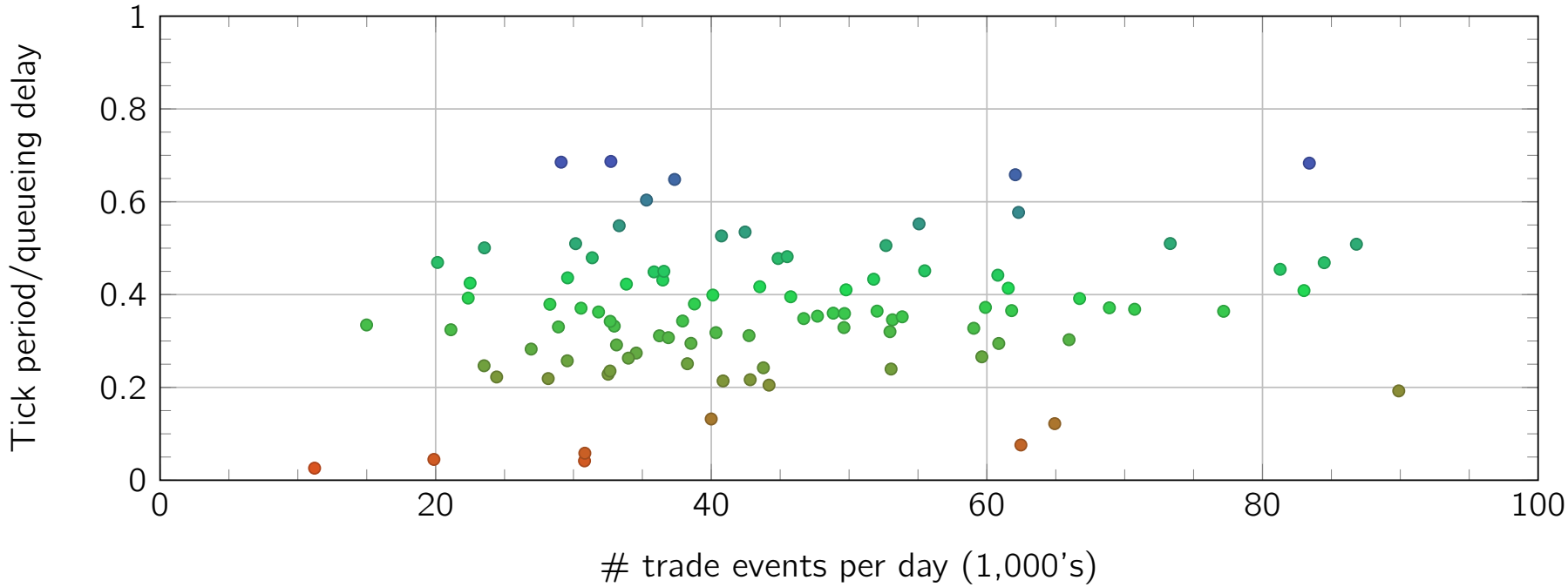


Interarrival times (log scale) (top of book)



- liquid stocks: # trades, # cancellations, # limit order arrivals are large
- # trades $\approx$ 1 order of magnitude less frequent than cancels or order arrivals

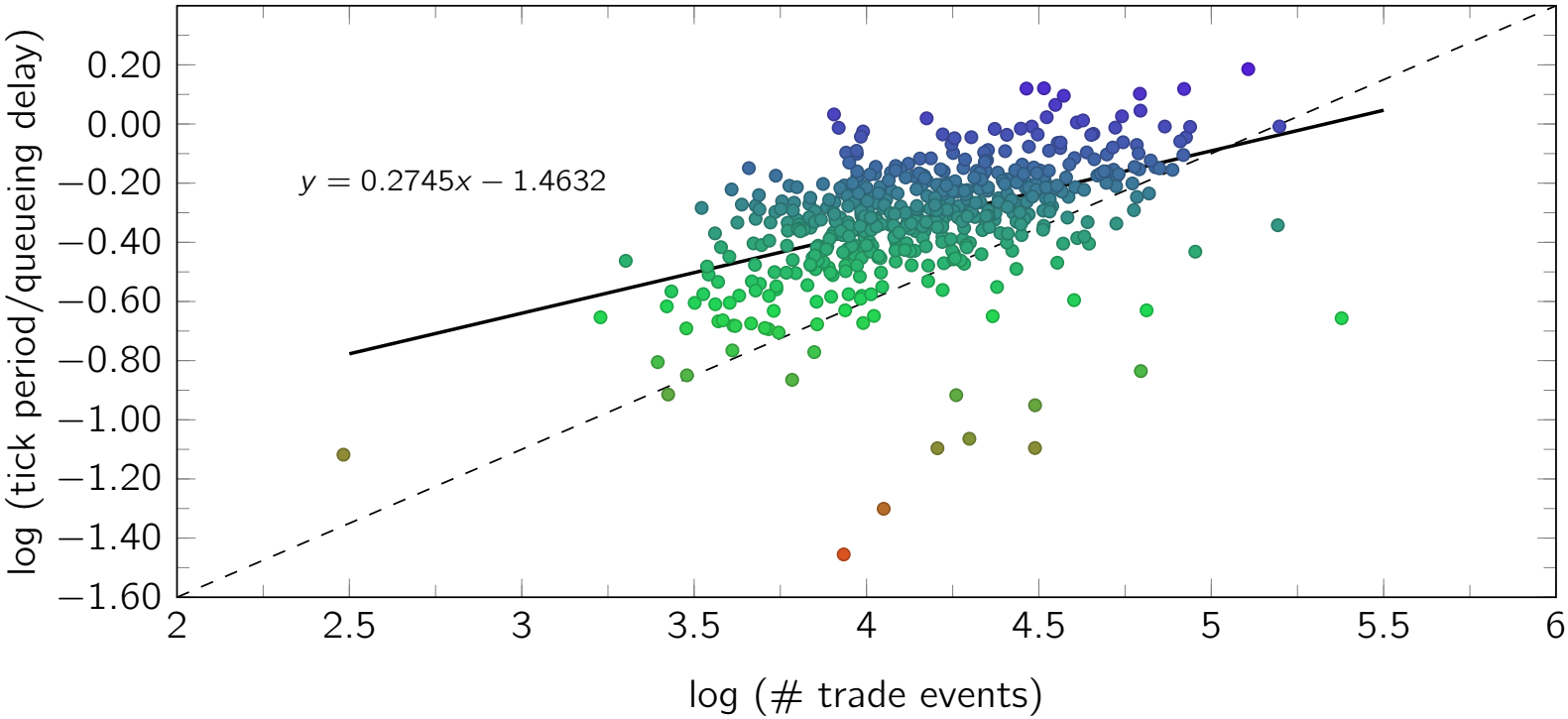
Tick period / queueing delay against # trade events



Tick period versus queueing delay: ratio against # trade events. (liquid names)

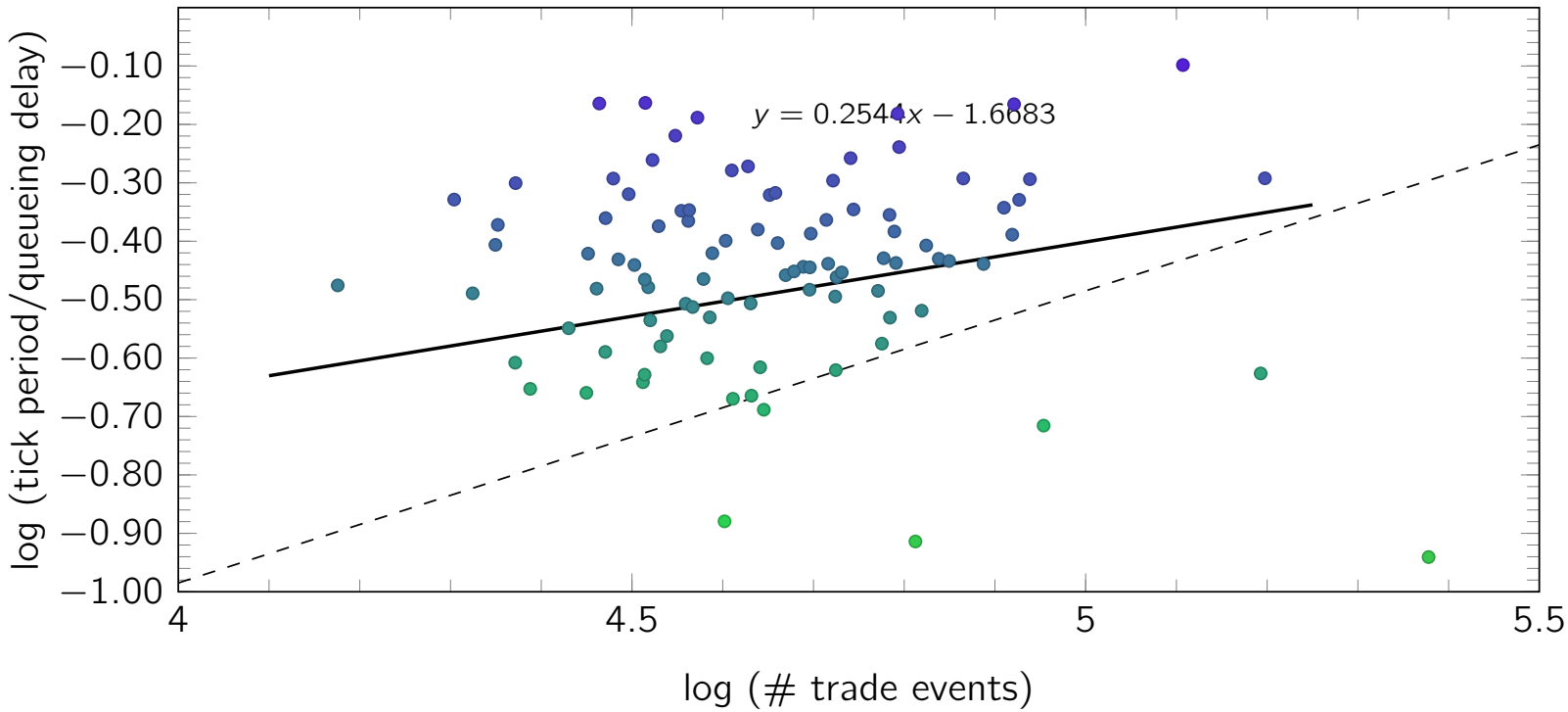
- tick period = avg time between changes in the mid-price
- tick period is on same (or smaller) order magnitude as queueing delay

Tick period versus queueing delay: log-log

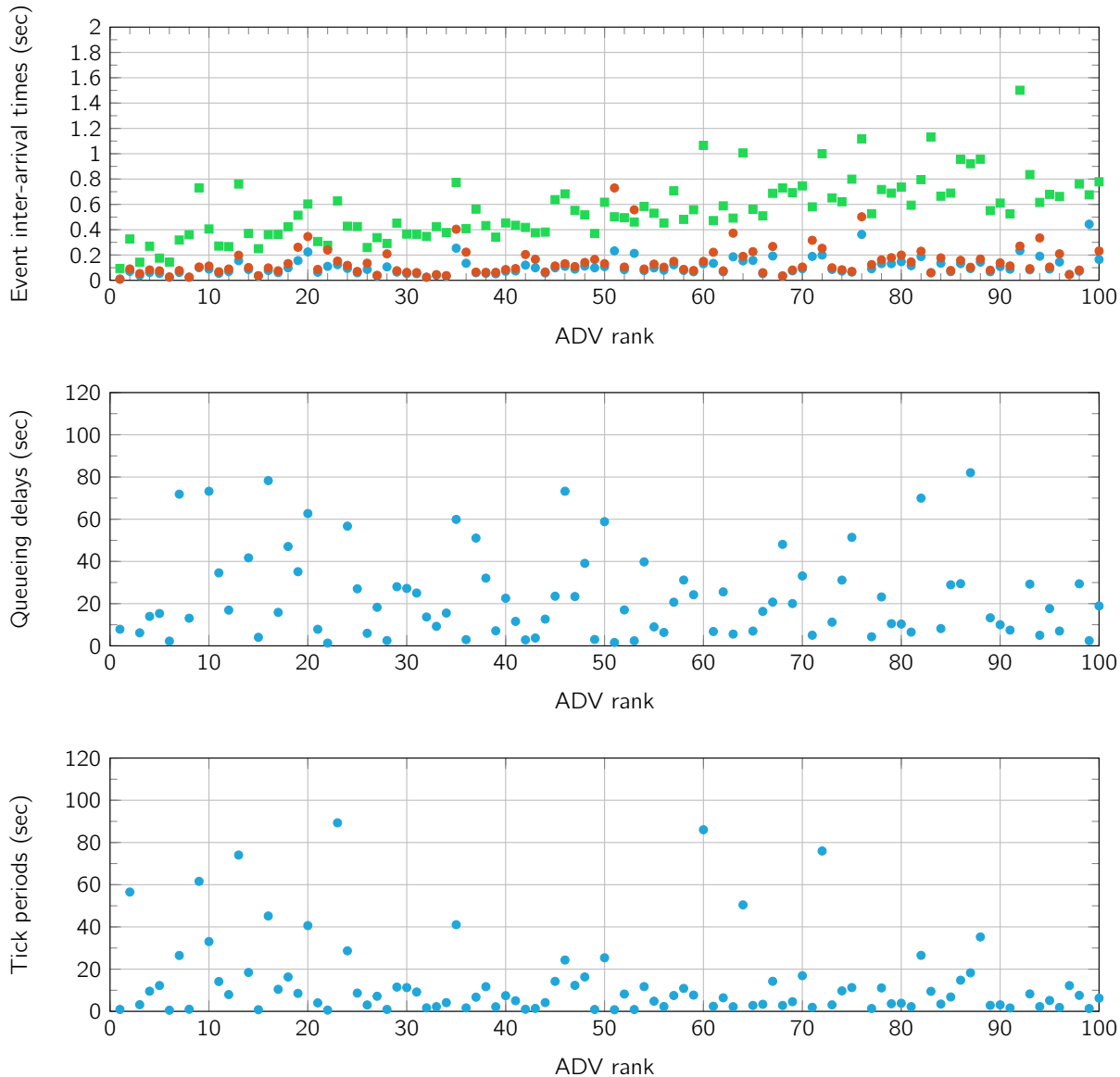


Tick period versus queueing delay: log-log, slope = 0.2745<0.5.

Tick period versus queueing delay (liquid names): log-log



Tick period versus queueing delay: log-log, slope = 0.2745<0.5.


Event inter-arrival times (sec)
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
10
20
30
40
50
60
70
80
90
100
ADV rank
Queueing delays (sec)
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
ADV rank
Tick periods (sec)
120
100
80
60
40
20
0
0
10
20
30
40
50
60
70
80
90
100
ADV rank

Variability of order arrival rates

	% obs. in $\pm 2\sigma_t$	% obs. in $\pm 3\sigma_t$	% obs. outside $\pm 3\sigma_t$
1 min	63.33%	79.23%	20.77%
3 min	32.56%	50.39%	49.61%
5 min	27.27%	35.06%	64.94%
10 min	13.16%	31.58%	68.42%

- table checks if $\mu_{t+1} \in$ intervals $\mu_t \pm k\sigma_t$ for $k = 2, 3$
- (λ, μ) exhibit significant differences in the time scale of 3 - 5 minutes
- cf. top 100 names (by ADV): average queueing delay = 61 sec (more on this later on)

Some observations

- Event data:

$$\lambda \gg \mu \quad \text{and} \quad \text{cancellation flow} \gg \mu$$

- significant cancellation volume to balance order flow at top of book
- price changes on the same time-scale as queueing delays
- event arrival rates fluctuate at slightly slower time scale than queueing delays
- heterogeneous trading behavior may impact order flow dynamics
- fragmentation affects delay estimates and cancellation behavior

2. Limit order book (LOB) as a queueing system

   - time/price priority & LOB as a multi-class queueing system
   - events:
     - limit order arrivals
     - trade executions (service completions)
     - cancellations
   - motivating questions:
     - delay estimation & heterogeneous order cancellation behavior
     - short-horizon optimal execution in the LOB & microstructure cost model
     - adverse selection
     - optimal order routing in a fragmented market structure
   - background on simple queueing models & their asymptotic behavior
     a quick view on time-scales

     (our focus today will be on "top-of-book")

Four motivating performance and control questions

1. Cancellation behavior and expected queueing delay at the top of book
   - **E**(delay) until an order gets filled as a fcn of model primitives
   - how does it depend on cancellation behaviors of different participants?
2. Optimal execution in a LOB and market impact, e.g.,
   - how to buy 5,000 shares of IBM over the next 3 min
   - estimated execution cost as fcn of real-time mkt conditions
   - microstructure model of market impact
3. Optimal order routing across LOB; fee/rebate tradeoffs; dynamics
   - tactical optimization of order routing decisions; money/delay tradeoffs
4. Stylized models of adverse selection as a fcn of queue position
   - value of queue position in AS costs

Where it all fits in the technology stack of an algo trading system. . .

- **Trade scheduling:** splits parent order into $\sim$ 5 min "slices"
  – relevant time-scale: minutes-hours
  – schedule follows user selected "strategy" (VWAP, POV, IS, . . .)
  – reflects urgency, "alpha," risk/return tradeoff
  – schedule updated during execution to reflect price, liquidity/. . .

- **Optimal execution of a slice ("micro-trader"):** further divides slice into child orders
  – relevant time-scale: seconds–minutes
  – strategy optimizes pricing and placing of orders in the limit order book
  – execution adjusts to speed of LOB dynamics, price momentum, ...

- **Order routing:** decides where to send each child order
  – relevant time-scale: $\sim$ 1–50 ms
  – optimizes fee/rebate tradeoff, liquidity/price, latency, etc.

separation of 2nd and 3rd steps mostly technological/historical artifact (should not be treated separately)

Overview of algorithmic trading and limit order book markets

2. Limit order book (LOB) as a queueing system

   - time/price priority & LOB as a multi-class queueing system
   - events:
     - limit order arrivals
     - trade executions (service completions)
     - cancellations
   - motivating questions:
     - delay estimation & heterogeneous order cancellation behavior
     - short-horizon optimal execution in the LOB & microstructure cost model
     - adverse selection
     - optimal order routing in a fragmented market structure
   - background on simple queueing models & their asymptotic behavior
     a quick view on time-scales

     (our focus today will be on "top-of-book")

Some basic building blocks from queueing theory

- $M/M/1$ system (Poisson limit and market order arrivals)
- $M/M/1+M$ with exponential patience clocks
- Basic facts for asymptotic behavior of $M/M/1$ and $M/M/1+M$

  regime we focus: (λ, μ) grow large

  – mean-field (fluid) models
  – diffusion models

$M/M/1$ queue

Model:

- arrivals (limit orders) $\sim$ Poisson rate λ
- service completions (market orders) $\sim$ Poisson rate μ (exponential service times, rate μ)
- single server; ∞ buffer; no cancellations

Steady-state probability distribution π of Markovian system:

$$\pi_n = (1-\rho)\rho^n \qquad \rho := \frac{\lambda}{\mu} < 1$$

Steady-state performance measures:

- Expected time in system $\mathbf{E}(W) = \dfrac{1}{\mu(1-\rho)} = \dfrac{1}{\mu-\lambda}$
- Expected number in system $\mathbf{E}(Q) = \dfrac{\rho}{1-\rho} = \dfrac{\lambda}{\mu-\lambda} = \lambda\mathbf{E}(W)$

$M/M/1+M$ queue (Erlang-A)

$M/M/1$ assumptions plus:

- exponential patience times, rate γ (iid)
- orders in queue cancel when the idiosyncratic wait time exceeds their patience

Steady-state probability distribution π:

$$\pi_n = \pi_0 \prod_{k=0}^{n-1} \frac{\lambda}{\mu + k\gamma} \qquad \pi_0 = \left(1 + \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} \frac{\lambda}{\mu + k\gamma}\right)^{-1}$$

Cancellation behavior:

- conditional on queue length $Q(t)$, cancellation intensity is uniform across queue positions $x < Q(t)$
- variant on cancellations.: residual patience > residual **E**(waiting time) (calculation needs queue position)

Mean-field (fluid) model: $M/M/1$

Liquid securities see significant event volume per minute. Suggests scaling:

$$\lambda^n = n\lambda \quad \text{and} \quad \mu^n = n\mu$$

$A^n(t) \sim$ Poisson rate λ^n and similarly for service completions $S^n(t)$

Strong approximation: $A^n(t) = n\lambda t + \sqrt{n\lambda}B(t) + O(\log(nt))$ a.s.

So:

- arrivals $\sim O(n)$
- trades $\sim O(n)$
- queue length $\sim O(n)$:

$$\bar{Q}^n(t) := \frac{1}{n}Q^n(t) \xrightarrow{a.s.} q(t) \quad \text{u.o.c.}$$

where $q(t)$ is a *deterministic* trajectory satisfying

$$\dot{q}(t) = \lambda - \mu$$

Mean-field (fluid) model: $M/M/1+M$

Scaling:

$$\lambda^n = n\lambda \qquad \mu^n = n\mu \quad \text{and} \quad \gamma^n = \gamma$$

- market grows large while order patience characteristics stay same

[Strong approximation + Gronwall's inequality + CMT (for reflection map):]

$$\bar{Q}^n(t) := \frac{1}{n} Q^n(t) \xrightarrow{a.s.} q(t) \text{ u.o.c.}$$

where

$$\dot{q}(t) = \lambda - \mu - \gamma q(t)$$

Transient of mean-field (fluid) model of $M/M/1+M$

$$\dot{q}(t) = \lambda - \mu - \gamma q(t)$$

ODE solution:

$$q(t) = \frac{\lambda - \mu}{\gamma}\left(1 - e^{-\gamma t}\right) + q(0)e^{-\gamma t}$$

- If $\lambda - \mu > 0$ (as in trade data), $q(t) \to \dfrac{\lambda - \mu}{\gamma} =: q_\infty$

  q_∞ = equilibrium depth (outflow = trades + cancellation = inflow)

- If $\lambda \leq \mu$, $q(t) \to 0$.

Heavy-traffic (diffusion) model: $M/M/1$ approximating diffusion

Scaling:

$$\lambda^n = n - \beta\sqrt{n}, \quad \mu^n = n \quad (\text{so that } \lambda^n \approx \mu^n),$$

Flow imbalance:

$$N^n(t) = (A^n(t) - S^n(t)) = -\beta\sqrt{n}t + \sigma\sqrt{n}B(t) + O(\log(nt))$$

$O(\sqrt{n})$ stochastic imbalance of Poisson flows, leads to $O(\sqrt{n})$ queue lengths

$$\hat{Q}^n(t) := \frac{Q^n(t)}{\sqrt{n}} \Longrightarrow \hat{Q}(t) = \text{reflected Brownian motion.}$$

$$d\hat{Q}(t) = -\beta dt + \sigma dB(t) + dL(t) \quad (\beta > 0)$$

$L(t) =$ local time at the origin; in LOB analogy, $L(t)$ fires when price moves

$\hat{N}(t) = -\beta t + \sigma B(t), \quad L(t) = \sup_{\{0 \le s \le t\}} \hat{N}^-(s) \quad (x^- = min(0, x))$

Heavy-traffic (diffusion) model: $M/M/1$ performance approximations

$\hat{Q}(\infty) \sim \exp(-2\beta/\sigma^2)$

- queue lengths:

$$\mathbf{E}(Q^n) = \frac{\rho^n}{1-\rho^n} = O(\sqrt{n})$$

- waiting times: $\sqrt{n}$ queue length, trades arrive at order n, so

$$\mathbf{E}(W^n) = \frac{\mathbf{E}(Q^n)}{\mu^n} = O\left(\frac{1}{\sqrt{n}}\right)$$

- how often does the queue gets depleted: τ^n is the length of busy periods

$$\mathbf{E}(\tau^n) \approx O(1) \quad \text{the natural time scale of the limiting RBM}$$

(regenerative cycles of RBM)

- time scale separation: $\mathbf{E}(\tau^n) \gg \mathbf{E}(W^n)$

$$\mathbf{E}(\tau^n) \approx \sqrt{n}\mathbf{E}(W^n)$$

Heavy-traffic (diffusion) model: $M/M/1+M$, $\lambda^n - \mu^n = \beta\sqrt{n}$

Scaling:

$$\lambda^n = n + \beta\sqrt{n}, \quad \mu^n = n \quad \text{and} \quad \gamma^n = \gamma$$

Similar to $M/M/1$ in heavy traffic:

$$\hat{Q}_n(t) := \frac{1}{\sqrt{n}} Q^n(t) \Rightarrow \hat{Q}(t) \quad \text{(reflected O-U process)}$$

where

$$d\hat{Q}(t) = \left(\beta - \gamma\hat{Q}(t)\right) dt + \sigma dB(t) + dL(t)$$

- stable queue due to cancellations (drift $-\gamma Q(t)$)
- cancellation volume $\approx O(\sqrt{n}) \ll \lambda^n$
- $\hat{Q}(\infty) \sim$ truncated Normal dist.
- time scale separation: $\mathbf{E}(\tau^n) \gg \mathbf{E}(W^n) \ldots \mathbf{E}(\tau^n) \approx \sqrt{n}\mathbf{E}(W^n)$

A different heavy-traffic regime: $M/M/1+M$, $\lambda^n \gg \mu^n$

Scaling:

$$\lambda^n = n\rho, \quad \mu^n = n \quad \text{and} \quad \gamma^n = \gamma \quad (\rho > 1)$$

- $O(n)$ imbalance between order arrivals and trades
- balanced through $O(n)$ cancellations
- proportional cancellation flow $\gamma Q^n(t)$, suggests $Q^n(t) = O(n)$
- indeed fluid path dominates behavior:

$$Q^n(t) \approx nq(t) + \sqrt{n}(\text{stochastic fluctuations}) + O(\log(nt))$$

  – for large t, $Q^n(t)/n \approx q_\infty$, where $\rho - 1 = \gamma q_\infty$

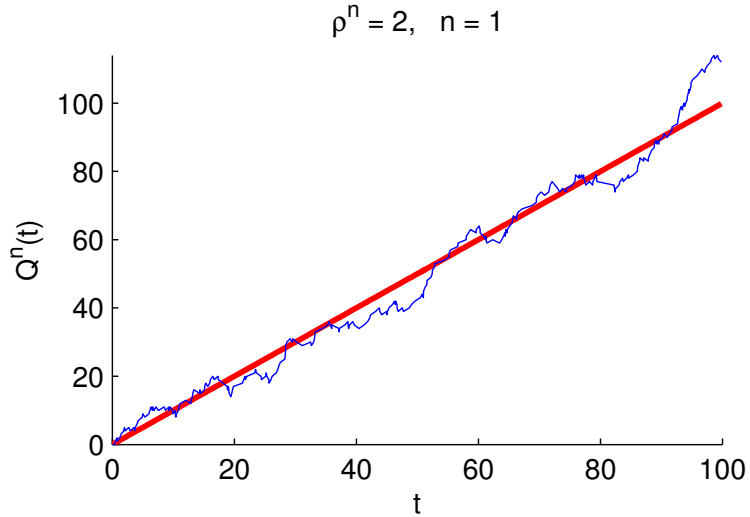
  – $\mathbf{E}(W) = O(1)$

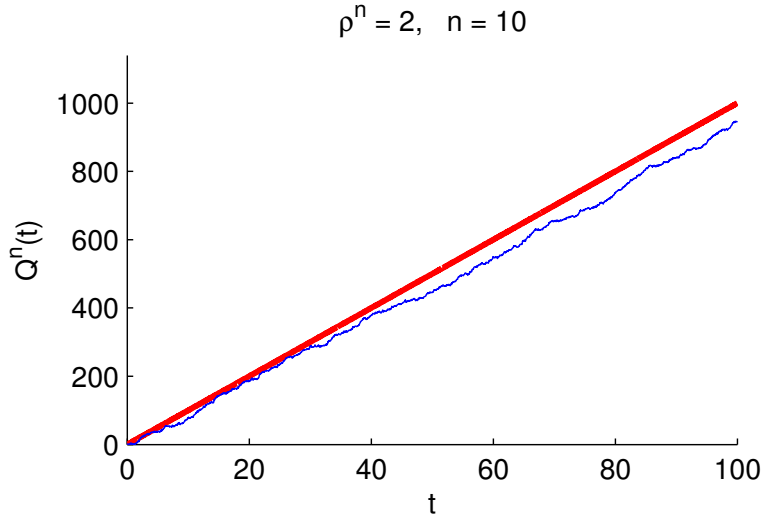
  – fluid paths cannot generate price changes (no queue depletions)
  . . . price changes triggered by changes in rate parameters

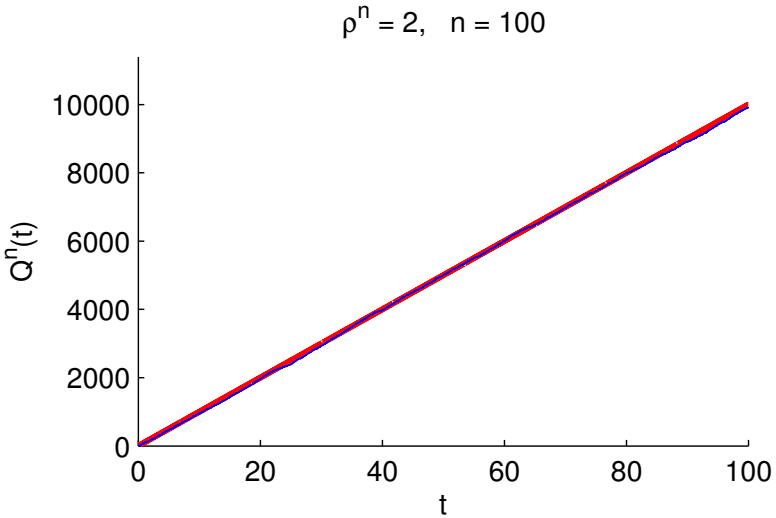
Quick observations

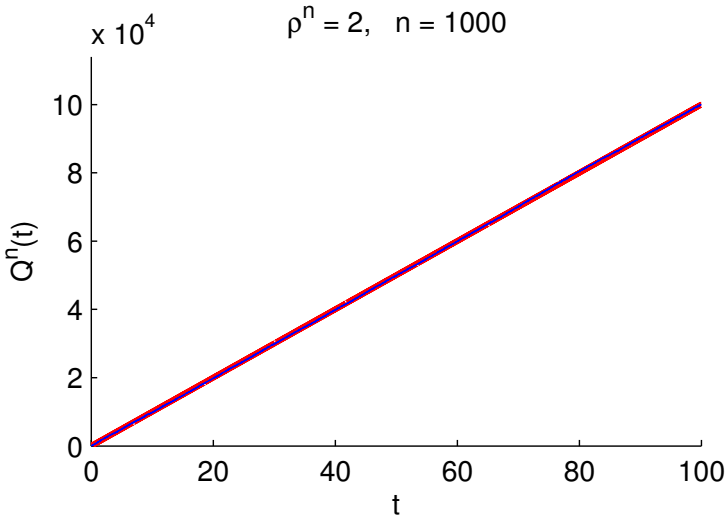
- data for liquid stocks suggests $\mathbf{E}W \approx \mathbf{E}\tau$
  - "heavy-traffic" diffusion models ($M/M/1$ or $M//M/1+M$) may not be appropriate to study queueing effects
  - queueing delays appear instantaneous in these diffusion models
  - also, arrival rates fluctuate on time scale of queueing delays
- data: cancellation volume seems to indicate queues of $O(n)$ where n= scale of the system (e.g., speed)
  - $O(\sqrt{n})$ variability of Poisson arrival flows "small" viz $O(n)$ queues
  - arrival rate fluctuations may yield $O(n)$ variability on order arrival processes
- non-exponential patience, e.g., $M/M/1+GI$ similar qualitative results

$M/M/1$, $\rho > 1$, Fluid Scale $O(n)$

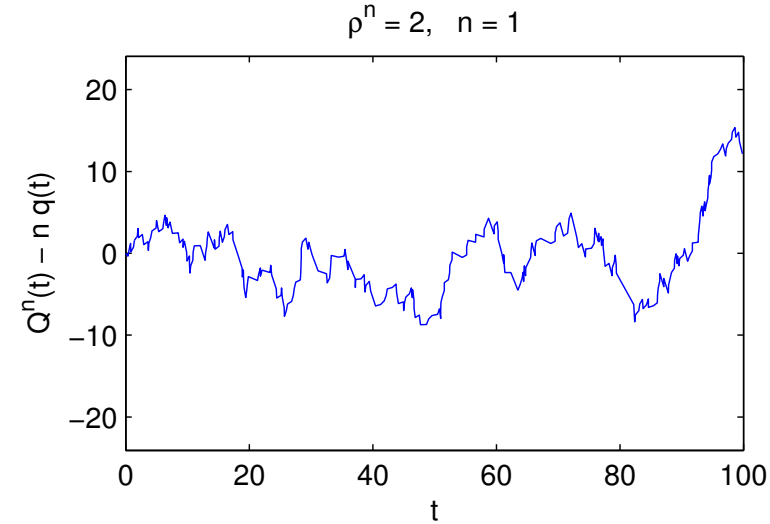


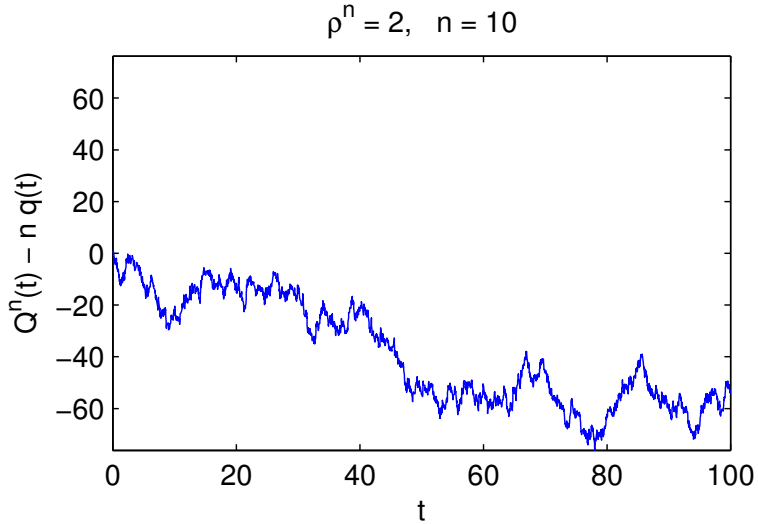


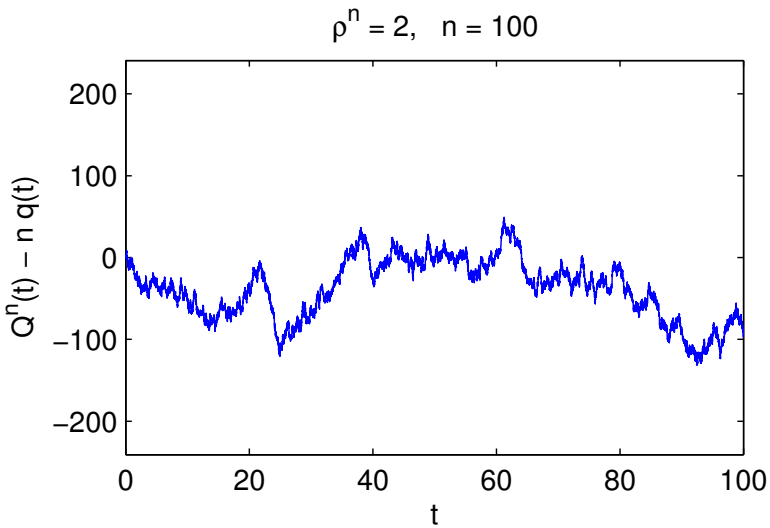


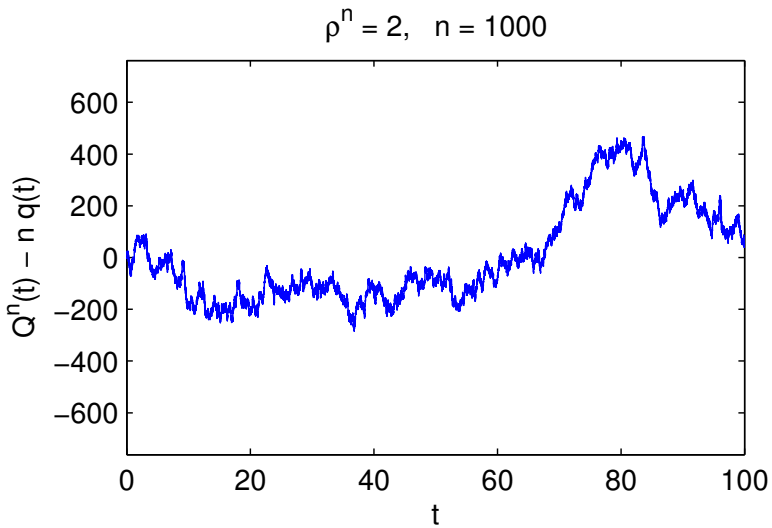


$M/M/1$, $\rho > 1$, Diffusion Scale $O(\sqrt{n})$

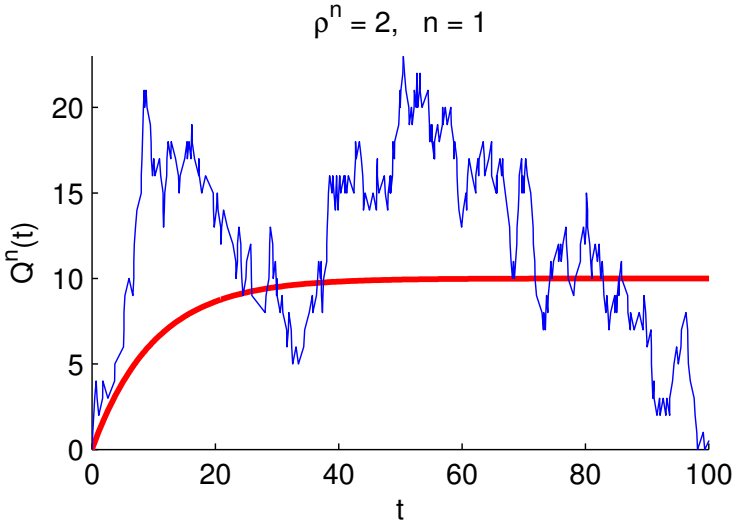


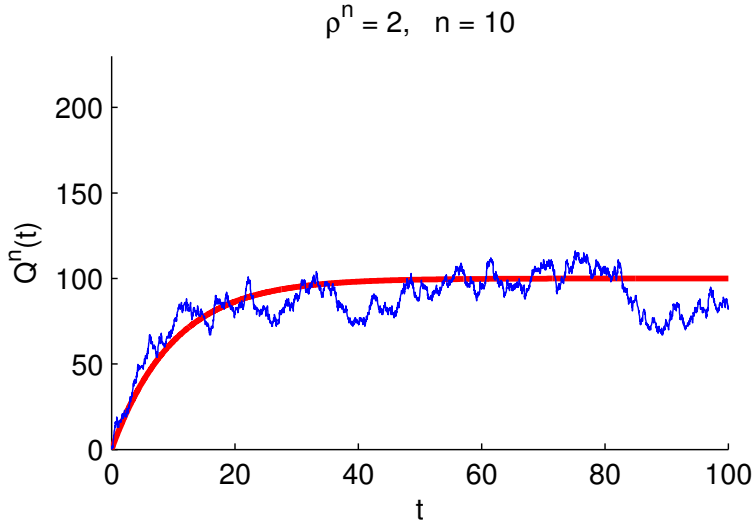


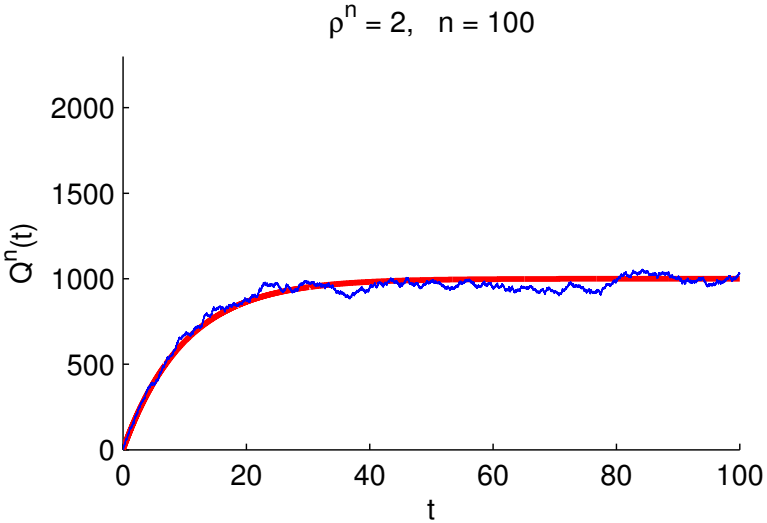


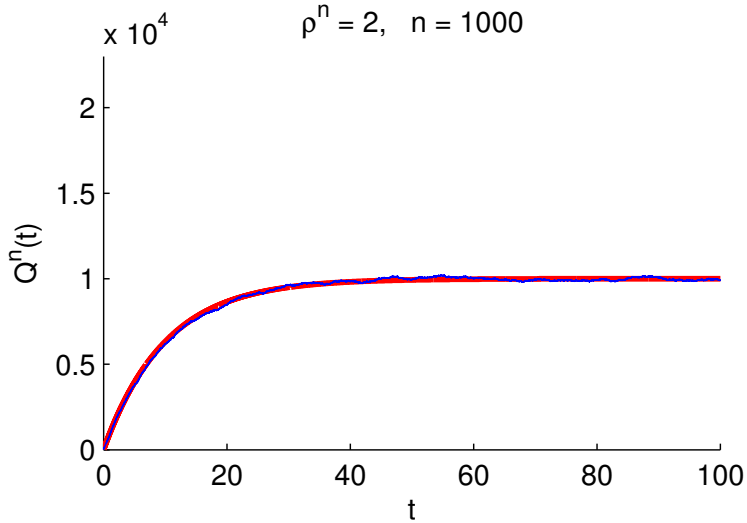


$M/M/1+M$, $\rho > 1$, Fluid Scale $O(n)$

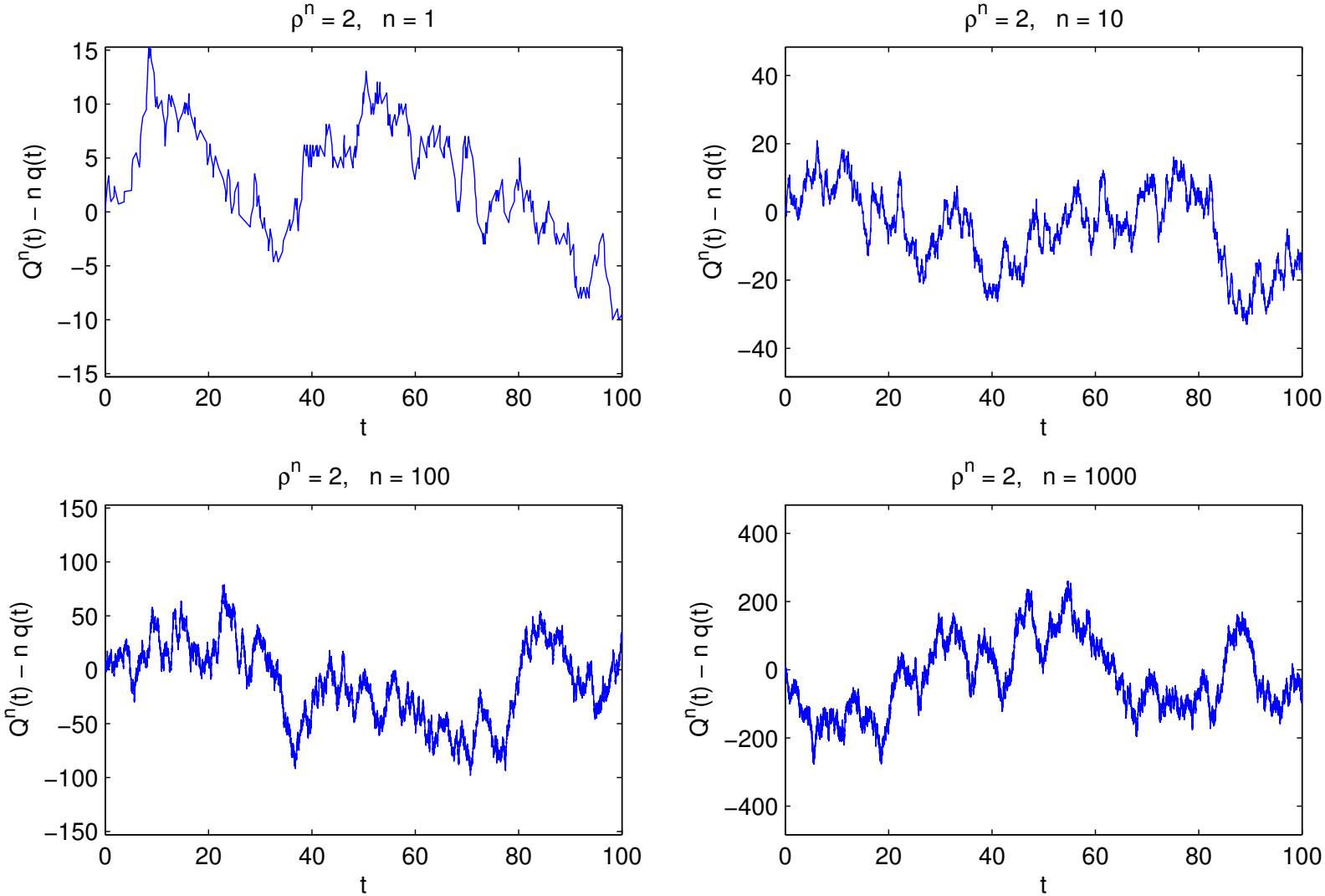




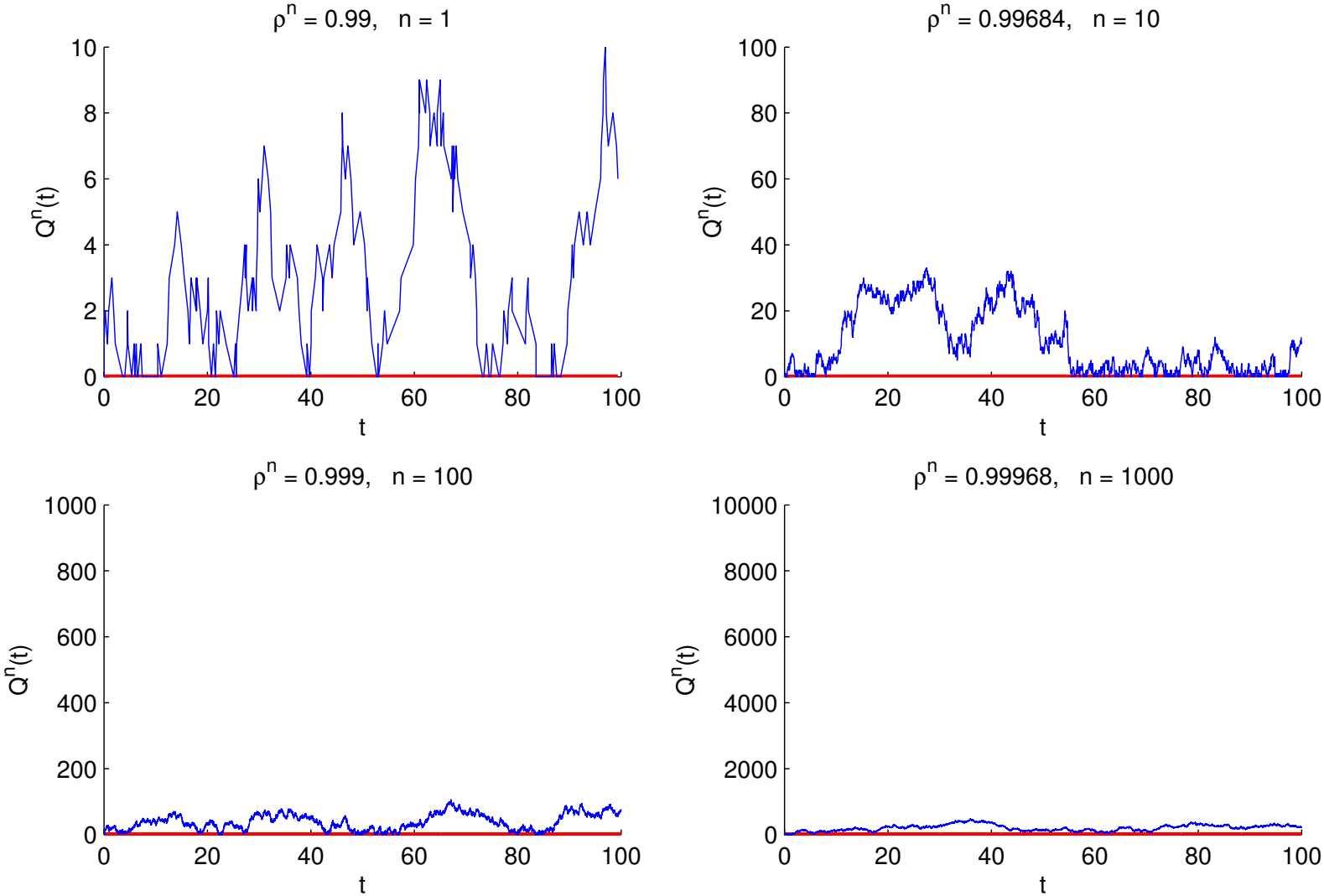




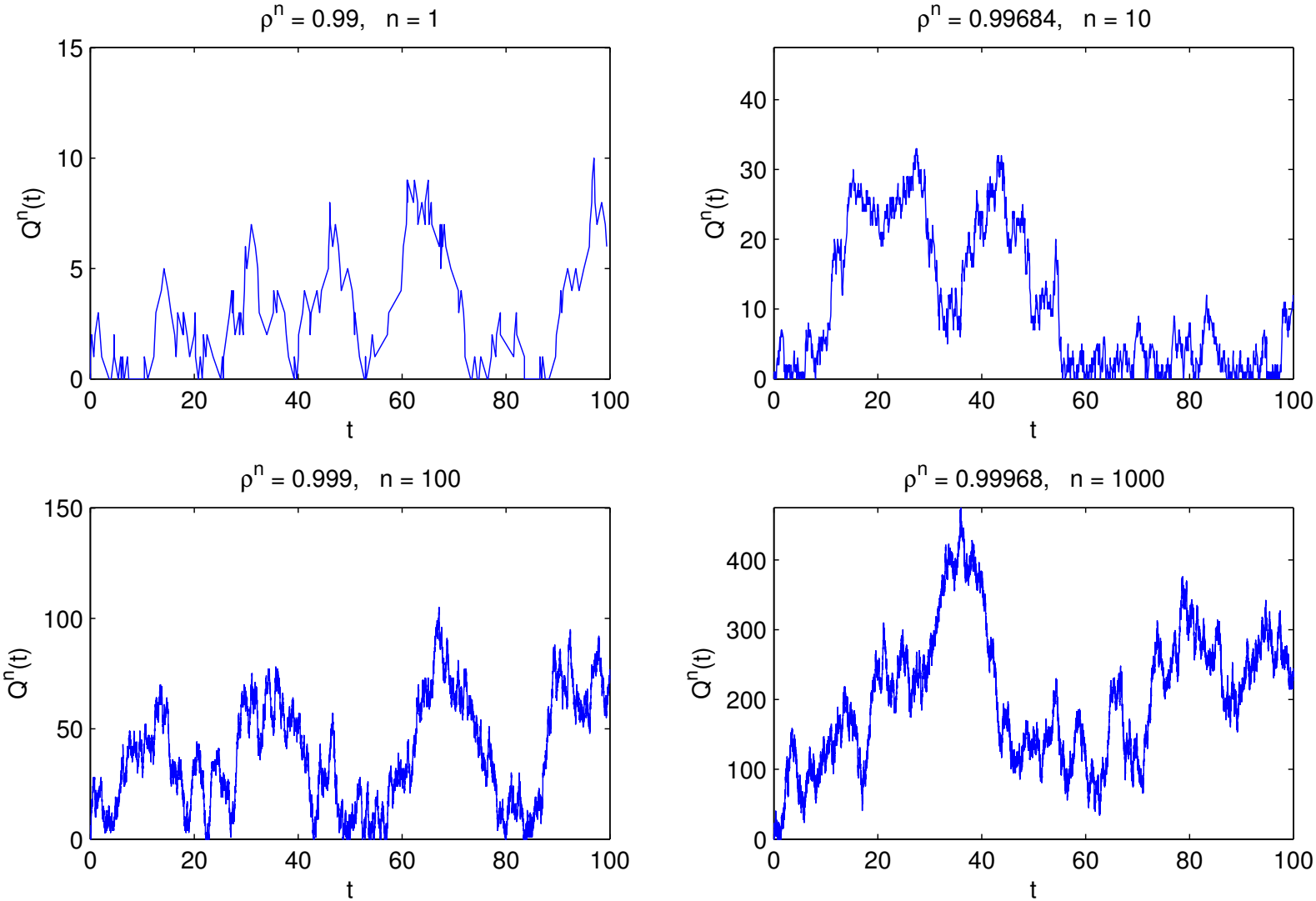
$M/M/1+M$, $\rho > 1$, Diffusion Scale $O(\sqrt{n})$



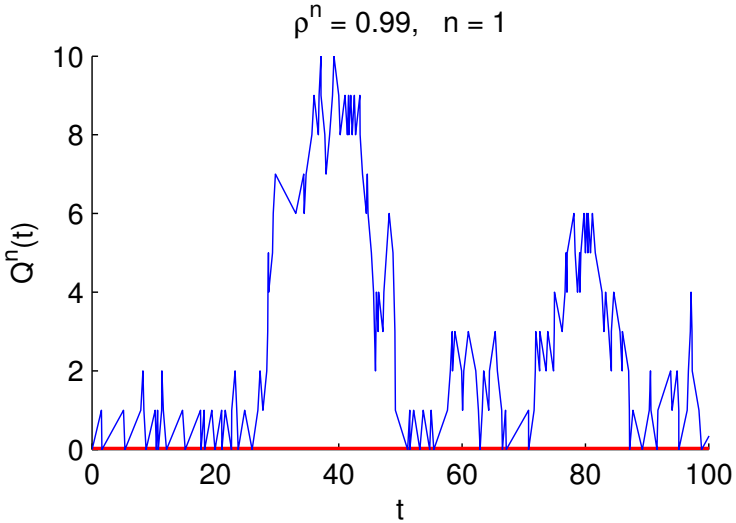
$M/M/1$, $\rho \approx 1$, Fluid Scale $O(n)$

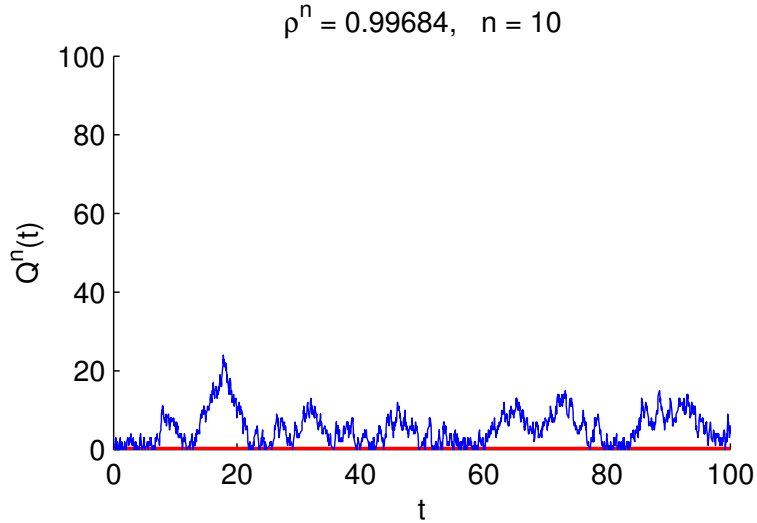


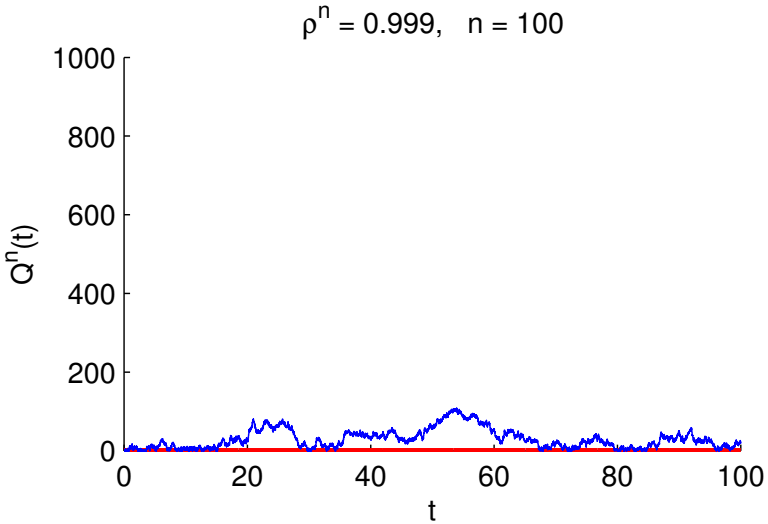
$M/M/1$, $\rho \approx 1$, Diffusion Scale $O(\sqrt{n})$

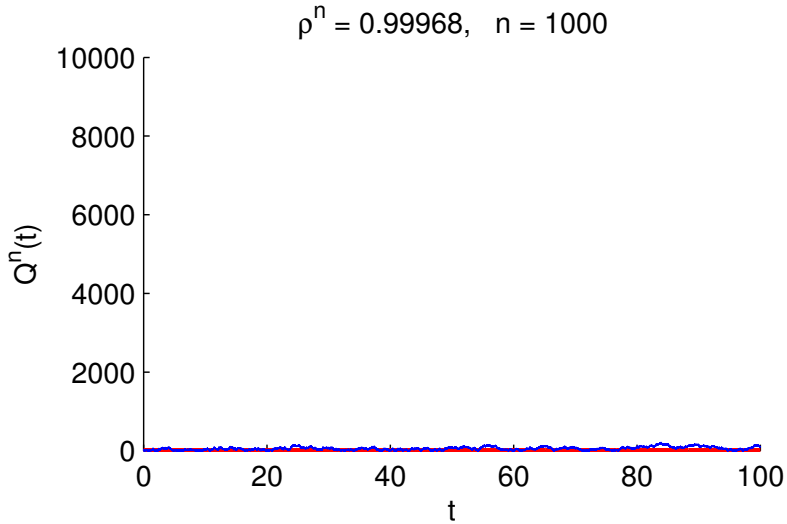


$M/M/1+M$, $\rho \approx 1$, Fluid Scale $O(n)$

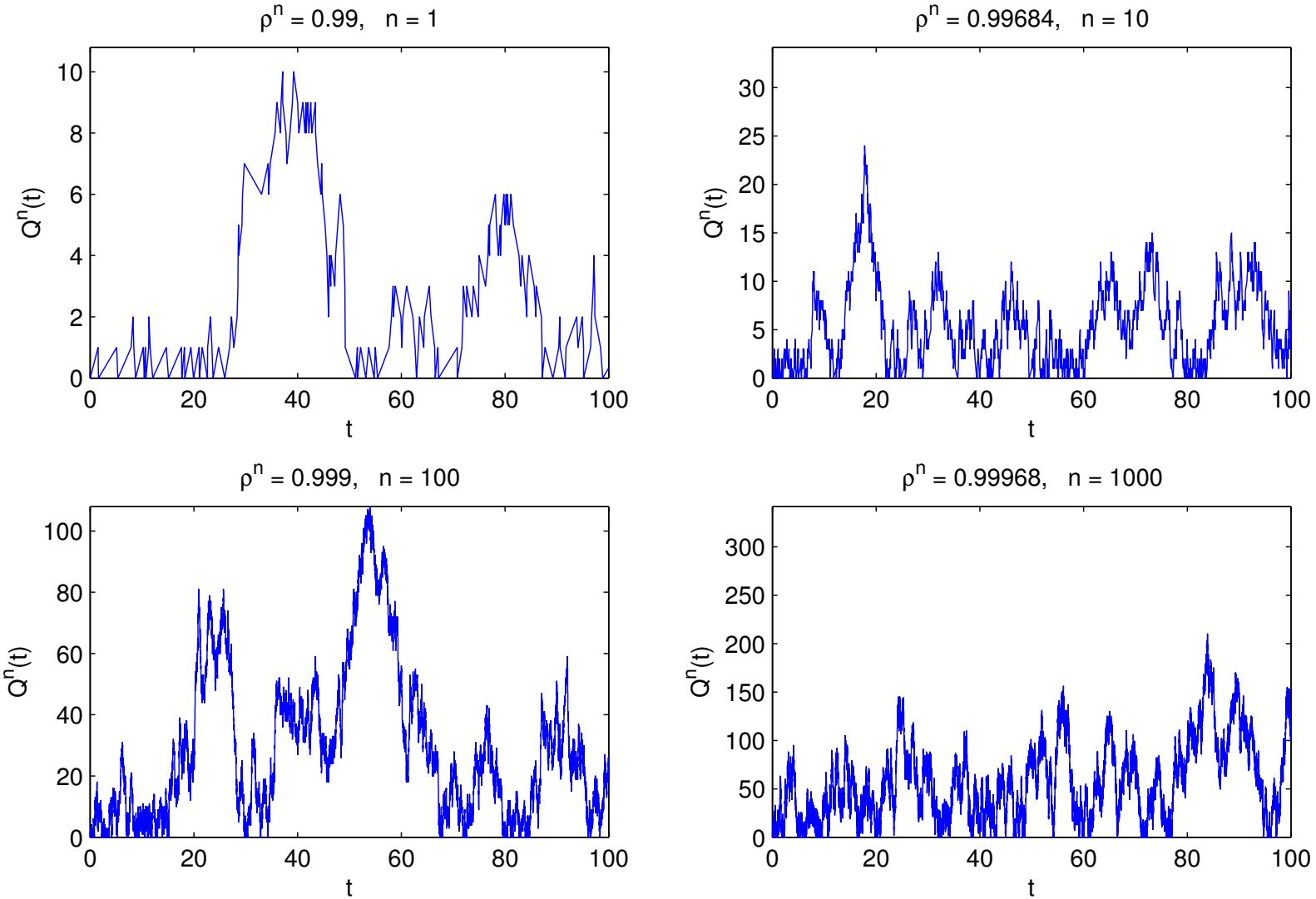




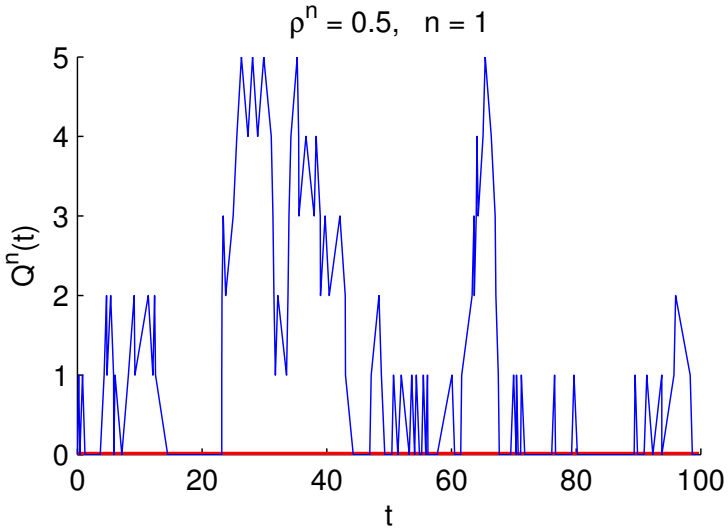


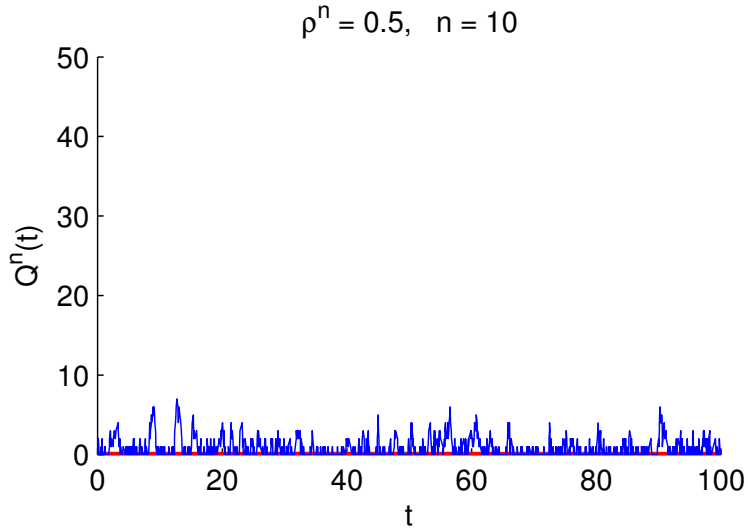


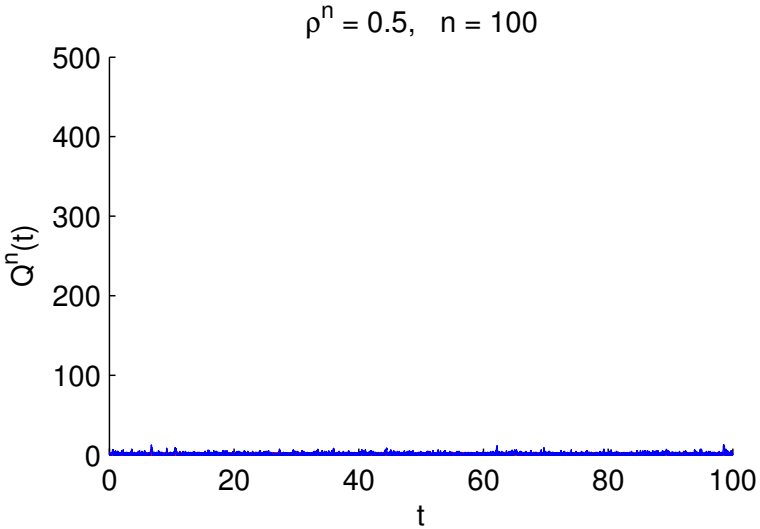
$M/M/1+M$, $\rho \approx 1$, Diffusion Scale $O(\sqrt{n})$

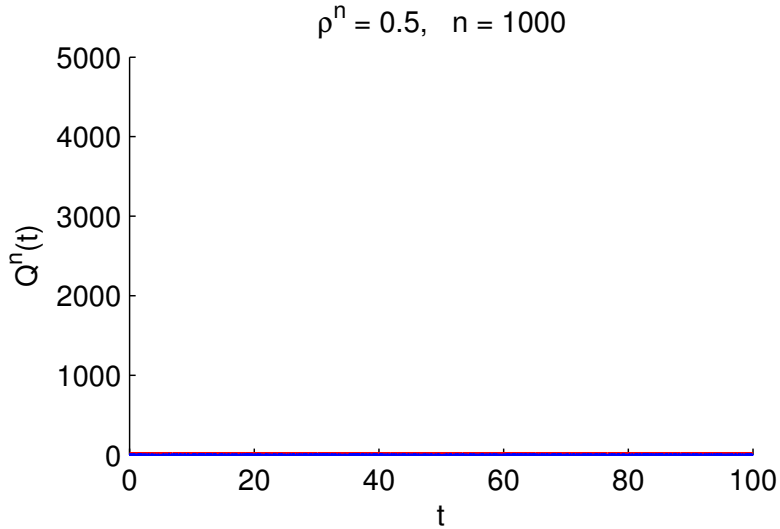


$M/M/1$, $\rho < 1$, Fluid Scale $O(n)$

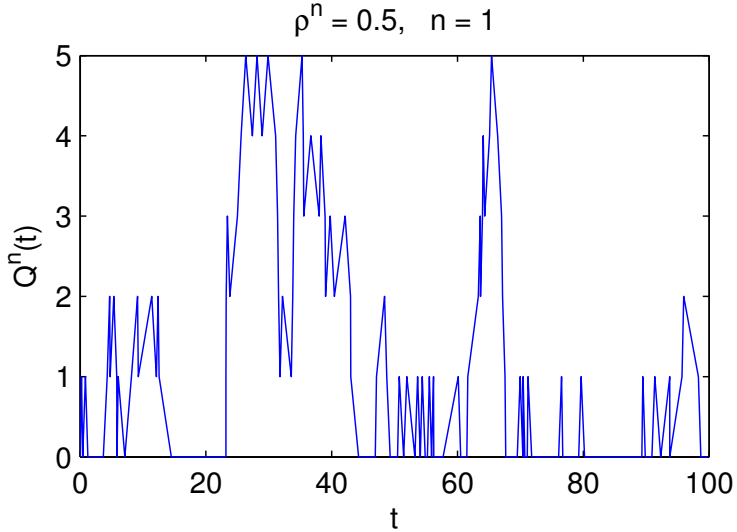


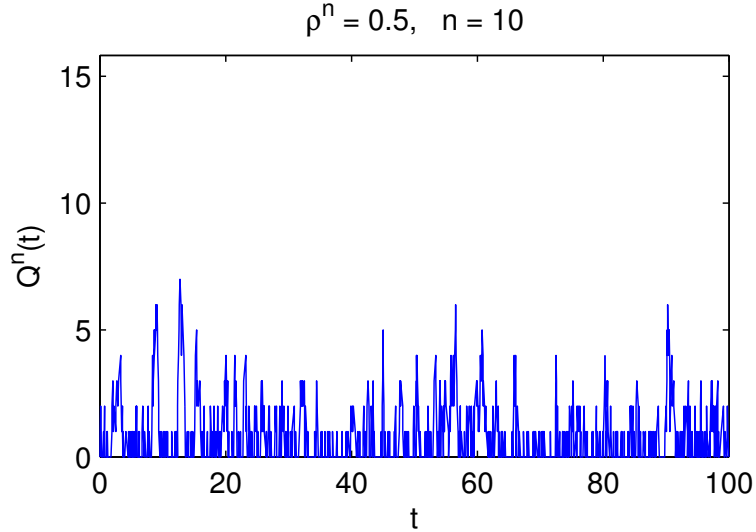


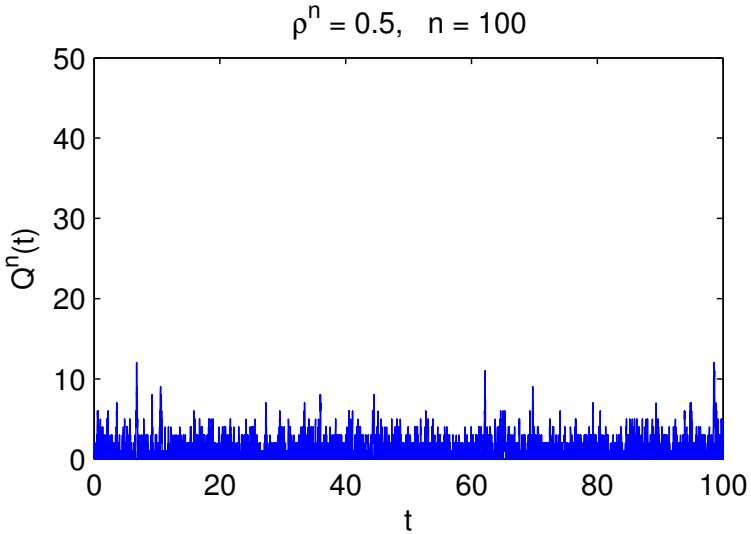


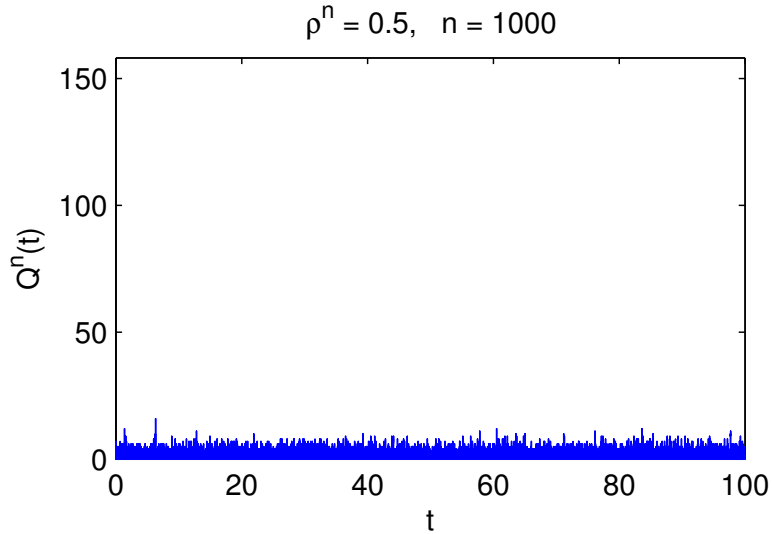


$M/M/1$, $\rho < 1$, Diffusion Scale $O(\sqrt{n})$









Quick recap

- LOB can be modeled as a multiclass queueing system
- data analysis:

$$\lambda \gg \mu \quad \text{and} \quad \text{cancellations} \gg \mu$$

- large scale queues can be approximated via tractable ODE or diffusion

  their analysis generate insights

  – expected waiting times
  – cancellation dynamics
  – ...

Next: study 4 problems encountered in analysis and trade execution in LOB

Outline

- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

Deterministic (mean-field) models of LOB dynamics

3. Transient dynamics, cancellations, and queue waiting times

   - single type of order flow
     - waiting time & equilibrium depth
     - queue position as fcn of elapsed waiting time
   - a view in the data
     - realized delays vs. delay estimates & cancellation flows
   - two types of order flow
     a) algo flow: exponential cancellations
     b) MM flow: event driven arrivals; state dependent cancellations
     - waiting time; depth; queue position as fcn of elapsed waiting time
   - back to the data

Motivating question #1

- **estimation of expected delay until a limit order gets filled**
- **related questions:**
  - **estimate queue position while in queue**
  - **estimate residual delay until an order gets filled while in queue**

– relevant in deciding when to place limit orders taking into account scheduling objective

– routing of orders across exchanges
(that may differ in their expected delays)

– input to understanding adverse selection

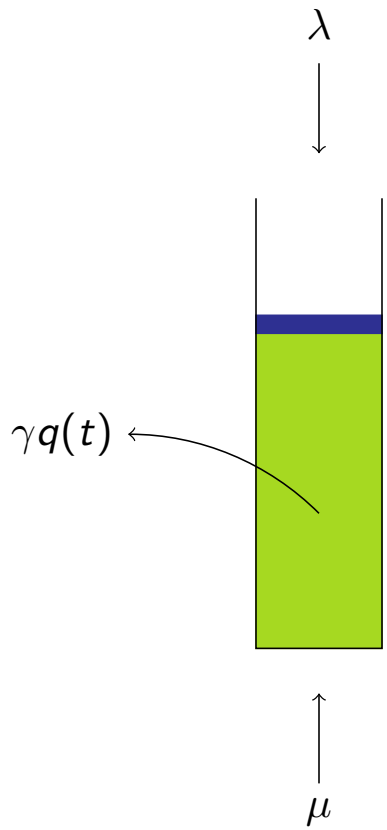
Two different estimates of delay in getting a fill

- Naive estimate (no cancellations, $\gamma = 0$):

$$w^0 = \frac{q(0)}{\mu}$$

- Proportional cancellations:

$$w^1 = \frac{1}{\gamma} \log \left(1 + \frac{q(0)\gamma}{\mu} \right)$$



derivation of w^1 uses fluid model of $M/M/1+M$:

$$w^1 = \inf\{t \geq 0 : q(t) = 0\}$$

ODE: $\dot{q}(t) = -\mu - \gamma q(t)$

$$\Rightarrow \quad q(t) = -\frac{\mu}{\gamma} \left(1 - e^{-\gamma t}\right) + q(0) e^{-\gamma t}$$

Queue position as fcn of sojourn time s

$x(s)$ = queue position s time units after posting infinitesimal (patient) order

- No cancellations:

$$q(s) = q(0) - \mu s$$

  – linear progress through the queue

- Proportional cancellations (exp. patience):

$$x(s) = q(0) - \lambda \int_0^s e^{-\gamma t} dt = \left(q(0) - \frac{\lambda}{\gamma} \right) + \frac{\lambda}{\gamma} e^{-\gamma s}$$

  – non-linear movement thru queue; impatient traders cancel early

Residual delay as fcn of queue position at time s

$x(s)$ = queue position s time units after posting infinitesimal (patient) order

- No cancellations:

$$w^1(x(s)) = \frac{x(s)}{\mu}, \quad x(0) = q(0)$$

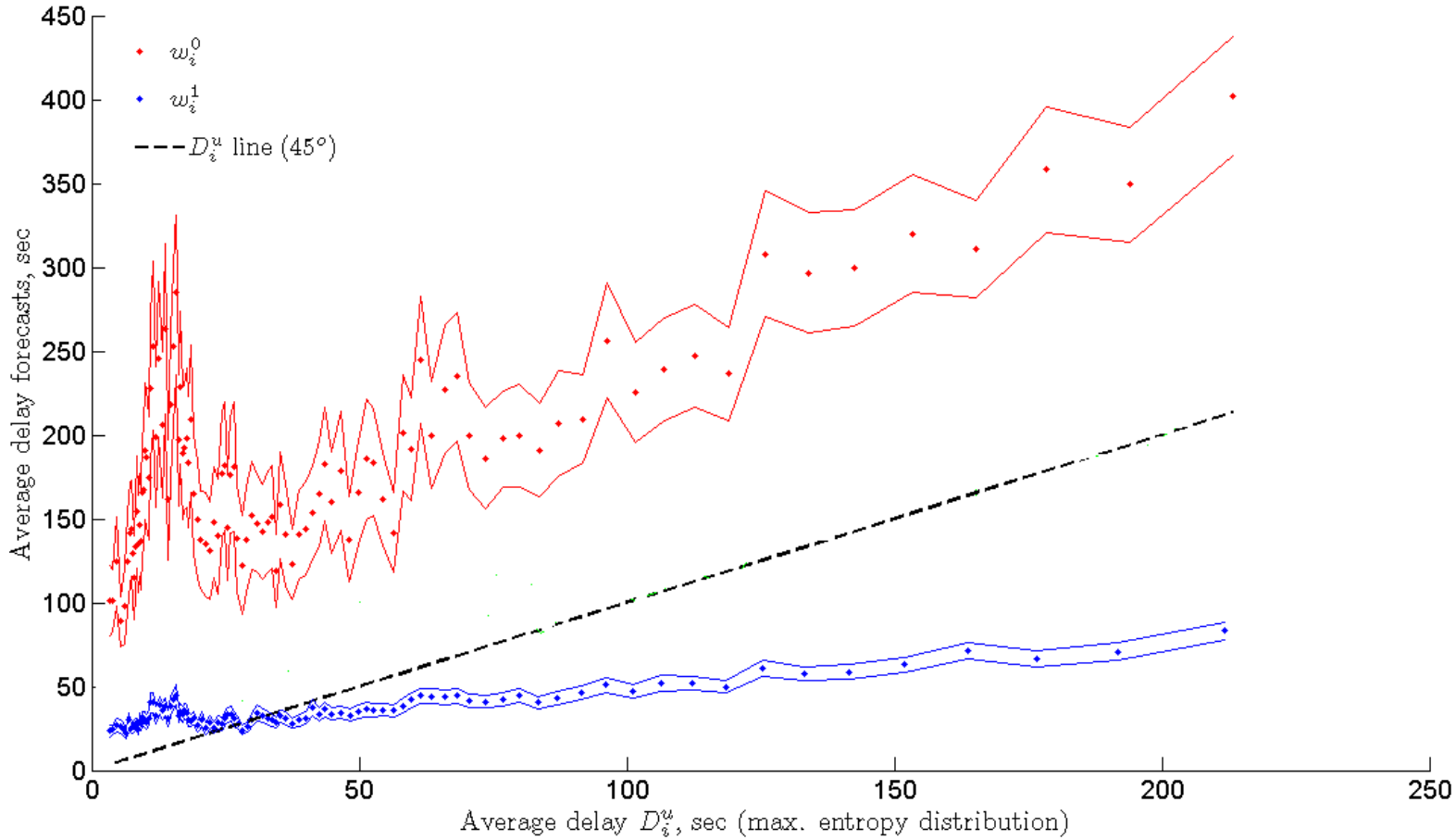
- Proportional cancellations (exp. patience):

$$w^2(x(s)) = \frac{1}{\gamma} \log\left(1 + \frac{x(s)\gamma}{\mu}\right) \quad x(0) = q(0)$$

Realized delays vs. estimates

- dataset: 325,000 algo limit orders, Mar-Apr 2012, ≈500 symbols
- fields: date, time (ms), exchange, symbol, buy/sell, parent strategy (e.g., VWAP), outcome, waiting time (till execution or cancellation)
- we estimated model parameters using trailing 3 minute statistics (TAQ)
- filtered symbols with too few points, to end with 109,000 orders, 268 symbols
- uncensored delay observations (the data set was censored due to cancels (65% of orders)

Realized delays vs. estimates (sample: 325,000 algo orders Mar-Apr 2012)



Realized limit order delays D^u (x axis) compared to delay estimates with proportional cancelations (blue), or no cancelations (red). Realized delays uncensored (max entropy).

Treatment of cancellations seems relevant to accuracy of delay estimates

- $\sim$ 80% of orders get cancelled
- disregarding cancellations seems too drastic of a simplification
- exponential patience / proportional cancellations appear too optimistic

Alternate model: constant (state-independent) cancellation intensity

$$\dot{q}(t) = \lambda - \mu - \eta, \quad 0 > \eta \geq \lambda - \mu$$

- $v(s) = \mu + (x(s)/q(0))\eta$ = speed of moving through queue after s time

$$\begin{aligned} x(s) &= q(0) - \int_0^s v(t)dt \\ &= q(0) - \mu s - \int_0^s (\eta/q(0))x(t)dt \;\Rightarrow\; \dot{x}(s) = -\mu - (\eta/q(0))x(s) \end{aligned}$$

It follows that

$$w^2 = \inf\{t \geq 0 : x(t) = 0\} = \cdots = \frac{q(0)}{\eta} \log\left(\frac{\lambda}{\mu}\right)$$

- If queue is stable, then $\eta \geq \lambda - \mu$. Set $\eta = \lambda - \mu$.

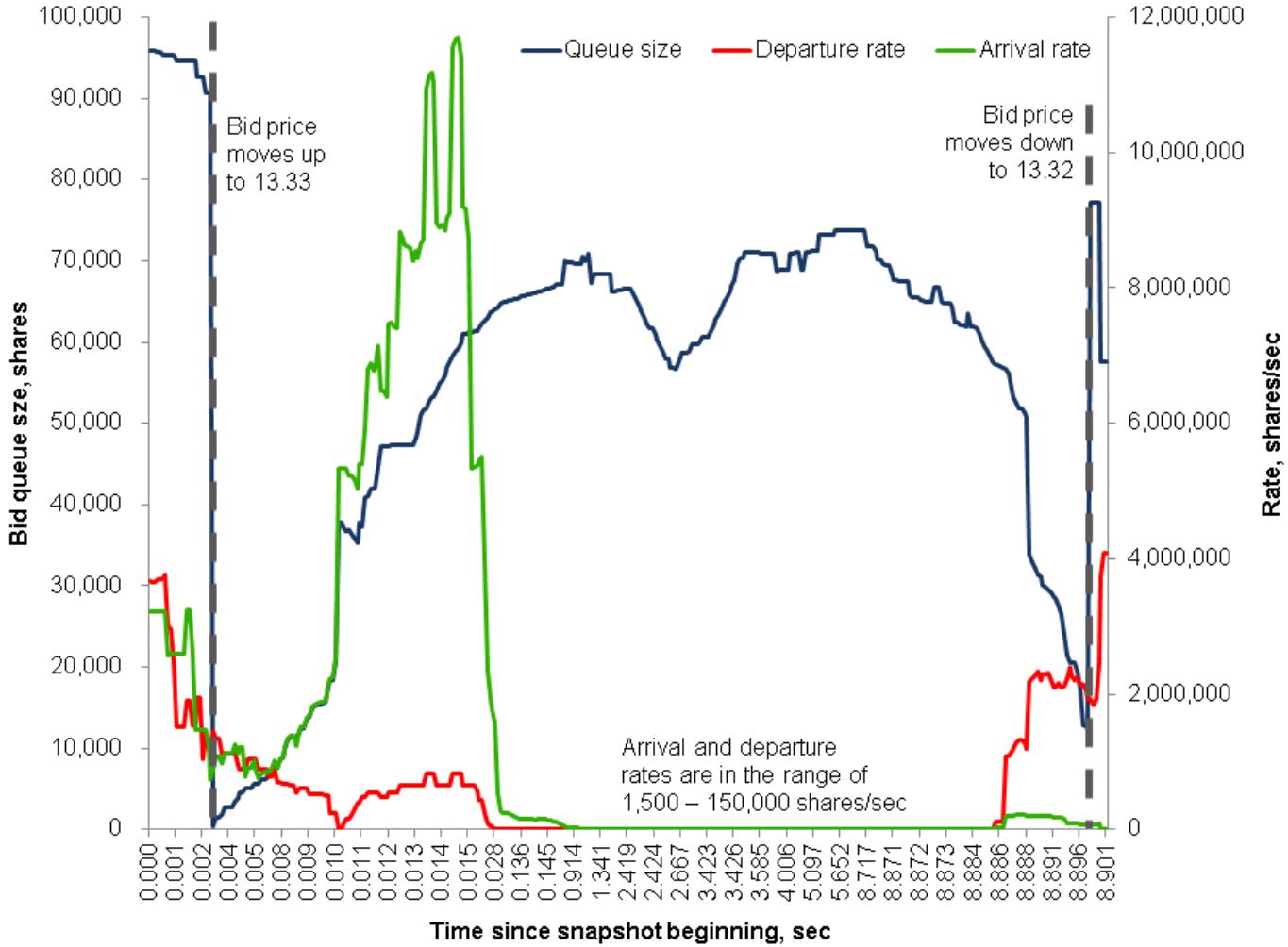
$M/M/1+M$: equilibrium depth q_∞ s.t. $\lambda = \mu + \gamma q_\infty \;\Rightarrow\; \eta = \gamma q_\infty$.

$$\text{if } q(0) = q_\infty, \quad w^1 = \frac{1}{\gamma} \log\left(1 + \frac{\gamma q(0)}{\mu}\right) = \frac{q(0)}{\eta} \log\left(\frac{\lambda}{\mu}\right) = w^2$$

if $q(0) < q_\infty$, then $w^2(q(0)) < w^1(q(0))$ (and vice versa)

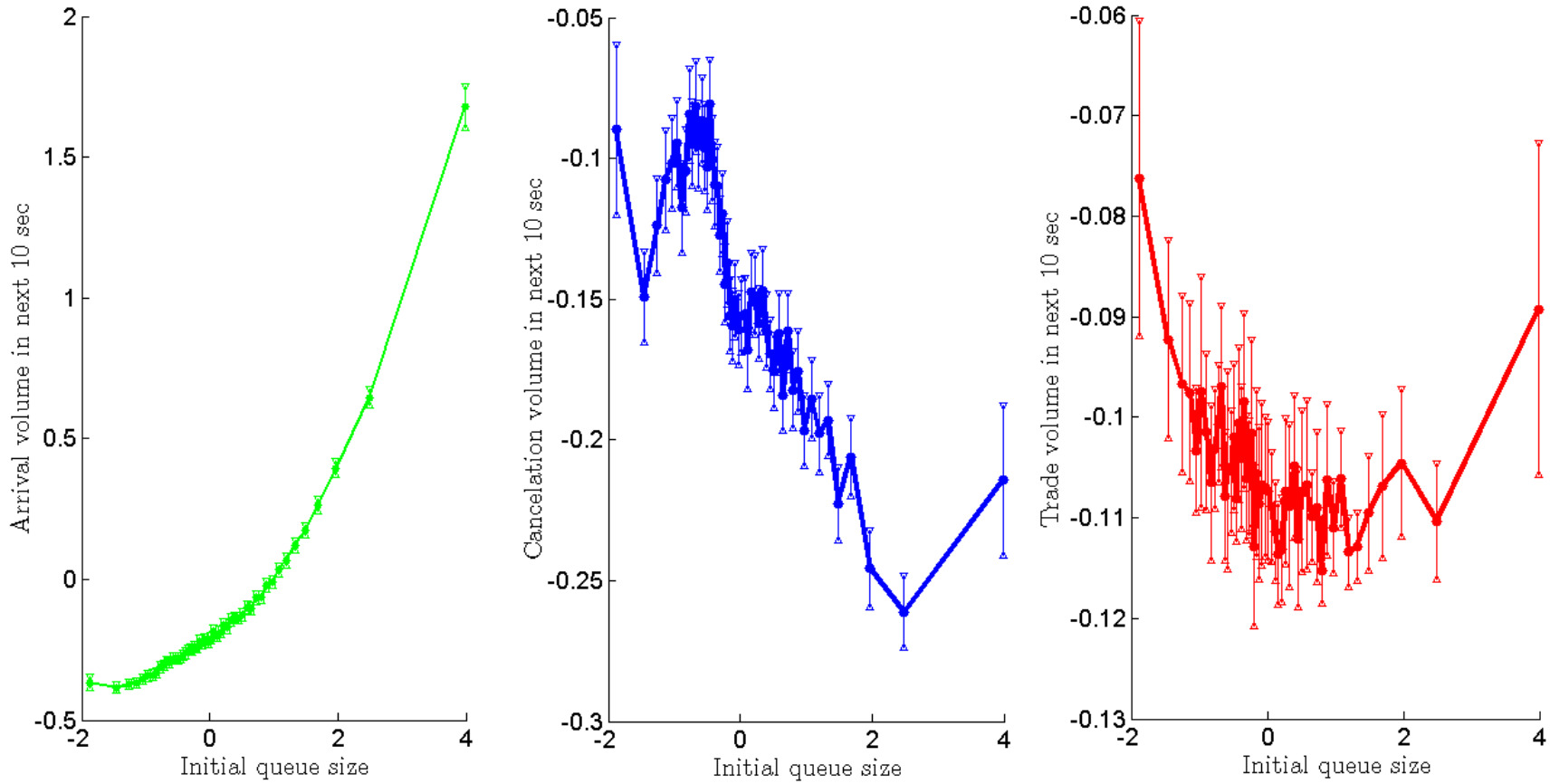
need more nuanced model to estimate cancellation effect on delay

Event and queue dynamics over a single price change



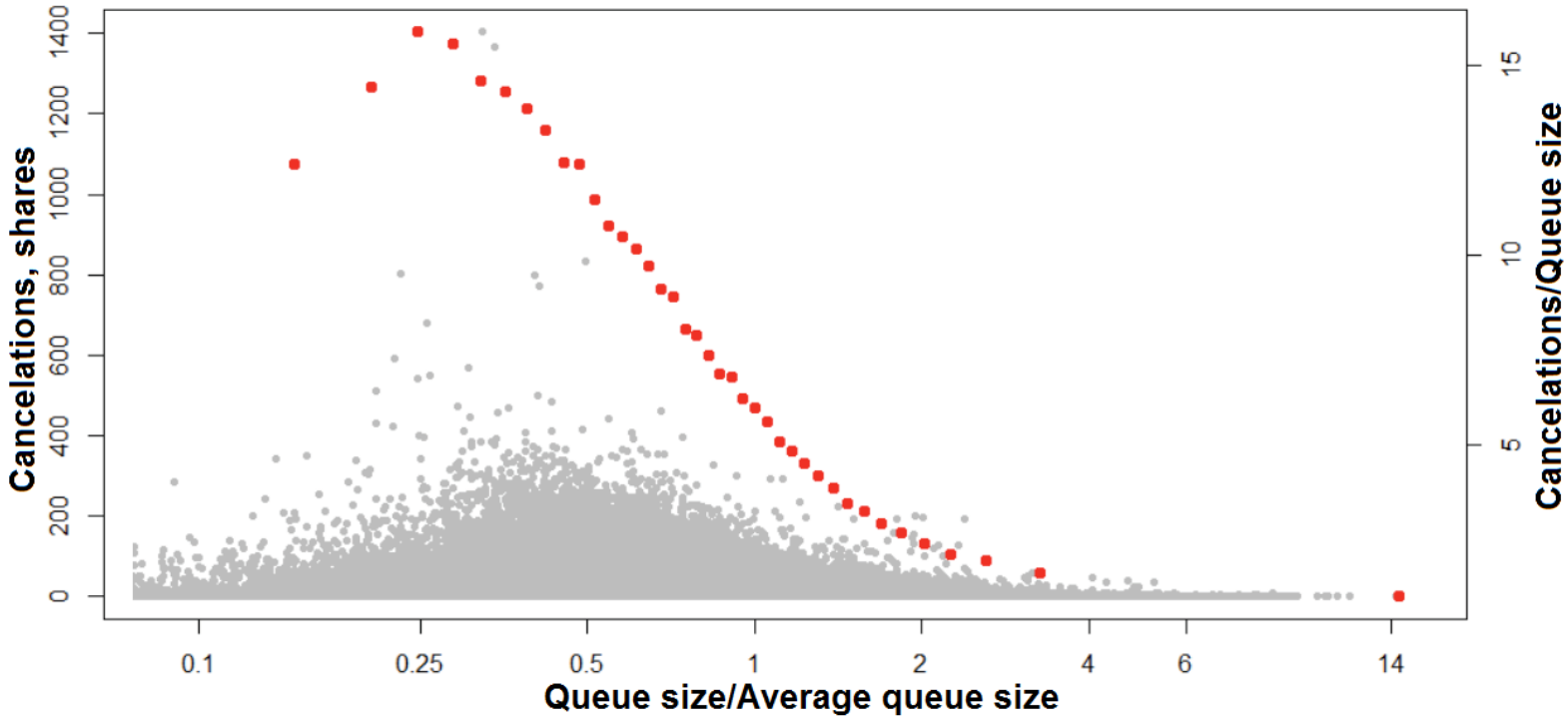
Trading episode in BAC stock on 6/18/2013 starting around 11:30:39

Order flows depend on LOB state



Standardized arrival and cancelation volumes for DJIA stocks - more orders cancel from small queues, less orders arrive to small queues.

Cancelations depend on LOB state



State-dependent cancelations - more orders cancel from small queues. (grey: (cancellations in δt intervals) (in shares); red: (cancellations in $\delta t)/Q_t$)

- exp. patience $\Rightarrow$ proportional cancellation model $\approx \gamma Q_t \delta t$
  $\Rightarrow$ (cancellations in $\delta t)/Q_t \approx \gamma$ (i.e., constant)
- data shows normalized cancellation intensity $\nearrow$ as normalized queue size $\downarrow$

Bursty event behavior & cancellation mechanism

Observations:

- Event rates increase when queues are small (and likely to get depleted)
- Cancellations also increase when queues are small
  - why?
  - does it matter in estimating delays and in order placement?

Deterministic (mean-field) models of LOB dynamics

3. Transient dynamics, cancellations, and queue waiting times

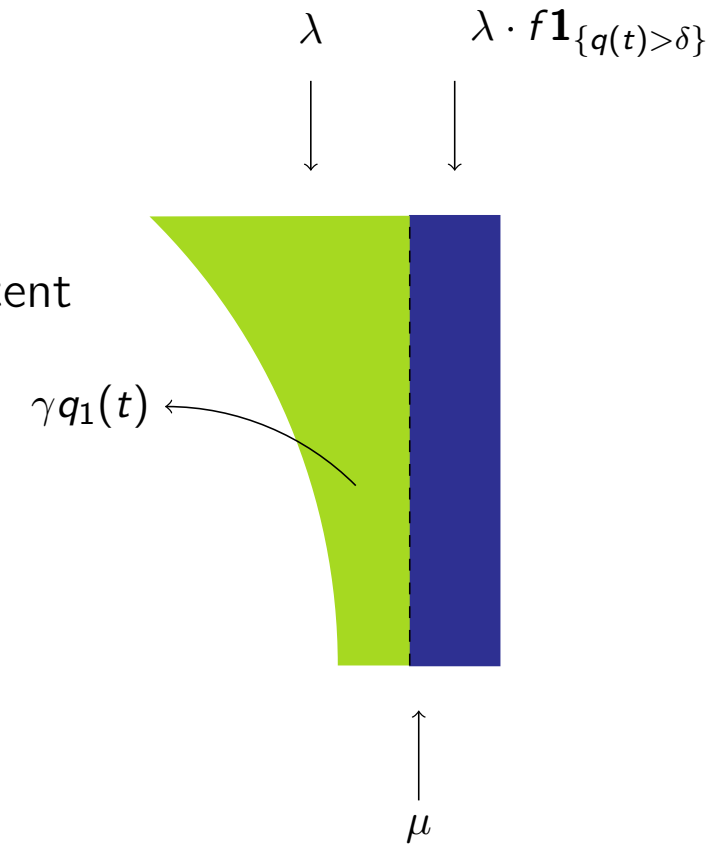
   - single type of order flow
     - waiting time & equilibrium depth
     - queue position as fcn of elapsed waiting time
   - a view in the data
     - realized delays vs. delay estimates & cancellation flows
   - two types of order flow
     a) algo flow: exponential cancellations
     b) MM flow: event driven arrivals; state dependent cancellations
     - waiting time; depth; queue position as fcn of elapsed waiting time
   - back to the data

Limit order FIFO queue with two types of order flow

- Type-1 orders (algorithmic flow):
  - Arrive (join the queue) according to a Poisson process with rate λ
  - Cancel according to finite deadlines $\sim \exp(\gamma)$
- Type-2 orders (MM) - event driven:
  - Join right after any other order joins, with probability F, as long as the queue length $q(t) > \theta$
  - Cancel all orders immediately whenever $q(t) \leq \theta$
    – for simplicity assume that θ is common across all type 2 orders
- Market orders arrive according to a Poisson process with rate μ.
- Intuition:
  - when $q(t)$ is small, a cascade of type-2 order cancelations is likely
  - when $q(t)$ is large, type-2 orders increase depth and waiting times ("order crowding")

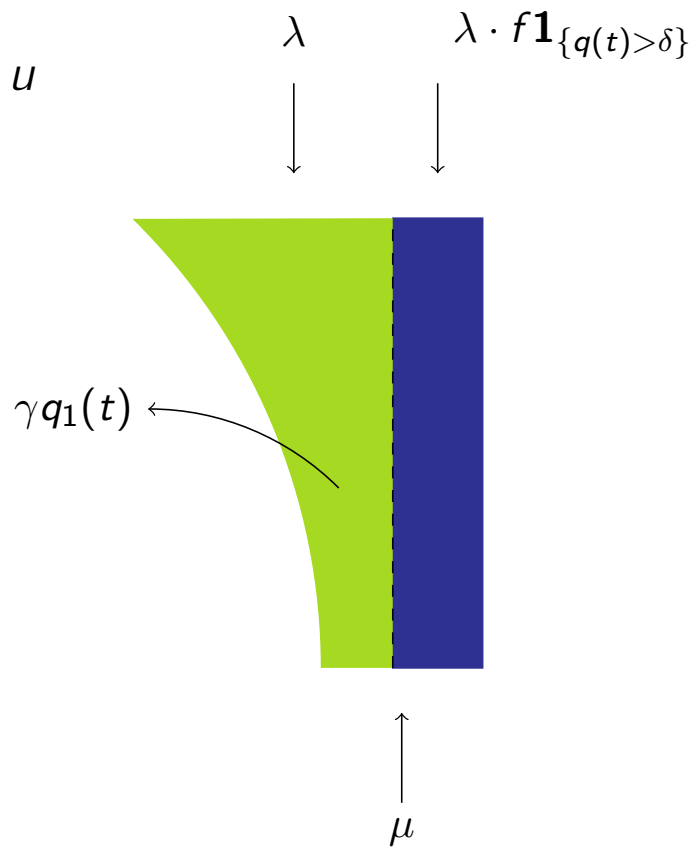
Associated fluid model

- $q_i(t)$ = type i orders, $i = 1, 2$
- $q(t) = q_1(t) + q_2(t)$ = total queue content
- cancellation behavior:
  – type 1: $-\gamma q_1(t) \quad \Leftarrow \quad \xi \sim \exp(\gamma)$
  – type 2: all $q_2(t)$ cancels if $q(t) \leq \theta$
- $\alpha(t)$ = % of μ that trades with type 1



Queue density as a function of sojourn time

- $\zeta_i(t,u)$ = type i density at time t of age u
- $q_i(t,y) = \int_0^y \zeta_i(t-u,u)du$
  (type i content at t of age $\leq y$)
- $\dot{\zeta}_1(t,u) = -\gamma\zeta_1(t,u)$
- $\tau(t)$ = age of HOL orders (= delay)
- $\alpha(t) = \dfrac{\zeta_1(t,\tau(t))}{\zeta_1(t,\tau(t)) + \zeta_2(t,\tau(t))}$
- $q_i(t) = q_i(t,\tau(t))$



Fluid model dynamics

- Total queue dynamics

$$\begin{aligned}
\dot{q}_1(t) &= \lambda - \gamma q_1(t) - \alpha(t)\mu \\
\dot{q}_2(t) &= (\lambda f - (1-\alpha(t))\mu)\,\mathbb{1}_{\{q(t)>\theta\}} - q_2(t)\delta\left(\mathbb{1}_{\{q(t)>\theta\}}\right) \\
\alpha(t) &= \frac{\zeta_1(t-\tau(t),\tau(t))}{\zeta_1(t-\tau(t),\tau(t)) + \zeta_2(t-\tau(t),\tau(t))}
\end{aligned}$$

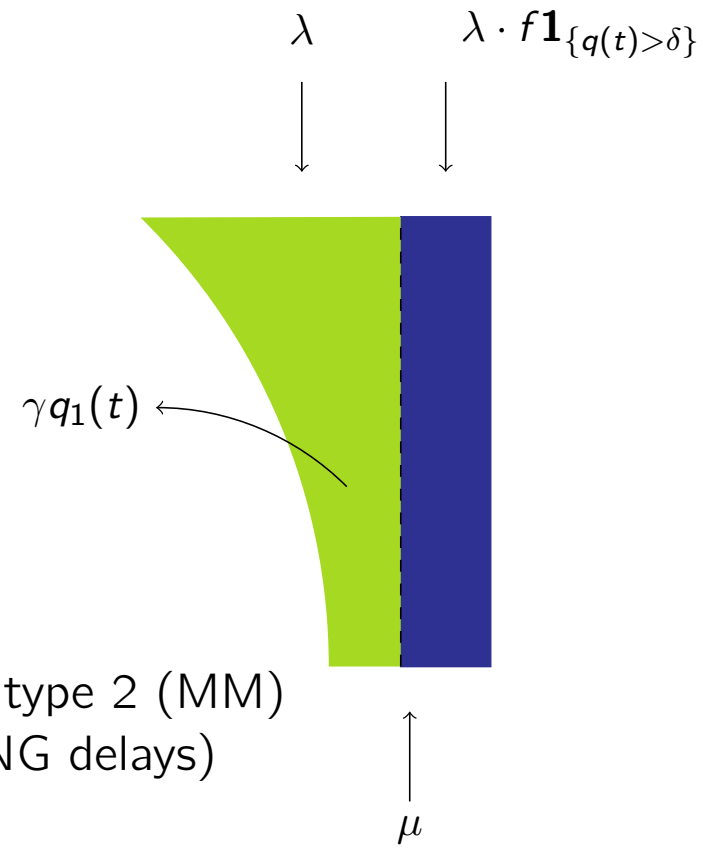
- Queue density dynamics

$$\begin{aligned}
&\zeta_1(t,0) = \lambda, t > 0 \\
&\zeta_2(t,0) = \lambda f\,\mathbb{1}_{\{q(t)>\theta\}}, t > 0 \\
&\frac{\partial \zeta_1(t,u)}{\partial u} = -\gamma\zeta_1(t,u), t \geq 0, 0 < u \leq \tau(t) \\
&\zeta_2(t,u) = \zeta_2(t-u,0)\mathbb{1}_{\left\{\min\limits_{0\leq v\leq u}(q(t-v))>\theta\right\}}, t \geq 0, 0 < u \leq \tau(t)
\end{aligned}$$

"Regular" queue profiles

a) large queue; b) type-2 doesn't cancel; c) exp. thinning of type-1

- $\zeta_1(t,u) = \lambda e^{-\gamma u}, \quad u \leq \tau(t)$
- $\zeta_2(t,u) = \lambda f, \quad u \leq \tau(t)$
- queue content takes intuitive shape
- in deep queues, most orders in front are type 2 (MM) (the only ones that survive/tolerate LONG delays)



Who trades & after how long?

- if queue is long, then it's profile must be regular
- once profile becomes regular, then it stays regular
  & profile always becomes regular after sufficient time
- fraction of trades against type 1 (algo) orders: $\alpha(t) = \dfrac{e^{-\gamma\tau(t)}}{e^{-\gamma\tau(t)} + f}$
- waiting time: $\tau(t) = \dfrac{1}{\gamma} \log\left(\dfrac{\lambda}{\lambda - \gamma q_1(t)}\right)$ for all $t > 0$

Queue composition in regular profiles

- the dynamics of type 1 simplify

$$\dot{q}_1(t) = \lambda - \gamma q_1(t) - \frac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}\mu$$

- $q_2(t) = \lambda f \tau(t) = \frac{\lambda f}{\gamma} \log\left(\frac{\lambda}{\lambda - \gamma q_1(t)}\right)$
- total queue length is determined by $q_1(t)$:

In steady state:

$$\alpha^* = 1 - \frac{\lambda f}{\mu} \quad \text{and} \quad \tau^* = \frac{1}{\gamma} \log\left(\frac{\lambda}{\mu - \lambda f}\right)$$

and

$$q_1^* = \frac{\lambda(1+f) - \mu}{\gamma} = \frac{\lambda - (\mu - \lambda f)}{\gamma}, \qquad q_2^* = \frac{\lambda f}{\gamma} \log\left(\frac{\lambda}{\mu - \lambda f}\right)$$

A new formula for delay in a LOB with heterogeneous order flow

- $w^3 = \frac{1}{\gamma} \log\left(\frac{\lambda}{\mu - \lambda f}\right)$
- $w^3 > w^1 \quad (w^1 = \frac{1}{\gamma} \log\left(\frac{\lambda}{\mu}\right)$ all algo flow)

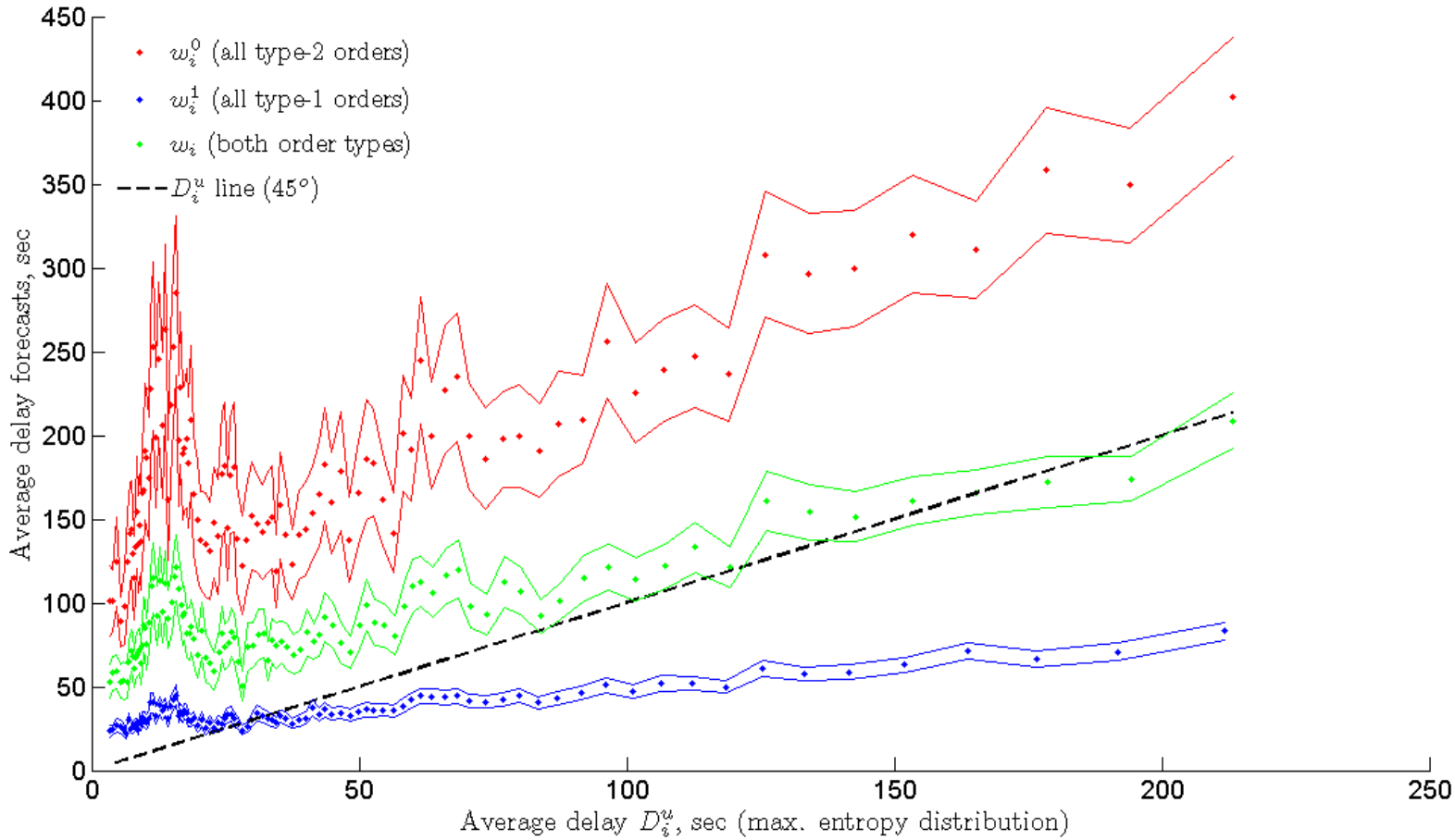
  – if $\lambda(1+f)/\mu = 5$ and $\lambda f \approx 3/4\mu$, then $w^3 \approx 2w^1$

- starting from an arbitrary IC and assuming profile is regular, w^3 is computed by solving the following system of differential equations:

$$\dot{p}_1(t) = -\gamma p_1(t) - \alpha(t)\mu \quad \text{and} \quad \dot{p}_2(t) = -(1-\alpha(t))\mu$$
$$\dot{q}_1(t) = \lambda - \gamma q_1(t) - \frac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}\mu$$
$$\alpha(t) = \frac{\lambda - \gamma q_1(t)}{\lambda(1+f) - \gamma q_1(t)}$$

IC $p_1(0) = q_1(0), p_2(0) = q_2(0)$ and TC $p_1(w^3) = p_2(w^3) = 0$.

Contrasting delay estimates against realized delays



Waiting times with one-type model w^1, with no cancelations w^0, with two-type model w compared to realized delays D^u.

Rough intuition

- need to estimate mixture of patient vs. inpatient orders
- incorporate "crowding" out effect of patient orders
- resulting delay estimate is not as pessimistic as q/μ (no cancellations)
- fragmentation . . . need delay estimates for each exchange

Flow heterogeneity has 1st order effect on LOB behavior

- Important to model heterogenous trade behaviors
  – order placement
  – cancellations
  – market orders
- possible explanation for anomalously long waiting times in large queues despite large cancellation rates (some orders never cancel, and in long queues only these orders survive)
- significant differences on state-dependent behavior across types of flow
  – MM/HFT flow sensitive to AS costs, primarily state-dependent policies
  – algo flow primarily driven by strategy participation considerations, mostly "timer-based"
- estimating state- & price-dependent event rates should deal with the above distinction

Followup question #4

- **probability that an order will get filled**
- **conditional probability that this will be an "adverse" fill**
- **estimate adverse selection costs as a fcn of queue position**

return to this tomorrow (towards the end of these slides)

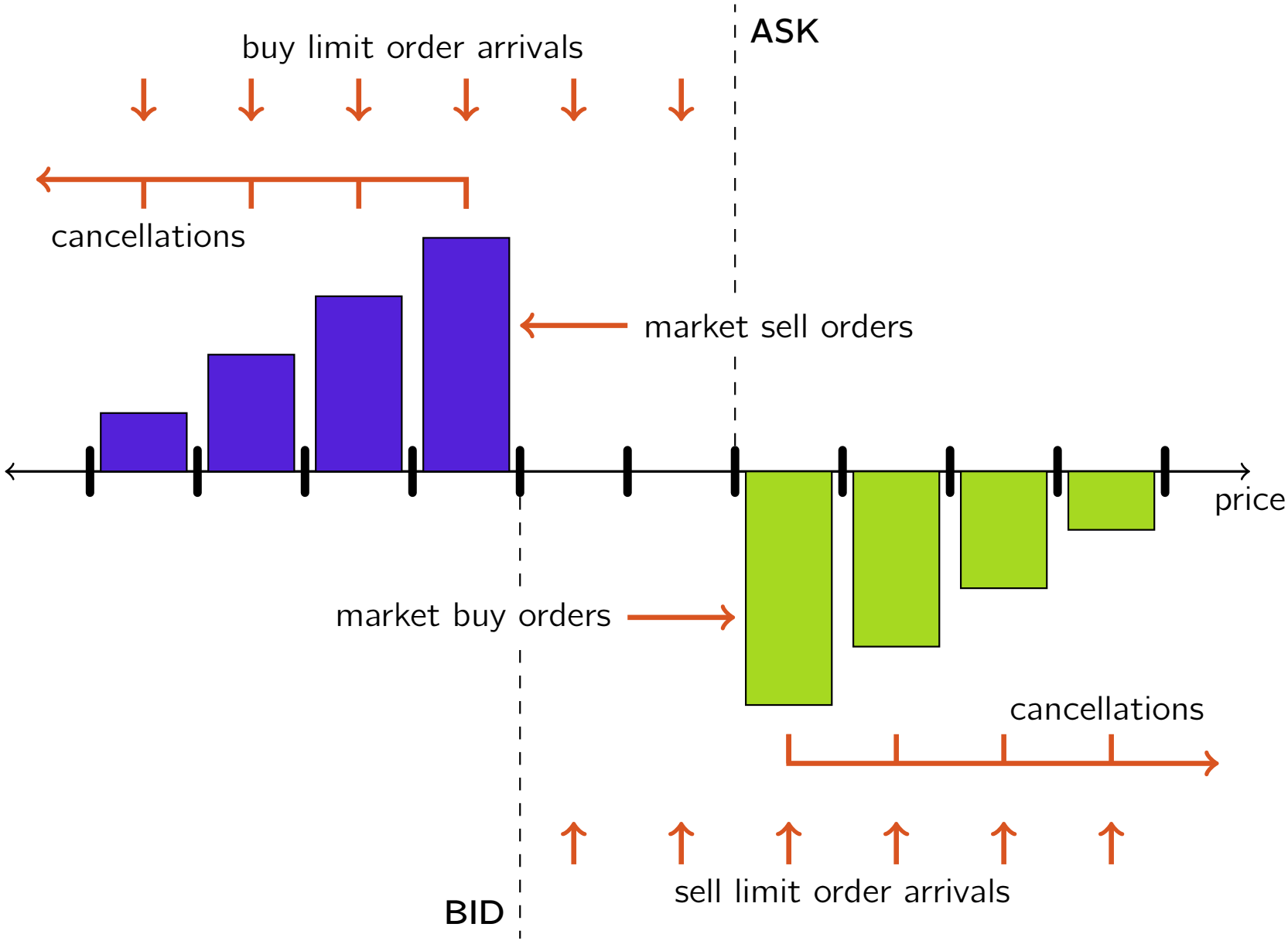
Outline

- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

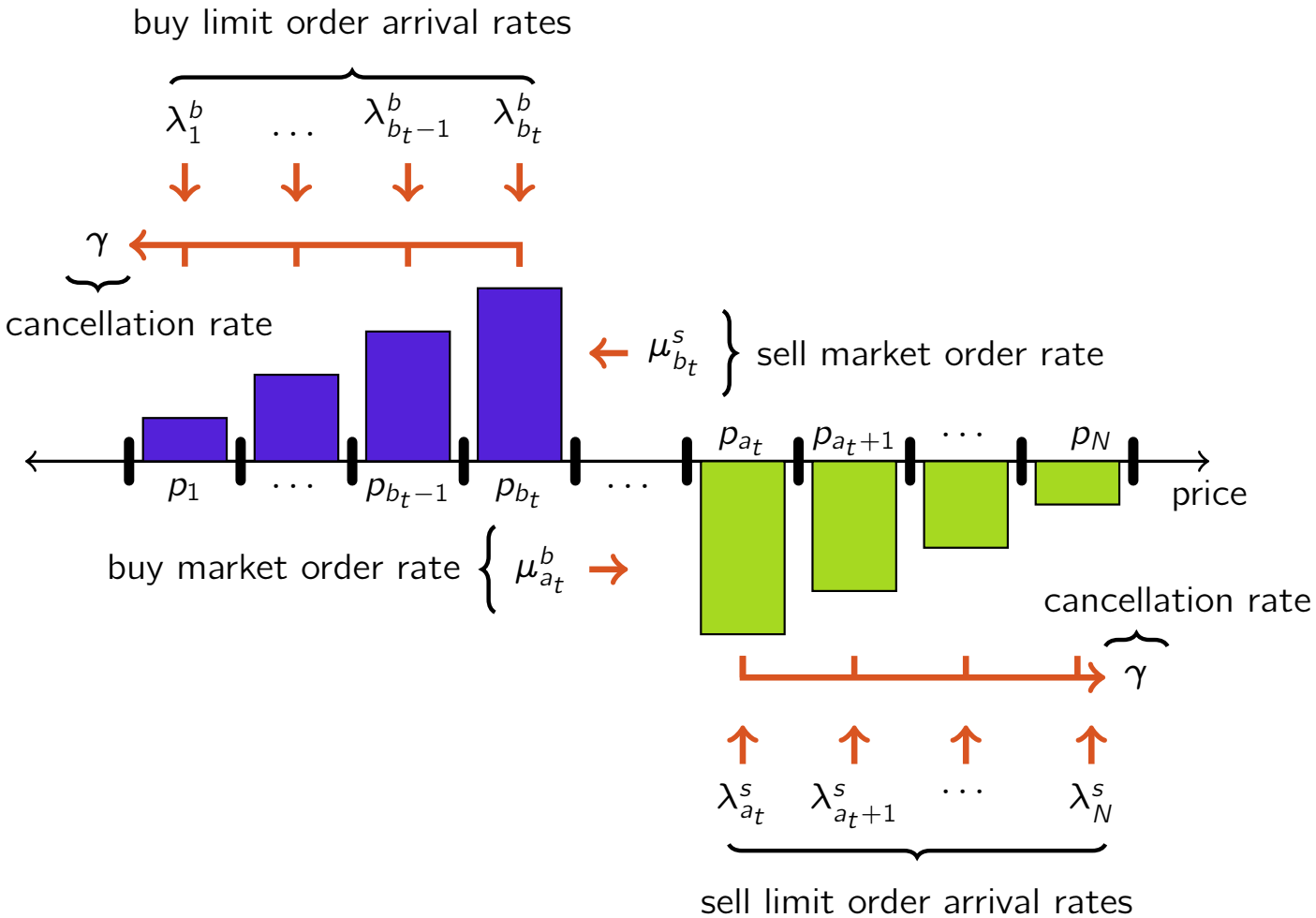
Deterministic (mean-field) models of LOB dynamics

4. Execution in a LOB and a microstructure model of market impact
   - formulate stylized optimal execution problem in LOB
   - characterize optimal execution policy & associated cost
   - a microstructure market impact model
   - calibration of the microstructure market impact model on trade data

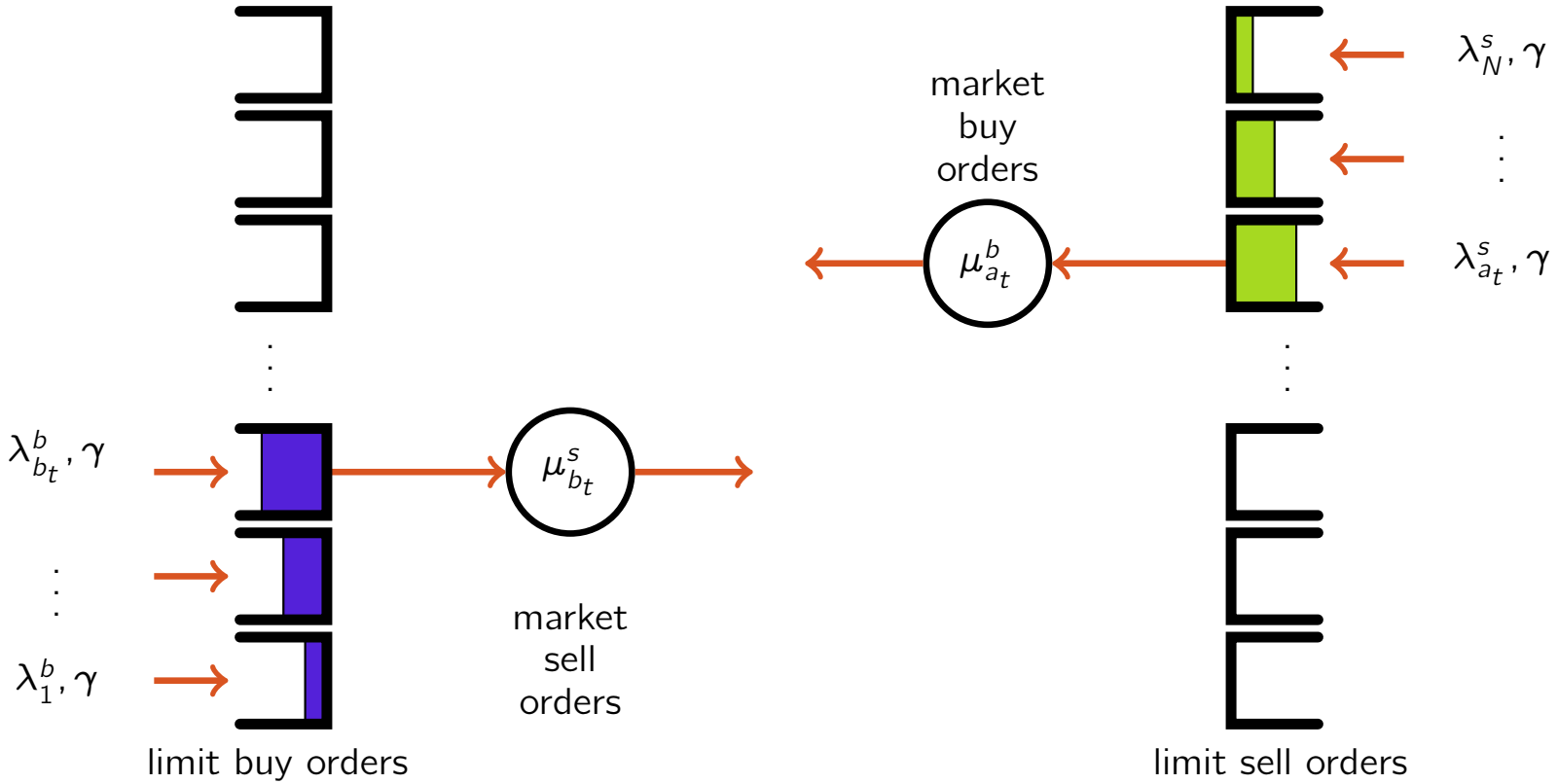
The Limit Order Book (LOB)



LOB: event driven (short-term) view



LOB re-drawn as a multi-class queueing network



Motivating question #2: Optimal execution in LOB & market impact cost

objective: how to buy C shares within time T at the lowest price

controls: how much, when, at what prices to trade

- trade with limit orders / market orders
- trade with block trades / continuously submitted trades (rate upper bounded by κ_i)

  – T is same order of magnitude as the queueing delays ($\approx$ 1 - 5 min)

  – microstructure of the LOB impact execution policy and resulting costs

  – we focus on a stylized execution problem (tractable)
  . . . to generate insight on impact cost drivers

LOB fluid model dynamics

$$\dot{Q}_i^b(t) = \lambda_i^b \cdot \mathbf{1}\{i \leq b_t\} - \mu_i^s \cdot \mathbf{1}\{i = b_t\} - \gamma Q_i^b(t),$$

$$\dot{Q}_i^s(t) = \lambda_i^s \cdot \mathbf{1}\{i \geq a_t\} - \mu_i^b \cdot \mathbf{1}\{i = a_t\} - \gamma Q_i^b(t).$$

Main assumptions

- $\lambda_i^b > \mu_i^s$ (motivated from earlier data analysis)
- constant bid-ask spread, no limit orders inside spread
- if price moves, limit orders *slide*, queue positions maintained

LOB behavior

b_t = best bid at time t; a_t = best ask at time t

- (best bid & best ask do not change) $b_t = b_0$, $a_t = a_0$, for all $t \geq 0$,
- $Q(t) \to q^*$ as $t \to \infty$

$$q_i^{*,b} := \begin{cases} \lambda_i^b/\gamma & \text{if } 1 \leq i < b_0, \\ \dfrac{\lambda_i^b - \mu_i^s}{\gamma} & \text{if } i = b_0, \\ 0 & \text{if } b_0 < i \leq N, \end{cases} \qquad q_i^{*,s} := \begin{cases} 0 & \text{if } 1 \leq i < a_0, \\ \dfrac{\lambda_i^s - \mu_i^b}{\gamma} & \text{if } i = a_0, \\ \lambda_i^s/\gamma & \text{if } a_0 < i \leq N, \end{cases}$$

– top of book queues equilibriate to balance arrivals with trades + cancellations
– other queues balance arrivals with cancellations
– cf. Gao, Dai, Dieker, Deng

Optimal execution policy

- **Limit orders**
  - at time $t = 0$, to the best bid b_0, submit limit orders $C_L = \min\left\{\mu^s_{b_0}\left(T - \frac{1}{\gamma}\log\left(1 + \frac{\gamma}{\mu_{b_0}}Q^0(0)\right)\right)^+, C\right\}$;

- **Market orders** $(\kappa_i = \kappa, \quad \forall i)$
  - at time $t = 0$, to the best ask a_0, submit block trade $\min\{C - C_L, Q^s_{a_0}(0)\}$;
  - for time $0 < t < T$, to the best ask a_0, continuously submit trade at rate κ, or until C is filled;
  - at time $t = T$, clean up with block trade, may deplete multiple queues at higher price levels.

Practical considerations

- avoid clean up trade, especially if this is a slice of a longer trade
- often times micro-trader does not have to complete C by T
- account of multiple exchanges in deciding how much and where to post
- do not post all limit order quantity in one block to avoid information leakage

- policy predicated on the following assumption:
  - trader can execute continuously with market orders at rate κ (presumably low)
  - $\kappa_i = \kappa$ for all price levels i
    one may expect supply to increase at higher price levels (more later)

- ...

Deterministic (mean-field) models of LOB dynamics

4. Execution in a LOB and a microstructure model of market impact

   - formulate stylized optimal execution problem in LOB
   - characterize optimal execution policy & associated cost
   - a microstructure market impact model
   - calibration of the microstructure market impact model on trade data

Execution cost

$$\overline{IS} := \frac{\text{Total cost}}{C} - p = s/2 - s \cdot \frac{S_L}{C} + \sum_{k=1}^{N-a_0} k\delta \cdot \frac{C_{a_0+k}}{C}.$$

Simplifications:

- disregard cancellations on the near side (limit order term)
  – $S_L = \min\left\{\mu^s_{b_0}(T - w^0)^+, C\right\}$, where $w^0 = Q_b(0)/\mu^s$
- clean up cost: the number of price levels needed to complete the trade is

$$n := \frac{\left(C - S_L - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s} \approx \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s}$$

For large C: $\overline{IS} = \alpha_0 + \alpha_1 \cdot C + \alpha_2 \cdot C^2$

– α_0 captures limit order offset, expect to be (-ve)
– α_2 captures effect of the additional price levels needed, expect to be (+ve)

Microstructure market impact model

- implementation shortfall of a buying order
  - benchmark on the aggressive side: $p_{0,\text{mid}} + s/2 = a_0$

$\overline{IS} := \bar{p} - p_{0,\text{mid}} = s/2 + (\text{limit order benefit}) + (\text{higher price level adjustment})$

- keep insightful structure, simplify the functional form

$$\overline{IS} = s/2 \underbrace{- \frac{\min\left\{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+, C\right\}}{C} \cdot s}_{\text{limit order benefit}} + \underbrace{\frac{\delta}{2} \cdot \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s} + \frac{\delta}{2}}_{\text{higher price level adjustment}} .$$

Microstructure market impact model

- implementation shortfall of a buying order
  benchmark on aggressive side: $p_{0,\text{mid}} + s/2 = a_0$

$$\overline{IS} = s/2 \underbrace{- \frac{\min\left\{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+, C\right\}}{C} \cdot s}_{\text{limit order benefit}} + \underbrace{\frac{\delta}{2} \cdot \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s} + \frac{\delta}{2}}_{\text{higher price level adjustment}}$$

- *(Effect of limit orders)* decreasing in $\mu^s_{b_0}$, T, increasing in $Q^b_{b_0}(0)$
- *(Effect of top-of-book market orders)* decreasing in κ and $Q^s_{a_0}(0)$
- *(Effect of higher price market orders)* decreasing in $\bar{Q}^s$
- increasing in s, δ
- if κ_i increase as p_i grows, then cost exhibits sub-linear growth
  if κ_i grow linearly in p_i, then cost grows like $\sqrt{C}$ for large C

Deterministic (mean-field) models of LOB dynamics

4. Execution in a LOB and a microstructure model of market impact
   - formulate stylized optimal execution problem in LOB
   - characterize optimal execution policy & associated cost
   - a microstructure market impact model
   - calibration of the microstructure market impact model on trade data

Proprietary trade data

realized trade stats: 5min slices for 2013/7-2013/9, $> 1,800$ securities traded

	JUL 2013	AUG 2013	SEP 2013
Sample Size			
5min Slices	27,760	30,054	29,226
Parent Orders	3,396	3,607	3,882
Distinct Securities	988	896	885
Characteristics			
Average Daily Volume (shares)	3,014,000	2,595,000	2,509,000
Size of 5min Slices (shares)	1,294	1,043	849
Average Queue Length	10,280	21,730	17,750
Realized Participation Rate	9.60%	9.40%	8.39%
Price ($)	46.80	38.16	41.41
Spread ($)	0.031	0.025	0.025
Daily Volatility	2.23%	1.90%	1.94%
Implementation Shortfall (bps)	3.04	3.09	3.48

Calibration of auxiliary model parameters

Three quantities not directly observable from data: continuous trading rate κ, equilibrium queue length $\bar{Q}^s$, effective tick size δ

- calibration of κ:
  1. postulate $\kappa = \theta \cdot \mu$, assume θ is the same on the bid and ask side
  2. identify slices that: a) queue length at far side less than 1/3 average length; b) no price change
  3. generate forecast for nominal trading rate μ
  4. θ estimated as average ratio of executed quantity to the nominal trading rate
- $\bar{Q}^s$ approximated by time-averaged queue length at top of the book over each 5min interval
- σ as a proxy for δ

Microstructure market impact model

- microstructure market impact model

$$\overline{IS} = s/2 \underbrace{- \frac{\min\left\{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+, C\right\}}{C} \cdot s}_{\text{limit order benefit}} + \underbrace{\frac{\delta}{2} \cdot \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s} + \frac{\delta}{2}}_{\text{higher price level adjustment}}$$

- linear regression:

$$IS = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot (R^L s^*) + \beta_3 \cdot (R^M \delta^*) + \beta_4 \cdot \delta^*$$

  - $R^L := \dfrac{\min\left\{C, \left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)^+\right\}}{C}$
  - $R^M := \dfrac{\left(C - Q^s_{a_0}(0) - \kappa T\right)^+}{\bar{Q}^s}$

In-sample regressions (ADV $\geq 300,000$ shares; POV $\in (1\%, 30\%)$)

Monthly linear regression results for microstructure market impact model

	JUL 2013	AUG 2013	SEP 2013
(intercept)			
coefficient	-0.6888***	-0.6941***	-0.5832**
std. error	0.1232	0.1140	0.1076
spread (bps): s^*			
coefficient	0.3187***	0.3905***	0.3950***
std. error	0.0069	0.0077	0.0070
limit order: $R^L s^*$			
coefficient	-0.3027***	-0.3415***	-0.3658***
std. error	0.0107	0.0100	0.0099
add. tick to pay: $R^M \sigma^*$			
coefficients	0.0991***	0.1480***	0.1486***
std. error	0.0234	0.0225	0.0348
tick size: σ^*			
coefficients	2.3238***	1.8508***	2.4290***
std. error	0.1098	0.0997	0.0996
R-squared	9.91%	10.62%	13.48%

Significance: *** p<0.001, ** p<0.01, * p<0.05

In-sample regressions

Monthly in-sample linear regression results for microstructure market impact model

- consistently good performance of our model, represented by high R^2 values
- coefficients of "micro-level" book variables are statistically significant
- signs of the coefficients are intuitive: limit order $-$, higher price order $+$

Cross-validation

- **Cross-Validation**

  (our "micro" model)

$$\nu^* = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot R^L s^* + \beta_3 \cdot R^M \delta^* + \beta_4 \cdot \delta^*$$

  (benchmark "macro" model)

$$\nu^* = \beta_0 + \beta_1 \cdot (\text{Percent of Market Vol.})^\alpha \, \sigma^* + \beta_2 \cdot \sigma^*$$

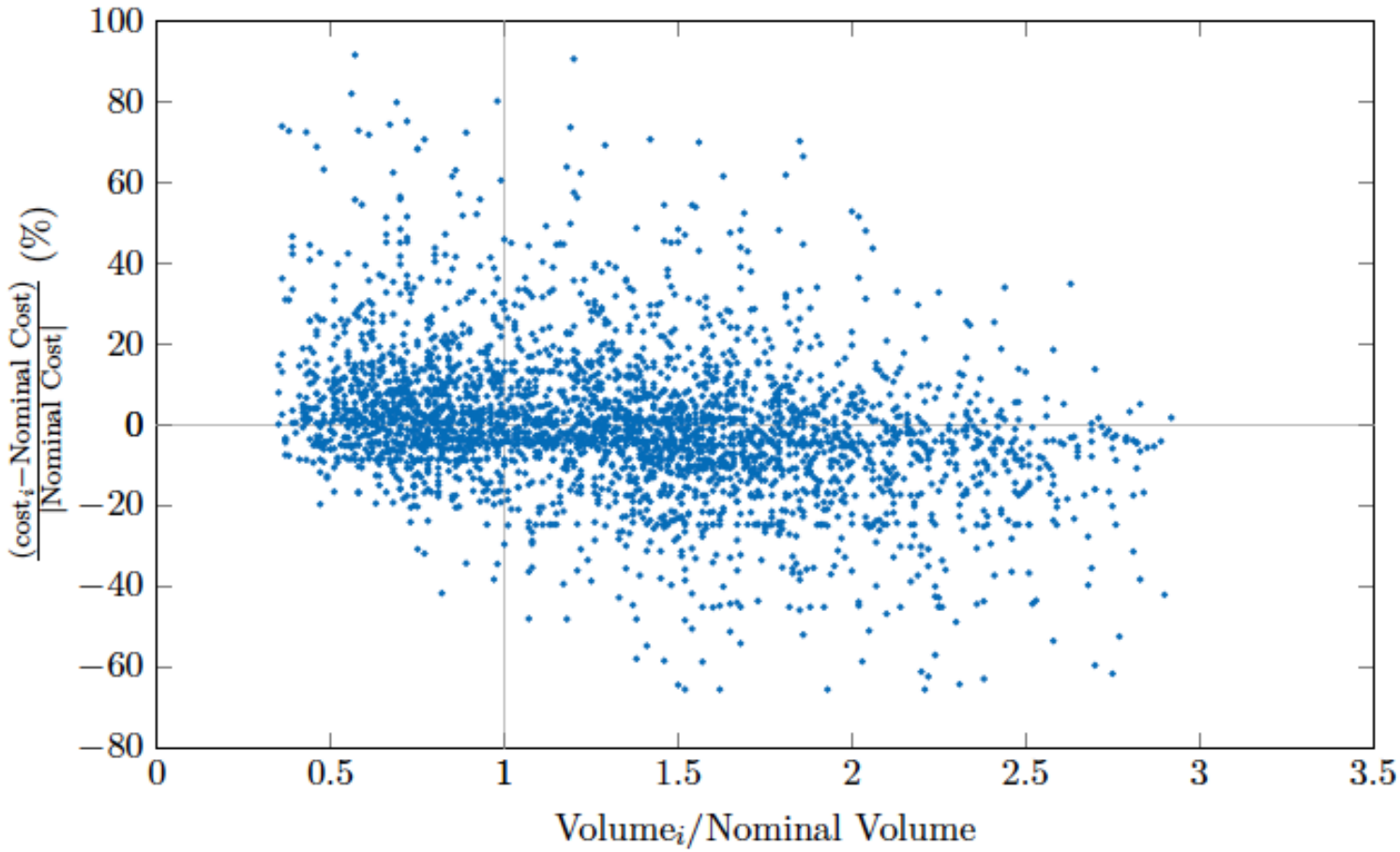
  out-of-sample R^2: our model 11% vs. benchmark models 3%

	Our Model	Linear	Square Root
avg. out-of-sample R^2	11.03%	3.11%	3.12%
relative improvement	0.00%	255%	254%

  - $\sigma(t)$ above; using daily σ reduces explanatory power by 1-2%
  - serial correlation: including 1 or 2 lagged residuals improves performance coefficients are stat. significant and have right signs

Simulated costs as microstructure variables are varied ($C = 3\times$Depth)

– randomly generated 4-tuples for (Q_b, Q_a, μ_b, μ_s)
– variables varied by a random multiplier in $(1/3, 1)$ w.p. .5 and $(1, 3)$ w.p. .5
– cost estimates vary by $\pm 60\%$ around "nominal" values



Robustness - order & security segmentation

Segmentation: by market participation rate

	micro model	*macro* model		sample size
		linear	square root	
Percent of market vol.				
[1%,10%]	8.82%	1.87%	1.89%	55,337
(10%,20%]	14.10%	5.34%	5.21%	19,974
(20%,30%]	15.08%	4.23%	4.24%	11,729
overall: [1%,30%]	11.03%	3.11%	3.12%	87,040

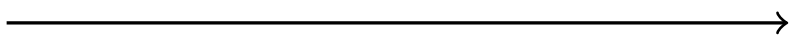
- *micro* model outperforms the *macro* benchmark models for all groups
- all models improve as the participation rate increases

Robustness - order & security segmentation

Segmentation: by (average daily volume, average queue length)

	Low Depth	Mid Depth	High Depth	Ultra Deep
Low ADV	6.26%	10.23%	17.14%	N/A
Mid ADV	5.38%	8.12%	12.62%	N/A
High ADV	N/A	5.56%	10.32%	24.84%

high volume



deep

- *micro* model outperforms the *macro* benchmark models for all groups
- model accuracy improves with queue length
- similar results when segmenting queue lengths in shares and dollars

Robustness - effect of nonlinearity

Simplification: remove the non-linearities

$$\nu^* = \beta_0 + \beta_1 \cdot s^* + \beta_2 \cdot \frac{\left(\mu^s_{b_0} T - Q^b_{b_0}(0)\right)}{C} \cdot s^* + \beta_3 \cdot \frac{\left(C - Q^s_{a_0}(0) - \kappa T\right)}{\bar{Q}^s} \cdot \delta^* + \beta_4 \cdot \delta^*$$

	micro w/o nonlinearity	*macro* linear	*macro* square root
avg. out-of-sample R^2	8.19%	3.11%	3.12%
relative improvement	0.00%	163%	163%

- may affect computational tractability in context of optimization
  e.g., stock selection, trade scheduling
- still significantly outperforms benchmark models

Robustness - effect of time horizon

Time horizon: 5min vs. 1min slices

model accuracy depends on the time horizon of the trade slices,
micro model has even better statistical fit for shorter-horizon slices

	Our Model	Linear	Square Root
avg. out-of-sample R^2	16.57%	2.67%	2.81%
relative improvement	0.00%	521%	490%

1-min horizon: order / stock segmentation

	Our Model	Linear	Square Root	Sample size
Percent of market vol.				
[1%,10%]	13.53%	0.94%	0.96%	73,166
(10%,20%]	19.24%	2.26%	2.26%	40,631
(20%,30%]	21.51%	3.59%	3.59%	19,830
overall: [1%,30%]	16.57%	2.67%	2.81%	133,627

		Low depth	Mid depth	High depth	Ultra deep	Overall
Our Model	Low ADV	12.18%	13.81%	23.12%	too few obs.	
	Mid ADV	9.41%	10.84%	18.78%	too few obs.	16.57%
	High ADV	too few obs.	3.91%	20.74%	28.98%	

Robustness - prediction vs. attribution

Prediction: pre-trade cost estimates
use information available at the beginning of the trading slice

Our Model		Linear		Square Root	
predictive	attributive	predictive	attributive	predictive	attributive
8.20%	11.07%	2.26%	2.82%	2.25%	2.84%

- the drop in explanatory power is more significant in *micro* model
- *micro* model still significantly outperforms the two benchmark models
- similar comparison when using historical forecasts (monthly averages)

Low ADV securities (ADV $\in (50K, 300K)$ shares, POV $\in (1\%, 30\%)$)

	JUL 2013	AUG 2013	SEP 2013
(intercept)			
coefficient	-0.1240	-0.0611	1.4911***
std. error	0.3093	0.2442	0.2147
spread (bps): s^*			
coefficient	0.3576***	0.3958***	0.2826***
std. error	0.0078	0.0072	0.0049
limit order: $R^L s^*$			
coefficient	-0.2829***	-0.2582**	-0.1753***
std. error	0.0123	0.0110	0.0093
add. tick to pay: $R^M \sigma^*$			
coefficients	0.7137***	0.5796***	0.5271***
std. error	0.1326	0.1499	0.1242
tick size: σ^*			
coefficients	1.1214***	0.6174***	1.2526***
std. error	0.2267	0.1972	0.1791
R-squared	25.02%	27.20%	21.56%

Significance: *** p<0.001, ** p<0.01, * p<0.05

	Micro model	Benchmark macro model		Mean predictor
		$\alpha = 1$	$\alpha = 0.5$	
Avg. out-of-sample R^2 (vs. predicted mean)	23.26%	4.72%	4.91%	0.00%
relative improvement	0.00%	393%	374%	Inf

Microstructure market impact model . . . closing comments

- depth of book information helps
- short-term price momentum predictions improve models predictions
- applications:
  - trade execution (short-term trade offs; opportunistic signals)
  - offers insight on dependence structure for cross-asset impact model volume, depth, volatility
  - short-term model useful in subsequently estimating impact decay & permanent price impact

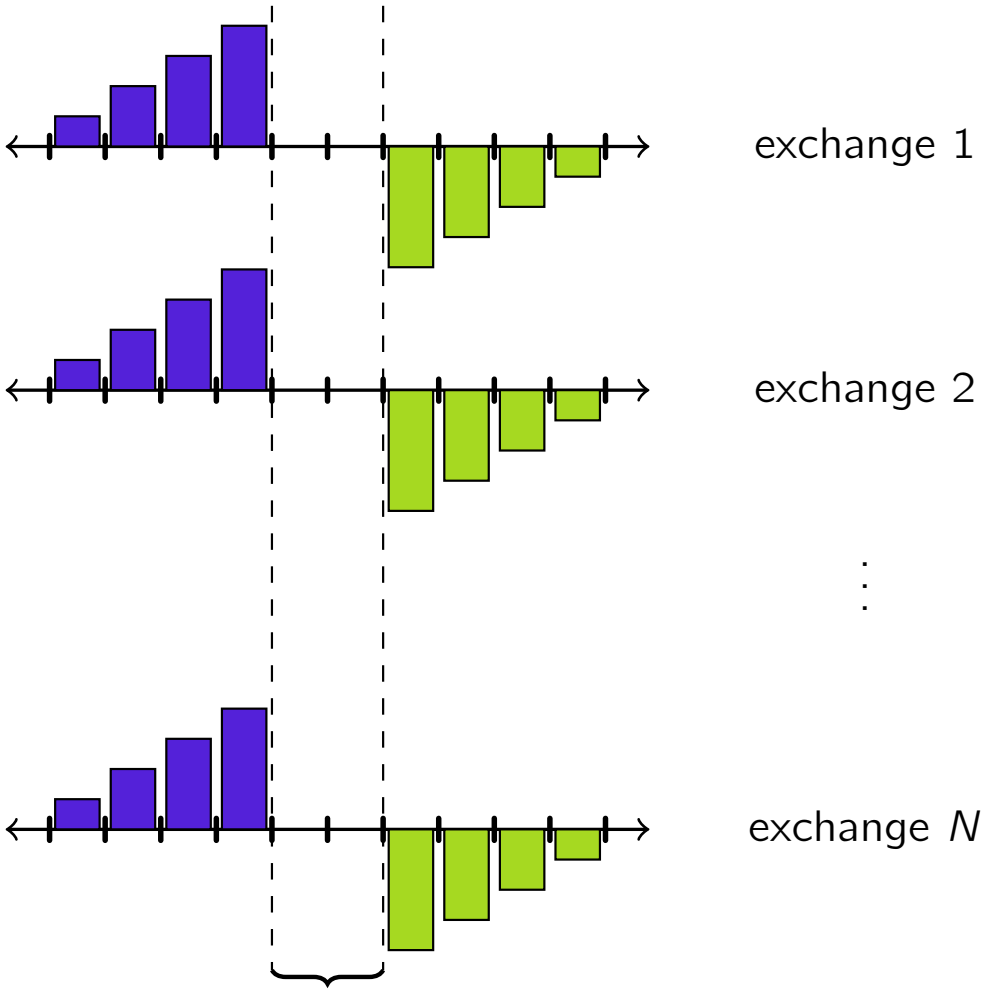
Outline

- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

5. Order routing in fragmented LOB markets

   - fragmentation & order routing decisions
   - mean-field analysis & state space collapse
   - quick look at some data
   - pointwise-stationary-fluid-model (PSFM) – a first glimpse

Multiple Limit Order Books



Price levels are coupled through protection mechanisms (Reg NMS)

We consider the evolution of:

- one side of the market
- the 'top-of-the-book', i.e., national best bid queues across all exchanges

Time Scales

Three relevant time scales:

- Events: order / trade / cancellation interarrival times ($\sim$ ms – sec)
- Delays: waiting times at different exchanges ($\sim$ sec – min)
- Rates: time-of-day variation of flow characteristics ($\sim$ min – hrs)

Order placement decisions depend on queueing delays in LOBs (our focus)

- assume constant arrival rates of limit orders and trades
- order sizes are small relative to overall flow over relevant time scale
- overall limit order and trade volumes are high

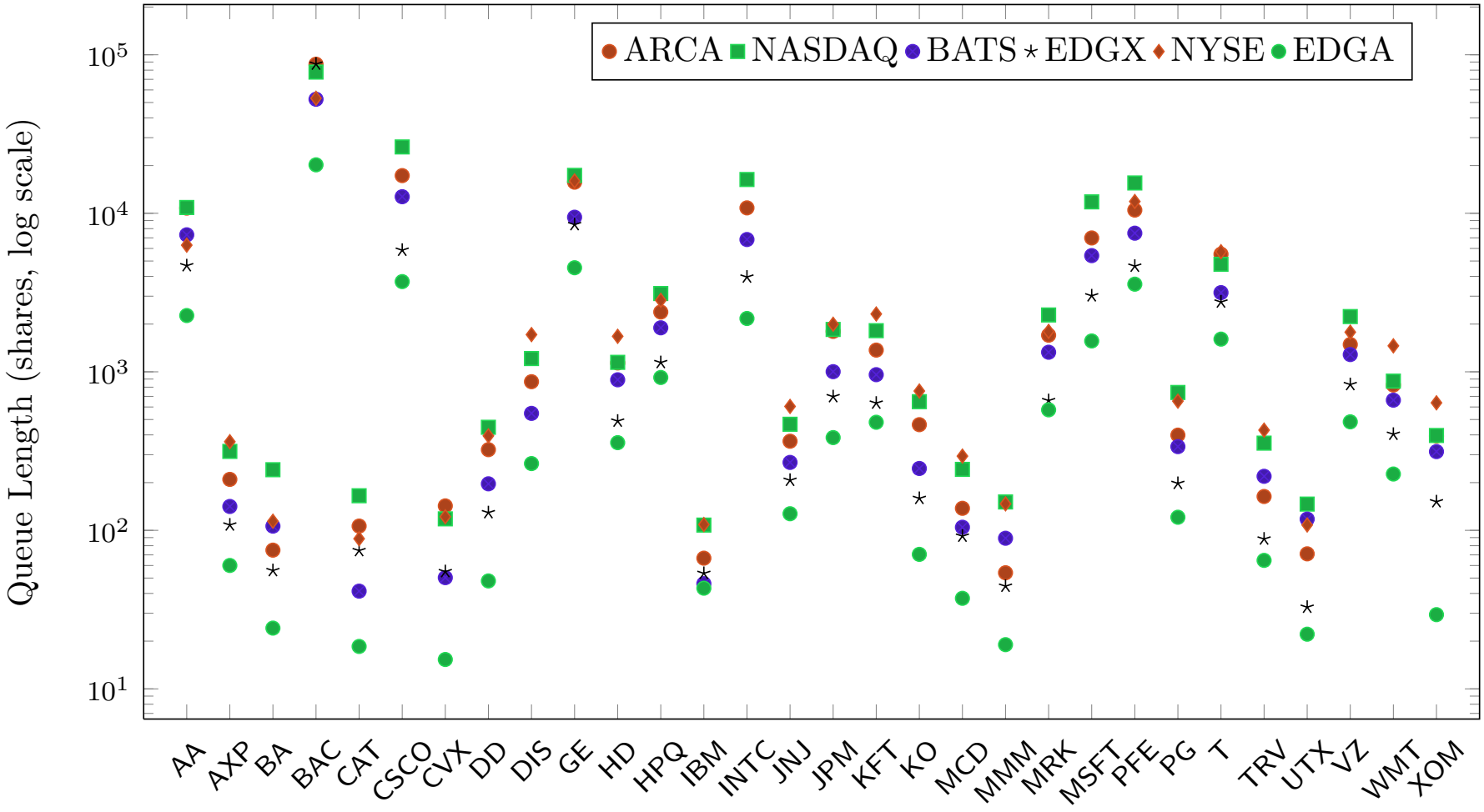
We will consider a variation of the problem of execution in a LOB that "incorporates" the fragmented nature of markets

DJIA 30: Summary statistics – Sept 2011

	Symbol	Listing Exchange	Price		Average Bid-Ask Spread	Volatility	Average Daily Volume
			Low ($)	High ($)	($)	(daily)	(shares, $\times 10^6$)
Alcoa	AA	NYSE	9.56	12.88	0.010	2.2%	27.8
American Express	AXP	NYSE	44.87	50.53	0.014	1.9%	8.6
Boeing	BA	NYSE	57.53	67.73	0.017	1.8%	5.9
Bank of America	BAC	NYSE	6.00	8.18	0.010	3.0%	258.8
Caterpillar	CAT	NYSE	72.60	92.83	0.029	2.3%	11.0
Cisco	CSCO	NASDAQ	14.96	16.84	0.010	1.7%	64.5
Chevron	CVX	NYSE	88.56	100.58	0.018	1.7%	11.1
DuPont	DD	NYSE	39.94	48.86	0.011	1.7%	10.2
Disney	DIS	NYSE	29.05	34.33	0.010	1.6%	13.3
General Electric	GE	NYSE	14.72	16.45	0.010	1.9%	84.6
Home Depot	HD	NYSE	31.08	35.33	0.010	1.6%	13.4
Hewlett-Packard	HPQ	NYSE	21.50	26.46	0.010	2.2%	32.5
IBM	IBM	NYSE	158.76	180.91	0.060	1.5%	6.6
Intel	INTC	NASDAQ	19.16	22.98	0.010	1.5%	63.6
Johnson & Johnson	JNJ	NYSE	61.00	66.14	0.011	1.2%	12.6
JPMorgan	JPM	NYSE	28.53	37.82	0.010	2.2%	49.1
Kraft	KFT	NYSE	32.70	35.52	0.010	1.1%	10.9
Coca-Cola	KO	NYSE	66.62	71.77	0.011	1.1%	12.3
McDonalds	MCD	NYSE	83.65	91.09	0.014	1.2%	7.9
3M	MMM	NYSE	71.71	83.95	0.018	1.6%	5.5
Merck	MRK	NYSE	30.71	33.49	0.010	1.3%	17.6
Microsoft	MSFT	NASDAQ	24.60	27.50	0.010	1.5%	61.0
Pfizer	PFE	NYSE	17.30	19.15	0.010	1.5%	47.7
Procter & Gamble	PG	NYSE	60.30	64.70	0.011	1.0%	11.2
AT&T	T	NYSE	27.29	29.18	0.010	1.2%	37.6
Travelers	TRV	NYSE	46.64	51.54	0.013	1.6%	4.8
United Tech	UTX	NYSE	67.32	77.58	0.018	1.7%	6.2
Verizon	VZ	NYSE	34.65	37.39	0.010	1.2%	18.4
Wal-Mart	WMT	NYSE	49.94	53.55	0.010	1.1%	13.1
Exxon Mobil	XOM	NYSE	67.93	74.98	0.011	1.6%	26.2

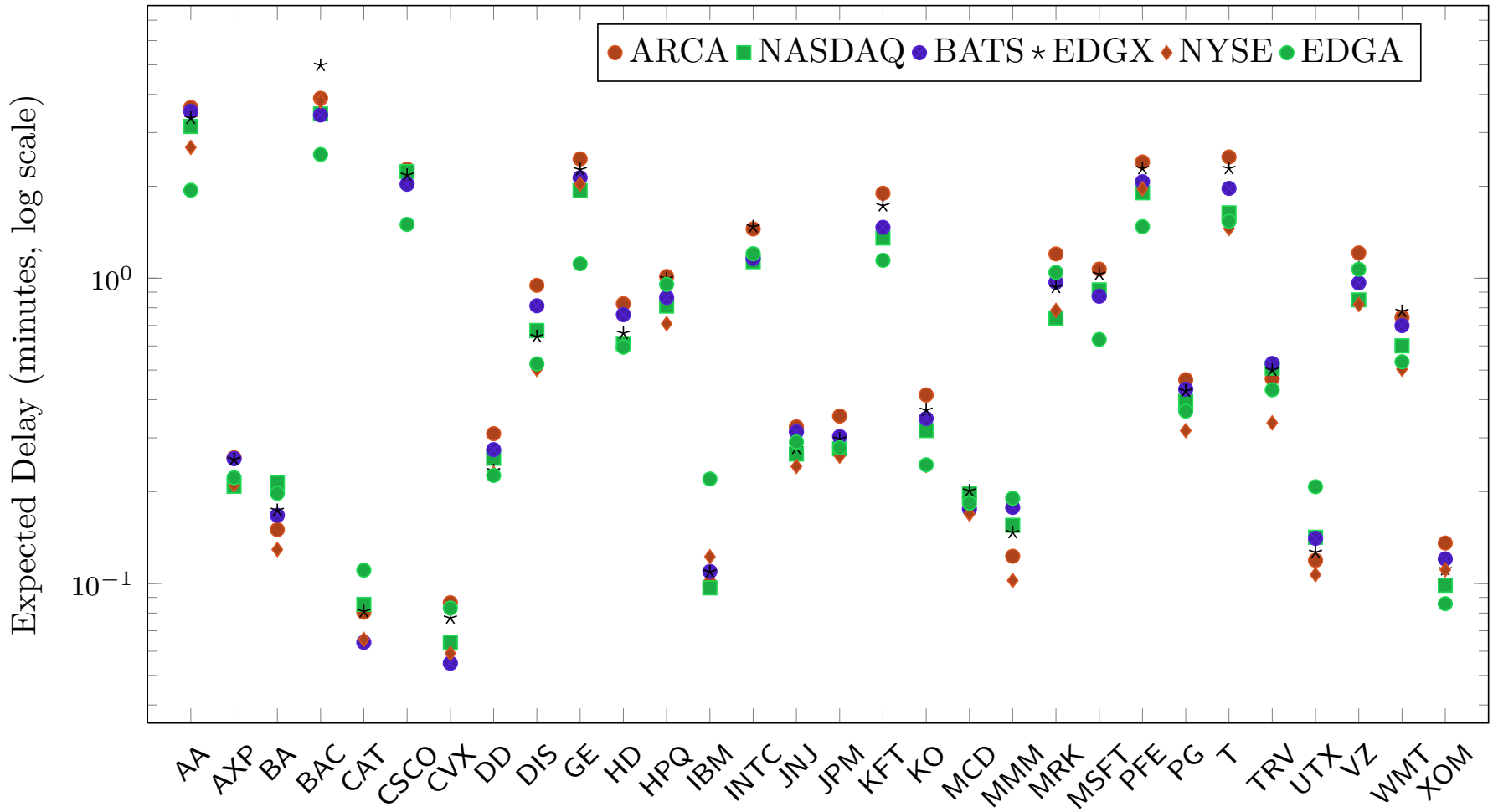
Table 1: Descriptive statistics for the 30 stocks over the 21 trading days of September 2011. All statistics except the average bid-ask spread were retrieved from Yahoo Finance; the average bid-ask spread is a time average computed from our TAQ data set. The daily volatility is computed from closing prices over the period in question.

DJIA 30: Expected Queue Lengths – Sept 2011



(b) Average queue length (number of shares at the NBBO) across stocks and exchanges.

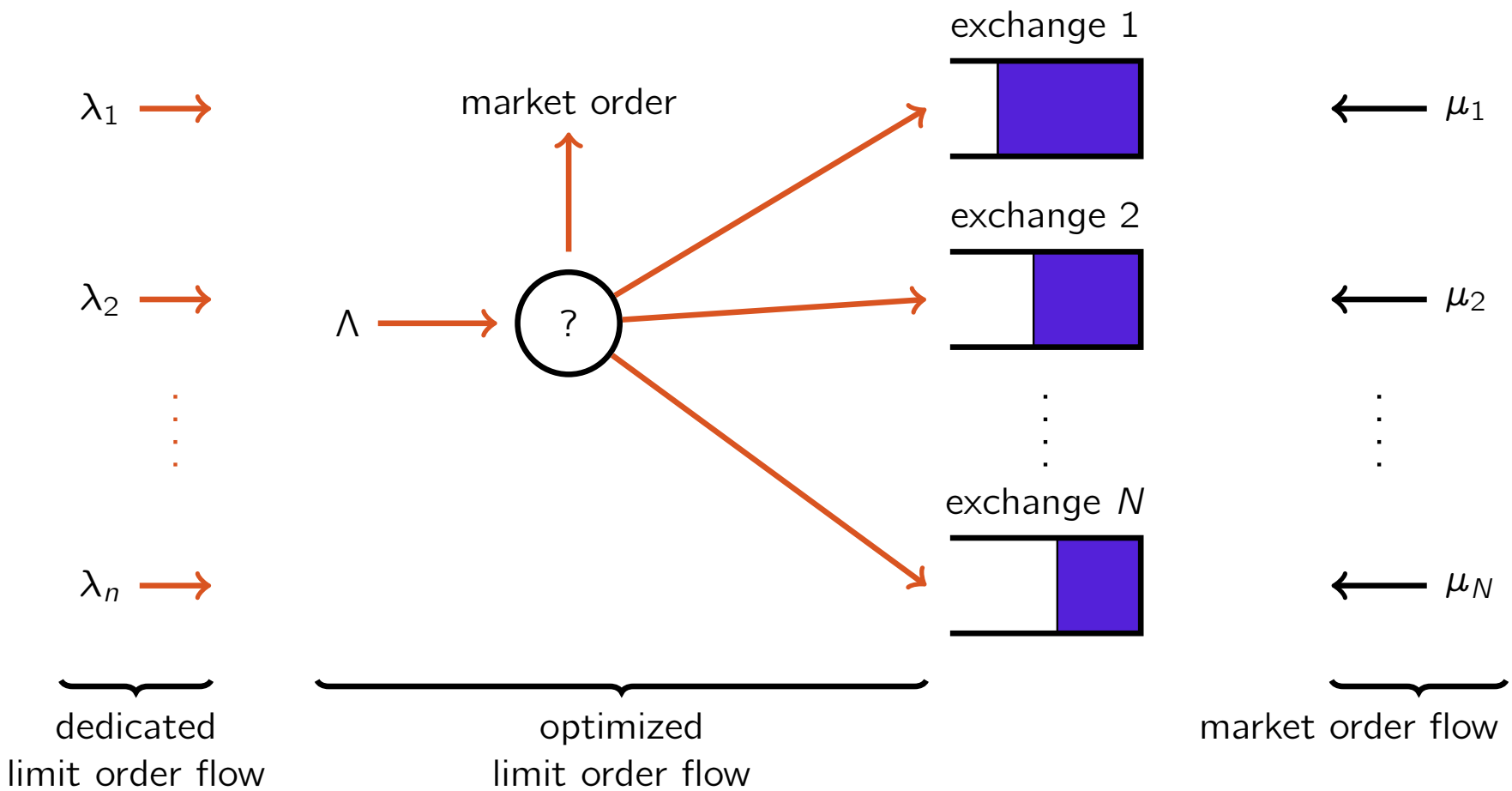
DJIA 30: Expected Delays – Sept 2011



(a) Average expected delay across stocks and exchanges.

One-sided Multi LOB Fluid Model

Fluid model: Continuous & deterministic arrivals of infinitesimal traders



Problem #3: "Fragmented market" version of LOB execution problem

In a fragmented market, a trader had multiple exchanges to choose from. They differ wrt

- Expected delay ($\approx$ 1 to 1000 seconds), $\mathbf{P}$(fill in t time)
- Rebates for limit orders ($\approx -\$0.0002$ to $\$0.0030$ per share) & fees for mkt orders
- Other factors that affect decision such as short-term alpha signals, estimates of adverse selection, tiering agreements with exchanges (similar \$ value as rebates, in general state dependent)

$$\max_{X_k} \sum_k \mathbf{E}(Y_k|X_k, T) r_k - (f+s) \cdot (C - \sum_k \mathbf{E}(Y_k|X_k, T)$$

where
– X_k = quantity to get posted at exchange K (at top of book – good?)
– Y_k = quantity that trades at exchange k up to time T
– simplifying mkt order problem to a clean up trade
– trading algorithms typically not allowed to post more quantity that C
– formulation limit orders are posted at $t = 0$

Order routing in fragmented market – cont.

$$\max_{X_k} \sum_k \mathbf{E}(Y_k|X_k, T) r_k - (f+s) \cdot (C - \sum_k \mathbf{E}(Y_k|X_k, T)$$

BUT – cannot post all quantity at $t = 0$ in practice; make incremental decisions

– randomize posting times across LOBs

– post so as to spread out execution profile

– posting decisions tend to be "dynamic" i.e., revisited in $[0, T]$ (especially for inverted exchanges)

– avoid clean up trade

– ...

Time vs. money tradeoff

despite the many caveats, previous problem captures time vs. money tradeoff

- time: trade now with a market order or sooner in a less congested LOB
- money: trade in a high rebate exchange and also avoid paying the spread
- incentives are such that most institutional flow tries to be patient

study a simpler problem for each trader, but allow many traders to participate

– each trader submits an infinitesimal order
– we consider the flow and mkt equilibrium across agents
– agents are heterogeneous wrt T (delay tolerance)
– leverage work on "economics of queues & congestion"

aim for structural insights & tractable model of fragmented market equilibrium (not tactical)

The Limit Order Placement Decision

Factors affecting limit order placement:

- Expected delay ($\approx$ 1 to 1000 seconds)
- Rebates ($\approx -\$0.0002$ to $\$0.0030$ per share)
- Other factors that affect decision such as short-term alpha signals, estimates of adverse selection, tiering agreements with exchanges (similar \$ value as rebates, in general state dependent)

$\tilde{r}_i := r_i + (\text{other factors}) =$ "effective rebate" $\qquad \text{ED}_i =$ expected delay

Traders choose to route their order to exchange i given by

$$\text{argmax}_i \quad \gamma \tilde{r}_i \; - \; \text{ED}_i$$

- $\gamma \sim F$ i.i.d. across traders, captures delay tolerance / rebate tradeoff $\Rightarrow \; \gamma \sim \; 10^1$ to 10^4 seconds per \$0.01
- allows choice amongst Pareto efficient $(\tilde{r}_i, \text{ED}_i)$ pairs
- Implicit option for a market order: $r_0 \ll 0$, $\text{ED}_0 = 0$

The Market Order Routing Decision

- Market orders execute immediately, no queueing or adverse selection
- Market orders incur fees ($\approx r_i$)
- Natural criterion is to route order according to

$$\operatorname{argmin}_i \quad \{\ r_i \ : \ Q_i > 0,\ \ i = 1, \ldots, N\ \}$$

Routing decision differs from "fee minimization" due to

- Order sizes are not infinitesimal; may have to be split across exchanges
- Latency to exchange introduces notion of $\mathbf{P}(\text{fill})$ when Q_i are small
- Not all flow is "optimized", or has other economic considerations
- Traders avoid "clearing" queues to avoid increased price slippage

The Market Order Routing Decision

Attraction Model: Bounded rationality and model intricacies motivate fitting a probabilistic model of the form

$$\mu_i(Q) := \mu \frac{f_i(Q_i)}{\sum_j f_j(Q_j)}$$

- $f_i(\cdot)$ captures "attraction" of exchange i:

  $$\uparrow \text{ in } Q_i \text{ and } \downarrow \text{ in } r_i$$

- these slides will use:

  $$f_i(Q_i) := \beta_i Q_i$$

  (we imagine $\beta_i \sim 1/r_i$)

5. Order routing in fragmented LOB markets

   - fragmentation & order routing decisions
   - mean-field analysis & state space collapse
   - quick look at some data
   - pointwise-stationary-fluid-model (PSFM) – a first glimpse

Transient Dynamics & Flow Equilibrium

- Given (Λ, μ) study mean-field approximation of coupled LOBs
- **Dynamics:** Coupled ODEs describe $\dot{Q}(t)$ dynamics
- **Convergence:** $Q(t) \to Q^*$ as $t \to \infty$ (γ-dist. sufficiently decreasing tail)

Fluid Model Equilibrium

$\pi_i(\gamma)$ = fraction of type γ investors who send orders to exchange i

An equilibrium (π^*, Q^*) must satisfy

(i) *Individual rationality:* for all γ, $\pi^*(\gamma)$ optimizes

$$\max_{\pi(\gamma)} \quad \pi_0(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^{N} \pi_i(\gamma)\left(\gamma\tilde{r}_i - \frac{Q_i^*}{\mu_i(Q^*)}\right)$$

$$\text{subject to} \quad \pi(\gamma) \geq \mathbf{0}, \quad \sum_{i=0}^{N} \pi_i(\gamma) = 1.$$

(ii) *Flow balance:* for all $1 \leq i \leq N$,

$$\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma) = \mu_i(Q^*)$$

Workload

- $W := \sum_{i=1}^{N} \beta_i Q_i$ is the workload, a measurement of aggregate available liquidity
- $W \neq$ total market depth, also accounts for time
- $\text{ED}_i = Q_i/\mu_i = \left(\sum_j \beta_j Q_j\right)/(\mu\beta_i) = W/(\mu\beta_i)$
- Workload is a sufficient statistic to determine delays

Fluid model equilibrium – rewritten wrt W

(π^*, W^*) satisfy

(i) *Individual rationality:* for all γ, $\pi^*(\gamma)$ optimizes

$$\max_{\pi(\gamma)} \quad \int_0^\infty \left(\pi_0(\gamma)\gamma\tilde{r}_0 + \sum_{i=1}^N \pi_i(\gamma)\left(\gamma\tilde{r}_i - \frac{W^*}{\mu\beta_i}\right)\right) dF(\gamma)$$

$$\text{subject to} \quad \pi(\gamma) \geq \mathbf{0}, \quad \sum_{i=0}^N \pi_i(\gamma) = 1.$$

(ii) *System-wide flow balance:*

$$\sum_{i=1}^N \left(\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma)\right) = \mu$$

if and only if (π^*, Q^*) is an equilibrium, where

$$Q_i^* := \left(\lambda_i + \Lambda \int_0^\infty \pi_i^*(\gamma)\, dF(\gamma)\right) \frac{W^*}{\mu\beta_i}$$

Fluid Model Equilibrium

System-wide flow balance: Most impatient investors (i.e., $\gamma \leq \gamma_0$) should choose market orders

$$\sum_{i=1}^{N} \lambda_i + \Lambda\big(1 - F(\gamma_0)\big) = \mu \quad \Longrightarrow \quad \gamma_0 = F^{-1}\left(1 - \frac{\mu - \sum_{i=1}^{N} \lambda_i}{\Lambda}\right)$$

Incentive compatibility:

$$\max_{i \neq 0} \gamma(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i} \leq 0 \quad \text{for all} \quad \gamma \leq \gamma_0$$

This is implied by the marginal indifference condition

$$\max_{i \neq 0} \gamma_0(\tilde{r}_i - \tilde{r}_0) - \frac{W^*}{\mu\beta_i} = 0$$

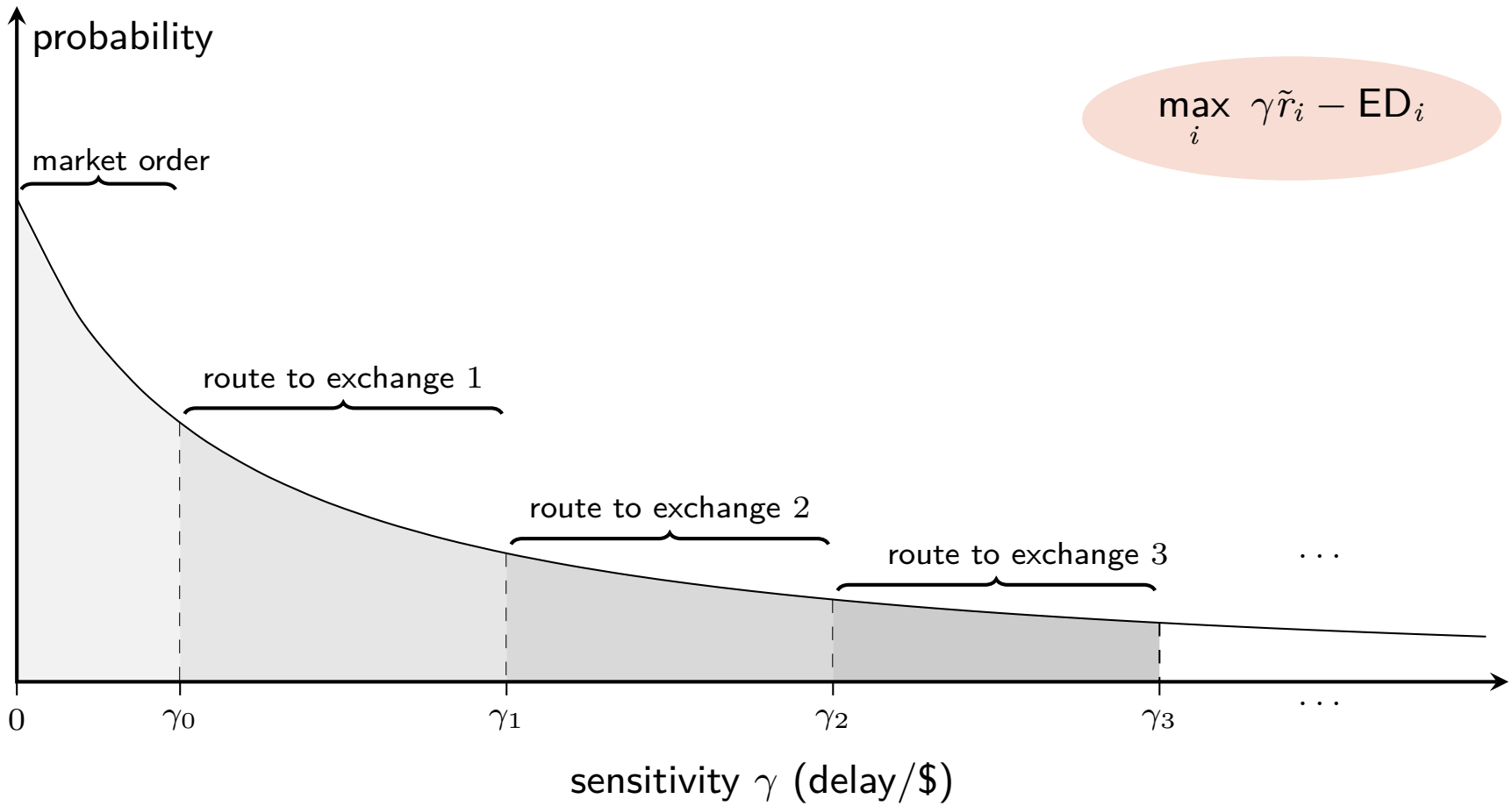
Under mild conditions, W^* is the equilibrium workload if and only if

$$W^* = \gamma_0 \mu \max_{i \neq 0} \beta_i(\tilde{r}_i - \tilde{r}_0)$$

Price-delay sensitivity & choice

Assume that $\tilde{r}_i \uparrow$ and that $\beta_i \downarrow$. Then,

$$\mathsf{ED}_0 = 0 < \mathsf{ED}_{\min} = \mathsf{ED}_1 < \mathsf{ED}_2 < \cdots < \mathsf{ED}_N = \mathsf{ED}_{\max}$$



5. Order routing in fragmented LOB markets

   - fragmentation & order routing decisions
   - mean-field analysis & state space collapse
   - quick look at some data
   - pointwise-stationary-fluid-model (PSFM) – a first glimpse

Empirical Results

- NYSE TAQ data, millisecond timestamps
- Stocks: DJIA 30 – Sept 2011
- 6 main exchanges (95% of “lit" volume)
- Analysis uses time-averaged 60 min slices from 9:45am - 3:45pm $\times$ 20 days

	Exchange Code	Rebate ($ per share, $\times 10^{-4}$)	Fee ($ per share, $\times 10^{-4}$)
BATS	Z	27.0	28.0
DirectEdge X (EDGX)	K	23.0	30.0
NYSE ARCA	P	21.0†	30.0
NASDAQ OMX	T	20.0†	30.0
NYSE	N	17.0	21.0
DirectEdge A (EDGA)	J	5.0	6.0

Table 2: Rebates and fees of the 6 major U.S. stock exchanges during the September 2011 period, per share traded. †Rebates on NASDAQ and ARCA are subject to “tiering”: higher rebates than the ones quoted may be available to traders that contribute significant volume to the respective exchange.

Estimation of market order routing model (β's)

Reduced form "attraction" model for market order arrival rates:

$$\mu_i^{(s,j)}(t) = \mu^{(s,j)}(t) \; \frac{\beta_i^{(j)} Q_i^{(s,j)}(t)}{\sum_{i'=1}^{N} \beta_{i'}^{(j)} Q_{i'}^{(s,j)}(t)},$$

where $\beta_i^{(j)}$ is the attraction coefficient for security j on exchange i.

Estimation procedure:

- Measure $\mu_i^{(s,j)}(t)$, $\mu^{(s,j)}(t)$ and $Q_{i'}^{(s,j)}(t)$
- estimate β_i^j via non-linear regression

DJIA 30: Market order routing model (β's) – Sept 2011

	Attraction Coefficient					
	ARCA	NASDAQ	BATS	EDGX	NYSE	EDGA
Alcoa	0.73	0.87	0.76	0.81	1.00	1.33
American Express	1.19	1.08	0.99	0.94	1.00	0.94
Boeing	0.95	0.67	0.81	0.74	1.00	0.73
Bank of America	0.94	1.04	1.01	0.77	1.00	1.43
Caterpillar	0.82	0.78	1.13	0.70	1.00	0.58
Cisco	0.95	1.00	1.06	0.98	-	1.45
Chevron	0.70	0.93	1.17	0.65	1.00	0.75
DuPont	0.90	0.98	0.98	1.03	1.00	1.00
Disney	0.69	0.88	0.78	0.88	1.00	1.04
General Electric	0.79	1.01	0.94	0.73	1.00	1.63
Home Depot	0.76	0.98	0.79	0.84	1.00	1.02
Hewlett-Packard	1.04	1.04	1.02	0.68	1.00	0.82
IBM	1.25	1.20	1.20	1.05	1.00	0.54
Intel	0.83	1.00	0.96	0.84	-	1.04
Johnson & Johnson	0.80	0.94	0.86	0.92	1.00	0.77
JPMorgan	0.78	0.99	0.93	0.84	1.00	0.91
Kraft	0.72	0.89	0.83	0.73	1.00	1.06
Coca-Cola	0.68	0.84	0.79	0.76	1.00	0.88
McDonalds	0.90	0.86	1.03	0.82	1.00	0.82
3M	0.89	0.67	0.62	0.66	1.00	0.57
Merck	0.68	1.01	0.83	0.90	1.00	0.81
Microsoft	0.83	1.00	1.02	0.95	-	1.41
Pfizer	0.84	1.01	0.96	0.87	1.00	1.29
Procter & Gamble	0.79	0.89	0.88	0.89	1.00	0.89
AT&T	0.62	0.94	0.75	0.59	1.00	1.00
Travelers	0.80	0.69	0.69	0.84	1.00	0.80
United Tech	1.18	0.89	0.79	0.87	1.00	0.53
Verizon	0.77	0.95	0.88	0.72	1.00	0.85
Wal-Mart	0.72	0.88	0.79	0.71	1.00	0.91
Exxon Mobil	0.89	1.13	0.97	0.89	1.00	1.35

Table 4: Estimates of the attraction coefficients β_i from nonlinear regression. Note that the attraction coefficient of the listing exchange is normalized to be 1.

State Space Collapse I

Under our model,

$$\mathrm{ED}_{i,t} = \frac{Q_{i,t}}{\mu_{i,t}} = \frac{W_t}{\mu_t} \cdot \frac{1}{\beta_i}$$

Therefore, the vector of expected delays

$$\vec{\mathrm{ED}}_t := \left(\frac{Q_{1,t}}{\mu_{1,t}}, \ldots, \frac{Q_{N,t}}{\mu_{N,t}} \right)$$

should have a low effective dimension.

State Space Collapse I – PCA output

	% of Variance Explained			% of Variance Explained	
	One Factor	Two Factors		One Factor	Two Factors
Alcoa	80%	88%	JPMorgan	90%	94%
American Express	78%	88%	Kraft	86%	92%
Boeing	81%	87%	Coca-Cola	87%	93%
Bank of America	85%	93%	McDonalds	81%	89%
Caterpillar	71%	83%	3M	71%	81%
Cisco	88%	93%	Merck	83%	91%
Chevron	78%	87%	Microsoft	87%	95%
DuPont	86%	92%	Pfizer	83%	89%
Disney	87%	91%	Procter & Gamble	85%	92%
General Electric	87%	94%	AT&T	82%	89%
Home Depot	89%	94%	Travelers	80%	88%
Hewlett-Packard	87%	92%	United Tech	75%	88%
IBM	73%	84%	Verizon	85%	91%
Intel	89%	93%	Wal-Mart	89%	93%
Johnson & Johnson	87%	91%	Exxon Mobil	86%	92%

Table 3: Results of PCA: how much variance in the data can the first two principle components explain.

State Space Collapse II

Under our model,

$$\mathrm{ED}_{i,t} = \frac{Q_{i,t}}{\mu_{i,t}} = \frac{W_t}{\mu_t} \cdot \frac{1}{\beta_i}$$

So:

$$\mathrm{ED}_{i,t} = \frac{\beta_j}{\beta_i} \cdot \mathrm{ED}_{j,t},$$

predicts linear pairwise relation between delay estimates across exchanges.

Test cross-sectionally (similar results within stocks; ARCA as benchmark exchange):

	Dependent Variable: $\mathrm{ED}_{\text{exchange}}$				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	$6.96 \times 10^{-4***}$	$1.27 \times 10^{-3***}$	-1.02×10^{-4}	$-4.60 \times 10^{-4***}$	$9.42 \times 10^{-4***}$
	(1.14×10^{-4})	(1.09×10^{-4})	(2.02×10^{-4})	(1.60×10^{-4})	(1.05×10^{-4})
Rescaled $\mathrm{ED}_{\mathrm{ARCA}}$	0.92^{***}	0.89^{***}	0.97^{***}	0.98^{***}	0.87^{***}
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
R^2	85%	87%	76%	77%	79%
Note:					$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

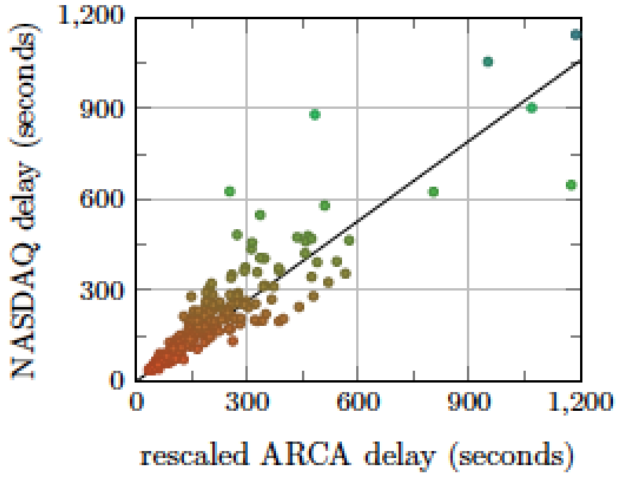
Table 5: Linear regressions of the expected delay on a particular exchange, versus that of the benchmark exchange (ARCA) rescaled by the ratio of the attraction coefficients of the two exchanges.

DJIA 30: Pairwise delay regressions – Sept 2011

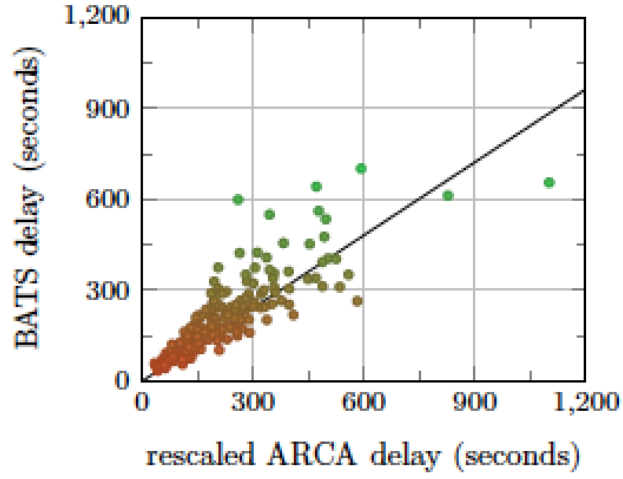
	NASDAQ		BATS		EDGX		NYSE		EDGA	
	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2	Slope	R^2
Alcoa	0.85	0.83	0.95	0.93	0.90	0.76	0.72	0.88	0.50	0.91
American Express	0.53	0.66	0.69	0.68	0.68	0.60	0.53	0.64	0.56	0.62
Boeing	1.29	0.91	1.01	0.86	1.12	0.85	0.77	0.90	1.22	0.81
Bank of America	0.84	0.92	0.82	0.90	1.28	0.84	1.01	0.77	0.63	0.86
Caterpillar	0.97	0.91	0.77	0.89	0.94	0.75	0.76	0.91	1.19	0.80
Cisco	0.97	0.95	0.86	0.93	0.95	0.90	-	-	0.63	0.90
Chevron	0.72	0.92	0.61	0.92	0.83	0.84	0.65	0.92	0.87	0.78
DuPont	0.78	0.95	0.85	0.93	0.69	0.83	0.67	0.94	0.65	0.86
Disney	0.66	0.95	0.82	0.92	0.65	0.87	0.46	0.91	0.50	0.86
General Electric	0.77	0.96	0.83	0.94	0.90	0.82	0.81	0.94	0.43	0.94
Home Depot	0.71	0.96	0.88	0.95	0.77	0.90	0.70	0.95	0.70	0.92
Hewlett-Packard	0.75	0.93	0.79	0.93	0.94	0.86	0.64	0.91	0.89	0.88
IBM	0.92	0.92	1.07	0.91	1.05	0.78	1.18	0.92	2.05	0.90
Intel	0.72	0.92	0.73	0.93	1.01	0.85	-	-	0.83	0.89
Johnson & Johnson	0.73	0.92	0.88	0.87	0.76	0.86	0.65	0.91	0.74	0.86
JPMorgan	0.76	0.96	0.83	0.95	0.81	0.90	0.71	0.96	0.74	0.92
Kraft	0.58	0.85	0.65	0.85	0.81	0.80	0.49	0.87	0.44	0.73
Coca-Cola	0.74	0.97	0.83	0.95	0.88	0.87	0.54	0.94	0.53	0.83
McDonalds	0.94	0.93	0.89	0.94	0.99	0.78	0.81	0.90	0.87	0.86
3M	1.07	0.82	1.27	0.87	1.02	0.75	0.71	0.88	1.24	0.72
Merck	0.57	0.92	0.77	0.92	0.73	0.82	0.62	0.93	0.83	0.88
Microsoft	0.85	0.92	0.80	0.95	0.99	0.77	-	-	0.59	0.95
Pfizer	0.74	0.92	0.83	0.94	0.92	0.87	0.78	0.92	0.58	0.92
Procter & Gamble	0.83	0.88	0.93	0.93	0.91	0.80	0.63	0.94	0.73	0.90
AT&T	0.61	0.90	0.72	0.89	0.92	0.79	0.55	0.93	0.58	0.88
Travelers	0.97	0.90	1.03	0.91	1.03	0.79	0.62	0.90	0.84	0.87
United Tech	1.11	0.92	1.07	0.91	1.04	0.84	0.79	0.91	1.37	0.61
Verizon	0.64	0.94	0.75	0.93	0.82	0.85	0.63	0.92	0.85	0.85
Wal-Mart	0.78	0.95	0.91	0.94	0.99	0.89	0.63	0.94	0.68	0.87
Exxon Mobil	0.70	0.97	0.86	0.97	0.78	0.84	0.79	0.92	0.61	0.89

Table 5: Linear regressions of the expected delays of each security on a particular exchange, versus that of the benchmark exchange (ARCA).

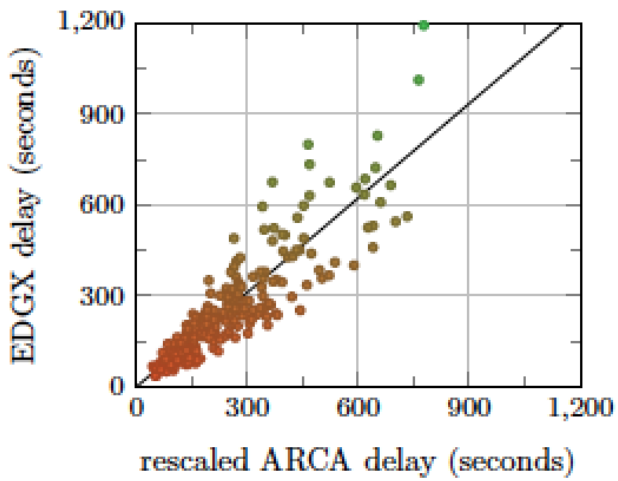
State Space Collapse II – BAC scatter plots



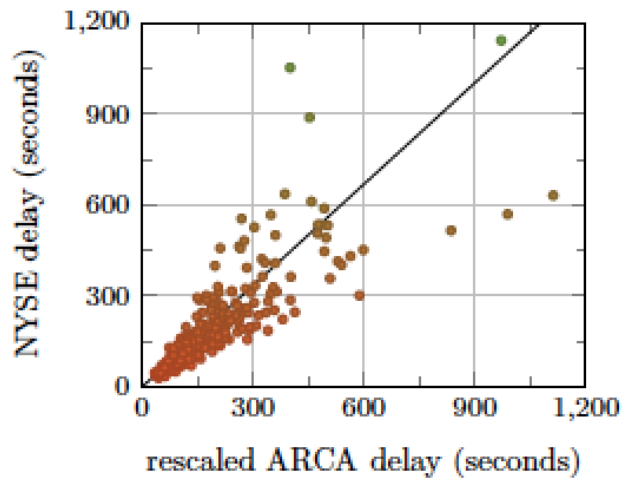
(a) slope $= 0.88$, intercept $= 6 \times 10^{-3}$, $R^2 = 84\%$



(b) slope $= 0.80$, intercept $= 9 \times 10^{-3}$, $R^2 = 79\%$



(c) slope $= 1.04$, intercept $= 9 \times 10^{-4}$, $R^2 = 71\%$



(d) slope $= 1.11$, intercept $= -4 \times 10^{-3}$, $R^2 = 63\%$

State Space Collapse III

Under our model,

$$\widehat{\mathrm{ED}}_t = \frac{W_t}{\mu_t} \cdot \left(\frac{1}{\beta_1}, \ldots, \frac{1}{\beta_N} \right)$$

How much of the variability of ED is explained by $\widehat{\mathrm{ED}}$? For each security j,

$$R_*^2 := 1 - \frac{\mathrm{Var}\left(\left\| \mathrm{ED}^{(s,j)}(t) - \widehat{\mathrm{ED}}^{(s,j)}(t) \right\| \right)}{\mathrm{Var}\left(\left\| \mathrm{ED}^{(s,j)}(t) \right\| \right)},$$

– $\mathrm{Var}(\cdot)$ = sample variance, averaged over all time slots t and both sides of market.

– R_*^2 close to 1, most variability of expected delays is explained via $(W_t/\mu_t)(1/\beta_i)$

DJIA 30: Residuals wrt SSC delay estimates – Sept 2011

	R^2_*		R^2_*		R^2_*
Alcoa	75%	Home Depot	87%	Merck	78%
American Express	64%	Hewlett-Packard	77%	Microsoft	80%
Boeing	75%	IBM	63%	Pfizer	79%
Bank of America	80%	Intel	82%	Procter & Gamble	80%
Caterpillar	58%	Johnson & Johnson	83%	AT&T	77%
Cisco	87%	JPMorgan	88%	Travelers	67%
Chevron	67%	Kraft	79%	United Tech	47%
DuPont	82%	Coca-Cola	81%	Verizon	79%
Disney	78%	McDonalds	74%	Wal-Mart	85%
General Electric	82%	3M	62%	Exxon Mobil	81%

Is it delays or queue lengths that drive routing decisions? I – PCA

$$\vec{\mathrm{ED}}_t := \left(\frac{Q_{1,t}}{\mu_{1,t}}, \ldots, \frac{Q_{N,t}}{\mu_{N,t}} \right) \quad \text{or} \quad \vec{Q}_t := (Q_{1,t}, \ldots, Q_{N,t})$$

have a low effective dimension.

- **delay:** 1st PC (& 2nd PC) explains 83% (90%) of variance on average.
- **queue length:**
  - 1st PC (& 2nd PC) explains 65% (78%) of variance on average.
  - less consistent, very low for some names

	% of Variance Explained	
	One Factor	Two Factors
Boeing	52%	66%
Caterpillar	31%	51%
Chevron	38%	59%
IBM	27%	53%
United Tech	39%	55%
Exxon Mobil	54%	69%

Delay or queue lengths: II – linear dependency

$$\mathrm{ED}_{i,t} = \frac{\beta_{\text{benchmark}}}{\beta_i} \cdot \mathrm{ED}_{\text{benchmark,t}} \quad \text{or} \quad Q_{i,t} = \beta \cdot Q_{\text{benchmark,t}}$$

- **delay:** linear regressions have average R^2 value 61%.
- **queue length:** linear regressions have average R^2 value 21%.

II – Linear regression tables (normalize by median on ARCA)

	Dependent Variable: Q_{exchange}				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	0.84^{***}	0.39^{***}	0.25^{***}	0.57^{***}	0.05^{***}
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)
Q_{ARCA}	0.74^{***}	0.45^{***}	0.29^{***}	0.96^{***}	0.24^{***}
	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)
R^2	19%	20%	13%	26%	26%
Note:					$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

	Dependent Variable: $\text{ED}_{\text{exchange}}$				
	NASDAQ OMX	BATS	DirectEdge X	NYSE	DirectEdge A
Intercept	0.27^{***}	0.28^{***}	0.24^{***}	0.28^{***}	0.36^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Rescaled ED_{ARCA}	0.70^{***}	0.72^{***}	0.72^{***}	0.63^{***}	0.60^{***}
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
R^2	70%	70%	52%	60%	52%
Note:					$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Welfare implications of fragmentation?

Some background (service completion process centrally controlled):

- Mkt homogeneous wrt delay preferences:
  - offering only one (delay, price) pair is welfare (rev max) optimal
  - $\max_i \gamma r_i + \mathbf{E}(D_i)$ yields same choice for all agents of they have same γ

- Mkt heterogeneous wrt delay preferences:
  - optimal to offer multiple (delay, price) options
  - welfare optimal to price the externality cost (wrt delay inflicted on others)
  - rev max solution also differentiated but more complex (involves the idea of "damaged goods")

Welfare implications of fragmentation?

Q: What about in studying trade execution in a fragmented market where the service process (aka. market orders) are not controlled by an optimizing central planner?

- Requires more nuanced analysis that models order generating process
- e.g., is rebate capture a useful trading strategy that is incentivized through rebate differences?

Parameter variability & Pointwise-Stationary-Fluid-Model

	% obs. in $\pm 2\sigma_t$	% obs. in $\pm 3\sigma_t$	% obs. outside $\pm 3\sigma_t$
1 min	63.33%	79.23%	20.77%
3 min	32.56%	50.39%	49.61%
5 min	27.27%	35.06%	64.94%
10 min	13.16%	31.58%	68.42%

- (λ, μ) exhibit significant differences in the time scale of > 5 minutes
- cf. top 100 names (by ADV): average queueing delay = 61 sec
- PSFM: after every rate change, FM converges to new SS; viewed in slower time scale of parameter changes, FM moves from one equilibrium state to another

Outline

- May 5: Overview of algorithmic trading and limit order book markets
  1. Overview of algorithmic trading
  2. Limit order book as a queueing system
- May 6: Deterministic (mean-field) models of LOB dynamics
  3. Transient dynamics, cancellations, and queue waiting times
  4. Execution in a LOB and a microstructure model of market impact
- May 7: Order routing and stochastic approximations of LOB markets
  5. Order routing in fragmented LOB markets
  6. Stochastic approximations of a LOB
- References

6. Stochastic approximations of LOB dynamics

   - recap of background information in asymptotic behavior of queueing models
   - a simple model of adverse selection
   - quick remarks on diffusion model of LOB top of book & PSFM
   - questions

Recap of basic building blocks from queueing theory

- $M/M/1$ system (Poisson limit and market order arrivals)
- $M/M/1+M$ with exponential patience clocks
- Basic facts for asymptotic behavior of $M/M/1$ and $M/M/1+M$

  regime we focus: (λ, μ) grow large

  – mean-field (fluid) models
  – diffusion models

Heavy-traffic (diffusion) model: $M/M/1$ approximating diffusion

Scaling:

$$\lambda^n = n - \beta\sqrt{n}, \quad \mu^n = n \quad (\text{so that } \lambda^n \approx \mu^n),$$

Flow imbalance:

$$N^n(t) = (A^n(t) - S^n(t)) = -\beta\sqrt{n}t + \sigma\sqrt{n}B(t) + O(\log(nt))$$

$O(\sqrt{n})$ stochastic imbalance of Poisson flows, leads to $O(\sqrt{n})$ queue lengths

$$\hat{Q}^n(t) := \frac{Q^n(t)}{\sqrt{n}} \Longrightarrow \hat{Q}(t) = \text{reflected Brownian motion.}$$

$$d\hat{Q}(t) = -\beta dt + \sigma dB(t) + dL(t) \quad (\beta > 0)$$

$L(t) =$ local time at the origin; in LOB analogy, $L(t)$ fires when price moves

$\hat{N}(t) = -\beta t + \sigma B(t), \quad L(t) = \sup_{\{0 \le s \le t\}} \hat{N}^-(s) \quad (x^- = min(0, x))$

Heavy-traffic (diffusion) model: $M/M/1$ performance approximations

$\hat{Q}(\infty) \sim \exp(-2\beta/\sigma^2)$

- queue lengths:

$$\mathbf{E}(Q^n) = \frac{\rho^n}{1-\rho^n} = O(\sqrt{n})$$

- waiting times: $\sqrt{n}$ queue length, trades arrive at order n, so

$$\mathbf{E}(W^n) = \frac{\mathbf{E}(Q^n)}{\mu^n} = O\left(\frac{1}{\sqrt{n}}\right)$$

- how often does the queue gets depleted: τ^n is the length of busy periods

$$\mathbf{E}(\tau^n) \approx O(1) \quad \text{the natural time scale of the limiting RBM}$$

(regenerative cycles of RBM)

- time scale separation: $\mathbf{E}(\tau^n) \gg \mathbf{E}(W^n)$

$$\mathbf{E}(\tau^n) \approx \sqrt{n}\mathbf{E}(W^n)$$

Heavy-traffic (diffusion) model: $M/M/1+M$, $\lambda^n - \mu^n = \beta\sqrt{n}$

Scaling:

$$\lambda^n = n + \beta\sqrt{n}, \quad \mu^n = n \quad \text{and} \quad \gamma^n = \gamma$$

Similar to $M/M/1$ in heavy traffic:

$$\hat{Q}_n(t) := \frac{1}{\sqrt{n}} Q^n(t) \Rightarrow \hat{Q}(t) \quad \text{(reflected O-U process)}$$

where

$$d\hat{Q}(t) = \left(\beta - \gamma\hat{Q}(t)\right) dt + \sigma dB(t) + dL(t)$$

- stable queue due to cancellations (drift $-\gamma Q(t)$)
- cancellation volume $\approx O(\sqrt{n}) \ll \lambda^n$
- $\hat{Q}(\infty) \sim$ truncated Normal dist.
- time scale separation: $\mathbf{E}(\tau^n) \gg \mathbf{E}(W^n) \ldots \mathbf{E}(\tau^n) \approx \sqrt{n}\mathbf{E}(W^n)$

A different heavy-traffic regime: $M/M/1+M$, $\lambda^n \gg \mu^n$

Scaling:

$$\lambda^n = n\rho, \quad \mu^n = n \quad \text{and} \quad \gamma^n = \gamma \quad (\rho > 1)$$

- $O(n)$ imbalance between order arrivals and trades
- balanced through $O(n)$ cancellations
- proportional cancellation flow $\gamma Q^n(t)$, suggests $Q^n(t) = O(n)$
- indeed fluid path dominates behavior:

$$Q^n(t) \approx nq(t) + \sqrt{n}(\text{stochastic fluctuations}) + O(\log(nt))$$

  – for large t, $Q^n(t)/n \approx q_\infty$, where $\rho - 1 = \gamma q_\infty$

  – $\mathbf{E}(W) = O(1)$

  – fluid paths cannot generate price changes (no queue depletions)
  . . . price changes triggered by changes in rate parameters

6. Stochastic approximations of LOB dynamics

   - recap of background information in asymptotic behavior of queueing models
   - a simple model of adverse selection
   - diffusion model of LOB top of book
   - pointwise-stationary-fluid-model (PSFM)

Motivating question #4

- **probability that an order will get filled**
- **conditional probability that this will be an "adverse" fill**
- **estimate adverse selection costs as a fcn of queue position**

Adverse selection

The issue:

- for a limit order to get filled, a trader must decide to cross the spread
- that action may convey information about the price (that may move adversely)
- more likely to get filled by a large trade if at the back of the queue
- large trades often indicate future price movements

The role of queue position:

- front of queue . . . less waiting time, higher probability of a fill, could trade against small counter order
- back of queue . . . higher waiting time, smaller probability of a fill, likely to trade against a large (informed) trade
  . . . higher probability that you may regret trading at that price

Simplified model of price changes

i. stochastic fluctuations in queue lengths that lead to occasional queue depletions

   – when queue is depleted, with probability $1 - \alpha$ is bounces back up, and
   – with probability α the price changes

ii. flow imbalance "detected" by MM

   – MM maintain a noisy measure of flow imbalance between natural interest to buy and sell
   – MM cancel orders or trade aggressively when flow imbalance becomes significant
   – in part, MM cancel to avoid AS by filling orders immediately prior to a price change
   – typically 1-2 ticks and do not require lots of volume to trade

iii. block trades (informed investors)

   – price change as a result of a block of volume traded

Simplified model of price changes - II

i. stochastic fluctuations in queue lengths that lead to occasional queue depletions

   – when queue is depleted, with probability $1 - \alpha$ is bounces back up, and
   – with probability α the price changes

   - unlikely in liquid & deep securities
   - λ, μ imbalance is $O(n)$, queues are $O(n)$ but stochastic fluctuations are $O(\sqrt{n})$
   - disregard this effect in the sequel

Simplified model of price changes - III

ii. flow imbalance "detected" by MM

- Poisson with rates κ_1^+ and κ_1^-
- rates could be state dependent (not here)

iii. block trades (informed investors)

- Poisson with rates κ_2^+ and κ_2^-

So:

– study superposition of Poisson flows

– if we model magnitude of price change, we get compound Poisson's

– could also model volume of block trades, again compound Poisson

Setup for adverse selection calculation

- given queue position x, $d = \mathbf{E}(W(x))$ = expected delay until the x^{th} order in queue will get filled

- events to keep track (convention: $+ve$ jump pushes price away (no fill)):

  $\mathbf{P}(\text{fill}) = \mathbf{P}(\text{no jumps in } [0, d] \text{ or 1st jump occurs in } [0, d] \text{ and is -ve})$

  $\mathbf{P}(\text{AS fill}) = \mathbf{P}(\text{1st jump occurs in } [0, d] \text{ and is -ve})$

  $\mathbf{P}(\text{no fill}) = \mathbf{P}(\text{1st jump occurs in } [0, d] \text{ and is +ve})$

  – above calculations depend on $d = \mathbf{E}(W(x))$ and adjust in real-time as fcn of $x(s)$

Probability of an adverse fill

- probability of a fill:

$$\mathbf{P}(\text{fill}) = \frac{\kappa^-}{\kappa} + \frac{\kappa^+}{\kappa} e^{-d\kappa}$$

- probability of an adverse fill (due to a jump):

$$\mathbf{P}(\text{adverse fill}) = (1 - e^{-d\kappa})\frac{\kappa^-}{\kappa}$$

and

$$\mathbf{P}(\text{adverse fill}|\text{fill}) = \frac{\kappa^-(1 - e^{-d\kappa})}{\kappa^- + \kappa^+ e^{-d\kappa}}$$

- probability of no fill:

$$\mathbf{P}(\text{no fill}) = (1 - e^{-d\kappa})\frac{\kappa^+}{\kappa}$$

- above consider a "race" between +ve and -ve jumps over the duration d
- AS often measures price moves within Δ after fill (similarly)

Including trading volume considerations & fragmentation

- fragmentation:
  - queue position: order exchanges by fee (lowest to largest)
  - cheaper exchanges placed ahead; more expensive placed behind
  - Q_{fr} = depth in front of order; Q_{beh} = depth behind order
  - Q_{oth} is the depth on the other side of the book
- consider jump size distribution:
  - Fill: no jump in $[0, d]$ or (-ve) jump of size $\geq Q_{fr}$
  - AS: (-ve) jump of size $\geq Q_{fr} + Q_{beh}$
  - No Fill: (+ve) jump of size Q_{oth}
- intuitive results:
  - side of next price move depends on relative sizes of bid and ask queues
  - duration of race depends on d, a function of our queue position
  - AS $\downarrow$ as $Q_{fr} \downarrow$ and as $Q_{beh} \uparrow$
  - "ubiquitous" queue imbalance seems to emerge

Measurements of adverse selection costs (Moallemi and Yuan, 2014)

- table below measures some of these quantities on Nasdaq ITCH dataset (incl. order IDs)
- their model is related to previous slides (but not exactly match our discussion)
- Nasdaq rebate = \$.29; spread = 1 tick

Symbol	Order Value		Fill Probability		Adverse Selection		Order Value at the Front	
	Model (ticks)	Backtest (ticks)	Model	Backtest	Model (ticks)	Backtest (ticks)	Model (ticks)	Backtest (ticks)
BAC	0.14	0.14	0.62	0.60	0.57	0.57	0.36	0.31
CSCO	0.08	0.07	0.63	0.59	0.68	0.68	0.24	0.21
GE	0.08	0.09	0.62	0.60	0.67	0.65	0.19	0.23
F	0.13	0.15	0.65	0.64	0.60	0.53	0.24	0.23
INTC	0.11	0.09	0.64	0.61	0.63	0.56	0.28	0.23
PFE	0.12	0.11	0.63	0.58	0.62	0.61	0.16	0.21
PBR	-0.03	-0.04	0.57	0.53	0.85	0.89	0.03	0.03
EMM	0.07	0.08	0.63	0.63	0.69	0.64	0.21	0.15
EFA	0.03	0.04	0.57	0.53	0.74	0.73	0.06	0.09

(Results averaged over August 2013)

6. Stochastic approximations of LOB dynamics

   - recap of background information in asymptotic behavior of queueing models
   - a simple model of adverse selection
   - quick remarks on diffusion model of LOB top of book & PSFM
   - questions

Stochastic approximation of LOB dynamics (A)

Two natural alternatives:

- A. Diffusion model:
  - $O(\sqrt{n})$ queue lengths
  - $O(\sqrt{n})$ stochastic fluctuations due to flow imbalance
  - $\mathbf{E}(W) = O(1/\sqrt{n})$ and $\mathbf{E}(\tau) = O(1)$
  - specifically: queueing delays not visible in diffusion model (snapshot principle)
  - τ random variables depends on queue sizes
  - cancellations: if proportional to Q, then $O(\sqrt{n})$
  - cancellations: if $-\eta\delta t$, then $O(n)$ but delay estimates suffer

- time scale separation:

  very short queueing delays vs. moderate price change periods

Stochastic approximation of LOB dynamics (B)

Two natural alternatives:

- B. Pointwise-Stationary-Fluid-Model:
  - order n imbalance in λ, μ, results in $O(n)$ queues
  - natural stochastic fluctuations $O(\sqrt{n})$ not important
  - $\mathbf{E}(W) = O(1)$
  - mean field transient converges to stationary state (no price changes)
  - "slower time scale:" model parameters λ_t, μ_t vary stochastically

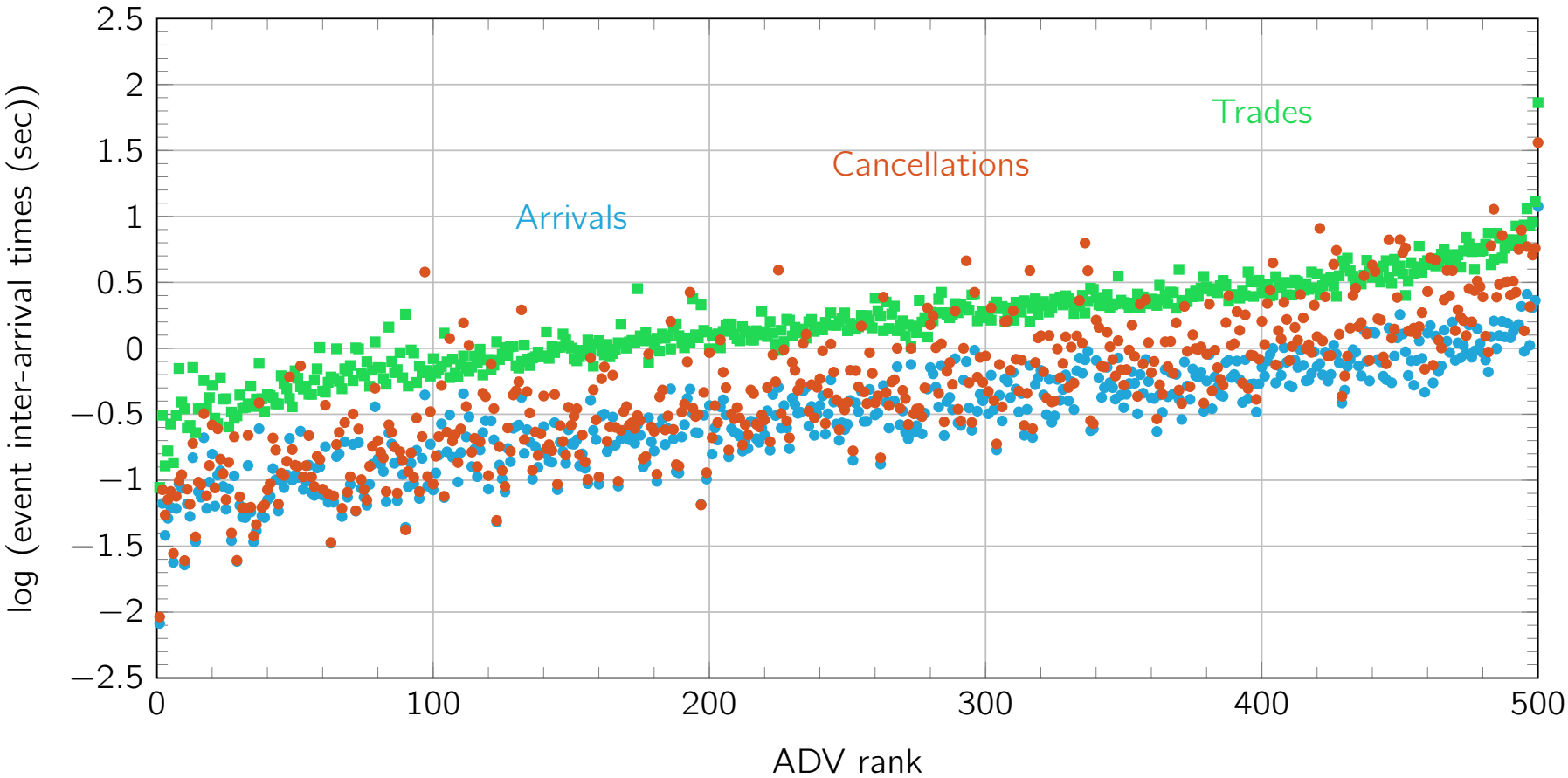
$$\mu^n(t) = n\mu(t/a_n), \quad \text{where } a_n \to \infty,\ a_n/n \to 0 \text{ as } n \to \infty$$

  - on λ_t, μ_t time scale, price changes, and LOB state changes
  - needs exogenous models of:
    a) λ_t, μ_t random evolution (drive volatility in state and price)
    b) price moves at time epochs where parameters change

- time scale separation:

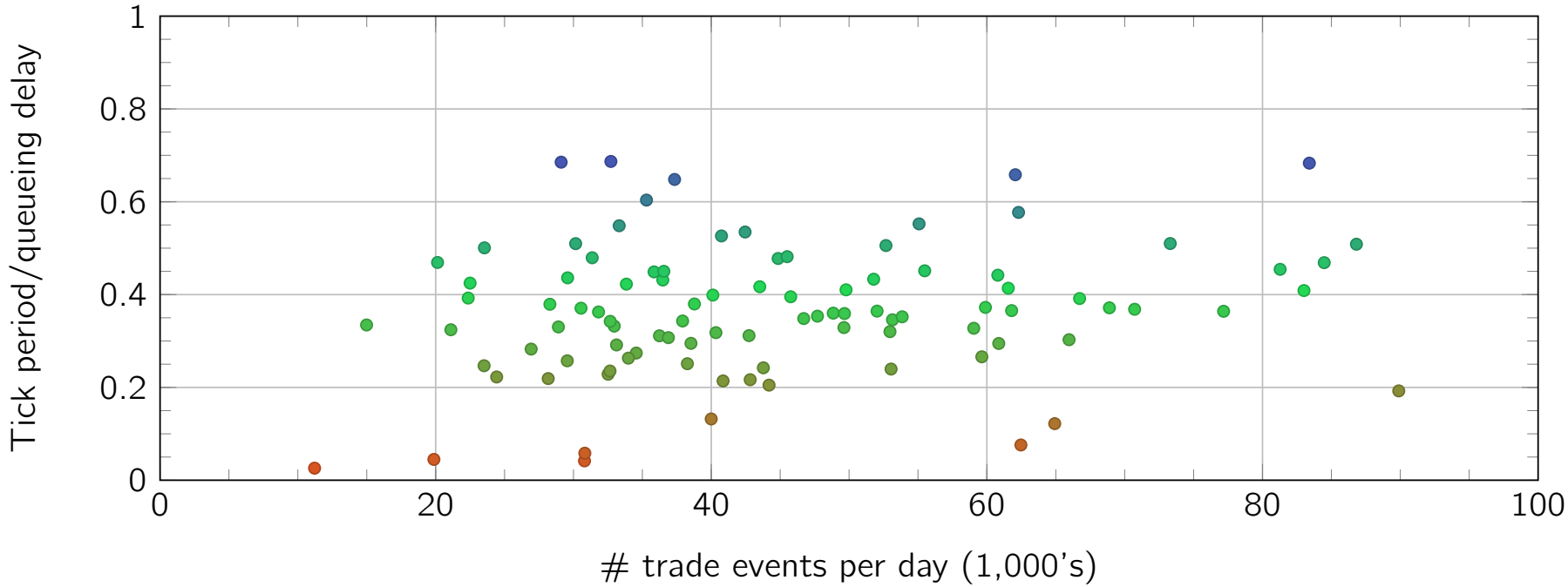
moderate queueing delays vs. longer price change periods

Interarrival times (log scale) (top of book)



- liquid stocks: # trades, # cancellations, # limit order arrivals are large
- # trades $\approx$ 1 order of magnitude less frequent than cancels or order arrivals

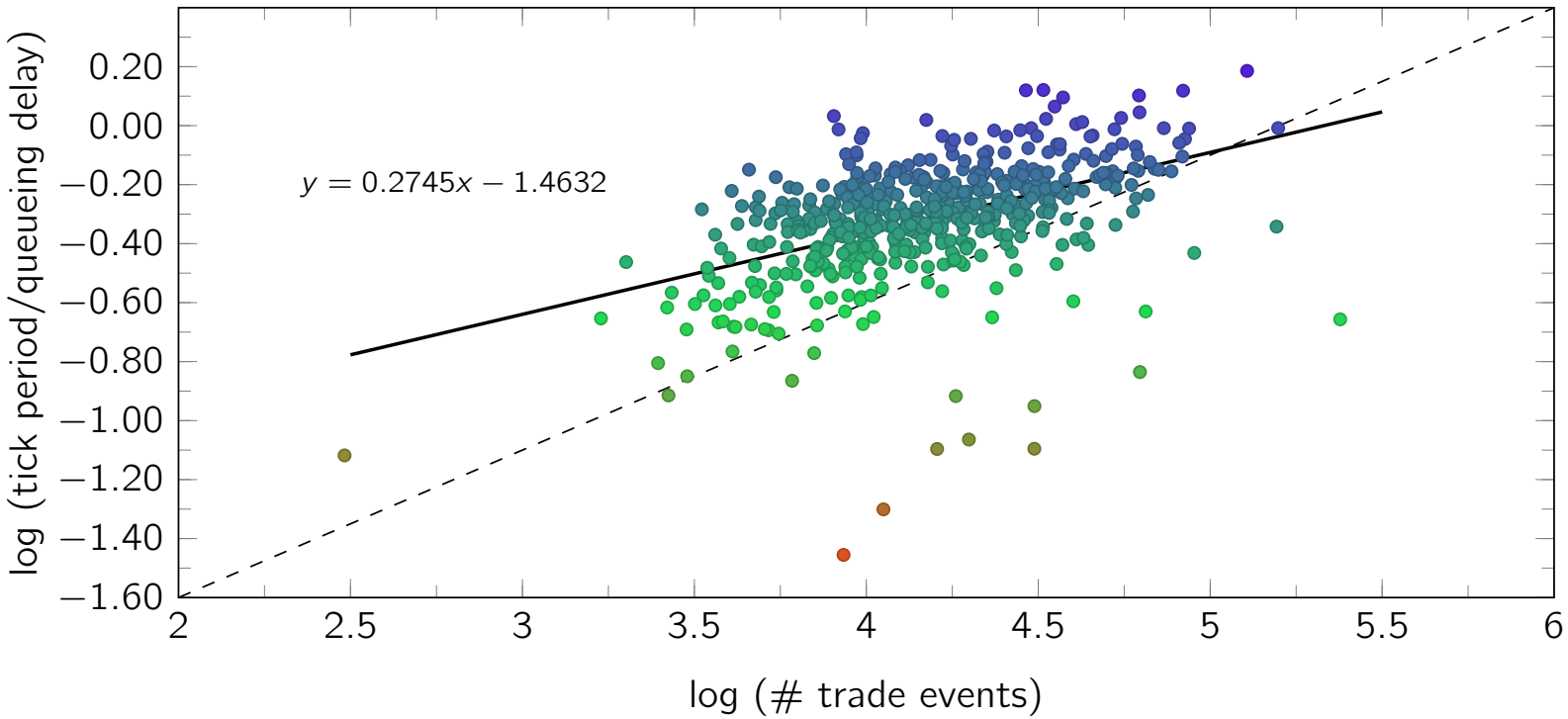
Tick period / queueing delay against # trade events



Tick period versus queueing delay: ratio against # trade events. (liquid names)

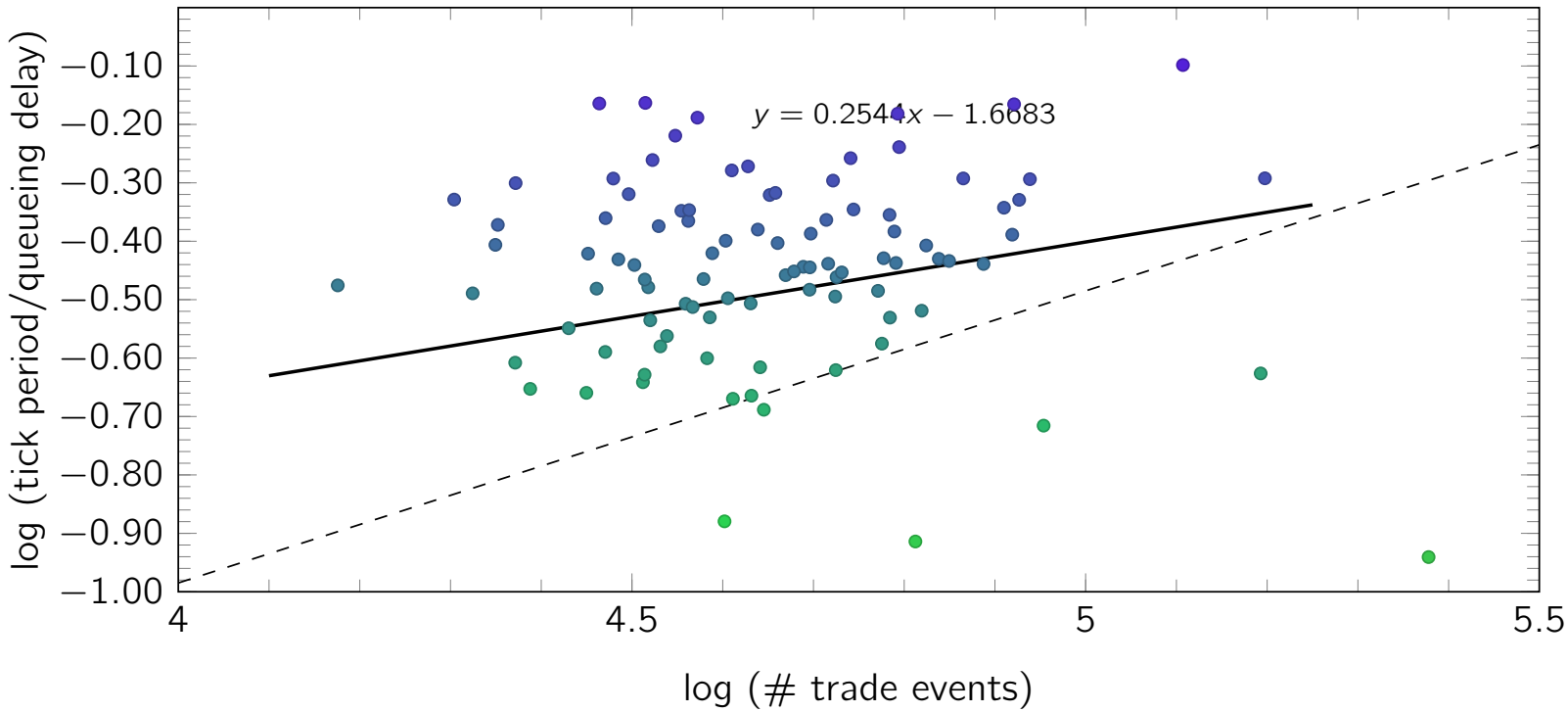
- tick period = avg time between changes in the mid-price
- tick period is on same (or smaller) order magnitude as queueing delay

Tick period versus queueing delay: log-log

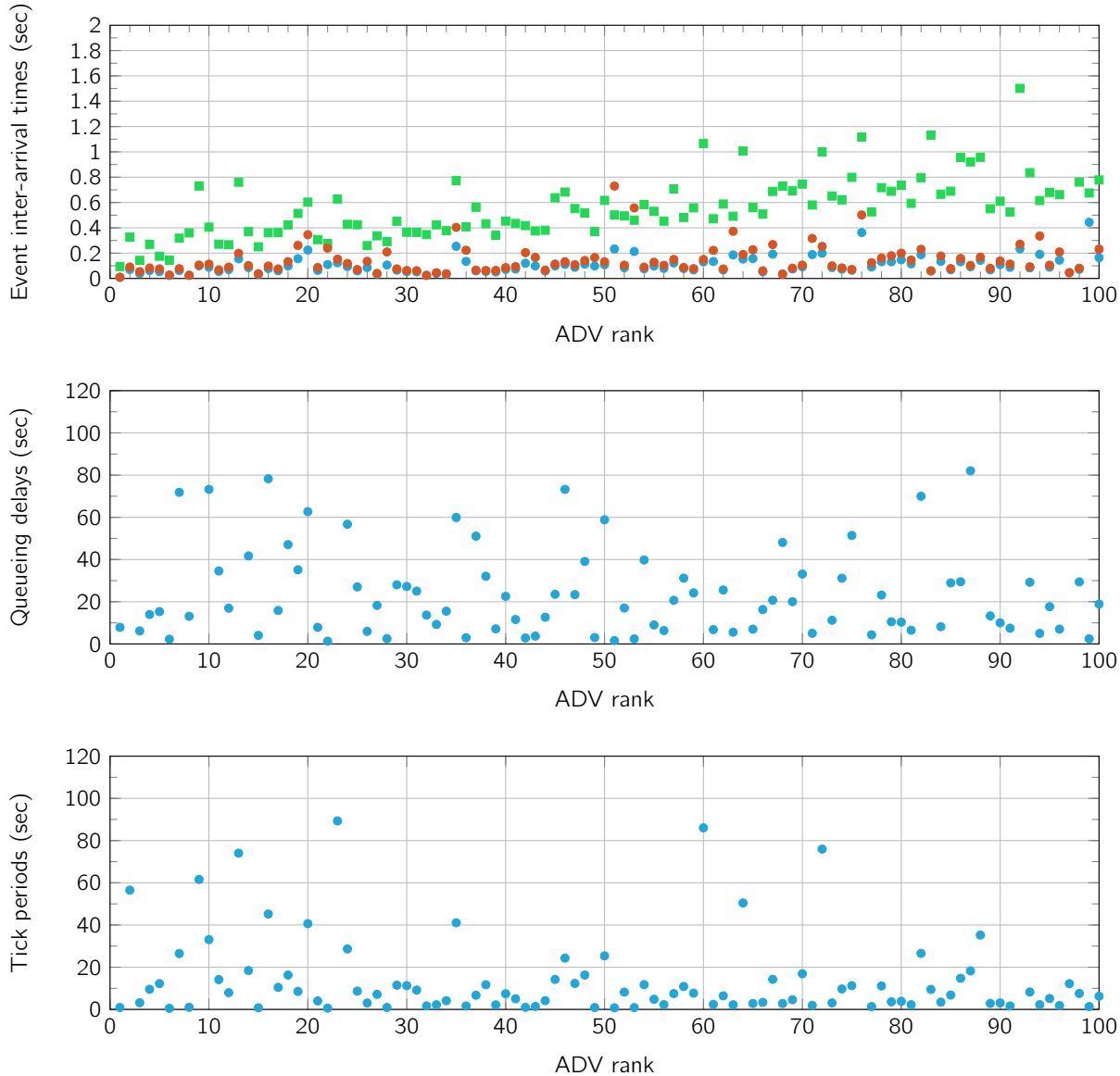


Tick period versus queueing delay: log-log, slope = 0.2745<0.5.

Tick period versus queueing delay (liquid names): log-log



Tick period versus queueing delay: log-log, slope = 0.2745<0.5.


Event inter-arrival times (sec)
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0 10 20 30 40 50 60 70 80 90 100
ADV rank
Queueing delays (sec)
120
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100
ADV rank
Tick periods (sec)
120
100
80
60
40
20
0
0 10 20 30 40 50 60 70 80 90 100
ADV rank

Variability of order arrival rates

	% obs. in $\pm 2\sigma_t$	% obs. in $\pm 3\sigma_t$	% obs. outside $\pm 3\sigma_t$
1 min	63.33%	79.23%	20.77%
3 min	32.56%	50.39%	49.61%
5 min	27.27%	35.06%	64.94%
10 min	13.16%	31.58%	68.42%

- table checks if $\mu_{t+1} \in$ intervals $\mu_t \pm k\sigma_t$ for $k = 2, 3$
- (λ, μ) exhibit significant differences in the time scale of 3 - 5 minutes
- cf. top 100 names (by ADV): average queueing delay = 61 sec

PSFM – setup

Slow time scale processes: 1st order variability of rate processes

$$\mu^n(t) = n\mu(t/a_n) \quad \text{and} \quad \lambda^n(t) = n\lambda(t/a_n)$$

where $a_n \to \infty$, $a_n/n \to 0$ as $n \to \infty$

e.g., $\mu(t), \lambda(t)$ are affine diffusions

$$d\mu(t) = \alpha_1(\bar{\mu}(t) - \mu(t))dt + \sigma_1\sqrt{\mu(t)}dB_1(t)$$

and

$$d\lambda(t) = \alpha_2(\bar{\lambda}(t) - \lambda(t))dt + \sigma_2\sqrt{\lambda(t)}dB_2(t)$$

Fast time scale transient: λ, μ appear stationary,

$$Q(s) \to f(\lambda - \mu)$$

Queue process in slow time scale: $dQ(t)$ in terms of $\frac{1}{\gamma}(d\lambda(t) - d\mu(t))$
(reflected at $Q(t) = 0$ and re-initialized at depth of bid-1 or ask+1 depths)

Wrap-up

- queueing dynamics of LOB seem crucial in tactical trading decisions:
  - timing order placement
  - order routing
  - adverse selection
  - cancellation behavior
- short-term market design & regulation initiatives should consider short time scale view of LOB and their impact
  - on short-term trading strategies
  - depth & AS
  - transaction costs
- interesting application domain for stochastic networks

QUESTIONS?

also email: c.maglaras@gsb.columbia.edu

References

Disclaimer: the list of references that follows is incomplete.

A few specific remarks:

- not referenced the extensive finance empirical literature on market microstructure
- not referenced most of the papers on market microstructure theory
- apart from a handful of papers, I have not referenced the econophysics literature on limit order books
- queueing papers on LOB either focus on descriptive issues or tactical decision making. I have only referenced a couple of the recent papers that strive to characterize the shape of the LOB. I have not reviewed most of the literature on double-sided queues, apart from referencing Kendall's early paper on the topic.
- I have provided a very limited set of references on stochastic networks. In addition the book by Chan and Yao could serve as a reference text, introductory texts on "Queueing systems" could provide background on simple queueing systems such as $M/M/1$ and the $M/M/1+M$.
- not referenced the extensive OR literature on the "economics of queues" or "queues with strategic users" that relate to the problems of order placement, order routing, and cancellation.
- not referenced the literature on PS queues that is related to prorata market structures.
- ...

References: background on limit order book markets

- Parlour, C.A., Price dynamics in limit order markets. Review of Financial Studies, 1998, 11, 789-816.
- E. Smith, J. D. Farmer, L. Gillemot, S. Krishnamoorthy, Statistical theory of the continuous double auction, Quantitative Finance, 2003, 3, 481-514.

  (good starting point for the "process flow" view of LOB; queueing per se is not in that paper)
- Parlour, C. and Seppi, D.J., Limit order markets: a survey. In Proceedings of the Handbook of Financial Intermediation and Banking, edited by A. Thakor and A. Boot, 2008, Elsevier.

  (good reference article on LOB, including a short overview of microstructure facts and the finance literature)
- M. D. Gould, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison. Limit order books. Quantitative Finance, 13:1709-1742, 2013.
- Ioanid Rosu. A dynamic model of the limit order book. Review of Financial Studies, 22:4601-4641, 2009.

References on queueing models of limit order books

- R. Cont, S. Stoikov, and R. Talreja. A stochastic model for order book dynamics. Operations Research, 58:549-563, 2010.

  (first paper to explicitly study the queueing dynamics of the LOB and make some tactical predictions of its queueing behavior; used exact analysis)

- R. Cont and A. De Larrard. Price dynamics in a markovian limit order market. SIAM Journal of Financial Mathematics, 4(1):1-25, 2013.

- R Cont, Adrien de Larrard (2011) Order book dynamics in liquid markets: limit theorems and diffusion approximations, working paper.

  (this paper derived and studied a diffusion model (in the conventional heavy-traffic regime) (cf. slides 78-79, 82, 209))

- S. Stoikov, M. Avellaneda, and J. Reed. Forecasting prices from level-i quotes in the presence of hidden liquidity. Algorithmic Finance, Forthcoming, 2011.

- C. Maglaras, C. Moallemi, H. Zheng, Queueing Dynamics and State Space Collapse in Fragmented Limit Order Book Markets, working paper, 2013.

- M. Toke. The order book as a queueing system: average depth and influence of the size of limit orders. Quantitative Finance, 14, 2014.

- J. Blanchet and X. Chen. Continuous-time modeling of bid-ask spread and price dynamics in limit order books. Working paper, 2013.

- P. Lakner, J. Reed, and S. Stoikov. High frequency asymptotics for the limit order book. Working paper, 2013.

- X. Gao, J. G. Dai, A. B. Dieker, and S. J. Deng. Hydrodynamic limit of order book dynamics. http://arxiv.org/pdf/ 1411.7502.pdf, 2014.

- F. Kelly and E. Yudovina, A Markov model of a limit order book: thresholds, recurrence and trading strategies, working paper, 2014.

References on order routing

- C. Maglaras, C. Moallemi, H. Zheng, Queueing Dynamics and State Space Collapse in Fragmented Limit Order Book Markets, working paper, 2013.

  (the order routing of infinitesimal traders and the resulting coupling across LOBs is based on this paper)

- R. Cont and A. Kukanov. Optimal order placement in limit order markets. Working paper, 2013.

  (the high-level order routing problem borrows from Cont-Kukanov)

- X. Guo, A. De Larrard, and Z. Ruan. Optimal placement in a limit order book. Working paper, 2013.

- G. Sofianos, J. Xiang, and A. Yousefi. Smart order routing: All-in shortfall and optimal order placement. Goldman Sachs, Equity Executions Strats, Street Smart, 42, 2011.

- T. Foucault and A. J. Menkveld. Competition for order flow and smart order routing systems. Journal of Finance, 63:119-158, 2008.

Some references on modeling market impact with microstructure variables

- Kyle, A. S., and A. A. Obizhaeva. 2011. arket Microstructure Invariants: Theory and Implications of Calibration.orking paper. Available at http://ssrn.com/abstract=1978932.
- Cont, A Kukanov, S Stoikov (2014) The price impact of order book events, Journal of Financial Econometrics Vol 12, No 1, 47-88
- C. Maglaras, C. Moallemi, H. Zheng (2015), Optimal execution in a limit order book and an associated microstructure market impact model, working paper, Columbia University.

  (the background for slides 123-152)
- Z. Eisler, J. Bouchaud, and J. Kockelkoren. The price impact of order book events: market orders, limit orders and cancellations. Quantitative Finance, 12(9):1395-1419, 2012.

Background references on market impact

- G. Huberman and W. Stanzl. Price manipulation and quasi-arbitrage. Econometrica, 74(4):1247-1276, 2004.
- J. Gatheral. No-dynamic-arbitrage and market impact. Quantitative Finance, 10(7):749-759, 2010.
- R. Almgren, C. Thum, E. Hauptmann, and H. Li. Direct estimation of equity market impact. Risk, July 2005.
- V. Rashkovich and A. Verma. Trade cost: Handicapping on PAR. Journal of Trading, 7(4), 2012.
- George C. Chacko, Jakub W. Jurek, and Erik Stafford. The price of immediacy. The Journal of Finance, 63(3):1253-1290, 2008. ISSN 1540-6261.
- C. Moallemi, M. Saglam, and M. Sotiropoulos. Short-term predictability and price impact. Working paper, 2014.
- J. Farmer, A. Gerig, F. Lillo, and S. Mike. Market efficiency and the long-memory of supply and demand: Is price impact variable and permanent or fixed and temporary? Quantitative Finance, 6(2):107-112, 2006.
- J.-P. Bouchaud, J. D. Farmer, and F. Lillo. How markets slowly digest changes in supply and demand. In Handbook of Financial Markets: Dynamics and Evolution, pages 5756. Elsevier: Academic Press, 2008.
- Bouchaud, J.-P. 2010. rice Impact In R. Cont (ed.), Encyclopedia of Quantitative Finance, Chichester (UK): Wiley.
- Plerou, V., P. Gopikrishnan, X. Gabaix, and H. Stanley. 2002. Quantifying Stock-Price Response to Demand Fluctuations. Physical Review E 66: 027104.
- Potters, M., and J. Bouchaud. More Statistical Properties of Order Books and Price Impact. Physica A, 324: 133-140.

Few references on adverse selection

- C. Moallemi and K. Yuan, A model of queue position valuation, working paper, 2015.
  (the motivation and structure of the adverse selection discussion in these slides is based on Moallemi-Yuan)
- S. Skouras and D. Farmer, The value of queue priority, working paper, 2013.
- B. Biais, F. Declerck, S. Moinas, Fast trading and prop trading, workign paper, 2014.
- D. Jeria and G. Sofianos, Passive orders and natural adverse selection. Goldman Sachs, Equity Executions Strats, Street Smart, 33, 2008.

Some references on stochastic networks

- Kurtz, T. G. (1977/78) Strong approximation theorems for density dependent Markov chains. Stochastic Processes Appl. 6, 223-240.
- Glynn, P. W. (1990) Diffusion Approximations. In Heyman, D. and Sobel, M. (eds.), Stochastic Models, volume 2 of Handbooks in OR MS, 145-198. North-Holland.
- A. Mandelbaum and G. Pats. State-dependent queues: approximations and applications. In F. Kelly and R. Williams, editors, Stochastic Networks, volume 71, pages 239-282. Proceedings of the IMA, 1995.
- J. M. Harrison. Balanced fluid models of multiclass queueing networks: a heavy traffic conjecture. In F. Kelly and R. Williams, editors, Stochastic Networks, volume 71, pages 1-20. Proceedings of the IMA, 1995.
- M. Bramson. State space collapse with applications to heavy-traffic limits for multiclass queueing networks. QUESTA, 30:89-148, 1998.
- Whitt, W. 2003. How multiserver queues scale with growing congestion-dependent demand. Oper. Res. 51(4) 531-542.
- Kendall DG (1951) Some problems in the theory of queues. J. Roy. Statist.Soc. B(13):151-185.

References on optimal execution (small subset)

Trade scheduling (not including papers on VWAP / TWAP, although some interesting work exists):

- D. Bertsimas and A. W. Lo. Optimal control of execution costs. Journal of Financial Markets, 1: 1-50, 1998.
- R. Almgren and N. Chriss. Optimal control of portfolio transactions. Journal of Risk, 3:5-39, 2000.
- R. Almgren. Optimal execution with nonlinear impact functions and trading-enhanced risk. Applied Mathematical Finance, 10:1-18, 2003.
- G. Huberman and W. Stanzl. Optimal liquidity trading. Review of Finance, 9:165-200, 2005.
- A. Alfonsi, A. Fruth, and A. Schied. Optimal execution strategies in limit order books with general shape functions. Quantitative Finance, 10:143-157, 2010.
- A. Obizhaeva and J. Wang. Optimal trading strategy and supply/demand dynamics. Working paper, 2006.

References on market microstructure

- R. Roll. Orange juice and weather. American Economic Review, pages 861-880, 1984
- L. R. Glosten and P. R. Milgrom. Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics, 14(1):71-100, 1985.
- A. S. Kyle. Continuous auctions and insider trading. Econometrica, 53:1315-1335, 1985
- L. Glosten. Is the electronic limit order book inevitable? Journal of Finance, 49 (4):1127-1161, 1994.
- J. Hasbrouck. Empirical market microstructure: The institutions, economics and econometrics of securities trading. Oxford University Press, Oxford, 2007.