L06: Market Microstructure & Price Discovery

Extended Slides – The Hidden Mechanics of Markets

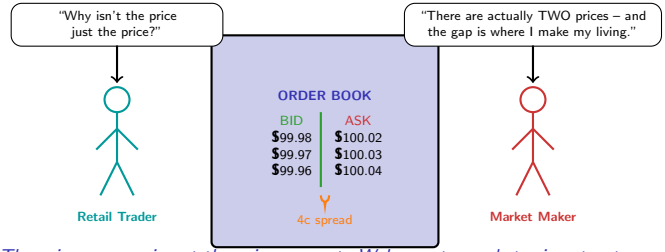
Digital Finance

What Will You Be Able to Do After This Lecture?

1. Reconstruct a limit order book from a message feed and compute real-time imbalance, depth, and mid-price dynamics
2. Decompose the bid-ask spread into order processing, inventory holding, and adverse selection components using the Huang-Stoll and Glosten-Milgrom models
3. Estimate Kyle's lambda from trade-and-quote data and interpret information asymmetry across stocks of different sizes
4. Evaluate how dark pools and venue fragmentation affect price discovery, execution quality, and systemic risk
5. Quantify the impact of HFT on market quality using latency distributions, market-maker P&L, and tick-size experiments

---

Five objectives: order book mechanics (1), spread theory (2), information asymmetry (3), venue structure (4), and HFT impact (5). Mathematical models with empirical evidence and 12 data visualizations.



The price you see is not the price you get. Welcome to market microstructure.

How Does a Limit Order Book Match Buyers and Sellers?

Limit Order Book (LOB) – a queue of resting orders at each price level, matched by **price-time priority**:

Order type	Mechanism	Risk	Who uses it
Limit order	Rests in the book at a specified price; executes only at that price or better	Non-execution risk: may never fill	Patient traders, market makers
Market order	Executes immediately against the best resting limit orders	Price impact: "walks the book" if size exceeds depth at best	Impatient traders, informed traders
Marketable limit	Limit order priced at or through the opposite side	Hybrid: immediate fill with price protection	Algo engines
Iceberg / reserve	Displays only a fraction; replenishes from hidden reserve	Detection risk: pattern recognition by HFT	Institutional desks

- **Price priority:** Best-priced orders fill first. A bid at \$100.01 fills before a bid at \$100.00.
- **Time priority:** Among orders at the same price, the earliest order fills first (FIFO).
- **Mid-price:** $m = (\text{best bid} + \text{best ask})/2$. The spread = best ask − best bid.

The LOB is the central mechanism of price formation in electronic markets. Price-time priority ensures fairness: best price wins, and among equals, earliest arrival wins.

Can the Shape of the Order Book Predict Where Prices Are Heading?

Order book imbalance (OBI) at the best level:

$$\text{OBI} = \frac{V_1^{\text{bid}} - V_1^{\text{ask}}}{V_1^{\text{bid}} + V_1^{\text{ask}}} \in [-1, +1]$$

Weighted imbalance across L levels (Cont, Kukanov & Stoikov, 2014):

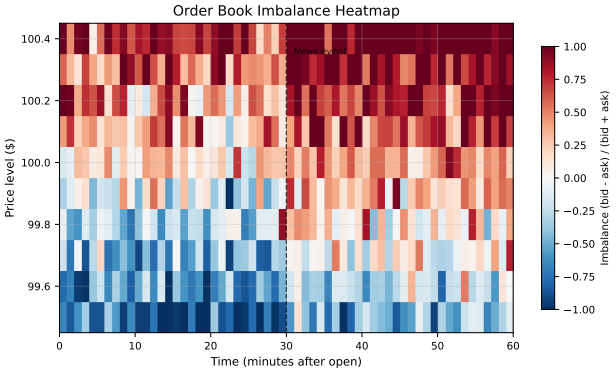
$$\text{WOBI} = \frac{\sum_{\ell=1}^{L} w_\ell \, V_\ell^{\text{bid}} - \sum_{\ell=1}^{L} w_\ell \, V_\ell^{\text{ask}}}{\sum_{\ell=1}^{L} w_\ell \, (V_\ell^{\text{bid}} + V_\ell^{\text{ask}})}, \quad w_\ell = e^{-\alpha(\ell-1)}$$

Worked example: Best 3 levels: Bid volumes $= [500, 300, 200]$, Ask volumes $= [200, 400, 350]$, $\alpha = 0.5$.

- Weights: $w_1 = 1.00$, $w_2 = 0.607$, $w_3 = 0.368$
- Weighted bid: $500(1.00) + 300(0.607) + 200(0.368) = 500 + 182.1 + 73.6 = 755.7$
- Weighted ask: $200(1.00) + 400(0.607) + 350(0.368) = 200 + 242.8 + 128.8 = 571.6$
- WOBI $= (755.7 - 571.6)/(755.7 + 571.6) = 184.1/1327.3 = +0.139$ (bid-heavy $\Rightarrow$ price likely to rise)
- **Empirical fact:** OBI predicts short-term price moves with $R^2 \approx 0.05$–0.15 at the 1-second horizon.

Order book imbalance is the single most predictive feature for short-horizon price changes. Cont et al. (2014) show weighted imbalance across multiple levels outperforms best-level imbalance.

What Does Order Book Imbalance Look Like Through Time?



- **Blue** = bid-heavy (buying pressure); **red** = ask-heavy (selling pressure)
- Each row is a price level; each column is a time snapshot
- The vertical dashed line marks a **news event** that shifts imbalance toward buying
- Before the event: balanced book. After: persistent bid-side imbalance across multiple levels
- HFT firms monitor this heatmap in real time to adjust quotes
- The imbalance gradient across price levels reveals the **depth of conviction**

Source: Simulated LOB data based on Cont, Kukanov & Stoikov (2014) order flow dynamics. Imbalance regime shifts after news events are a well-documented empirical phenomenon.

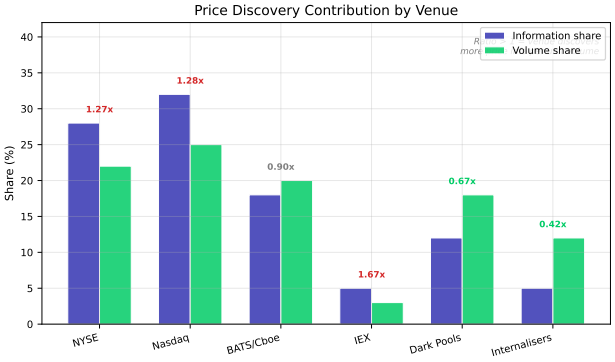
Can You Rebuild an Order Book from a Raw Message Feed?

```python
from collections import defaultdict
class OrderBook:
    def __init__(self):
        self.bids = defaultdict(float)
        self.asks = defaultdict(float)
    def update(self, side, price, qty):
        book = self.bids if side=='B' else self.asks
        if qty == 0: book.pop(price, None)
        else: book[price] = qty
    def best_bid(self):
        return max(self.bids) if self.bids else 0
    def best_ask(self):
        return min(self.asks) if self.asks else 1e9
    def spread(self):
        return self.best_ask() - self.best_bid()
    def imbalance(self, levels=1):
        bids = sorted(self.bids, reverse=True)
        asks = sorted(self.asks)
        bv = sum(self.bids[b] for b in bids[:levels])
        av = sum(self.asks[a] for a in asks[:levels])
        return (bv-av)/(bv+av) if bv+av else 0
ob = OrderBook()
ob.update('B', 99.98, 500)
ob.update('B', 99.97, 300)
ob.update('A', 100.02, 200)
ob.update('A', 100.03, 400)
print(f"Spread: ${ob.spread():.2f}")
print(f"OBI: {ob.imbalance():.3f}")
```

- The `OrderBook` class maintains bid/ask sides as price $\rightarrow$ quantity dictionaries
- `update()` processes each message: qty zero deletes a level
- `imbalance()` computes OBI across the top L levels
- In production, ITCH/OUCH feeds deliver $\sim$50K messages/sec per symbol
- Extends to multi-level depth, order-ID tracking, and queue position estimation

Every exchange publishes a message feed (ITCH, OUCH, PITCH). Reconstructing the order book from this feed is the first step in any microstructure analysis or trading system.

Which Venue Actually Discovers the True Price?



- **Information share** (Hasbrouck, 1995) measures each venue's contribution to the common efficient price
- Nasdaq and NYSE contribute the most information per unit volume – their quotes lead price changes
- **Ratio** > 1: venue discovers more price information than its volume share implies
- Dark pools have low information share per unit volume – trades are less informative
- IEX's speed bump protects quotes but limits price discovery contribution
- Internalisers handle retail flow with minimal information content

Source: Hasbrouck (1995) information share methodology applied to US equities. Lit exchanges dominate price discovery; dark pools contribute volume but less information.

How Does Each Trade Permanently Shift the Price?

Linear price impact model (Kyle, 1985):

$$\Delta p_t = \lambda \cdot \text{OFI}_t + \varepsilon_t$$

where $\text{OFI}_t = \sum_i q_i \cdot \mathbb{1}[\text{buyer-initiated}] - \sum_i q_i \cdot \mathbb{1}[\text{seller-initiated}]$ is order flow imbalance and λ is **Kyle's lambda** (price impact per unit of net order flow).

Interpretation of λ:

- High λ = illiquid market: each trade moves the price a lot (high information asymmetry)
- Low λ = liquid market: the book absorbs order flow with minimal price concession

Worked example: Over a 5-minute interval, 8 buyer-initiated trades totalling 5,000 shares and 5 seller-initiated trades totalling 3,000 shares. Price moves from \$50.00 to \$50.06.

- $\text{OFI} = 5{,}000 - 3{,}000 = 2{,}000$ shares net buy
- $\Delta p = \$0.06$, so $\lambda = 0.06/2{,}000 = 3 \times 10^{-5}$ \$/share
- In basis points per \$1M: $\lambda \times (10^6/50.00) \times 10{,}000 = 6$ bps/\$1M
- **Cross-sectional fact:** $\lambda \propto 1/\sqrt{\text{market cap}}$ (see Kyle's lambda chart)

Kyle's lambda is the cornerstone of market microstructure: it connects order flow to price changes. Empirically, lambda scales as the inverse square root of market cap.

Why Must the Spread Exist Even Without Any Operating Costs?

Glosten-Milgrom (1985) sequential trade model:
A market maker faces informed traders (fraction μ) who know the true value V, and uninformed traders $(1-\mu)$ who trade randomly. The maker's quotes must protect against adverse selection:

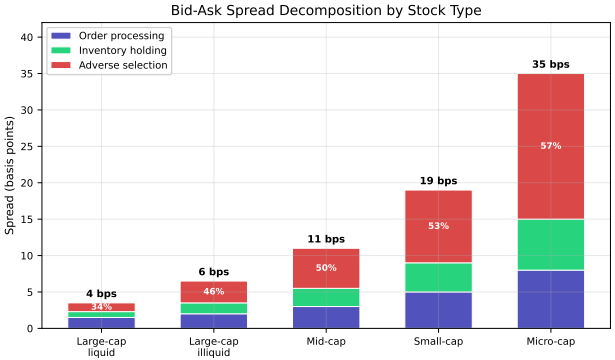
$$\text{ask} = E[V \mid \text{buyer arrives}] = V_H \cdot P(V = V_H \mid \text{buy}) + V_L \cdot P(V = V_L \mid \text{buy})$$
$$\text{bid} = E[V \mid \text{seller arrives}] = V_H \cdot P(V = V_H \mid \text{sell}) + V_L \cdot P(V = V_L \mid \text{sell})$$

Worked example: $V_H = \$52$, $V_L = \$48$, prior $P(V_H) = 0.5$, informed fraction $\mu = 0.3$.

- $P(\text{buy} \mid V_H) = \mu + (1-\mu)/2 = 0.3 + 0.35 = 0.65$; $\quad P(\text{buy} \mid V_L) = (1-\mu)/2 = 0.35$
- By Bayes: $P(V_H \mid \text{buy}) = \frac{0.5 \times 0.65}{0.5 \times 0.65 + 0.5 \times 0.35} = \frac{0.325}{0.500} = 0.65$
- Ask $= 52 \times 0.65 + 48 \times 0.35 = 33.80 + 16.80 = \$50.60$
- Similarly: Bid $= 52 \times 0.35 + 48 \times 0.65 = \$49.40$
- Spread $= 50.60 - 49.40 = \$1.20$ (purely from adverse selection – no operating cost!)
- **If $\mu = 0$:** Ask $=$ Bid $= \$50.00$ (zero spread). The spread exists *because* informed traders exist.

Glosten-Milgrom proves that the spread must be positive whenever informed traders exist, even with zero operating costs. The spread is the market's "insurance premium" against being picked off.

How Does the Composition of the Spread Change Across Stock Types?



- **Order processing** (blue): fixed cost of matching – relatively constant across stocks
- **Inventory holding** (green): increases with volatility and illiquidity
- **Adverse selection** (red): the dominant component for small and micro-caps
- For micro-caps, adverse selection is **57%** of the total spread – informed traders dominate
- For large-cap liquid stocks, order processing is the largest component (43%)
- **Policy implication:** Tick size reduction benefits large-caps (reduces processing floor) but may harm small-caps (reduces market-making incentive)

Source: Huang & Stoll (1997) decomposition methodology. Adverse selection dominates for illiquid stocks; order processing dominates for liquid ones.

How Do Econometricians Separate the Three Spread Components?

Huang-Stoll (1997) realized spread decomposition:
After a trade at price P_t (with trade direction $D_t = \pm 1$), the subsequent price change decomposes as:

$$P_{t+1} - P_t = \alpha S D_t/2 + (1-\alpha) S D_t/2 \cdot (\rho D_{t-1}/D_t) + \varepsilon_t$$

Simplified three-way split:

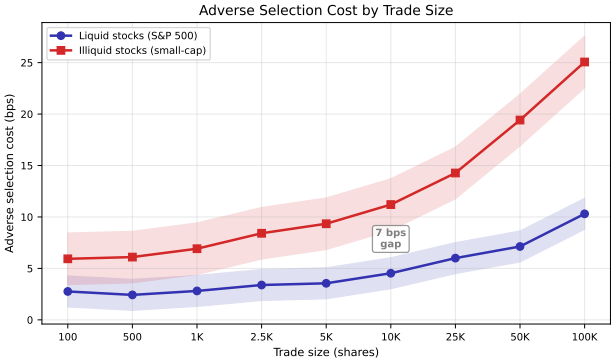
$$\frac{S}{2} = \underbrace{\pi}_{\text{adverse selection}} + \underbrace{\phi}_{\text{inventory}} + \underbrace{\theta}_{\text{order processing}}, \quad \pi + \phi + \theta = S/2$$

Estimation from data: Regress post-trade price changes on trade direction:

- **Realized spread** $= 2(P_t - m_{t+\Delta}) \cdot D_t$, where $m_{t+\Delta}$ is the midpoint Δ minutes later
- **Adverse selection** $= 2(m_{t+\Delta} - m_t) \cdot D_t$ (permanent price impact of the trade)
- **Verification:** Quoted spread = Realized spread + Adverse selection
- **Example:** Quoted spread = 4 bps. Realized spread = 1.5 bps. Adverse selection = 2.5 bps. The maker earns only 1.5 bps per round trip after losing 2.5 bps to informed traders.

The Huang-Stoll decomposition uses post-trade price changes to separate the spread into economic components. The key insight: the "realized" spread the maker earns is far less than the quoted spread.

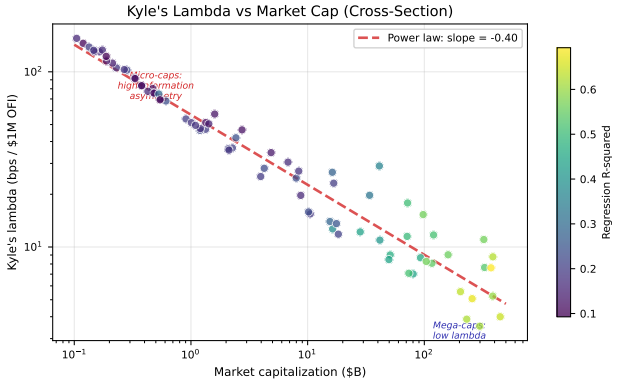
Why Do Larger Trades Carry More Information About the True Value?



- Adverse selection cost rises with trade size because **informed traders prefer large orders**
- The relationship is concave (square-root): doubling size does not double the cost
- **Liquid stocks** (blue): flat curve – depth absorbs large orders efficiently
- **Illiquid stocks** (red): steep curve – information asymmetry is amplified
- The gap between curves widens with size: **large trades in illiquid names are the most toxic**
- Market makers use trade-size-dependent quoting to protect against this

Source: Glosten-Milgrom adverse selection framework calibrated to US equity data. Informed traders "select" against the market maker by submitting larger orders.

How Does Information Asymmetry Scale with Company Size?



- Each dot is a stock; color intensity shows regression R^2
- **Power law:** $\lambda \propto \text{mcap}^{-0.5}$ (the fitted slope)
- Micro-caps: $\lambda \sim$ 50–100 bps/**$**1M OFI – each dollar of net buying moves the price substantially
- Mega-caps: $\lambda \sim$ 1–5 bps/**$**1M OFI – deep books absorb order flow
- Higher R^2 (brighter points) for liquid stocks: the Kyle model fits better when noise is lower
- **Practical use:** λ calibrates pre-trade cost estimates for execution algorithms

Source: Kyle (1985) regression on 5-minute intervals for 80 US equities. Lambda scales as the inverse square root of market cap – a robust empirical regularity.

Can You Estimate the Adverse Selection Component from Trade Data?

```python
import numpy as np
def spread_decomposition(prices, mids,
                         directions, lag=5):
    """Huang-Stoll spread decomposition.
    prices: trade prices
    mids: mid-prices at trade time
    directions: +1 buy, -1 sell
    lag: minutes for realized spread
    """
    n = len(prices) - lag
    quoted = 2 * np.abs(prices[:n] - mids[:n])
    realized = (2 * (prices[:n] - mids[lag:lag+n])
                * directions[:n])
    adverse = (2 * (mids[lag:lag+n] - mids[:n])
               * directions[:n])
    return {
        "quoted_bps": np.mean(quoted)*1e4/np.mean(mids),
        "realized_bps": np.mean(realized)*1e4/np.mean(mids),
        "adverse_bps": np.mean(adverse)*1e4/np.mean(mids),
        "as_fraction": np.mean(adverse)/np.mean(quoted)}
np.random.seed(42)
N = 1000; mid = 50 + np.cumsum(0.001*np.random.randn(N))
d = np.random.choice([-1,1], N)
p = mid + d * 0.02 + 0.005 * np.random.randn(N)
r = spread_decomposition(p, mid, d)
for k,v in r.items(): print(f"{k}: {v:.3f}")
```

- **Quoted spread:** distance from trade price to mid at execution time
- **Realized spread:** what the maker actually earns (trade price vs. future mid)
- **Adverse selection:** the permanent price impact (future mid vs. current mid)
- The `as_fraction` measures what fraction of the spread the maker loses to informed flow
- Typical values: 40–60% for most stocks
- Lag choice matters: 5 minutes is standard; shorter lags underestimate adverse selection

This Huang-Stoll estimator decomposes the spread from trade-and-quote data. The adverse selection fraction tells us how much of the spread is "information rent" paid to informed traders.

What Is the Market Maker's Fundamental Optimization Problem?

Avellaneda-Stoikov (2008) – a market maker maximizes expected terminal wealth with inventory penalty:

$$\max_{\delta^a, \delta^b} \; E\big[W_T - \gamma\, q_T^2\, \sigma^2 (T-t)/2\big]$$

Reservation price (the maker's "true" valuation given inventory q):

$$r(t) = s - q\,\gamma\,\sigma^2 (T-t)$$

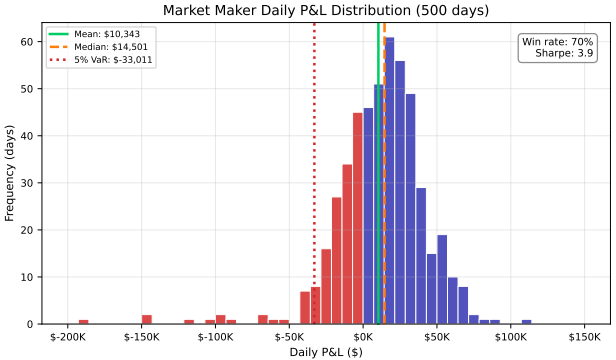
Optimal half-spread:

$$\delta^*(t) = \frac{1}{2}\left[\gamma\sigma^2 (T-t) + \frac{2}{\gamma}\ln\Big(1+\frac{\gamma}{k}\Big)\right]$$

- When $q > 0$ (long), $r < s$: the maker lowers quotes to encourage selling (shed inventory)
- When $q < 0$ (short), $r > s$: the maker raises quotes to encourage buying
- Higher γ (risk aversion) $\Rightarrow$ wider spread and more aggressive inventory management
- Higher k (order arrival intensity) $\Rightarrow$ tighter spread (more chances to earn the spread)
- As $T - t \to 0$ (end of day): spread tightens, reservation converges to mid (less risk remaining)

The Avellaneda-Stoikov model is the theoretical foundation of modern electronic market making. The reservation price shifts away from mid to incentivize inventory-reducing trades.

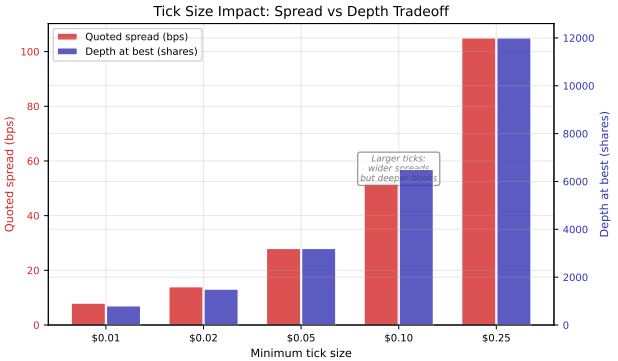
What Does a Market Maker's P&L Actually Look Like?



- The distribution is **right-skewed with a fat left tail**: many small wins, occasional large losses
- **Blue bars** = profitable days; **red bars** = losing days
- Win rate ∼75% is typical: makers earn the spread most days but lose big on adverse selection events
- The **5% VaR** shows the worst-case daily loss at the 95th percentile
- Citadel Securities reported a **losing day in less than 1% of trading days** – survivorship bias
- **Key risk:** The left tail comes from correlated informed flow (news, earnings, macro shocks)

Source: Simulated Avellaneda-Stoikov market maker over 500 trading days. The P&L distribution confirms the theory: many small gains from the spread, rare large losses from adverse selection.

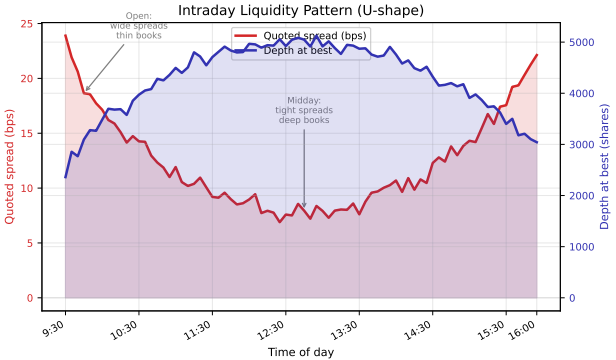
Does Making the Minimum Price Increment Smaller Help or Hurt Traders?



- Larger tick sizes mechanically **widen** the quoted spread (red bars) – the spread cannot be smaller than one tick
- But larger ticks also **deepen** the book (blue bars) – makers queue at each level because the priority rent is higher
- This is the **tick-size tradeoff**: narrow ticks benefit small orders (lower cost) but thin the book (less depth)
- SEC Tick Size Pilot (2016–2018) tested $0.05 ticks for small-caps
- Result: spreads widened but depth improved – the net effect on execution quality was ambiguous
- **EU:** MiFID II mandates venue-specific tick tables

Source: SEC Tick Size Pilot Program data (2016–2018). The tradeoff is fundamental: smaller ticks reduce visible cost but thin the book. There is no free lunch in tick-size policy.

Why Is Liquidity Worst When You Need It Most?



- **Red line:** Spread follows a U-shape – widest at open and close, tightest at midday
- **Blue line:** Depth is the mirror image – deepest at midday, thinnest at extremes
- At the open: overnight information creates uncertainty $\Rightarrow$ makers widen quotes
- At the close: portfolio rebalancing and MOC orders create urgency $\Rightarrow$ spreads widen
- **Midday:** Low information arrival, balanced flow $\Rightarrow$ tightest spreads
- **Implication:** VWAP algorithms that trade proportional to volume partially exploit this pattern

The U-shaped intraday pattern is one of the most robust empirical facts in microstructure. Spreads are widest when uncertainty is highest (open, close) and tightest when information flow is lowest (midday).

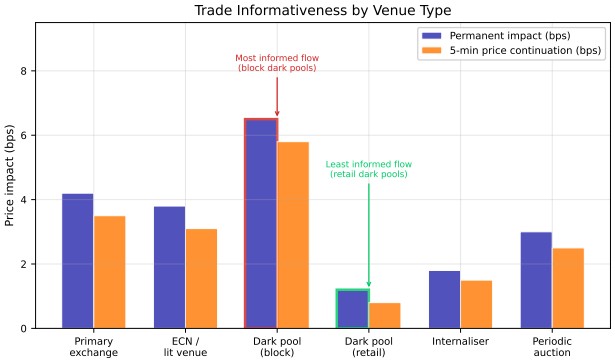
Can You Simulate a Market Maker's Quoting Strategy?

```python
import numpy as np
def simulate_mm(T=390, sigma=0.003,
                gamma=0.1, k=1.5, q0=0):
    """Avellaneda-Stoikov market maker."""
    s = 50.0 # mid-price
    q, pnl = q0, 0.0
    inv_path, spread_path = [], []
    for t in range(T):
        tau = T - t
        r = s - q * gamma * sigma**2 * tau
        delta = (gamma * sigma**2 * tau
                 + 2/gamma * np.log(1+gamma/k))
        bid, ask = r - delta/2, r + delta/2
        # Random arrivals
        if np.random.rand() < 0.3:
            q -= 100; pnl += ask * 100
        if np.random.rand() < 0.3:
            q += 100; pnl -= bid * 100
        s += sigma * np.random.randn()
        inv_path.append(q)
        spread_path.append(delta)
    pnl += q * s # mark-to-market
    return pnl, inv_path, spread_path
pnl, inv, sp = simulate_mm()
print(f"PnL: ${pnl:,.0f} Final q: {inv[-1]}")
print(f"Avg spread: ${np.mean(sp):.4f}")
```

- Reservation price r shifts below mid when long, above when short
- Spread δ widens with remaining time and risk aversion
- Customer buys (hits ask) $\Rightarrow$ maker sells: q decreases, PnL increases
- Customer sells (hits bid) $\Rightarrow$ maker buys: q increases, PnL decreases
- End-of-day mark-to-market converts final inventory to P&L
- **Extension:** Add adverse selection by making arrival probability conditional on price direction

This simulator implements the core Avellaneda-Stoikov dynamics. Production systems add adverse selection modeling, multi-asset inventory, and venue-specific fee optimization.

Do All Trading Venues Attract Equally Informed Order Flow?



- **Block dark pools** have the highest permanent impact – institutional "block" trades carry strong information signals
- **Retail dark pools** have the lowest impact – uninformed flow is valuable to market makers ("payment for order flow")
- Primary exchanges sit in between: a mix of informed and uninformed flow
- **Periodic auctions** (new EU venue type) aggregate orders to reduce adverse selection
- This ordering explains why Citadel pays billions for retail order flow: it is *non-toxic*
- **Regulatory tension:** SEC proposed banning PFOF; makers argue it improves retail execution

Source: Permanent price impact measured as 5-minute post-trade price continuation. Block dark pool trades are the most informative; retail internalized trades are the least.

Why Would Anyone Want to Trade in the Dark?

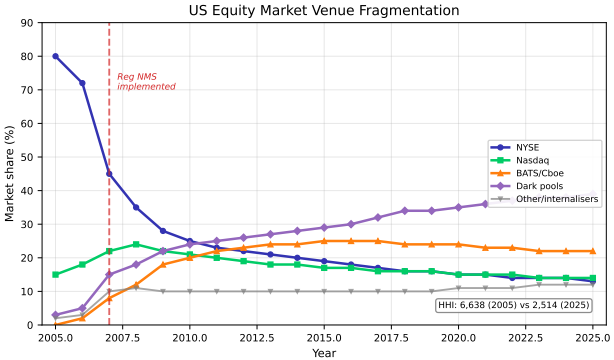
Dark pool – a trading venue that does not display orders before execution. Three types:

Type	Matching rule	Advantage	Risk
Midpoint crossing	Matches at NBBO midpoint	Zero spread cost; no information leakage from displayed quotes	Low fill probability; may reference stale NBBO
Block crossing	Minimum size threshold (e.g. 10K shares)	Natural institutional matching; minimal market impact	Very low fill rate; adverse selection if minimum size is gamed
Broker-dealer internal	Internaliser matches against own inventory	Fast execution; price improvement vs. NBBO	Conflict of interest; maker profits from spread

- **Why dark?** A displayed order to sell 1M shares would signal supply and move the price before execution.
- **Trade-off:** Dark pools reduce pre-trade transparency (less information leakage) at the cost of post-trade price discovery contribution.
- **Regulation:** SEC Reg ATS requires dark pools to report trades; MiFID II imposes double volume caps (DVC) on dark trading.

Dark pools exist because large institutional orders need protection from information leakage. The regulatory challenge is balancing this need against the public good of price transparency.

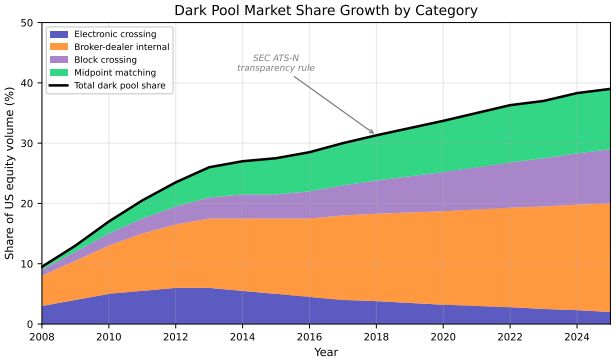
How Did US Equity Trading Go from Two Venues to Fifty?



- Before Reg NMS (2007): NYSE had ∼80% market share
- After Reg NMS: the Order Protection Rule forced venues to route to the best price, enabling competition
- NYSE's share collapsed from 80% to 13% in 20 years
- Dark pools grew from 3% to 39% of volume
- The HHI (market concentration index) fell from 6,600+ to 2,400 – a massive deconcentration
- **Debate:** Fragmentation improves competition (lower fees) but complicates best execution and creates systemic interconnectedness

Source: SEC market share data, 2005–2025. Reg NMS (2007) was the catalyst: by mandating order routing to the best price across venues, it inadvertently created the fragmented landscape we see today.

How Much of the Market Has Gone Dark?



- Total dark pool share has grown from ∼10% (2008) to ∼39% (2025)
- **Broker-dealer internal** (orange) is the largest and fastest-growing category – driven by retail order flow internalization
- **Midpoint matching** (green) is the second largest – institutional demand for zero-spread execution
- **Electronic crossing** networks are declining – replaced by newer venue types
- The SEC ATS-N transparency rule (2018) required dark pools to disclose operational details
- **EU DVC:** MiFID II caps dark trading at 4% per venue and 8% across all venues (per instrument)

Source: FINRA ATS transparency data. Nearly 40% of US equity volume now executes in the dark, raising ongoing debates about price discovery and market quality.

How Do You Measure Whether Fragmentation Helps or Hurts?

Herfindahl-Hirschman Index (HHI) for venue concentration:

$$\text{HHI} = \sum_{m=1}^{M} s_m^2 \times 10{,}000, \quad s_m = \text{market share of venue } m$$

Effective spread – the true cost paid by a trader:

$$\text{ES}_t = 2 \cdot D_t \cdot (P_t - m_t), \quad D_t = \begin{cases} +1 & \text{buyer-initiated} \\ -1 & \text{seller-initiated} \end{cases}$$

Empirical relationship (O'Hara & Ye, 2011):

- More fragmentation (lower HHI) $\Rightarrow$ **lower** effective spreads (more competition)
- But also: more fragmentation $\Rightarrow$ **lower** price discovery efficiency (information dispersed)
- **Worked example:** 5 venues with shares $[0.30, 0.25, 0.20, 0.15, 0.10]$
- $\text{HHI} = (0.09 + 0.0625 + 0.04 + 0.0225 + 0.01) \times 10{,}000 = 2{,}250$ (moderately concentrated)
- Compare: monopoly HHI $= 10{,}000$; perfectly fragmented (100 equal venues) HHI $= 100$
- **US equities 2025:** HHI $\approx 2{,}400$ – similar to the airline industry

O'Hara & Ye (2011) found that fragmentation reduces trading costs but complicates price discovery. The optimal level of fragmentation remains an open research question.

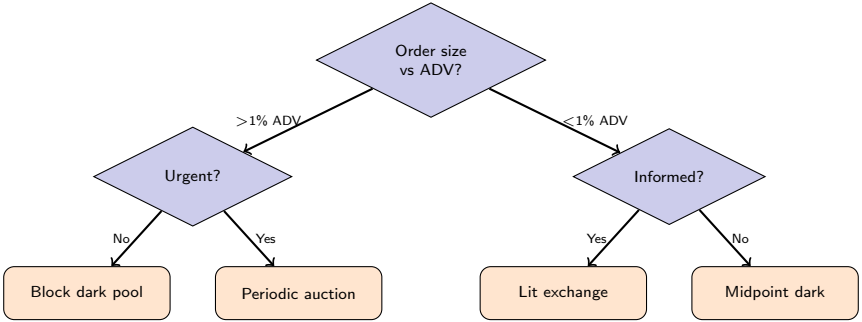
Can You Simulate Whether a Dark Pool Improves Execution?

```python
import numpy as np
def dark_vs_lit(qty, mid, spread_lit,
                fill_prob_dark=0.3,
                adv_sel_dark=0.0002):
    """Compare dark pool vs lit execution."""
    # Lit exchange: certain fill at ask
    lit_cost = qty * (mid + spread_lit/2)
    lit_bps = (spread_lit/2) / mid * 1e4
    # Dark pool: uncertain fill at mid
    n_trials = 10000
    dark_costs = []
    for _ in range(n_trials):
        if np.random.rand() < fill_prob_dark:
            # Filled at mid + adverse selection
            c = qty * (mid + adv_sel_dark)
        else:
            # Not filled -> go to lit (delay cost)
            delay = mid * 0.0003 # 3bps drift
            c = qty * (mid + spread_lit/2 + delay)
        dark_costs.append(c)
    dark_avg = np.mean(dark_costs)
    dark_bps = (dark_avg/qty - mid)/mid * 1e4
    return {"lit_bps": lit_bps,
            "dark_bps": dark_bps,
            "savings": lit_bps - dark_bps}
r = dark_vs_lit(10000, 50.0, 0.04)
for k,v in r.items(): print(f"{k}: {v:.2f}")
```

- **Lit exchange:** certain fill at the ask (half-spread cost)
- **Dark pool:** fill at midpoint with probability p (zero spread cost), but with adverse selection risk
- If not filled in dark: must go to lit market with a **delay cost** (price has drifted)
- The tradeoff: spread savings vs. non-fill risk $\times$ delay cost
- Dark pools are optimal when p is high and delay cost is low (patient orders)
- **Key result:** Dark routing saves $\sim$1–3 bps for large, patient institutional orders

The dark-vs-lit decision is a classic explore-exploit problem: dark pools offer better prices (midpoint) but uncertain fills. The breakeven fill probability depends on delay cost and spread savings.

How Should a Trader Choose Between Lit, Dark, and Periodic Auction?



- **Large + patient** → Block dark pool (minimize information leakage, accept low fill rate)
- **Large + urgent** → Periodic auction (aggregate demand, reduce adverse selection)
- **Small + informed** → Lit exchange (price priority, immediate execution)
- **Small + uninformed** → Midpoint dark pool (save the spread, no information to leak)
- Production SOR combines this logic with real-time fill probability estimates and fee optimization

Venue selection depends on two dimensions: order size (relative to ADV) and information content. Large uninformed orders benefit most from dark pools; small informed orders belong on lit exchanges.

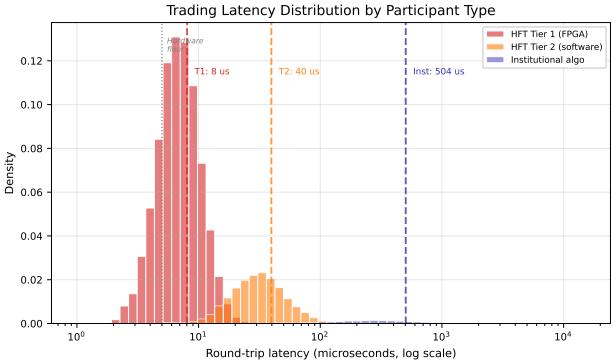
What Exactly Do High-Frequency Traders Do All Day?

Strategy	Mechanism	Edge source	Holding period	Social value
Market making	Quote bid/ask, earn spread	Speed to cancel stale quotes; inventory management	Seconds to minutes	Positive: provides liquidity
Latency arbitrage	Pick off stale quotes across venues	Speed advantage (co-location, FPGA, microwave)	Microseconds	Debated: tax on slow traders
Statistical arb	Exploit short-term mean reversion in correlated assets	Speed + signal processing	Seconds to hours	Positive: corrects mispricings
Momentum ignition	Submit aggressive orders to trigger stop-losses, profit from cascade	Illegal if intentional	Milliseconds	Negative: manipulative

- **Market making** is the largest HFT category by volume ($\sim$60% of HFT flow)
- **Latency arbitrage** generates $\sim$**$**5B/year globally (Aquilina, Budish & O'Neill, 2022)
- **Budish et al. (2015)** propose frequent batch auctions (every 100ms) to eliminate the latency arms race
- The IEX speed bump (350μs delay) is a partial solution: protects displayed quotes without eliminating continuous trading

HFT encompasses at least four distinct strategies with different risk profiles and social welfare implications. The regulatory challenge is to preserve market making while curbing latency arbitrage.

How Fast Is Fast in Today's Markets?



- **Tier 1 HFT** (red, FPGA): median $\sim 8\mu$s – approaching the hardware floor
- **Tier 2 HFT** (orange, software): median $\sim 40\mu$s – co-located but software-based
- **Institutional algo** (blue): median $\sim 500\mu$s – cloud/data center execution
- The gap between Tier 1 and institutional is $>60\times$ – this is the **latency tax**
- Log scale reveals the heavy right tail: even HFT has occasional slow messages (GC pauses, network congestion)
- **Speed of light:** NY-to-Chicago is 3.9ms; microwave towers cut this to 4.1ms one-way

Source: Simulated latency data calibrated to Aquilina et al. (2022) empirical measurements. The three-tier structure reflects real market participant categories.

How Much Is a Microsecond Worth in Dollar Terms?

Latency arbitrage occurs when a fast trader observes a price change on venue A before it propagates to venue B, picking off the stale quote:

$$\text{Rev}_{\text{daily}} = N_{\text{races}} \times P(\text{win}) \times E[|\Delta p|] \times \bar{Q}$$

Aquilina, Budish & O'Neill (2022) estimated on London Stock Exchange:

- $\sim$20% of HFT profits come from latency arbitrage ("races")
- Races occur at a rate of $\sim$1 per second for liquid stocks
- The winner earns $\sim$0.4 bps per race; the loser is picked off

Global scale estimate:

- US equity volume: $\sim$\$500B/day. If 0.1% is contestable via latency arb:
- Contestable flow: $500B \times 0.001 = \$500\text{M/day}$
- At 0.4 bps profit: $500\text{M} \times 0.00004 = \$20{,}000\text{/day/stock}$
- Across 3,000 liquid stocks: $\$60\text{M/day} = \$15\text{B/year}$ (upper bound)
- Aquilina et al. estimate $\sim$\$5B/year globally (more conservative)
- **Key question:** Is this \$5B a pure tax on slower participants, or does it fund socially useful liquidity provision?

Latency arbitrage is a pure speed race. Aquilina et al. (2022) provide the first rigorous measurement: roughly \$5B/year globally, or about 0.4 bps per contestable trade.

Could Batch Auctions Eliminate the Socially Wasteful Speed Race?

Budish, Cramton & Shim (2015) propose replacing continuous trading with **frequent batch auctions**:
Mechanism: Instead of continuous price-time priority, aggregate orders over discrete intervals (e.g., 100ms) and clear at a single price:

$$p^* = \arg\max_p \ \min\big(D(p), S(p)\big)$$

where $D(p)$ = cumulative demand at price $\leq p$, $S(p)$ = cumulative supply at price $\geq p$.
Why this helps:

- **Eliminates latency arbitrage:** All orders within the batch window are treated equally – speed advantage within the window is worthless
- **Preserves price discovery:** Prices still update every 100ms – fast enough for virtually all economic purposes
- **Deepens the book:** Traders compete on *price*, not *speed* – the order book becomes thicker at the best level
- **Reduces arms race spending:** No incentive to invest in sub-millisecond infrastructure

Counter-argument: Continuous markets provide immediate execution for urgent orders. Batch intervals create execution delay.

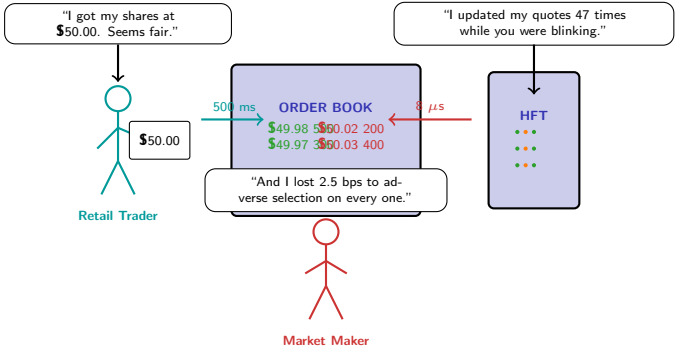
Budish et al. (2015) argue that the continuous limit order book is "flawed by design" because it rewards speed over price. Frequent batch auctions fix this while preserving sub-second price discovery.

Has HFT Made Markets Better or Worse for Everyone Else?

Dimension	Evidence: HFT helps	Evidence: HFT hurts
Spreads	Tighter quoted spreads (Hendershott et al., 2011)	Effective spreads may not improve proportionally
Depth	More quotes at best bid/ask	"Phantom liquidity" – quotes cancel before execution
Volatility	Lower intraday volatility in normal times	Amplifies volatility in stress (Flash Crash 2010)
Price discovery	Faster incorporation of information	Mostly from public info; private info still via institutions
Fairness	Lower costs for retail via PFOF	Latency tax on institutional investors
Systemic risk	No systemic event directly caused by HFT	Interconnectedness creates correlated failure modes

- **Net assessment:** HFT market making is net positive for market quality. Latency arbitrage is net negative (wasteful arms race).
- **The challenge:** They are performed by the same firms, making regulation difficult.
- **Best proposal:** Frequent batch auctions (Budish) eliminate latency arb while preserving MM incentives.

The academic consensus: HFT market making improves liquidity metrics; latency arbitrage is a socially wasteful tax. The two activities are entangled, making surgical regulation difficult.



The price looks simple. The plumbing underneath is anything but.

References and Further Reading

1. Kyle, A. (1985). Continuous auctions and insider trading. *Econometrica*, 53(6), 1315–1335.
2. Glosten, L. & Milgrom, P. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100.
3. Huang, R. & Stoll, H. (1997). The components of the bid-ask spread. *Review of Financial Studies*, 10(4), 995–1034.
4. Cont, R., Kukanov, A. & Stoikov, S. (2014). The price impact of order book events. *Journal of Financial Econometrics*, 12(1), 47–88.
5. Avellaneda, M. & Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3), 217–224.
6. O'Hara, M. & Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics*, 100(3), 459–474.
7. Budish, E., Cramton, P. & Shim, J. (2015). The high-frequency trading arms race. *Quarterly Journal of Economics*, 130(4), 1547–1621.
8. Aquilina, M., Budish, E. & O'Neill, P. (2022). Quantifying the high-frequency trading "arms race". *Quarterly Journal of Economics*, 137(1), 493–564.

Eight foundational references: information asymmetry (1–2), spread decomposition (3), order book dynamics (4), market making (5), fragmentation (6), and HFT debate (7–8).