What Is in High-Frequency Price Pressure?*

Bart Zhou Yueshen†

August 15, 2020

* This paper benefits tremendously from discussions with Sergei Glebkin, Naveen Gondhi, Shiyang Huang, Vincent van Kervel, John Kuong, and Marcin Zamojski. There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

† INSEAD; b@yueshen.me; 1 Ayer Rajah Avenue, Singapore 138676.

What Is in High-Frequency Price Pressure?

Abstract

Trades' transitory price impact is found to be negative in high-frequency data: a market buy (sell) order on average pushes contemporaneous price pressure down (up). This counterintuitive yet robust pattern is explained through a model where liquidity providers do not always quote competitively. Lacking competitive quote updates, prices fail to immediately reflect new information, forcing price pressure to move against and offset permanent price changes. Other novel yet counterintuitive model predictions also find support in data: A stock's price pressure persistence and its return volatility are both *lower* when its liquidity providers are less competitive.

Keywords: price pressure, price impact, high-frequency data

There are no competing financial interests that might be perceived to influence the analysis, the discussion, and/or the results of this article.

1 Introduction

Asset prices fluctuate around their fundamentals. Such fluctuations arise because, among many other reasons, liquidity providers "press" prices up for buys and down for sells to make profit, to cover operation and funding costs, to compensate for the risk of bearing nonzero inventories, and to induce future trades to offset undesirable holdings. Properties of such "price pressures" are of great interest both academically and in practice for they reflect important market qualities, like market efficiency, liquidity, volatility, trading costs, and competition.

This paper examines price pressures in high-frequency trading data. A standard decomposition of intraday prices finds that changes in price pressures are *negatively* correlated with contemporaneous liquidity demand, i.e., order flows. In other words, there is negative (contemporaneous) transitory price impact—market or marketable buy (sell) orders on average are followed by immediate reductions (increases) in price pressures. Section 2 documents the prevalence and robustness of this pattern. The estimates of the said correlation are negative for every one of 400 randomly selected stocks on every trading day in 2018, under a variety of model specifications.

Standard theories tend to suggest the opposite: Liquidity providers want to tilt prices *in the same direction* as the liquidity demand (the order flow), even absent of information, simply to cover the costs of bearing the inventories.[1] That is, theoretically and intuitively, price pressures—non-fundamental price fluctuations—should be positively correlated with contemporaneous trades. The contrast with the finding from high-frequency data suggests that there is some other mechanism at work. What might it be—what is in high-frequency price pressure?

Section 3.1 provides a tentative explanation through a stylized limit order book model to study liquidity providers' optimal quoting. The key model friction is that in a very short period of time, liquidity providers might not be able to react to trades timely—they might be inactive. In such cases,

[1] Such inventory costs can arise from risk aversion, as in Grossman and Stiglitz (1980), Hellwig (1980), Grossman and Miller (1988); or from market power, as in Kyle (1989). More explicitly, liquidity providers' inventory management problem has been studied in Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), Madhavan and Smidt (1993), and more recently Hendershott and Menkveld (2014). Bruche and Kuong (2019) show that a moral hazard problem between market makers and their financiers can also result in costly inventory holdings.

the price does not immediately adjust after a possibly informed trade; that is, the observed total (permanent plus transitory) price impact is zero. Yet, by Bayesian updating, this trade still yields a permanent price impact, moving the fundamental in the trade's direction. The price pressure then must offset the change in the fundamental, for otherwise the observed price would not remain the same. The resulting transitory price impact—the change in the price pressure—is the negative of the permanent price impact and, hence, in the opposite direction of the contemporaneous trade.[2]

When instead liquidity providers are active, quoting competitively, the new price shall immediately incorporate both the trade's information and liquidity providers' required compensation (due, e.g., to inventories), just like what standard theories suggest. Combining these two scenarios probabilistically, the equilibrium change in the price pressure becomes a weighted average between liquidity providers' required compensation and the *negative* permanent price impact. The contemporaneous transitory price impact therefore can indeed be negative, as found in the data.

One might wonder why liquidity providers would be "inactive" in a modern limit order market, where the posting and modification of limit orders are performed by fast speed computers. Effectively monitoring market conditions in real time is very challenging, because advancement in speed technology is a double-edged sword. On the one hand, traders can process information faster and more efficiently. On the other, they generate more market events (trades, order submission and revision, etc.) that others need to process. The net effect is not clear, especially because the latter effect is amplified by the number of traders gaining speed. For example, if each of n traders is now able to submit one more order, then every trader in the market needs to parse $(n-1)$—not just one—more messages. Thus, the more traders gaining speed, the heavier is the burden of tracking the market for everyone. Further, product complexity and market fragmentation also add to the cost of monitoring. O'Hara (2015) cites Berman (2014) for an example of strenuous market

[2] As a numerical example, suppose the current midquote is at \$10.00. A market buy order moves the price up permanently by ¢1. With active liquidity providers, the midquote immediately updates to \$10.01. When they are inactive, however, the price does not change, not at least immediately, and the total (contemporaneous) price impact in this case is zero. Since the permanent price impact is ¢1, the transitory price impact must be -¢1, yielding a negative correlation with the buy order flow.

making across 14 exchange-traded products linked to gold, involving 91 distinct pairs of arbitrage relationships that must be monitored continuously in time. It is onerous to effectively monitor and react to, in real time, all events like quotes, trades, and news flashing on all these manifold marketplaces trading the same or related assets. Therefore, with capacity constraint, inaction of liquidity providers in a short period of time is likely.

Further, there can be *endogenous* inaction. This happens when there is lack of (quote) competition among liquidity providers. If a liquidity provider's pre-trade quotes become more profitable post-trade, she has no incentive to modify these quotes, even though they might be indicating the wrong, "stale" fundamental. Lacking sufficient competition, such "strategic" inaction might be optimal, serving as another reason, other than limitied monitoring capacity, for the negative (contemporaneous) transitory price impact.

Section 3.2 builds on the above intuition and develops a flexible structural model. It shows that the price pressure can in general be written as a weighted average between (i) liquidity providers' private value due, e.g., to inventory considerations and (ii) a price distortion component due to stochastic inaction. The component (i) is likely to be relatively persistent, because it takes time to mean revert a nonzero inventory position. Instead, component (ii) is likely to be of extremely low persistence, because there are "snipers" who aim at such price distortions and compete to eliminate them quickly (Budish, Cramton, and Shim, 2015). This general structural model thus yields a novel prediction: When liquidity providers are more active or competitive, component (i) will dominate, making price pressure *more persistent*.

A second novel prediction follows the general structural model. When liquidity providers are more active or competitive, their private value component (i) reflects more often in the observed price. Hence, the fluctuation in (i) also manifests more often in return volatility. That is, the model predicts *higher volatility* with more active or competitive liquidity providers.

These two novel predictions challenge the conventional wisdom: Should not more competitive liquidity providers mean-revert price pressure to zero more quickly, thus lowering price pressure

persistence? Should they not quote more competitively, thus reducing noise and also volatility? This conventional line of reasoning has not taken into account the pricing incentive of imperfectly competitive liquidity providers. A monopolistic liquidity provider has little incentive to keep quotes constantly in sync with the fundamental. In fact, she would like to keep the quotes stale, so long as their executions are profitable. In the extreme case, when the monopolistic prices do not move, there is zero persistence and no volatility. Therefore, when competing liquidity providers become more active, the prices change more often (to reflect the fundamental and the private values), raising both the persistence of price pressure and the realized volatility.

Section 4 takes the two predictions to the data and finds support for both. Given liquidity providers' limited capacity to maintain competitive quotes (Corwin and Coughenour, 2008), news events about *other* stocks are used as (negative) shocks to a stock's liquidity provider activeness and competition. In other words, liquidity providers with limited capacity can be "distracted" from covering a no-news stock by other stocks' news. When this happens, their activeness and competition in the distracted stock will drop, manifesting in the reduction in price pressure persistence and in realized volatility. The data agree: When a stock itself has no news, every 1,000 entries of other stocks' news reduce the stock's price pressure persistence by about a quarter percentage points and reduce its realized volatility by about three basis points.

Further evidence is found in support of the channel of liquidity providers' limited capacity. The reductions in price pressure persistence and in realized volatility are most significant when the distracted stock is medium or small, doubling the effects' magnitudes in large stocks. This is consistent with the limited capacity argument. Even with many other stocks' news, large stocks have a lot more activity than medium and small stocks. Therefore, liquidity providers are more likely to keep their capacity in large stocks than in small ones, given the same distraction. In addition, the reductions almost exclusively arise from news about stocks in other industries. That is, when the news is about a different company but in the same industry, liquidity providers do not get distracted. Indeed, they should not, because such same-industry news might also have direct or

indirect effects on all stocks in this industry.

Overall, in answering what is in high-frequency price pressure, the empirical evidence echoes the tentative explanation given by the theory: (the lack of) liquidity providers' activeness or competition plays an important role. Such inaction or lack of competition helps explain the negative contemporaneous transitory price impact. It also reveals why more active or competitive liquidity providers do *not* make price pressure mean-revert more quickly or reduce volatility. The analysis contributes to the better understanding of high-frequency price-trade dynamics.

The price pressure literature and contribution

Various event studies have examined the existence and the magnitudes of price pressure under different settings. They focus on the price dynamics around surges of demand due to non-fundamental reasons like block trades (Kraus and Stoll, 1972; Scholes, 1972), changes of index constituents (Shleifer, 1986; Harris and Gurel, 1986), second-hand information in "Dartboard" (Barber and Loeffler, 1993), merges and acquisitions (Mitchell, Pulvino, and Stafford, 2004), tax (Gibson, Safieddine, and Titman, 2000; Jin, 2006), fund flows (Coval and Stafford, 2007; Ben-Rephael, Kandel, and Wohl, 2011), and clientile demand (Greenwood and Vayanos, 2010, 2014).

A second strand of the literature, without focusing on surges of non-fundamental liquidity demand, makes use of liquidity providers' inventory data. Typically, the observed asset price is decomposed into a random walk (fundamental) and a stationary component (price pressure). The latter is then associated with the inventory changes of (selected) liquidity providers. Examples include Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), Madhavan and Smidt (1991, 1993), Hasbrouck and Sofianos (1993), Madhavan and Sofianos (1998), Hendershott and Seasholes (2007), Menkveld (2013), and Hendershott and Menkveld (2014). Most of these studies use low-frequency inventory data (daily), with the exception of Menkveld (2013), who studies one high-frequency trader's intraday cross-market making.

Data of liquidity providers' inventories are usually proprietary. A third strand of the literature,

therefore, circumvents this data issue by distinguishing liquidity demand and supply according to trades' aggressive sides, using methods of Lee and Ready (1991) and the alike. The signed order flows—the liquidity demand—can then be associated with the random walk component in the price to study trade informativeness, toxicity, or adverse-selection risk; or be associated with the price pressure component to study liquidity providers' order processing and inventory management costs. Examples include Glosten and Harris (1988), Brennan and Subrahmanyam (1996), Sadka (2006), Brogaard, Hendershott, and Riordan (2014), and Chordia, Green, and Kottimukkalur (2018), among many others.

The main focus of this third strand of the literature has largely been on how order flows relate to the random walk component in price—trades' informativeness, illiquidity due to adverse-selection, and price discovery. In doing so, some studies have also reported negative (contemporaneous) transitory price impacts, though often in passing, as in Sadka (2006), Brogaard, Hendershott, and Riordan (2014), and Chordia, Green, and Kottimukkalur (2018). Following the same methodology, this paper instead focuses on price pressure and contributes to a deeper understanding of it in the high-frequency data. In the same vein, the theory developed in Section 3 also intentionally focuses more on how liquidity providers' inventory cost reflects in price and less on the information and learning aspects.

2 Stylized features of high-frequency price pressure

This section examines some stylized features of high-frequency price pressure, through the lens of a state space model commonly used in the literature.

A structural framework. Consider a data set of a security's intraday trading. Trades are sequentially indexed by $k \in \{1, 2, \ldots\}$, each with timestamp t_k and signed size x_k. The series $\{x_k\}$ will also be referred to as the order flow, modeled as a stationary autoregressive process of the form $(1 - A(L))x_k = x_k^*$, where $A(L)$ is some lag polynomial guaranteeing the stationarity, and $\{x_k^*\}$ is a

white noise process capturing order flow innovations.[3]

Write p_k as the log midquote observed just *before* the $(k+1)$-th trade. This timing convention allows the price p_k to be understood as reflecting all information up to and including the k-th trade. The price p_k is decomposed into a random walk m_k and a stationary residual s_k:

$$
\begin{aligned}
(1) \qquad \text{observed price:} \qquad & p_k = m_k + s_k; \\
\text{hidden efficient price:} \qquad & m_k = m_{k-1} + \lambda x_k^* + \mu_k; \\
\text{hidden price pressure:} \qquad & (1 - \phi(L))s_k = (\psi_0 + \psi(L))x_k + \nu_k.
\end{aligned}
$$

The innovation terms $\{\mu_k\}$ and $\{\nu_k\}$ are two white noises, capturing the respective changes in m_k and in s_k that are unrelated to trades $\{x_k\}$. That is, $\text{cov}[x_k, \mu_k] = \text{cov}[x_k, \nu_k] = 0$. The random walk m_k is assumed to be tracking the fundamental value of the asset, hence the name "efficient price." It has persistent innovations $\{\lambda x_k^* + \mu_k\}$. Instead, the stationary s_k mean-reverts to zero, with its innovations diminishing according to the autoregressive structure $\phi(L)$, hence the name "price pressure." The order flow $\{x_k\}$ affects both $\{m_k\}$ and $\{s_k\}$:[4]

- The "permanent price impact" is reflected by λ. Note that only the surprise x_k^* but not the full x_k appears because, for m_k to be consistent with the interpretation of efficient price, only the unexpected trading can affect it—the expected part would have been reflected already.
- The "transitory price impact" is reflected in the structure $\psi_0 + \psi(L)$. The parameter ψ_0

[3] In this paper, a lag polynomial $A(L)$ always starts with the first order; i.e., $A(L) = A_1 L + A_2 L^2 + ...$ with $A(0) = 0$. Therefore, an autoregressive moving-average (ARMA) process $\{y_k\}$ can be written as $(1 - A(L))y_k = (1 + B(L))\varepsilon_k$, where ε_k is some white noise, $A(L)$ captures the autoregressive structure, and $B(L)$ captures the moving-average structure.

[4] Although modeled separately as an autoregression of itself, $(1 - A(L))x_k = x_k^*$, the order flow is actually allowed to endogenously respond to m_k and s_k. To see this, consider a general linear structure of $x_k = a(L)x_k + b(L)\Delta m_k + c(L)s_k + x_k^*$, where the lag polynomials starting with lag one. That is, x_k also responds to *lagged* efficient price changes $\{\Delta m_{k-1}, \Delta m_{k-2}, ...\}$ and *lagged* price pressures $\{s_{k-1}, s_{k-2}, ...\}$. Together with (1), this forms a familiar vector-autoregression (VAR) structure. Using Δm_k and s_k from (1), hence, $x_k = \left(a(L) + \frac{(\psi_0 + \psi(L))c(L)}{1 - \phi(L)}\right)x_k + \lambda b(L)x_k^* + \left(b(L)\mu_k + \frac{c(L)}{1 - \phi(L)})\nu_k\right) + x_k^*$, which is a combination of lagged x_k, lagged x_k^*, lagged uncorrelated noises μ_k and ν_k, and the innovation x_k^*. Wold representation theorem ensures that there is an equivalent autoregressive form of $(1 - A(L))x_k = x_k^*$ for some $A(L)$. That is, within linear frameworks, it is without loss of generality to model the order flow $\{x_k\}$ separately. The direct effects of (past) prices on x_k, i.e., the estimates of $a(L)$ and $b(L)$, are implicit in $A(L)$. The key restriction is that in the system of $\{x_k, \Delta m_k, s_k\}$, only x_k *contemporaneously* affects the other two states, not the other way—a standard assumption in VAR models of price-trade dynamics (see, e.g., Chapter 9 of Hasbrouck, 2007).

captures the *contemporaneous* effect, i.e., how the order flow x_k affects s_k *immediately*. The lag polynomial $\psi(L)$ describes how s_k might react to previous trades.

Estimation. There are three sets of structural parameters: (i) the contemporaneous price impacts—both permanent and transitory—$\{\lambda, \psi_0\}$; (ii) the lagged transitory price impacts $\psi(L)$; and (iii) the persistence of the price pressure $\phi(L)$. For parsimony considerations, the polynomials $\phi(L)$ and $\psi(L)$ are typically chosen to be of low orders. For example, in Sadka (2006), $\phi(L) = 0$, i.e., the price pressure is serially correlated only via $\{x_k\}$. In Brogaard, Hendershott, and Riordan (2014) and Chordia, Green, and Kottimukkalur (2018), $\phi(L) = \phi_1 L$ and $\psi(L) = \psi_0$. To estimate these parameters, one typically first obtains the order flow innovations $\{x_k^*\}$ by estimating a (sufficiently long) autoregression of $x_k \sim A(L)x_k + x_k^*$. Then, taking $\{x_k^*\}$ as given, the parameters can be estimated by maximum likelihood (ML) under some assumptions about the joint distribution of $\{\mu_k, \nu_k\}$. An alternative estimation technique, using generalized method of moments (GMM), is provided in Appendix A. The advantage of the GMM approach is that it only makes use of the structure (1), without additional distribution assumptions on $\{\mu_k, \nu_k\}$ (apart from that they are white noises uncorrelated with the order flow $\{x_k\}$).

Data. The sample is a panel of 400 stocks randomly selected from the S&P 1500 index over the 251 trading days in 2018. All trades and prices during trading hours (9:30 to 16:00, Eastern Standard Time) are collected from Daily Trade and Quote (DTAQ) database. To facilitate comparison across stocks and days, order flows are measured in US$10,000 and price changes in basis points (bps). For a stock-day to be a valid sample, it is required to have a minimum of 500 trades and a minimum of 10 price changes.

Results. The GMM approach is applied to each stock-day in the sample and the estimation results are presented in Table 1. Panel (a) reports the results for the most parsimonious specification of $\phi(L) = \phi_1 L$ and $\psi(L) = \psi_0$. On average, every surprise order flow of $10,000 moves the efficient price permanently by about $\lambda \approx 4.6$ bps. The contemporaneous transitory price impact is *negative* at $\psi_0 \approx -3.8$ bps per $10,000, and it decays at a rate of $\phi_1 \approx 60\%$ from one trade to the next. The

				Percentiles						
	Unit	Mean	Std Dev	1%	5%	25%	50%	75%	95%	99%
(a) Price pressure specified as: $(1-\phi_1 L)s_k = \psi_0 x_k + v_k$										
λ	bps/$10,000	4.64	7.84	0.06	0.24	0.77	2.01	5.04	17.52	44.41
ψ_0	bps/$10,000	-3.79	6.56	-37.29	-14.40	-4.05	-1.61	-0.62	-0.20	-0.05
ϕ_1	%	60.36	9.76	35.16	44.71	54.07	60.25	66.58	77.03	83.72
Convergence: 98,594 stock-days out of 98,955, or 99.6%										
(b) Price pressure specified as: $(1-\phi_1 L-\phi_2 L^2)s_k = \psi_0 x_k + v_k$										
λ	bps/$10,000	5.01	8.69	0.06	0.26	0.82	2.15	5.37	18.90	49.20
ψ_0	bps/$10,000	-4.15	7.39	-41.75	-15.76	-4.34	-1.73	-0.66	-0.22	-0.05
ϕ_1	%	59.17	9.16	36.54	44.97	53.29	58.76	64.63	75.41	83.11
ϕ_2	%	4.64	4.89	-10.13	-2.41	2.61	4.60	6.68	10.00	19.18
Convergence: 98,691 stock-days out of 98,955, or 99.7%										
(c) Price pressure specified as: $(1-\phi_1 L)s_k = (\psi_0+\psi_1 L)x_k + v_k$										
λ	bps/$10,000	4.88	8.33	0.04	0.25	0.81	2.12	5.30	18.30	47.10
ψ_0	bps/$10,000	-4.04	7.10	-40.14	-15.28	-4.30	-1.71	-0.66	-0.20	-0.03
ψ_1	bps/$10,000	0.28	0.88	-1.69	-0.20	0.03	0.09	0.29	1.41	4.42
ϕ_1	%	66.17	13.17	8.95	47.51	61.53	67.88	73.46	81.69	87.76
Convergence: 97,262 stock-days out of 98,955, or 98.3%										
(d) Price pressure specified as: $(1-\phi_1 L)s_k = \psi_F \text{sign}[x_k] + \psi_0 x_k + v_k$										
Efficient price specified as: $\Delta m_k = \lambda_F \cdot (\text{sign}[x_k] - \mathbb{E}_{k-1}[\text{sign}[x_k]]) + \lambda x_k^* + \mu_k$										
λ_F	bps/$10,000	0.09	0.12	-0.03	-0.00	0.02	0.05	0.12	0.34	0.59
λ	bps/$10,000	3.57	5.89	0.04	0.19	0.61	1.57	3.96	13.54	33.19
ψ_F	bps/$10,000	-0.80	0.57	-2.67	-1.93	-1.11	-0.64	-0.36	-0.17	-0.10
ψ_0	bps/$10,000	-0.44	1.56	-8.37	-2.53	-0.37	-0.06	0.01	0.43	2.25
ϕ_1	%	45.10	8.89	23.86	30.00	39.02	44.98	51.04	59.94	66.70
Convergence: 98,762 stock-days out of 98,955, or 99.8%										

Table 1: Estimated structural parameters across stock-days. This table reports the summary statistics of the estimated parameters of the structural model (1). In Panels(a)-(c), the efficient price is modeled the same as $m_k = m_{k-1} + \lambda x_k^* + \mu_k$, while the lags of the polynomials $\phi(L)$ and $\psi(L)$ vary for the price pressure s_k in the three panels. Panel (d) adds to Panel (a) the fixed-effects of order flows as in Sadka (2006).

actual available stock-day observations are slightly short of the possible maximum of 400×251 because some stock-days have too few trades or price changes and because the estimation algorithm does not always converge. The overall convergence rate is very high, above 99%.

The negative ψ_0 seems very counterintuitive. Consider the quoting of a liquidity provider, whose inventory is currently at some optimal level. After a market buy order, the inventory drops below the optimal level and the deviation is costly to the liquidity provider (due to, e.g., inventory costs or risk-aversion). Therefore, the "correct" reaction should be to tilt her quotes *up*, attracting future market sell orders while deterring future buys. That is, a positive (negative) price pressure should follow a market buy (sell), suggesting $\psi_0 > 0$. The extant theories of trading share such a prediction. Examples include Grossman and Stiglitz (1980), Hellwig (1980), Grossman and Miller (1988), Kyle (1989), Glosten (1994), Sandås (2001), Vayanos and Wang (2012), etc. The underlying intuition is that the cost for liquidity providers to take the order flow requires compensation and such compensation is always opposite to the direction of the order flow. Dealers' inventory management problem has been studied in Amihud and Mendelson (1980), Ho and Stoll (1981, 1983), and more recently Hendershott and Menkveld (2014).

What the negative ψ_0 is not. The negative ψ_0 should not be interpreted as order flows trading against current price pressure. Such an interpretation would imply a *lagged* negative correlation between s_k and x_{k+1}, according to the timing convention used here. Further, even a trade is to correct the price pressure, its *contemporaneous* transitory price impact should still be positive: For example, if a buy order flow $x_{k+1} > 0$ successfully "corrects" a negative price pressure $s_k < 0$, it follows that $s_{k+1} > s_k$ or, on average, $\text{cov}[x_{k+1}, \Delta s_{k+1}] > 0$, hence also $\psi_0 > 0$.

It is possible that, at times, some liquidity providers use market(able) orders to aggressively rebalance their inventories. Such aggressive trades, however, cannot explain the negative ψ_0. This is because the limit order trader executing against such a market order will suffer from the inventory cost, and then the same intuition as before applies, leading to $\psi_0 > 0$. In other words, whenever a liquidity provider turns aggressive with market orders to rebalance her inventory, she effectively

becomes a liquidity demander.

Robustness. It is worth emphasizing the robustness of the finding of $\psi_0 < 0$. It can be seen from Panel (a) that even the 99% quantile of ψ_0 estimate is still negative. For the avoidance of doubt, no restriction on the sign or the magnitude of ψ_0 (or of λ) is imposed in the numerical procedure. The only constraint is $-1 < \phi_1 < 1$ to ensure the stationarity of the price pressure. In Panels (b) and (c), two additional specifications of s_k are considered, varying the lags in $\phi(L)$ and in $\psi(L)$. Regardless of the specification, $\psi_0 < 0$ remains robust across the board. Panel (d) adds order flows' sign-dependent fixed effects as in Glosten and Harris (1988) and in Sadka (2006). While about a quarter of the stock-days see $\psi_0 > 0$ under this specification, the fixed effect of sign$[x_k]$, captured by the coefficient ψ_F is negative across the sample.

The robustness of $\psi_0 < 0$ is also collaborated by the extant literature employing similar empirical frameworks. For example, negative estimates of transitory price impact can be found in Table 1 of Sadka (2006), Tables 2-4 of Brogaard, Hendershott, and Riordan (2014), and Table 9 of Chordia, Green, and Kottimukkalur (2018). The latter two also provide evidence of such negative estimates also when the data are aggregated by clock time intervals, which is also found by the current paper but omitted for brevity. The finding is unlikely driven by order flow signing accuracy issues as Brogaard, Hendershott, and Riordan (2014) report the same using NASDAQ HFT data with exact buy-sell indicators. (These papers do not focus on the transitory price impact in price pressure but more on the permanent price impact or the total price impact.)

What is in high-frequency price pressure? The empirical evidence poses a puzzle: Why does price pressure appear to move against order flow? Section 3 develops an equilibrium model to explain the phenomenon. The model also yields additional predictions regarding the persistence of price pressure $\phi(L)$ and the return volatility. These novel predictions are then taken into data in Section 4, which finds support for the theory.

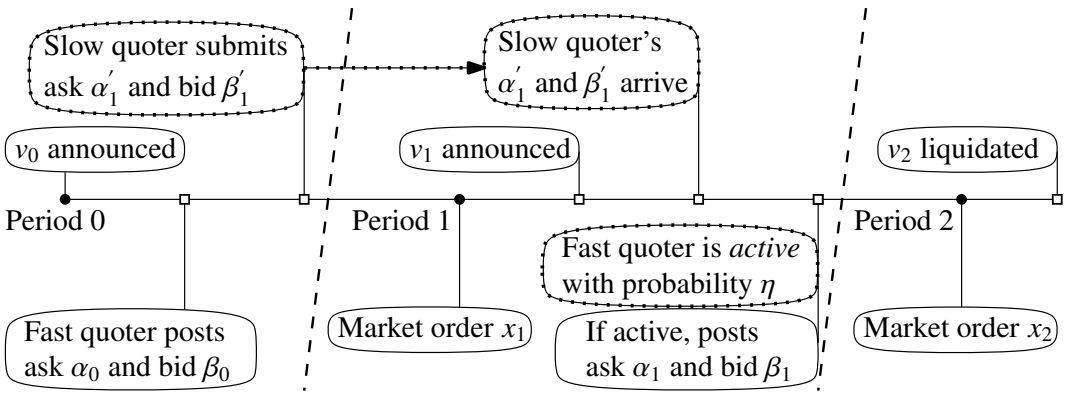

Figure 1: Timeline of the two-trade game. There are thee periods, $k \in \{0, 1, 2\}$. Periods 1 and 2 are marked by the arrival of the market orders, x_1 and x_2, which execute against the prevailing best quotes. The asset's fundamental value v_k evolves over time and liquidates at v_2 at the end of the game. There are two types of competitive quoters differing in speed. The fast quoter posts limit orders in period 0 but is only able to do so—being "active"—in period 1 with probability η. The slow quoter submits limit orders period 0 but these orders only arrive in period 1, hence "slow."

3 Model

This section sheds light on the above stylized facts through the lens of an equilibrium model. Section 3.1 studies a minimalistic version to provide an explanation to the negative contemporaneous transitory price impact. A general structural model is laid down in Section 3.2, yiedling novel empirical predictions.

3.1 A two-trade equilibrium model

3.1.1 Model setup

The model is an extension of Glosten and Milgrom (1985). Figure 1 gives an overview. There are two market orders x_k, $k \in \{1, 2\}$, which cut the timeline into three periods of $\{0, 1, 2\}$. The focus lies in period 1, where the equilibrium midquote will be shown to follow the same structure as in the empirical framework (1).[5] Two key model elements are highlighted in the dotted boxes: (i)

[5] It is worth emphasizing that (at least) two trades are needed to speak to the structural model (1). Without the late order flow x_2, there is no reason for anyone to quote at $t = 1$. The early order flow x_1 needs to be associated the $t = 1$ quotes to study price impacts. Then the $t = 0$ quotes must also be endogenously determined to accommodate x_1.

the slow quoter's orders arrive one period late; and (ii) the fast quoter is active (able to quote) in period 1 only with probability η. These two features enrich x_1's contemporaneous transitory price impact ψ_0, making it possibly negative. The model details are described below.

The risky asset. A risky asset is traded in a limit order market. Each unit of the asset will pay off v_2 units of the numéraire consumption good at the end of period 2. The evolution of $\{v_k\}$—the fundamental value—follows a random walk:

$$v_{t+1} = v_t + \Delta v_{t+1} \text{ for } t \in \{0, 1\},$$

where the innovation is $\Delta v_t \in \{\pm\sigma\}$ equally likely. The parameter σ captures the fundamental volatility of the asset. The unconditional expectation $v_0 := \mathbb{E}[v_2]$ is publicly announced at the beginning of the game. Similarly, v_1 is publicly announced right after the first trade x_1. Below it will be shown that such announcements match the non-trade component μ_k in the structural model.

The market orders. A liquidity demander arrives at the beginning of each period $k \in \{1, 2\}$, before v_k is announced. He demands liquidity by submitting a market order for unmodeled urgency reasons. He values the asset as the sum of the common payoff v_2 and a private value $\kappa\sigma u_k$, where κ (> 0) measures *the magnitude* of the private value as a multiple of the volatility σ; and

$$u_k = I_k \frac{\Delta v_k}{\sigma} + (1 - I_k)(-1)^{\varepsilon_k} \tag{2}$$

determines *the sign* of the private value. Specifically, I_t and ε_t are independent Bernoulli draws with success rates $\mathbb{E}[I_t] = \tau \in (0, 1)$ and $\mathbb{E}[\varepsilon_t] = \frac{1}{2}$, respectively. In words, the expression (2) says that with probability τ, the private value $\kappa\sigma u_k = \kappa\Delta v_k$ is driven by information;[6] or, with probability $(1 - \tau)$, driven by non-fundamental reasons summarized in $\kappa\sigma \cdot (-1)^{\varepsilon_k}$. The random variables $\{\Delta v_k, I_k, \varepsilon_k\}$ are pair-wise independent and i.i.d. over time.

Based on his valuation $\mathbb{E}[v_2 + \kappa\sigma u_k | v_{k-1}, u_k]$ and the prevailing bid and ask quotes (discussed

[6] The liquidity demander's private value can be correlated with the fundamental for various reasons. For example, he might have other income correlated with the risky asset. He might have a leveraged position in the asset and the private signal thus amplifies the expected profit or the loss.

below), the liquidity demander optimally chooses between a market buy order $x_k = 1$ or a sell $x_k = -1$. That is, the market orders are restricted to be of one unit.

The limit orders. There is a fast and a slow liquidity provider, or "quoter" for simplicity. They represent many competitive limit order traders of respective types, so that they post limit orders at prices that make them just indifferent between trading or not. Section 3.1.4 discusses the competitiveness in more detail.

- The "slow" (representative) quoter is slow in that it takes time for her submitted orders to appear in the limit order book. Specifically, in this two-trade model, she submits ask $\alpha_1^{'}$ and bid $\beta_1^{'}$ at $k = 0$, but these orders are added to the order book *after* the market order x_1.
- The "fast" (representative) quoter can post ask α_k and bid β_k *before* the next market order x_{k+1}. She can do so, however, only when she is *active*, which is denoted by $F_k = 1$ (and $F_k = 0$ for being inactive). In this two-trade model, it is assumed that she is always active at $k = 0$, i.e., $F_0 = 1$ deterministically, for otherwise there will be no quotes to accommodate x_1. Instead, F_1 is a Bernoulli draw, independent of all other random variables, with $\mathbb{E}[F_1] = \eta \in [0, 1]$.

Both the fast and the slow quoters bear quadratic inventory costs. Thus, holding y units of the risky asset, the fast quoter receives a total terminal payoff of $v_2 y - \frac{\gamma}{2} y^2$ with marginal cost γ (> 0). Similarly, holding $y^{'}$ units, the slow quoter receives $v_2 y^{'} - \frac{\gamma^{'}}{2} y^{'2}$ with marginal cost $\gamma^{'}$ (> 0). The initial inventories of both quoters are normalized to zero, i.e., $y_0 = y_0^{'} = 0$.

As a standard simplifying assumption (e.g., Foucault, 1999), a posted limit order lives only for one period. That is, every period-k limit order immediately expires if not executed by x_{k+1}.

Equilibrium. Three sets of endogenous variables characterize the equilibrium: (i) the liquidity demanders' optimal x_k as a function of the their valuation $\mathbb{E}[v_2 + \kappa\sigma u_k | \, v_{k-1}, u_k]$ and the state of order book $\{\alpha_{k-1}, \beta_{k-1}, \alpha_{k-1}^{'}, \beta_{k-1}^{'}\}$; (ii) the competitive slow quotes $\{\alpha_1^{'}, \beta_1^{'}\}$ as functions of the common value v_0, the slow quoter's initial inventory $y_0^{'}$, and the optimal market order strategy x_2; and (iii) the competitive fast quotes $\{\alpha_k, \beta_k\}$ as functions of the common value v_k, the fast quoter's inventory y_k, the optimal market order strategy x_{k+1}, and the slow quotes $\{\alpha_1^{'}, \beta_1^{'}\}$ if $k = 1$.

Parameter assumptions. The following conditions on the parameters are imposed:

$$\kappa > 1 + \frac{\gamma'}{2\sigma} \text{ and } \gamma' > 2\sigma + 3\gamma \tag{3}$$

The first inequality ensures that the private value magnitude κ is large enough to avoid uninteresting cases of no-trade (ruling out $x_k = 0$). The second inequality ensures a wider bid-ask spread from the slow than from the fast, for otherwise there will be no variation in period 1 quotes (the tighter slow quoters would prevail irrespective of whether the fast was active). This is achieved by assuming that the slow quoter's marginal inventory cost γ' is sufficiently higher than that of the fast quoter. Intuitively, $\gamma' > \gamma$ reflects the more severe adverse-selection risk faced by the slow quoter than by the fast quoter (two periods vs. one period).

3.1.2 Equilibrium analysis

This subsection solves the three sets of equilibrium objects. Consider the market orders first. A period-$k \in \{1, 2\}$ liquidity demander values the asset at

$$\mathbb{E}[v_2 + \kappa\sigma u_k |\, u_k, v_{k-1}] = v_{k-1} + \mathbb{E}[\Delta v_k |\, u_k] + \kappa\sigma u_k = v_{k-1} + (\tau + \kappa)\sigma u_k \tag{4}$$

where the last equality follows the conditional expectation

$$\mathbb{E}[\Delta v_k |\, u_k] = \mathbb{P}[I_k = 1]\mathbb{E}\left[\Delta v_k \middle| u_k = \frac{\Delta v_k}{\sigma}\right] + \mathbb{P}[I_k = 0]\mathbb{E}[\Delta v_k] = \tau\sigma u_k. \tag{5}$$

For notation simplicity, denote by

$$q_k := (\tau + \kappa)\sigma u_k \tag{6}$$

the "valuation wedge" between the liquidity demander and other market participants (who have not seen u_k). The demander compares his valuation (4) to the prevailing best bid b_{k-1} and ask a_{k-1} (endogenously determined below). Under the one-unit size constraint, the optimal market order is

$$x_k = \mathbb{1}[v_{k-1} + q_k > a_{k-1}] - \mathbb{1}[v_{k-1} + q_k < b_{k-1}], \tag{7}$$

where $\mathbb{1}[\cdot]$ is the indicator function. (The equilibrium bid and ask will not cross, i.e., $a_{t-1} > b_{t-1}$.)

The competitive slow quoter breaks even when executes against $x_2 = 1$, in which case her inventory level changes to $y_2^{'} = y_0^{'} - x_2 = -1$ (with $y_0^{'} = 0$). That the ask $\alpha_1^{'}$ is executed means $v_1 + q_2 > \alpha_1^{'}$ according to Equation (7). Therefore, the break-even ask must satisfy

$$\alpha_1^{'} + \mathbb{E}\left[\left(y_0^{'} - 1\right)v_2 - \frac{\gamma^{'}}{2}\left(y_0^{'} - 1\right)^2 \middle| v_1 + q_2 > \alpha_1^{'}\right] = \mathbb{E}\left[y_0^{'}v_2 - \frac{\gamma^{'}}{2}{y_0^{'}}^2 \middle| v_1 + q_2 > \alpha_1^{'}\right], \tag{8}$$

As is common in limit order book models, the equilibrium $\alpha_1^{'}$ is a fixed-point jointly solving (7) and (8). The same analysis applies to the bid $\beta_1^{'}$ and is omitted here. The proof to Proposition 1 gives the details of the fixed-point analysis.

In period-$k \in \{0, 1\}$, the fast quoter (if active) solves a similar problem, except (i) that she now observes v_k in period $k \in \{0, 1\}$ and (ii) that her inventory y_k might not be zero (as $y_1 = y_0 - x_1$). Her break-even ask α_k, trading against $x_{k+1} = 1$, must satisfy

$$\alpha_k + \mathbb{E}\Big[(y_k - 1)v_{k+1} - \frac{\gamma}{2}(y_k - 1)^2 \big| v_k + q_{k+1} > \alpha_k, v_k\Big] = \mathbb{E}\Big[y_k v_{k+1} - \frac{\gamma}{2}{y_k}^2 \big| v_k + q_{k+1} > \alpha_k, v_k\Big]. \tag{9}$$

The above also holds true for period $k = 0$ because the expected continuation value is zero due to competition—she is indifferent between trading at $k = 0$ or not, and if she does not trade, she continues to derive zero utility from her zero inventory. The detailed solution is again deferred to the proof of Proposition 1.

Proposition 1 (Equilibrium of the two-trade game). *There is an equilibrium, under assumption (3), where the slow quoter submits limit orders at $k = 0$ according to*

$$\alpha_1^{'} = v_0 + \tau\sigma + \frac{\gamma^{'}}{2} \;\; \textit{and} \;\; \beta_1^{'} = v_0 - \tau\sigma - \frac{\gamma^{'}}{2}; \tag{10}$$

the fast quoter posts limit orders at $k \in \{0, 1\}$ according to

$$\alpha_k = v_k + \tau\sigma + \frac{\gamma}{2} - \gamma y_k \;\; \textit{and} \;\; \beta_k = v_k - \tau\sigma - \frac{\gamma}{2} - \gamma y_k; \tag{11}$$

and the period-$k \in \{1, 2\}$ market order is $x_k = u_k$.

Several features of the equilibrium are worth highlighting. First, the market order trader always trades, i.e., $x_k \neq 0$, because his private value magnitude κ is ensured to be sufficiently large by assumption (3). This helps focus on the more interesting scenario where there are trades. Second, as

the market order $x_k = u_k$ fully reveals the private value, it is easy to compute the trades' permanent price impact—or the adverse-selection cost. Following Equation (5),

$$\mathbb{E}[\Delta v_k | x_k] = \mathbb{E}[\Delta v_k | u_k] = \tau\sigma u_k = \lambda x_k \tag{12}$$

where $\lambda := \tau\sigma$ is the permanent price impact. Third, the equilibrium quotes have fairly intuitive components. For example, the fast quoter first marks up the prevailing fundamental price v_k by the adverse-selection cost λx_{k+1}, on top of the marginal inventory cost $\frac{\gamma}{2}x_{k+1}$; and then, to account for the existing inventory cost, she tilts the quotes against the current inventory level by $-\gamma y_k$. In fact, when $y_k = y_k^{'} = 0$, the equilibrium quotes have the same form as in the structural mode by Madhavan, Richardson, and Roomans (1997). Finally, when the fast quoter is active in period 1, her quotes are tighter than the slow quotes: $\alpha_1^{'} > \alpha_1 > \beta_1 > \beta_1^{'}$. This is guaranteed by condition (3), which assumes $\gamma^{'}$ is sufficiently larger than γ to reflect the higher adverse-selection risk faced by the slow than by the fast (two periods vs. one period). The tighter fast quotes reflects the simple intuition that prices adjust timely only when fast quoters are active. This turns out to be a key feature necessary to produce the negative contemporaneous transitory price impact ψ_0.

3.1.3 The permanent and the transitory price impacts

Consider the equilibrium midquote p_k, the arithmetic average of the best ask and the best bid. Right before the first trade x_1, the prevailing midquote is simply $p_0 = \frac{1}{2}(\alpha_0 + \beta_0) = v_0$, following Equation (11) with $y_0 = 0$. The more interesting midquote is the prevailing p_1 right before the second trade x_2. Recall that F_1 indicates the activeness of the fast quoter. Then,

$$p_1 = \frac{F_1}{2}(\alpha_1 + \beta_1) + \frac{1 - F_1}{2}\left(\alpha_1^{'} + \beta_1^{'}\right) = F_1 \cdot (v_1 - \gamma y_1) + (1 - F_1)v_0 = v_1 + F_1\gamma x_1 - (1 - F_1)\Delta v_1.$$

The first equality holds because the fast quotes are tighter than the slow quotes; i.e., $\alpha_1 < \alpha_1^{'}$ and $\beta_1 > \beta_1^{'}$. The second equality uses the expressions in Proposition 1. The last equality obtains by substituting $v_0 = v_1 - \Delta v_1$ and $y_1 = y_0 - x_1 = -x_1$.

From an econometrician's point of view, the midquote p_1 can always be written as the sum of a

random walk efficient price m_1 and a stationary, zero-mean price pressure s_1:

$$p_1 = \overbrace{v_1}^{=:m_1,\text{ efficient price}} + \underbrace{F_1\gamma x_1 - (1-F_1)\Delta v_1}_{=:s_1,\text{ price pressure}}.$$

The econometrician, observing only the price and the trade $\{p_1, x_1\}$, only knows the structure of such a decomposition, but not its components. (Only the market participants observe the hidden state variables like v_1, Δv_1, and F_1.) As such, in order to spell out the price impacts, the econometrician projects the hidden efficient price m_1 and the price pressure s_1 onto the observed trade x_1. To do so, recall from Equation (12) that $\mathbb{E}[\Delta v_1 | x_1] = \lambda x_1$ and, hence,

$$\Delta m_1 = \Delta v_1 = \lambda x_1 + \mu_1 \tag{13}$$

where $\mu_1 := \Delta v_1 - \lambda x_1$ is uncorrelated with the order flow x_1 as, by law of iterated expectations, $\mathrm{cov}[\lambda x_1, \mu_1] = \lambda\mathbb{E}[x_1 \cdot (\Delta v_1 - \lambda x_1)] = \lambda\mathbb{E}[x_1 \cdot (\mathbb{E}[\Delta v_1 | x_1] - \lambda x_1)] = 0$. That is, from the econometrician's point of view, the efficient price innovation has a trade-related component λx_1 and a non-trade component μ_1. Therefore, writing $m_0 := v_0$, the efficient price becomes

$$m_1 = v_1 = v_0 + \Delta v_1 = m_0 + \lambda x_1 + \mu_1, \tag{14}$$

the same as in the structural model (1), with the order flow innovation $x_k^* = x_k = u_k$ being a white noise here. Similarly, the fast quoter's activeness F_1 can be decomposed as $F_1 = \eta + (F_1 - \eta)$, where $\mathbb{E}[F_1] = \eta \in [0, 1]$. The price pressure then becomes

$$\begin{aligned} s_1 &= F_1\gamma x_1 - (1-F_1)\Delta v_1 = (\eta\gamma x_1 - (1-\eta)\Delta v_1) + (F_1 - \eta)(\gamma x_1 + \Delta v_1) \\ &= \underbrace{(\eta\gamma - (1-\eta)\lambda)}_{=:\psi_0} x_1 + \underbrace{(F_1 - \eta)(\Delta v_1 + \gamma x_1) - (1-\eta)\mu_1}_{=:\nu_1}. \end{aligned} \tag{15}$$

That is, s_1 can also be decomposed into a trade-related component of $\psi_0 x_1$, where

$$\psi_0 := \eta\gamma - (1-\eta)\lambda \tag{16}$$

and a non-trade component ν_1 with $\mathrm{cov}[x_1, \nu_1] = 0$. Note that $s_1 = \psi_0 x_1 + \nu_1$ is a special case of the

structural model (1) with $\phi(L) = \psi(L) = 0$.

The parameter ψ_0 measures how much the order flow x_1 moves the price pressure s_1—the contemporaneous transitory price impact, as in the empirical framework (1). The equilibrium reveals that ψ_0 has two parts. When the fast quoter is active (probability η), the first part shows that the marginal inventory cost γ adds to ψ_0 positively. This part reflects the conventional intuition that price pressure moves in the trade's direction. The second part of $-(1-\eta)\lambda$ is new. It highlights that when the slow quotes prevail (probability $1-\eta$), the price pressure reacts *negatively* to the trade. This is because the slow quotes, submitted at $k = 0$, are unable to adjust to the new information brought by the trade x_1 (or the news μ_1). Therefore, when the latent efficient price m_1 moves along the trade by λx_1, the price pressure s_1 must move in the opposite direction by $-\lambda x_1$ to offset it.

This second component of $-(1-\eta)\lambda$ is the model's novel insight. It makes the parameter ψ_0 possibly negative, echoing the empirical findings from Section 2. To emphasize, this result arises because (i) the fast quoter might be inactive and, in such a case, (ii) the *stale* slow quotes prevail. Indeed, if $\eta \to 1$, the fast quoter reacting to each and every trade, then $\psi_0 \to \gamma > 0$ as only the fast quoter's marginal inventory cost γ matters.

3.1.4 The fast and the slow quotes

The stochastic alternation between the fast quotes $\{\alpha_1, \beta_1\}$ and the slow quotes $\{\alpha_1', \beta_1'\}$ is the key modeling element that gives rise to the possibly negative ψ_0. Because of its importance, this subsection dedicates some additional discussion to this construct.

The slow quotes essentially serve as the ceiling and the floor for the ask and the bid, respectively. They are sometimes referred to as the "absorbing" price bounds set by investor crowds, as in Seppi (1997), Parlour (1998), and Roşu (2009), among others. They do *not* need to be competitive as assumed in Section (3.1.1). The competitiveness is a simple way to pin down these quotes in equilibrium. The key economic restriction is that some trades (and news) occur in between the submission of these quotes and their arrival in the limit order book. As such, these prices are slow

and cannot adjust to the latest information of these in-between trades.[7]

The competitive fast quotes, on the other hand, set the floor and the ceiling for, respectively, the ask and the bid. These are the Bertrand prices that multiple high-frequency market makers would eventually achieve after repeatedly undercutting each other (see, e.g., Menkveld and Zoican, 2017).

The reality probably strides between the floors and the ceilings. With imperfect price competition, the eventual ask and bid respectively fall in $[\alpha_1, \alpha_1^{'}]$ and $[\beta_1^{'}, \beta_1]$. The two-trade game models such imperfect competition through a Bernoulli draw F_1. This can be microfounded by assuming $n \geq 2$ fast quoters, each of them has probability $\zeta \in (0, 1)$ to be able to quote (being active) for this asset. Such limited capacity to quote can arise from their fixed cost in processing data feeds, their opportunity costs (quoting for this asset rather than for another), or their inventory constraints (see, e.g., Duffie, 2010). Then, with probability $(1-\zeta)^n$ there will be no fast quoters active at period 1, making the slow quotes prevail. Or, with probability $n\zeta \cdot (1-\zeta)^{n-1}$, there will be only one fast quoter active and, in this case, again the slow quotes will prevail because the monopolistic active fast quoter will post ask and bid just inside the the slow quotes. Only with probability $\eta := 1 - (1-\zeta)^n - n\zeta \cdot (1-\zeta)^{n-1}$ will there be Bertrand competition.[8]

The above intuition also applies in richer models, where quoters can post limit orders for multiple periods ahead. In such settings, after a trade, a monopolist might not want to update her previously submitted quotes, as long as they are still profitable. That is, imperfect competition can cause *endogenous inaction* by quoters and such inaction, in turn, can result in $\psi_0 < 0$.

In short, the parameter η weighs the equilibrium quotes between the tighter competitive quotes $\{\alpha_1, \beta\}$ and the slacker monopolistic quotes $\{\alpha_1^{'}, \beta_1^{'}\}$. So far η has been mostly referred to as the "activeness" of the fast quoters. Following the discussion above, it can be more generally

[7] Neither do these slow quoters need to be always active. They can be absent from time to time, just like the fast quotes, but then the limit order book might sometimes be empty and admit no trade. Section 3.1 essentially assumes a large crowd of such slow traders so that almost surely some are active and can provide liquidity competitively.

[8] There are other setups that can lead to qualitatively similar equilibrium outcome, endogenizing η in various ways. For example, Roşu (2009) models the sequential undercutting of limit orders. As the arrival of market order is random, the prevailing ask randomly falls between a competitive lower bound and an exogenous upper bound. Jovanovic and Menkveld (2015, 2019) find that facing unknown number of price competitors, limit order traders strategically randomize their quotes within a price range.

understood as the *competitiveness* of (fast) liquidity providers in the limit order market.

The analysis also nests the friction of price discreteness as a special case. For example, suppose the current bid is at \$9.99 and ask at \$10.01, with midquote at \$10.00. If a market buy order has a permanent price impact of 2 bps, it moves the efficient price up by two-tenths of a cent ($¢0.2 = 0.0002 \times \$10.00$). Without price discreteness (no minimum tick size) and all else equal, such a change would have resulted in upward shifts in both the bid and the ask to \$9.992 and \$10.012 respectively. However, given a minimum tick size of one cent, the quotes will not move, keeping the same midquote at \$10.00. As a result, the 2 bps increase in the efficient price must be met by a negative contemporaneous transitory price impact of -2 bps. Such an example can be thought of as the corner case of $\eta \downarrow 0$; i.e., the fast is unable to undercut the slow due to price discreteness. This follows the intuition that price discreteness curbs price competition (Chao, Yao, and Ye, 2017; Yao and Ye, 2018). Indeed, Equation (16) implies that $\psi_0 \downarrow -\lambda$ when $\eta \downarrow 0$.

To sum up, as long as there are (i) slow quotes, *unable to respond to the latest trade* and (ii) *imperfectly competitive* fast quotes, the equilibrium s_1 will react to the order flow x_1 in two ways: either reflecting the price pressure (like inventory cost, γx_1) or *offsetting* the permanent price impact ($-\lambda x_1$). Section 3.2 below generalizes such intuition and builds a structural model to derive additional predictions.

3.2 A general structural model

Consider a standard trading data set of order flows $\{x_k\}$ and midquotes $\{p_k\}$ as described in Section 2. This subsection derives a general structural model for the price pressure s_k in the midquote p_k, building on the insight from the equilibrium model, and extracts additional empirical predictions.

Fundamental value. Each unit of the asset can be liquidated for v_T dollars (the numeraire) in some remote future T (e.g., at the closing auction). The current expectation of v_T is written as

$$m_k := \mathbb{E}_k[v_T],$$

where the expectation operator $\mathbb{E}_k[\cdot]$ emphasizes that it is conditioning on all public information by, and including, the k-th trade.

Market orders. The next market(able) order is denoted by x_{k+1}, which can take integer values: $x_{k+1} \in \{1, 2, ...\}$ for buys and $x_{k+1} \in \{-1, -2, ...\}$ for sells. It is common knowledge that x_{k+1} can be informative of v_T. In particular, the impact of the *first* unit of x_{k+1} is denoted as:

$$\mathbb{E}_k[v_T | \, x_{k+1} \geq 1\,] =: m_k^+ \geq m_k \geq m_k^- := \mathbb{E}_k[v_T | \, x_{k+1} \leq -1\,].$$

One can also introduce further notations for how larger x_{k+1} affects the learning of v_T. But for the purpose of formulating the midquote p_k, it is sufficient to look at the best ask and the best bid, for which only $m_k^{\pm}$ will be used.

Slow quotes. There are many buying and selling limit orders resting at some bid price $\beta_k^{'}$ and at some ask price $\alpha_k^{'}$, respectively, where $\beta_k^{'} < m_k < \alpha_k^{'}$. These limit orders constitute sufficient depths that are large enough to accommodate the next market order x_{k+1}. Importantly, these quotes do not change during the interval $[t_k, t_{k+1})$, because they are *previously* submitted by relatively *slow* trading crowds.

Fast quotes. There are $n_k \in \{0, 1, ...\}$ active (high-frequency) fast quoters who can post and revise (or cancel) limit orders during the time interval $[t_k, t_{k+1})$. The number of fast quoters, n_k, can be time-varying, due to their limited capacity (subject to, e.g., computation power constraints), latencies (transmission delays), inventory constraints, etc. When attentive, the fast quoters observe the same public information and agree on the common values m_k, m_k^+, and m_k^-. In addition, each fast quoter i may have a private value $w_{ik}^{\pm}$ for the marginal unit. Formally, the i-th fast quoters' reservation value for a buy market order is $\mathbb{E}_{ik}[v_T | \, x_{k+1} \geq 1\,] = m_k^+ + w_{ik}^+$; and that for a sell is

$\mathbb{E}_{ik}[v_T \,|\, x_{k+1} \leq -1] = m_k^- + w_{ik}^-$. In Section 3.1, such private values arise from inventory costs, but they can broadly represent risk-aversion, private information, disagreement, sentiment, etc.

Best bid and best ask. Characterizing the exact distribution of the best quotes is difficult (and need more assumptions). Instead, it is straightforward to establish their supports, which are sufficient for formulating the structural model. Consider the best ask a_k. Due to the "slow" selling crowd, it has a ceiling of $a_k \leq \alpha_k'$. The competition among the n_k fast quoters will likely drive the ask below α_k'. Letting $w_k^+ := \min\left\{\left\{\alpha_k' - m_k^+\right\} \cup \left\{w_{ik}^+\right\}_{i=1}^{n_k}\right\}$, then the equilibrium a_k is bounded from below by $m_k^+ + w_k^+$: No fast quoter is willing to undercut this lowest valuation. Therefore, the best ask to prevail before the next trade can be written as

$$a_k = (m_k^+ + w_k^+) \underbrace{+\zeta_k^+ \cdot \left(\alpha_k' - (m_k^+ + w_k^+)\right)}_{\text{markup due to imperfect competition}},$$

where $\zeta_k^+ \in [0, 1]$ is possibly time-varying (e.g., driven by n_k). The best bid b_k similarly follows as

$$b_k = (m_k^- + w_k^-) \underbrace{-\zeta_k^- \cdot \left((m_k^- + w_k^-) - \beta_k'\right)}_{\text{markdown due to imperfect competition}},$$

where $w_k^- := \max\left\{\left\{\beta_k' - m_k^-\right\} \cup \left\{w_{ik}^-\right\}_{i=1}^{n_k}\right\}$ and $\zeta_k^- \in [0, 1]$ is capturing the lack of competitiveness of fast quoters on the bid side, just like ζ_k^+ on the ask side.

Symmetry assumptions. It is helpful to assume some mild symmetry between buys and sells to simplify the subsequent derivations. First, assume that $m_k^+ - m_k = m_k - m_k^-$; i.e., the expected permanent price impact of a marginal buy is the same as that of a marginal sell in size. Second, assume that the fast quoters' price competitions are symmetric on the bid and the ask sides. One can then define $1 - \eta := \mathbb{E}[\zeta_k^+] = \mathbb{E}[\zeta_k^-] \in [0, 1]$ as the average markup (markdown) fraction. In other words, $\eta \in [0, 1]$ captures fast quoters' average competitiveness. In Section 3.1, the equilibrium model effectively assumes $\zeta_k^+ = \zeta_k^- = 1 - F_1$ as the same Bernoulli draw.

Midquote. The midquote p_k as the arithmetic mean between the best bid and the best ask is

$$\begin{aligned} p_k &= \frac{1}{2}\left[(m_k^+ + w_k^+) + (m_k^- + w_k^-)\right] + \frac{1}{2}\left[(\zeta_k^+ \alpha_k^{'} + \zeta_k^- \beta_k^{'}) - \zeta_k^+ \cdot (m_k^+ + w_k^+) - \zeta_k^- \cdot (m_k^- + w_k^-)\right] \\ &= \eta \cdot \left(m_k + \frac{1}{2}\left(w_k^+ + w_k^-\right)\right) + (1-\eta) \cdot \frac{1}{2}\left(\alpha_k^{'} + \beta_k^{'}\right) + z_k \end{aligned} \tag{17}$$

where the second line first makes use of the symmetry of $m_k^+ - m_k = m_k - m_k^-$ and then takes out the mean $(1-\eta)$ from both ζ_k^+ and ζ_k^-, leaving z_k to capture the zero-mean residuals.

Denote by $w_k := \frac{1}{2}(w_k^+ + w_k^-)$ the fast quoters' competitive "average" private value for handling the marginal unit of the asset. (In Section 3.1, $w_k = -\gamma y_k$ reflects how the fast quoters tilt their quotes against their existing inventory y_k; see Proposition 1.) Define also $c_k := \frac{1}{2}(\alpha_k^{'} + \beta_k^{'})$ as the center point implied by the slow crowds' limit orders. Note that a monopolistic fast quoter will set $\alpha_k \uparrow \alpha_k^{'}$ and $\beta_k \downarrow \beta_k^{'}$. Therefore, c_k is also the center point of the monopolistic ask $\alpha_k^{'}$ and the monopolistic bid $\beta_k^{'}$. Importantly, c_k can deviate from the efficient price m_k, because the *slow* crowds cannot not catch up to update their quotes timely. The deviation $(c_k - m_k)$ can thus be thought of as the *price distortion* due to imperfect quote competition.

The midquote p_k can then be written as

$$p_k = \eta \cdot (m_k + w_k) + (1-\eta)c_k + z_k, \tag{18}$$

which is a weighted average between the fast quoters' competitive midpoint $m_k + w_k$ and the monopolistic midpoint c_k, plus some zero-mean deviation z_k. The weight η captures the competitiveness of the fast quoters: When they are perfectly competitive, price competition a la Bertrand drives p_k to the competitive midpoint. When there is little competition, the monopolistic midpoint obtains.

Price pressure. Subtracting the efficient price m_k from (18) gives the price pressure

$$s_k := p_k - m_k = \eta w_k + (1-\eta)(c_k - m_k) + z_k, \tag{19}$$

which is a weighted average between the fast quoters' competitive private value w_k and the price distortion $(c_k - m_k)$, plus some zero-mean noise z_k.

Suppose the three hidden states $\{w_k, c_k, m_k\}$ can be linearly approximated by $w_k \approx (\gamma_0 +$

$\gamma(L))x_k$ + noise, $c_k \approx \delta(L)x_k$ + noise, and $m_k \approx m_{k-1} + \lambda x_k$ + noise. (Note that the *slow* quote midpoint c_k does *not* respond to the trade x_k contemporaneously; that is, the lag polynomial $\delta(L)$ satisfies $\delta(0) = 0$.) It follows that the contemporaneous transitory price impact ψ_0 still has the form as (16) in the equilibrium model:

$$\psi_0 := \mathbb{E}\left[\frac{\partial s_k}{\partial x_k}\middle| x_k, x_{k-1}, \ldots\right] = \eta\gamma_0 - (1-\eta)\lambda.$$

Thus, the finding still holds that ψ_0 can be negative, just like in the equilibrium model. In fact, the price pressure (15) from the equilibrium model is a special case of the general form (19). Proposition 1 shows that the private value w_1 arises only from inventory considerations with $w_1 = -\gamma y_1 = \gamma x_1$. The slow quotes' midpoint is $c_1 = \frac{1}{2}\left(\alpha_1' + \beta_1'\right) = v_0$. Hence, by Equations (13) and (14), the distortion is $c_1 - m_1 = c_1 - v_1 = -\Delta v_1$. The white noise residual $(F_1 - \eta)(\gamma x_1 + \Delta v_1)$ then becomes z_1 here.

One limitation of the simple two-trade equilibrium model is that it only allows the trade x_1 to affect the price pressure s_1 contemporaneously: Both w_1 and $(c_1 - m_1)$ are uncorrelated to past trades, as there are none. In the more general setting here, both the private value w_k and the distortion $(c_k - m_k)$ can be serially correlated:

- The private value w_k, reflecting the inventory level like in the equilibrium model, is likely a slow-moving *persistent* process, because it takes time for the fast quoters to mean revert their net inventory exposure to zero. See equilibrium models by, e.g., Ho and Stoll (1981, 1983) and Hendershott and Menkveld (2014); the latter also provides recent empirical evidence.
- The distortion $(c_k - m_k)$ can also be autocorrelated but its persistence is likely very *low*, because when lagging behind, the slow quotes become stale and very profitable to trade against, possibly by high-frequency "snipers" (Budish, Cramton, and Shim, 2015).

The contrast between the high-persistence w_k and the low-persistence $(c_k - m_k)$ suggests the following novel empirical prediction:

Prediction 1 (Price pressure persistence vs. fast quoter competitiveness). *When the fast quoters are more competitive (higher η), the price pressure s_k is more persistent.*

Intuitively, as the price pressure s_k weighs between the private value w_k and the price distortion $(c_k - m_k)$, so does its persistence. Since the price distortion $(c_k - m_k)$ cannot be very persistent, all else equal, when the fast quoters are more competitive (higher η), the price pressure s_k also tilts more to the high-persistence private value component w_k. Under such an assumption, the dynamic equilibrium model developed in Appendix C proves such monotonicity analytically.

Volatility. Another insight from the general structural model is that the competitiveness η also plays an important role in the asset's return volatility. Observe from the midquote expression (18) that a higher η amplifies the efficient price m_k, amplifies the private value w_k, but diminishes the slow midpoint c_k. That is, these three components' contribution to price—hence also to return and to volatility—is scaled up and down by the competitiveness η.

The net effect is ultimately an empirical question (explored in Section 4). Yet, intuitively, one can see that the amplification in m_k and the dampening in c_k will on average offset each other: The slow midpoint c_k is the efficient price m_k with some short lags. The changes in these two therefore should have similar magnitudes. When η increases, it amplifies a similar amount in m_k as it diminishes in c_k. As a result, the remaining amplification effect in w_k dominates:

Prediction 2 (Volatility vs. fast quoter competitiveness). *When the fast quoters are more competitive (higher η), the volatility is higher.*

In Appendix C, a dynamic equilibrium is shown to bear this property.

A remark on the two predictions. At first sight, both Predictions 1 and 2 might seem counterintuitive: When liquidity providers (the fast quoters in the context) are more competitive (or active), they should make the price more "efficient" (i) by making price pressure s_k mean-revert to zero more quickly, lowering price pressure persistence; and (ii) by quoting more competitively, reducing noise and hence volatility.

Such an intuition fails in that it has not taken into account the pricing incentive of imperfectly competitive liquidity providers. Consider a monopolistic fast quoter for example. She wants to quote a spread as wide as possible, i.e., at the *stale* bid and ask set by the slow quoters.[9] The resulting price series thus becomes also stale, fluctuating less to new trades and new information. As an extreme example, if the slow quotes are sufficiently wide and change very infrequently, the fast monopolist price will be like a constant, hence very impersistent and of little volatility. As competition intensifies (η increases), the price becomes more up-to-date, not only reflecting new information but also the persistent private value (w_k, e.g., inventory costs), adding to the volatility. Section 4 takes these two predictions to the data and finds consistent evidence.

Parameter identification. Given the importance of η in both Prediction 1 and 2, it is tempting to try to estimate it directly from the data. Unfortunately this is not possible under the current structural model. To see why, note from (19) that η loads on both the *hidden* private value w_k and the *hidden* distortion $(c_k - m_k)$. In order to identify η, therefore, one would need instruments for w_k and for $(c_k - m_k)$ to separate the two. However, the only observable explanatory variable is the order flow $\{x_k\}$ and it likely affects both w_k and $(c_k - m_k)$. It is not clear whether (or which) other observable data series, like order book depths and deep-book quotes, can be used as instruments to separate w_k and $(c_k - m_k)$.

The identification challenge can also be seen from the equilibrium model (Section 3.1), where the coefficient loading on the order flow x_k is found to be $\psi_0 = \eta\gamma - (1-\eta)\lambda$, mixing η with other primitive parameters. Appendix B further shows that the current structural model exactly replicates the structural framework (1). That is, absent of further assumptions, one can only identify ψ_0, $\psi(L)$, and $\phi(L)$ regarding the price pressure s_k, but not how η parametrizes these coefficients. Nevertheless, if the current structural model holds true in the data, Predictions 1 and 2 shall be

[9] One might wonder why a fast monopolist quoter might want to set the price at the slow bid and ask: Are these prices not subject to increased adverse-selection? Indeed, from time to time, the fast monopolist might want to deviate from quoting such prices due to new information. However, *on average*, the slow quotes are not subject to additional adverse-selection, because the slow quoters are also rational and they quote a wider spread in anticipation of possible adverse information in the future. See the equilibrium model in Section 3.1.

verifiable. Section 4 brings such evidence.

4 Testing the additional predictions

This section investigates Prediction 1 and 2. Section 4.1 constructs the left-hand side variables. Section 4.2 discusses the empirical strategy. The results are then presented in Section 4.3 and 4.4.

4.1 The left-hand side variables

Prediction 1 is about the persistence of price pressure s_k. Consider the general specification (1): $(1 - \phi(L))s_k = \psi(L)x_k + v_k$. Starting from a steady state where $s_{k-1} = 0$, given a unit impulse of order flow shock $x_k = 1$, one can compute the immediate impulse response s_k and, recursively, the T-step future response s_{k+T}. One way to measure the persistence of the price pressure is to compute the ratio of $|s_{k+T}/s_k|$. The larger is such "decay ratio", the more persistent is the price pressure. Under the most parsimonious specification of $s_k = \phi s_{k-1} + \psi x_k + v_k$, this decay ratio reduces to ϕ^T (for $T \geq 1$ and $\phi > 0$). As such, ϕ fully characterizes the price pressure persistence and can be used as the left-hand side variable for Prediction 1. The findings are robust to other price pressure specifications, which is not surprising because the first-order autoregression coefficient ϕ_1 clearly dominates as seen in Table 1.

Prediction 2 is about return volatility. For each stock-day in 2018, three volatility measures are constructed: (i) 5-minute realized volatility, (ii) 1-minute realized volatility, and (iii) trade-by-trade realized volatility. The 5-minute realized volatility is known for its accuracy in measuring the quadratic variation of the true return process (Liu, Patton, and Sheppard, 2015). Higher frequency measures introduce so-called microstructure noises, which could be important to reflect the role of private value w_k as illustrated in Section 3.2. Table 2 reports the summary statistics of these return volatility measures.

				Percentiles						
	Unit	Mean	Std Dev	1%	5%	25%	50%	75%	95%	99%
5-minute snapshots	bps/day	181.7	98.5	57.5	76.4	113.9	156.8	221.3	371.5	557.7
1-minute snapshots	bps/day	185.3	97.5	62.7	80.8	117.8	161.1	225.4	371.9	558.6
Trade-by-trade	bps/day	177.2	110.0	55.1	69.8	102.9	145.2	214.7	401.3	608.5

Table 2: Realized volatility across stock-days. This table reports the summary statistics of the realized volatility across stock-days. Three different frequencies are considered: 5-minute, 1-minute, and trade-by-trade. Specifically, for each stock-day, the intraday midquotes are snapshotted at the respective frequency and the realized variance is computed as the sum of the squares of the log differences of the midquote snapshots. The reported realized volatility is the square root of the realized variance, scaled to basis points.

4.2 Empirical strategy

To examine the predictions, ideally, one would like to perform a panel regression of the form

$$y_{id} \sim \eta_{id} + \text{controls},$$

where y_{id} is one of the left-hand side variables for stock i on day d in the sample, while the key right-hand side variable η_{id} is the fast quoters' competitiveness wanted by the predictions. Unfortunately, η is not observable (or identifiable from the structural model; see the discussion on p. 27). Therefore, one would need some events that affect η and, ideally, only η.

A naïve attempt. One candidate type of events that can affect η is news about stock i. When there is such news, anticipating more trading activity, liquidity providers will likely pay more attention to the stock (Corwin and Coughenour, 2008). This will intensify their price competition, i.e., raising η. News data are thus collected from RavenPack Equities Dow Jones Edition for all U.S. equities in 2018. Only news entries with RavenPack scores (relevance, novelty, and sentiment) are kept to ensure the substantiality of news. A news entry about stock i is associated to ϕ_{id} if its timestamp is after the closing of trading day $d-1$ and before the closing of trading d.

Simple regressions show that, indeed, when there are more news about a stock i, there are more quote messages in DTAQ data, the quoted spread is tighter, and the top of the book is deeper. This confirms the above intuition that (fast) liquidity providers are indeed more active and more competitive when there is news; i.e., higher η. Then, using the news count for stock i on day d as a proxy for η_{id}, it is found that both two left-hand side variables load positively on this proxy in the panel regression. That is, supporting both predictions, a stock's price pressure is more persistent and its realized volatility higher, when there is more news about it.

Potential confounders. The obvious problem is that such news events also affect the fundamentals of the stocks and change other aspects of trading, confounding the channel through which the left-hand side variables are affected. For example, consider Prediction 1, which builds on the price pressure expression (19): $s_k = \eta w_k + (1-\eta)(c_k - m_k) + z_k$. The persistence of s_k, ϕ, is roughly the average of the persistence of w_k and that of $(c_k - m_k)$, weighted by the competitiveness η. Suppose there is some news of a stock that grabs market participants' attention, intensifying the quote competition (higher η). All else equal, this would make s_k more persistent, raising ϕ. This is the "weight channel" underscored by Prediction 1. But there are other effects. Notably, the persistence of w_k and that of $(c_k - m_k)$ may also be affected. That is, an observed higher ϕ might be due to increases in the components, *instead of the weights*, in the weighted average. Such a "component channel" therefore confounds the wanted "weight channel."

Figure 2(a) graphs such confounding channels. The dependent variable ϕ is placed on the left-hand side of the graph, as in a regression. The channel of interest—the "weight channel" by Prediction 1—is the one from η to ϕ, marked with many arrows. However, the news will likely also affect ϕ through the persistence of w_k and of $(c_k - m_k)$—the "component channel." In addition, there are possibly other unknown channels, shown in the dashed-line arrows (though such mechanisms are unclear, absent of theory guidance).

Muting the potential confounders. A number of efforts are made to switch the confounding channels off. Figure 2(b) illustrates them.

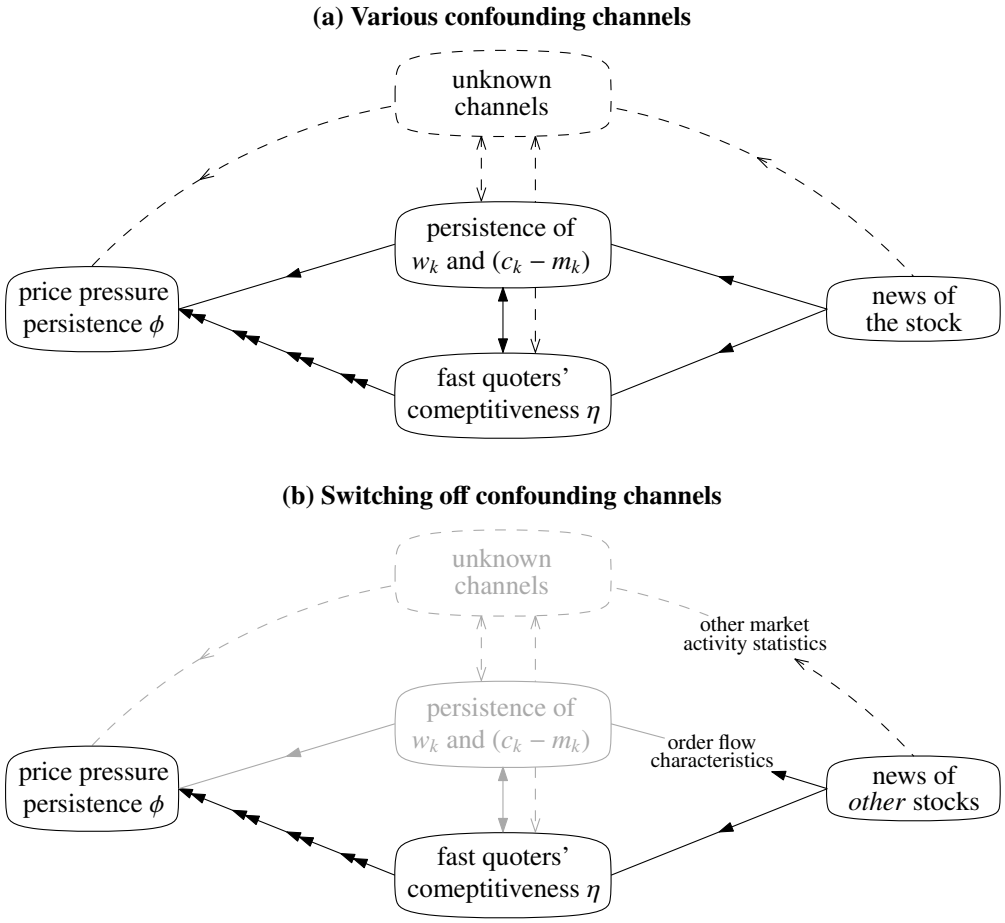

Figure 2: Empirical strategy for Prediction 1. Panel (a) highlights the various channels that might affect price pressure persistence ϕ when a news event occurs. In particular, the structural model (19) explicitly shows how both the fast quoters' competitiveness η (the weight) and the persistence in w_k and in $(c_k - m_k)$ (the components) could affect ϕ, as shown in the solid-line arrows. There are possibly also other unknown channels (not reflected in the structural model), which are shown in the dashed-line arrows. The channel of interest is the one through fast quoters' competitiveness η, marked with multiple arrows (Prediction 1). Panel (b) uses news of *other* stocks as a shock, together with order flow characteristics and other market activity statistics as controls, to switch the confounding channels off.

1. One can use news on day d but *not* about stock i as a (negative) shock to η_{id}. The validity of such shocks builds on two sides. On the one hand, such news shocks do go through the wanted "weight channel," under the assumption that the liquidity providers have *limited* attention or processing capacity (Corwin and Coughenour, 2008): When the news about a different stock j

grabs their limited attention, fewer will remain as attentive as they used to be to the no-news stock i. As a result, there is less price competition in stock i, making the slow quotes more likely to prevail and reducing η_{id}. On the other hand, since such news is not about stock i, its other impact—if any—on the fundamental demand and supply of i is likely to be minimum. This mitigates the concerns for the "unknown channels" and the "component channel."

2. Order flow characteristics are also controlled for, in order to further mitigate the concern of the "component channel." As shown in the equilibrium model in Section 3.1, the private value w_k arises from (fast) liquidity providers' inventory position, whose changes are the negative of order flows. When order flows become more one-sided, so will liquidity providers' inventory, making the private value w_k more persistent. Likewise, the persistence of the price distortion $(c_k - m_k)$ can also be driven by order flows, especially the "sniping" ones (Budish, Cramton, and Shim, 2015). When snipers eliminate stale quotes, they make the price distortion less persistent. To the extent that the persistence of w_k and that of $(c_k - m_k)$ are mainly driven by order flows, one can shut down this "component channel" by controlling for order flow characteristics.

   Specifically, the following three measures are computed for each stock-day and included as controls: (i) order flow volatility (standard deviation in $10,000); (ii) order flow persistence (first order autocorrelation); and (iii) absolute value of order imbalance (in dollars, as a ratio of the market capitalization). These are referred to as "Group I" controls.

3. To further mitigate concerns of the "unknown channels," a large number of other market activity statistics are included as controls. The idea is that the news shocks must impact ϕ via some market activity, like trading or quoting. Such activity always leave statistical traces in volume, volatility, bid-ask spread, order book depth, etc. By controlling such market activity, these "unknown channels" can therefore be closed.

   Specifically, these market activity characteristics include: (i) the other two structural parameters, λ and ψ; (ii) closing price; (iii) close-to-close holding-period return; (iv) time-weighted average quoted spread (in basis points relative to the midquote); (v) time-weighted average depth at best

quotes (in dollars); (vi) realized volatility (trade-by-trade, 1-minute, and 5-minute); (vii) trading volume (in dollar, as a ratio of the market capitalization); and (viii) minimum tick binding time (as a fraction of trading hours). These are the "Group II" controls.

4. To address concerns regarding trends, seasonality, and stock-specific patterns, a battery of fixed effects are included: (i) stock, (ii) week, (iii) month, (iv) day of week, and (v) industry (two-digit SIC). Also included are the lags of the control variables to control for trends.

Eventually, the above discussion motivates the following panel regression to examine Prediction 1:

$$\phi_{id} \sim \text{noNews}_{id} + \text{noNews}_{id} \times \#\text{News}_{-id} \text{ [+controls] [+fixed effects] [+lagged controls].} \tag{20}$$

The left-hand side variable ϕ_{id} is the price pressure persistence of stock i on day d, estimated from the structural model (1). The key right-hand side variable is the interaction term of

- noNews_{id}, a dummy equal to one if stock i has no news of its own on day d; and
- $\#\text{News}_{-id}$, the number of news on day d that are *not* about stock i.

The regression coefficient on this interaction term reflects the average impact of additional non-i news on a no-news day d for stock i. Table 3 presents the summary statistics of these two key variables. The count of $\#\text{News}_{-id}$ is measured in 1,000 news entries, which is approximately one standard deviation across the sample. Across the sample, there are about 78% of the stock-days that see no news about the stock itself. The count of others' news within these 78% stock-days are shown to be very similar to the unconditional sample, as seen in the last row.

The control variables include Group I and II discussed above, as well as their lags. Note that day fixed effects are *not* included because of multicollinearity. For example, all stocks that have no news on day d have the same interaction term $\text{noNews}_{id} \times \#\text{News}_{-id}$. The day-$d$ dummy, therefore, absorbs the effect of this interaction term.

				Percentiles						
	Unit	Mean	Std Dev	1%	5%	25%	50%	75%	95%	99%
noNews_{id}	Boolean	0.78	0.41	0.00	0.00	1.00	1.00	1.00	1.00	1.00
$\#\text{News}_{-id}$	1,000	2.54	1.06	0.63	1.31	1.80	2.19	3.22	4.78	5.47
$\#\text{News}_{-id} \mid \text{noNews}_{id}$	1,000	2.46	1.03	0.63	1.22	1.77	2.12	2.99	4.69	5.47

Table 3: Other stocks' news count. This table reports the summary statistics of the key right-hand side variables used in regression (20): noNews_{id} is a dummy equal to one if there is no news entry on day d about stock i; $\#\text{News}_{-id}$ is the number of news entries on day d that are *not* about stock i, counted in 1,000 entries.

4.3 Evidence for Prediction 1 (price pressure persistence)

Table 4 presents the results of the panel regression (20). Column (i) conveys the main finding through a simplest regression specification, without controls or fixed effects: When a stock i has no news on day d, for every 1,000 additional non-i news, its price pressure persistence ϕ_i *drops* by a quarter percentage points. This is consistent with the theory that "distracted" (fast) liquidity providers, subject to limited capacity or attention, quote less competitively. Such lack of competition makes lagged quotes more likely to prevail, reducing the price pressure persistence. The magnitude is statistically significant and economically meaningful. From Table 1(a), ϕ has a standard deviation of about 9.76 percentage points. Hence, the effect of other stocks' news is roughly 2.4% ($\approx$0.23/9.76), relative to the unconditional fluctuation in ϕ. Moving onto the other specifications, (ii)-(v) in Panel (a), it can be seen that this effect remains robust after introducing fixed effects, control variables, and their lags. The evidence is in strong support of Prediction 1.

Two further analyses are performed to provide additional evidence on the "weight channel" through the competitiveness η. First, consider two different stocks i and j. If neither of them has news on some day d, will their price pressure persistence ϕ_{id} and ϕ_{jd} be affected equally by other stocks' news? The answer probably depends on how "important" i is relative to j. If i is far more

	(a) Baseline				
	(i)	(ii)	(iii)	(iv)	(v)
noNews_{id}	-1.09***	-0.13	0.08	0.23	0.22
	(-2.67)	(-0.55)	(0.41)	(1.25)	(1.13)
$\text{noNews}_{id}\times\#\text{News}_{-id}$	-0.23**	-0.38***	-0.32***	-0.24***	-0.24***
	(-2.02)	(-4.97)	(-4.63)	(-3.92)	(-3.74)
Fixed effects	No	Yes	Yes	Yes	Yes
Group I controls	No	No	Yes	Yes	Yes
Group II controls	No	No	No	Yes	Yes
Three-day lagged controls	No	No	No	No	Yes

	(b) Stock size split			(c) News industry split		
	(vi)	(vii)	(viii)	(ix)	(x)	(xi)
noNews_{id}	0.04	0.19	0.17	0.08	0.23	0.22
	(0.21)	(1.05)	(0.90)	(0.41)	(1.25)	(1.13)
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isSmall}_i$	-0.37***	-0.29***	-0.30***			
	(-4.63)	(-3.86)	(-3.99)			
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isMedium}_i$	-0.34***	-0.24***	-0.23***			
	(-4.52)	(-3.66)	(-3.34)			
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isLarge}_i$	-0.24***	-0.17***	-0.15**			
	(-3.13)	(-2.59)	(-2.24)			
$\text{noNews}_{id}\times\#\text{News}_{-id}^{\neq\text{2digit-SIC}}$				-0.34***	-0.25***	-0.24***
				(-4.59)	(-3.83)	(-3.59)
$\text{noNews}_{id}\times\#\text{News}_{-id}^{\equiv\text{2digit-SIC}}$				0.09	-0.12	-0.23
				(0.19)	(-0.26)	(-0.51)
Fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Group I controls	Yes	Yes	Yes	Yes	Yes	Yes
Group II controls	No	Yes	Yes	No	Yes	Yes
Three-day lagged controls	No	No	Yes	No	No	Yes

Table 4: Effects of other stocks' news on price pressure persistence. This table shows the effects of other stocks' news on price pressure persistence ϕ under various regression specifications. The left-hand side variable is ϕ_{id} from the structural model estimates. The key right-hand side variable is the interaction between noNews_{id}, which is a dummy equal to one if there is no news entries related to stock i from RavenPack on day d; and $\#\text{News}_{-id}$, which is the total number of intraday news on day d that are *not* about stock i. Control group I includes order imbalance, order flow volatility, and order flow persistence. Control group II includes the other two structural parameters (λ and ψ), closing price, close-to-close return, bid-ask spread, depth, intraday volatility, volume, and minimum tick binding time. The fixed effects include dummies for stock, week, month day of week, and industry. The t-statistics reported in the brackets are based on stock-day double clustered standard errors. The superscripts *** and ** indicate statistical significance at 1% and at 5% respectively, based on two-sided t-tests.

important than j (e.g., i has a lot of trading activity unconditionally, while j has little), it is likely that η_{id} will be less affected than η_{jd}, because (fast) liquidity providers in i will pay attention to activity in i regardless of other stocks news. Those liquidity providers in j, on the other hand, might be more willing to divert their limited capacity to other news-struck stocks.

Table 4(b) exploits such "importance" heterogeneity across stocks by sorting the 400 sampled stocks into large, medium, and small according to whether they are in S&P 500 (large), S&P 400 (medium), or S&P 600 (small) indices. It amends regression (20) by splitting the interaction term $\text{noNews}_{id} \times \#\text{News}_{-id}$ with the size-tercile dummies. In other words, small, medium, and large stocks' ϕ reactions to other stocks' news are separately studied and compared. It can be seen that, regardless of the controls and the fixed effects, the drop in ϕ in small and medium stocks is always more salient than that in large stocks, consistent with the importance argument. Standard t-tests easily reject the null that the effects are equal between small and large stocks (or between medium and large stocks) at 99.9% confidence level.

Second, for a stock i that has no news on day d, does all non-i news affect ϕ_{id} equally? Different news' relevance to stock i is key. If the news is about a stock j that is remotely related to i, the liquidity providers in i should be more willing to divert some of their attention or capacity to stock j, as there is really nothing changed in i. On the other hand, if the news is also pertinent to i, then the liquidity providers in i should refrain from reallocating their capacity out of i.

Table 4(c) exploits such news relevance heterogeneity by decomposing the count of non-i news into those in the same industry of i and those different, using the companies' two-digit SIC codes: $\#\text{News}_{id} = \#\text{News}_{id}^{\neq \text{2digit-SIC}} + \#\text{News}_{id}^{\equiv \text{2digit-SIC}}$. The finding is robust across specifications that most of the drop in price pressure persistence ϕ_{id} arises only from distracting news from other industries. The effect of same-industry news is statistically indistinguishable from zero. This is again consistent with the news relevance argument. These additional evidence in Panel (b) and (c) lend additional support for Prediction 1.

	(a) 5-minute			(b) 1-minute			(d) Trade-by-trade		
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)
noNews_{id}	-3.35	-4.02*	-3.36	-4.14*	-4.87**	-4.16*	-4.09*	-4.86**	-4.11*
	(-1.44)	(-1.72)	(-1.45)	(-1.88)	(-2.20)	(-1.89)	(-1.79)	(-2.14)	(-1.81)
$\text{noNews}_{id}\times\#\text{News}_{-id}$	-3.00***			-3.11***			-2.89***		
	(-3.30)			(-3.60)			(-3.30)		
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isSmall}_{i}$		-3.76***			-3.86***			-3.83***	
		(-3.98)			(-4.24)			(-3.94)	
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isMedium}_{i}$		-3.18***			-3.41***			-2.97***	
		(-3.38)			(-3.91)			(-3.29)	
$\text{noNews}_{id}\times\#\text{News}_{-id}\times\text{isLarge}_{i}$		-1.63*			-1.61*			-1.34	
		(-1.74)			(-1.80)			(-1.58)	
$\text{noNews}_{id}\times\#\text{News}_{-id}^{\neq\text{2digit-SIC}}$			-3.15***			-3.33***			-3.18***
			(-3.35)			(-3.72)			(-3.51)
$\text{noNews}_{id}\times\#\text{News}_{-id}^{=\text{2digit-SIC}}$			1.19			3.23			5.51
			(0.28)			(0.73)			(1.04)
Fixed effects, controls, and lags	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 5: Effects of other stocks' news on realized volatility. This table shows the effects of other stocks' news on realized volatility under various regression specifications. The left-hand side variable is one of the realized volatility measure, as shown in Table 2. The key right-hand side variable is the interaction between noNews_{id}, which is a dummy equal to one if there is no news entries related to stock i from RavenPack on day d; and $\#\text{News}_{-id}$, which is the total number of intraday news on day d that are *not* about stock i. The fixed effects include dummies for stock, week, month day of week, and industry. The control variables include order imbalance, order flow volatility, and order flow persistence, the structural parameters (λ, ψ, and ϕ), closing price, close-to-close return, bid-ask spread, depth, intraday volatility, volume, and minimum tick binding time. The t-statistics reported in the brackets are based on stock-day double clustered standard errors. The superscripts ***, **, and * indicate statistical significance at 1%, 5%, and 10% respectively, based on two-sided t-tests.

4.4 Evidence for Prediction 2 (realized volatility)

For Prediction 2, the left-hand side variable in regression (20) is replaced with RV_{id}, which is one of the three realized volatility measures shown in Table 2:

$$\text{RV}_{id} \sim \text{noNews}_{id} + \text{noNews}_{id} \times \#\text{News}_{-id} \text{ [+controls] [+fixed effects] [+lagged controls].}$$

In addition, the realized volatility measures (vi) in Group II controls are removed and the price pressure persistence ϕ is added to (i).

Table 5 presents the findings for the three different realized volatility measures in Panels (a), (b), and (c) respectively. Columns (i), (iv), and (vii) report the baseline regression results. It can be seen that the key interaction term always has a significantly negative loading, strongly supporting Prediction 2. The effect is also economically sizable: A stock's realized volatility *drops* by about 3 basis points for every 1,000 news not about it. This is approximately 3% of the unconditional fluctuation (see the standard deviations in Table 2).

Columns (ii), (v), and (viii) report how small, medium, and large stocks' realized volatility respond differently to the same news distraction. The results indicate that the volatility drops are mainly from medium and small stocks, consistent with that the liquidity provbiders in large stocks do not reallocate their limited capacity upon other stocks' news. The effects on large stocks are both economically small (halving those of small and medium stocks) and statistically weak.

Finally, Columns (iii), (vi), and (ix) exploit the news' relevance and find that the drops in realized volatility exclusively arise from distracting news of stocks in other industries. This supports that a stock's liquidity providers only reallocate their capacity to news-struck stocks if those stocks are remotely related to the current stock. Indeed, the coefficients on same-industry news shocks are positive, suggesting that if the news is closely related, the realized volatility does increase, though statistically insignificantly.

5 Conclusion

This paper examine the price pressure component of intraday asset prices, especially how it is related to trades or order flows. Classic theories suggest that price pressure should move along the same direction as trades, because liquidity providers—limit order traders or quoters—require compensation to accommodate the order flow. But the data seems to suggest otherwise. Across a sample of 400 randomly selected stocks and under various structures of price pressure, the contemporaneous transitory price impact of trades is always negative. That is, market order buys

(sells) appear to reduce (increase) price pressure.

An equilibrium model is developed to explain this counterintuitive finding. The key novel model feature is that the fast liquidity provders are not always active or perfectly competitive and, when they are not, the midquotes do not immediately incorporate the latest information but reflect lagged "stale" prices. In such cases, the latent efficient price still moves in the same direction with the trade (permanent price impact). Yet, since the midquote is not immediately updated, the price pressure must offset the efficient price change, as if moving against the trade.

The model yields additional predictions regarding price pressure persistence and realized volatility. Perhaps surprisingly, when liquidity providers are more competitive, the model predicts more persistent price pressure and higher realized volatility. Using news from other stocks as (negative) shocks to the fast quoters' competition, the empirical evidence finds support for these predictions.

Appendix

A Estimating the structural model

There are three sets of parameters governing the structural model: the permanent price impact λ, the transitory price impact $\psi(L)$, and the price pressure persistence $\phi(L)$. They can be jointly estimated by generalized method of moments (GMM). Rearrange the price change between two trades Δp_k as

$$\Delta p_k = (1-L)(m_k + s_k) = \lambda x_k^* + \mu_k + \frac{1-L}{1-\phi(L)}(\psi(L)x_k + v_k) = \Delta \hat{s}_k + \lambda y_k^* + \mu_k + \frac{1-L}{1-\phi(L)} v_k,$$

where the last equality introduces a shorthand notation of

$$\Delta \hat{s}_k := \frac{1-L}{1-\phi(L)}\psi(L)x_k.$$

Note that $\Delta \hat{s}_k$ can be thought of as the projection of the true price pressure change Δs_k onto the order flows $\{x_k\}$. Subtracting $\Delta \hat{s}_k$ from Δp_k, one obtains a residual term that is only correlated with current order flow innovation x_k^* and uncorrelated with all past order flows x_{k-j}^* ($j \geq 1$). This leads

to sufficiently many moment conditions to identify $\{\lambda, \phi(L), \psi(L)\}$ as follows:

$$\mathbb{E}\left[\left(\Delta p_k - \Delta\hat{s}_k - \lambda x_k^*\right)x_k^*\right] = 0; \text{ and}$$
$$\mathbb{E}\left[(\Delta p_k - \Delta\hat{s}_k)x_{k-j}^*\right] = 0, \text{ for } j \geq 1.$$

To implement the above, the unobservable $\Delta\hat{s}_k$ needs to be be approximated recursively. For example, with $\phi(L) = \phi L$ and $\psi(L) = \psi$, then $\Delta\hat{s}_k = \psi \cdot (x_k - x_{k-1}) + \phi\Delta\hat{s}_{k-1}$, which can be iterated assuming initial values like $\Delta\hat{s}_0 = 0$ and $x_0 = 0$.

B Completing the general structural model

This appendix continues the derivation of the general structural model presented in Section 3.2. To recap, it has been shown that the midquote price p_k has the form (18) and the price pressure s_k has the form (19). There are four unobservable states underlying these expressions: $\{m_k, w_k, (c_k - m_k), z_k\}$. The observables are the trades $\{x_k\}$ and the midquotes $\{p_k\}$. Therefore, in order for an econometrician to be able to make association between these observables, additional structures are needed to tie these hidden states to trades $\{x_k\}$.

First, exactly like in (1), assume the efficient price m_k is assumed to follow a random walk whose increment has two components:

$$m_{k+1} = m_k + \lambda x_{k+1}^* + \mu_{k+1}. \tag{B.1}$$

The product λx_{k+1}^* captures trade-related effects, while the residual μ_{k+1} captures all other information unrelated to trades. This is also consistent with Equation (14) in the equilibrium model.

Second, the fast quoters' private value w_k is modeled as a moving average of

$$w_k = (\gamma_0 + \gamma(L))x_k + (1 + \varphi(L))\xi_k, \tag{B.2}$$

so that it is a stationary, zero-mean process. The right-hand side has a trade-related component $(\gamma_0 + \gamma(L))x_k$ and a non-trade residuals $(1 + \varphi(L))\xi_k$. The parameter γ_0 captures trades' contemporaneous impact on fast quoters' (marginal) private valuation, while $\gamma(L)$ allows possible lagged effects. In the equilibrium model in Section 3.1, the private value arises only from inventory costs with $w_k = -\gamma y_k$ (Proposition 1), which is a special case to the above with $\gamma(L) = 0$ and $\xi_k = 0$.

Third, similarly, one flexible representation for the price distortion of $(c_k - m_k)$ can be

$$(c_k - m_k) = (-\lambda + \delta(L))x_k^* + (1 + \vartheta(L))\epsilon_k. \tag{B.3}$$

As before, the price distortion is written into a trade-related component $(-\lambda + \delta(L))x_k^*$ and a non-trade component $(1 + \vartheta(L))\epsilon_k$. Recall that the order flow x_k has an autoregressive representation of

$(1 - A(L))x_k = x_k^*$. Therefore, (B.3) can be equivalently written as $(c_k - m_k) = (-\lambda + \hat{\delta}(L))x_k + (1 + \vartheta(L))\epsilon_k$, where $\hat{\delta}(L) = \lambda A(L) + (1 - A(L))\delta(L)$. The structure is flexible but embeds two economic restrictions on $\{c_k\}$, the center point of the slow quotes:

- As the slow crowds cannot update their $\alpha_k^{'}$ or $\beta_k^{'}$ during the interval $[t_k, t_{k+1})$, the center point $c_k = \frac{1}{2}\left(\alpha_k^{'} + \beta_k^{'}\right)$ does *not* respond to the trade x_k immediately. To reflect such slowness, the specification (B.3) restricts the contemporaneous effect of x_k on $(c_k - m_k)$ to be $-\lambda x_k^*$. That is, only $-m_k$ responds to x_k (according to B.1) but not c_k.
- Despite being slow, c_k should still track and not deviate unboundedly from m_k. That is, the moving average structures are assumed to make $(c_k - m_k)$ stationary.

In the equilibrium model in Section 3.1, Proposition 1 gives that $c_1 = \frac{1}{2}\left(\alpha_1^{'} + \beta_1^{'}\right) = v_0$. Hence, by Equations (13) and (14), the $t = 1$ distortion is $c_1 - m_1 = c_1 - v_1 = -\Delta v_1 = -\lambda x_1 - \mu_1$. This is a special case of (B.3) with $\delta(L) = \vartheta(L) = 0$ and $\epsilon_k = -\mu_k$.

Finally, the hidden $\{z_k\}$, capturing the zero-mean residuals in Equation (17), is assumed to be uncorrelated with trades $\{x_k\}$. This restriction effectively assumes that, on average, trades move the midquote p_k only through the efficient price m_k, the private value w_k, and the distortion $(c_k - m_k)$. In Section 3.1, Equation (15) implies that $z_1 = (F_1 - \eta)(\gamma x_1 + \Delta v_1)$, which is indeed uncorrelated with the trade x_1, because $(F_1 - \eta)$ is an independent white noise.

The structural model can now be completed. The specification (B.1) describes how the efficient price m_k relates to trades x_k. As $p_k = m_k + s_k$, it remains to express s_k in terms of trades x_k (and residual noises). By Equation (19), $s_k = \eta w_k + (1 - \eta)(c_k - m_k) + z_k$, or

$$s_k = \eta \cdot (\gamma_0 + \gamma(L))x_k + (1 - \eta)(-\lambda + \delta(L))x_k^* + [\eta \cdot (1 + \varphi(L))\xi_k + (1 - \eta)(1 + \vartheta(L))\epsilon_k + z_k].$$

Substituting in the autoregressive form of the order flow $x_k^* = (1 - A(L))x_k$, the price pressure s_k can be seen as the sum of a trade-related component

$$\begin{aligned} &\eta \cdot (\gamma_0 + \gamma(L)) + (1 - \eta)(-\lambda + \delta(L))(1 - A(L)) \\ &= (\eta\gamma_0 - (1 - \eta)\lambda)x_k + [\eta\gamma(L) + (1 - \eta)(\lambda A(L) + \delta(L)(1 - A(L)))]x_k =: \left(\psi_0 + \hat{\psi}(L)\right)x_k \end{aligned} \tag{B.4}$$

and a non-trade component

$$\eta \cdot (1 + \varphi(L))\xi_k + (1 - \eta)(1 + \vartheta(L))\epsilon_k + z_k =: \hat{v}_k. \tag{B.5}$$

Thanks to the stationarity of w_k and $(c_k - m_k)$ assumed through $1 + \varphi(L)$ and $1 + \vartheta(L)$, therefore, the non-trade component (B.5) is also stationary. Hence, by Wold theorem, it has an equivalent autoregressive representation of $(1 - \phi(L))\hat{v}_k = v_k$ for some white noise v_k. Therefore, summing

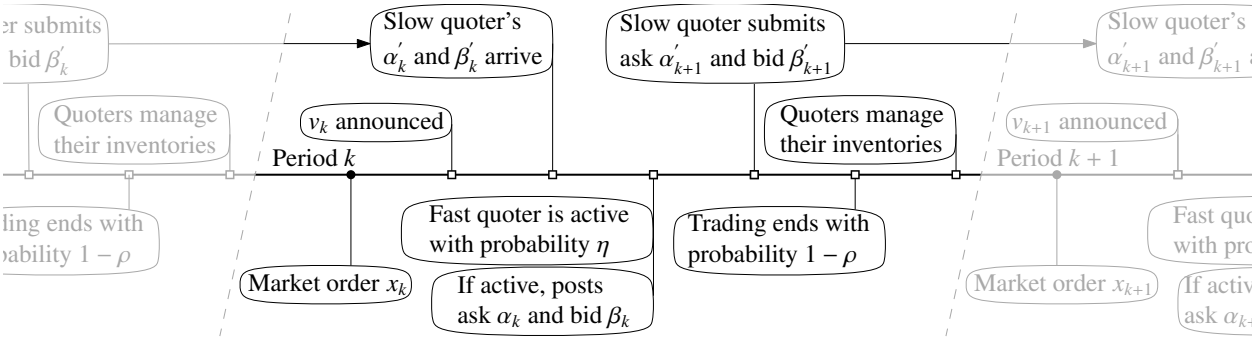

Figure 3: Timeline of the dynamic equilibrium model. This graph illustrates a typical period k. The period begins with the k-th market order x_k, followed by a public announcement of v_k. After the slow quoter's quotes—submitted in the previous period—arrive, the fast quoter submits new quotes if active. The slow quoter then submits for the next period. Finally, both quoters manage their inventories.

up (B.4) and (B.5), $s_k = (\psi_0 + \hat{\psi}(L))x_k + (1-\phi(L))^{-1}v_k$, or equivalently,

$$(1-\phi(L))s_k = (\psi_0 + \psi(L))x_k + v_k, \tag{B.6}$$

where $\psi(L) := -\psi_0\phi(L) + (1-\phi(L))\hat{\psi}(L)$. Combining the efficient price (B.1) and the pricing error (B.6), one can see that the derivation has exactly replicated the framework (1).

C A dynamic equilibrium model

This Appendix derives a dynamic equilibrium, built on the two-trade model studied in Section 3.1. The purpose is to provide an analytical foundation for Predictions 1 and 2, which have only been argued intuitively before.

Model setup. The setup follows Section 3.1.1. Figure 3 outlines the timing of events in a typical period. Compared to period 1 in Figure 1, there are two modifications. First, toward the end of each period, there is probability $1-\rho \in (0,1)$ that the trading ends, upon which the asset is liquidated at v_k and all agents consume. Only with probability ρ, the game continues to the next period.

Second, the quoters' inventory processes are enriched. Consider the fast quoter's inventory y_k for example. She bears inventory cost whenever $y_k \neq 0$. Therefore, she has incentive to, for example, to trade in related assets or through derivatives to reduce the net inventory exposure. Specifically, the following reduced-form is assumed to reflect such out-of-the-model hedging activity:

$$y_{k+1} = \varphi y_k - F_k x_{k+1},$$

where y_k is the inventory level right after the k-th market order, F_k is an indicator equal to one if

the fast quoter is active, and x_k is the market order. In words, the above says that the fast quoter's inventory decays, in between periods, at a rate of $\varphi \in (0,1)$, which reflects the (lack of) hedging efficiency. If the fast quoter is active in this period ($F_k = 1$), she accommodates the next order x_{k+1} and accumulates $-x_{k+1}$ units of the asset. Note that the parameter φ is not the price pressure persistence ϕ, which will be shown as an endogenous function of φ (and the two are very related).[10]

Likewise, the slow quoter's inventory management is modeled as $y'_{k+1} = \varphi' y'_k - (1 - F_k)x_{k+1}$, where $\varphi' \in (0,1)$ reflects the slow quoter's (lack of) hedging efficiency. It is further assumed that φ' is sufficiently lower than φ. This is an important assumption for the results on price pressure persistence and volatility below. The economic motivation is that the slow quoter represents a large pool of long-term investors with diverse intrinsic demand that aggregate to zero. As such, changes in y'_k can be quickly redistributed to those with appropriate needs across the pool. The remaining "unwanted" inventory position, as reflected in y'_k, thus mean reverts quickly to zero. (The key difference with the fast quoter is that the fast represents professional market makers who do not have intrinsic demand for the asset.) For simplicity, it shall be assumed that $\varphi' \to 0$ to facilitate the calculation of price pressure persistence and realized volatility.

The above autoregressive structures ensure that the quoters' inventories are bounded: $\hat{y} := \sup |y_k| < \infty$ and $\hat{y}' = \sup |y'_k| = 1$ for the fast and the slow, respectively. (Specifically, $\hat{y} = 1/(1-\varphi)$ as $F_k x_{k+1} \in \{-1, 0, 1\}$.) The following assumption

$$\text{(C.1)} \qquad \kappa > 1 + \frac{(1-\rho)\gamma'}{2\sigma} \quad \text{and} \quad \gamma' > \frac{2\sigma}{1-\rho} + \frac{1 + 2\varphi\hat{y}}{1 - \rho\varphi^2}\gamma$$

replaces assumption (3) to ensure that (i) the market order trader's private value is sufficiently large so that there is always trade in every period; and (ii) the fast quotes are tighter than the slow ones so that the price is alternating between the two.

Equilibrium. The equilibrium is very similar to the two-trade game and is given by the following proposition. The detailed analysis is deferred to its proof.

Proposition 2 (The dynamic equilibrium). *There is an equilibrium, under assumption* (C.1)*, where the slow quoter submits limit orders at $k-1$ according to*

$$\alpha'_k = v_{k-1} + \tau\sigma + \frac{\delta'\gamma'}{2} \quad \textit{and} \quad \beta'_k = v_{k-1} - \tau\sigma - \frac{\delta'\gamma'}{2}, \quad \textit{with} \quad \delta' = 1 - \rho;$$

[10] The main motivation for such a specification is to ensure that the inventory y_k is bounded. Indeed, without any hedging, i.e., when $\varphi = 1$, y_k would behave like a random walk with increment $-F_k x_k \in \{-1, 0, 1\}$. Then there is non-zero probability that the inventory will be too one-sided. In such cases, the the fast quoter's bid and ask will be tilted to that side, so much so that market order traders on this same side will not be able to trade. For example, when $y_k \to \infty$, the fast quoter will never want to buy, hence no market sell orders will be accommodated. In other words, there will always be some uninteresting periods with no trade. By imposing the autoregressive structure above, y_k is bounded. Together with assumption C.1 below, such uninteresting no-trade cases are eliminated.

the fast quoter posts limit orders at k according to

$$\alpha_k = v_k + \tau\sigma + \frac{\delta\gamma}{2} - \delta\gamma\varphi y_k \;\; \textit{and} \;\; \beta_k = v_k - \tau\sigma - \frac{\delta\gamma}{2} - \delta\gamma\varphi y_k, \;\; \textit{with} \;\; \delta' = \frac{1-\rho}{1-\rho\varphi^2};$$

and the period-k $\in \{1, 2\}$ *market order is* $x_k = u_k$.

Compared to Proposition 1, the key difference is that the markup (markdown) in the ask (bid) due to inventory is now scaled down with $\delta \in (0, 1)$ (for the fast, and δ' for the slow). This is because of the probability ρ of continuing to the next period and, in that cast, the inventory decay φ (for the fast, and φ' for the slow). Indeed, the scalar δ is decreasing in ρ but increasing in φ.

The midquote p_k and the price pressure s_k. By Proposition 2, the midquote follows to be

$$p_k = \frac{F_k}{2}(\alpha_k + \beta_k) + \frac{1-F_k}{2}\left(\alpha_k' + \beta_k'\right) = F_k \cdot (v_k - \delta\gamma\varphi y_k) + (1-F_k)v_{k-1}. \tag{C.2}$$

Subtracting the efficient price $m_k = v_k$ from p_k yields the price pressure

$$\begin{aligned} s_k =& p_k - v_k = -\delta\gamma\varphi F_k y_k - (1-F_k)\Delta v_k \\ =& \eta \cdot (-\delta\gamma\varphi y_k) + (1-\eta)(-\Delta v_k) + (F_k - \eta)(-\delta\gamma\varphi y_k - \Delta v_k) \end{aligned}$$

where the second line is a special case of (19) with $w_k = -\delta\gamma\varphi y_k$, $c_k - m_k = -\Delta v_k$, and the rest as the residual z_k. Note that

$$\mathbb{E}\left[\frac{\partial w_k}{\partial x_k}\Big| x_k, x_{k-1}, ...\right] = -\delta\gamma\varphi\mathbb{E}[-F_{k-1}] = \delta\gamma\varphi\eta =: \gamma_0 \tag{C.3}$$

is the (expected) contemporaneous impact of x_k on the fast quoter's private value, as assumed in (B.2). One can also rewrite s_k into the structural representation as in (1), where it has three components: the contemporaneous effect of x_k, the lagged effects of $\{x_{k-1}, x_{k-2}, ...\}$, and the residual uncorrelated with $\{x_k\}$. To do so, convert y_k to its moving average representation:

$$y_k = \varphi y_{k-1} - F_{k-1}x_k = -\eta x_k - \frac{\eta\varphi x_{k-1}}{1-\varphi L} - \frac{(F_{k-1} - \eta)x_k}{1-\varphi L}$$

where the last term is uncorrelated with order flows $\{x_k\}$. Also recall that $\Delta v_k = \lambda x_k + \mu_k$, where μ_k is uncorrelated with x_k. Plug these expressions into s_k above to get

$$s_k = \underbrace{(\eta\gamma_0 - (1-\eta)\lambda)}_{:=\psi_0} x_k + \underbrace{\frac{\eta\gamma_0\varphi L}{1-\varphi L}}_{=:\psi(L)} x_k + \nu_k \tag{C.4}$$

where ν_k collects all terms that are uncorrelated with $\{x_k\}$. This is a special case of the general s_k structure assumed in (1).

Price pressure persistence. The simplest way to examine the persistence of s_k is to study its response to an impulse in the order flow x_k. Suppose by period $k-1$ the system has entered a steady state with $s_{k-1} = 0$. An impulse of $x_k = 1$ is given to the system. It can be seen from the structural representation (C.4) that the immediate response at period k is $\psi_0 = \eta\gamma_0 - (1-\eta)\lambda$, while the $k+1$ response is $\eta\gamma_0\varphi$. Subsequently, the response decays at a constant rate of φ. Therefore, the persistence of s_k is determined by the first-period decay, which is at the rate of

$$\frac{\eta\gamma_0\varphi}{\eta\gamma_0 - (1-\eta)\lambda} = \frac{\delta\gamma\varphi^2\eta^2}{(\delta\gamma\varphi\eta + \lambda)\eta - \lambda},$$

which is monotone increasing in η. This confirms the intuition argued for Prediction 1 that the persistence of s_k can be roughly thought as a weighted average between the persistence of w_k and that of $c_k - m_k$. Indeed, the above first-period decay rate is zero when $\eta = 0$ (all the weights allocated to $c_k - m_k = \Delta v_k$, which has zero persistence) and increases to φ when $\eta = 1$ (all the weights allocated to $w_k \propto y_k$, which has persistence φ).

Volatility. To examine the volatility (or variance) $\mathrm{var}[\Delta p_k]$, let us first examine the price return, which, following Equation (C.2), can be written as

$$\Delta p_k = F_k \Delta v_k + (1 - F_{k-1})\Delta v_{k-1} - \gamma_0 \cdot (F_k y_k - F_{k-1} y_{k-1})$$

where γ_0 is proportional to η as defined in (C.3). This expression highlights the key intuition: The price return is a combination of the current efficient price change Δv_k, the change in the slow quotes $\Delta v_{k-1} (= \Delta c_k)$, and the change in the fast quoter's private value (reflected through the change in the inventory). As η increases, F_k (and F_{k-1}) is more likely to be one, which adds to the weight of Δv_k, increasing return volatility. But this is exactly offset by the decrease in the weight of Δv_{k-1}. Indeed, $\mathrm{var}[F_k \Delta v_k + (1 - F_{k-1})\Delta v_{k-1}] = \eta\sigma^2 + (1-\eta)\sigma^2 = \sigma^2$ is unaffected by η. The net effect of η on return volatility arises from the last term, i.e., the change of fast quoter's private value, which is scaled by $\gamma \propto \eta$. Hence, generally, when η increases, this last effect balloons, raising the return volatility. Proving the general result, however, is challenging, because $\mathrm{var}[\Delta v_k]$ after careful evaluation turns out to be a polynomial in η of degree 5:

$$\mathrm{var}[\Delta v_k] = \sigma^2 + 2\varphi\lambda\gamma_0\eta^2 + 2\left(\lambda + \frac{\gamma_0}{1-\varphi^2}\right)(1 - \varphi\eta)\gamma_0\eta^2$$

where $\gamma_0 = \delta\gamma\varphi\eta$. Nevertheless, the polar cases of $\eta \in \{0, 1\}$ can be compared easily: $\mathrm{var}[\Delta p_k]|_{\eta=1} = \sigma^2 + 2\gamma_0^2/(1+\varphi) + 2\gamma_0\lambda > \sigma^2 = \mathrm{var}[\Delta p_k]|_{\eta=0}$. Therefore, by continuity, there exists $\hat{\eta} > 0$ such that $\mathrm{var}[\Delta p_k]$ is monotone increasing locally in the range of $\eta \in (0, \hat{\eta})$, reflecting the aforementioned intuition. In extensive numerical exercise, it has been found that $\hat{\eta} \geq 1$ for a large range of parameters. In such cases, the monotonicity is everywhere for $\eta \in (0, 1)$.

D Proofs

Proposition 1

Proof. The proof proceeds in three steps: the slow quotes, the fast quotes, and then the market orders. The first two steps only consider the asks, omitting the symmetric analysis for the bids.

Slow quotes. Equation (8) simplifies to $\alpha_1^{'} = \mathbb{E}\left[v_2 \middle| v_1 + q_2 > \alpha_1^{'}\right] + \frac{\gamma^{'}}{2}$. Note that $v_1 + q_2$ can take four possible values of $\{v_0 \pm \sigma \pm (\tau + \kappa)\sigma\}$, which can be sorted as

$$v_0 + \sigma + (\tau + \kappa)\sigma > v_0 - \sigma + (\tau + \kappa)\sigma > v_0 + \sigma - (\tau + \kappa)\sigma > v_0 - \sigma - (\tau + \kappa)\sigma,$$

because assumption (3) implies $\tau + \kappa > 1$. Hence, the choice of $\alpha_1^{'}$ can be split into five regions. Below the inference of $\mathbb{E}[v_2 \big| v_1 + q_2 > \alpha_1^{'}]$ is examined for each range of $\alpha_1^{'}$ and then the implied $\alpha^{'} = \mathbb{E}[v_2 \big| v_1 + q_2 > \alpha_1^{'}] + \frac{\gamma^{'}}{2}$ is derived.

(i) $\alpha_1^{'} > v_0 + \sigma + (\tau + \kappa)\sigma$: This contradicts with the conditioning event of $v_1 + q_2 > \alpha_1^{'}$, as the maximum of $v_1 + q_2$ is $v_0 + \sigma + (\tau + \kappa)\sigma$. Hence, $\alpha_1^{'}$ cannot be in this range, in equilibrium.

(ii) $v_0 + \sigma + (\tau + \kappa)\sigma \geq \alpha_1^{'} > v_0 - \sigma + (\tau + \kappa)\sigma$: When $\alpha_1^{'}$ is in this range, $v_1 + q_2 > \alpha_1^{'}$ necessarily implies that $v_1 = v_0 + \sigma$ and $u_2 = 1$. From Equation (5), $\mathbb{E}[\Delta v_2 | \, u_2 = 1] = \tau\sigma$. It follows that $\alpha_1^{'} = v_0 + \sigma + \tau\sigma + \frac{\gamma^{'}}{2}$.

(iii) $v_0 - \sigma + (\tau + \kappa)\sigma \geq \alpha_1^{'} > v_0 + \sigma - (\tau + \kappa)\sigma$: When $\alpha_1^{'}$ is in this range, $v_1 + q_2 > \alpha_1^{'}$ implies that $\{\Delta v_1 = \pm\sigma, u_2 = 1\}$. Again it follows that $\mathbb{E}[\Delta v_2 | \, u_2 = 1] = \tau\sigma$. Since Δv_1 is orthogonal to u_2, the above implies that $\alpha_1^{'} = v_0 + \mathbb{E}[\Delta v_1] + \mathbb{E}[\Delta v_2 | \, u_2 = 1] + \frac{\gamma^{'}}{2} = v_0 + \tau\sigma + \frac{\gamma^{'}}{2}$.

(iv) $v_0 + \sigma - (\tau + \kappa)\sigma \geq \alpha_1^{'} > v_0 - \sigma - (\tau + \kappa)\sigma$: When $\alpha_1^{'}$ is in this region, it might be executed when $\{\Delta v_1 = \sigma, u_2 = 1\}$, when $\{\Delta v_1 = -\sigma, u_2 = 1\}$, or when $\{\Delta v_1 = \sigma, u_2 = -1\}$, each with unconditional probability $\frac{1}{4}$. By Equation (5), $\mathbb{E}[\Delta v_2 | \, u_2 = \kappa] = \tau\sigma$ and $\mathbb{E}[\Delta v_2 | \, u_2 = -\kappa] = -\tau\sigma$. Therefore, in this range, $\mathbb{E}[\Delta v_1 \big| v_1 + q_2 > \alpha_1^{'}] = (\frac{1}{4}\sigma - \frac{1}{4}\sigma + \frac{1}{4}\sigma)/(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}) = \frac{1}{3}\sigma$; and $\mathbb{E}[\Delta v_2 \big| u_2 + v_1 > \alpha_1^{'}] = (\frac{1}{4}\tau\sigma + \frac{1}{4}\tau\sigma - \frac{1}{4}\tau\sigma)/(\frac{1}{4} + \frac{1}{4} + \frac{1}{4}) = \frac{1}{3}\tau\sigma$. It follows that $\alpha_1^{'} = v_0 + \frac{1}{3}(1 + \tau)\sigma + \frac{\gamma^{'}}{2}$.

(v) $v_0 - \sigma - (\tau + \kappa)\sigma \geq \alpha_1^{'}$: When $\alpha_1^{'}$ is in this range, $u_2 + v_1 > \alpha_1^{'}$ always happens and there is no learning at all. It follows that $\mathbb{E}[\Delta v_1 + \Delta v_2 \big| u_2 + v_1 > \alpha_1^{'}] = 0$ and $\alpha_1^{'} = v_0 + \frac{\gamma^{'}}{2}$.

Region (i) has been ruled out. The $\alpha_1^{'}$ implied by each of the four remaining regions must be compatible with the respective support. For (iv), the implied $\alpha_1^{'} > v_0$ but this is incompatible with the upper bound of the support because $(\tau + \kappa) > 1$ by assumption (3). Again, the (v)-implied $\alpha_1^{'} > v_0$ and this is not compatible with the upper bound of the support. Both (ii) and (iii) are possible. However, under the implicit Bertrand competition, the lower ask implied by (iii) will prevail. Therefore, the equilibrium $\alpha_1^{'}$ is the one in Region (iii), i.e., $\alpha_1^{'} = v_0 + \tau\sigma + \frac{\gamma^{'}}{2}$.

Fast quotes. The break-even condition (9) simplifies to $\alpha_k = \mathbb{E}[v_2 | v_k + q_{t+1} > \alpha_k, v_k] + \frac{\gamma}{2} - \gamma y_k$. Note that given v_k, $v_k + q_{k+1}$ can take two possible values of $\{v_k \pm (\tau + \kappa)\sigma\}$. These two values cut the possible values of α_k into three regions.

(i) $\alpha_k > v_k + (\tau + \kappa)\sigma$: This contradicts with the conditioning event of $v_k + q_{k+1} > \alpha_k$. Hence, the equilibrium α_k cannot be in this range.

(ii) $v_k + (\tau + \kappa)\sigma \geq \alpha_k > v_k - (\tau + \kappa)\sigma$: In this range, $v_k + q_{k+1} > \alpha_k$ necessarily implies $u_{k+1} = 1$. By Equation (5), therefore, $\mathbb{E}[\Delta v_{k+1} | u_{k+1} = 1] = \tau\sigma$, implying $\alpha_k = v_k + \tau\sigma + \frac{\gamma}{2} - \gamma y_k$.

(iii) $\alpha_k \leq v_k - (\tau + \kappa)\sigma$: In this range, $v_k + q_{k+1} > \alpha_k$ always happens. Then there is no learning about Δv_{k+1}. As a result, $\alpha_k = v_k + \frac{\gamma}{2} - \gamma y_k$.

Region (i) has been ruled out. The region (ii) implied α_k is indeed admissible in the support of (ii) under assumption (3). For (iii) to support an equilibrium, it requires $\alpha_k = v_k + \frac{\gamma}{2} - \gamma y_k \leq v_k - (\tau + \kappa)\sigma$, or $\kappa \leq -\frac{\gamma}{2\sigma} + \frac{\gamma}{\sigma} y_t - \tau$. Since $y_t \in \{-1, 0, 1\}$ in this two-trade game, the least binding requirement would be $\kappa \leq \frac{\gamma}{2\sigma} - \tau$, which violates assumption (3). To sum up, the equilibrium α_k can only be in Region (ii), i.e., $\alpha_k = v_k + \tau\sigma + \frac{\gamma}{2} - \gamma y_k$.

Market orders. Given the above equilibrium quotes, the optimal market order x_k in (7) can be simplified. In particular, under assumption (3), it can be easily verified that $u_k = 1$ implies $v_{k-1} + q_k > a_{k-1}$ and $u_k = -1$ implies $v_{k-1} + q_k < b_{k-1}$, for any equilibrium ask a_{k-1} or bid b_{k-1} stated in the proposition. That is, $x_k = u_k$. □

Proposition 2

Proof. To begin with, conjecture that the market order follows $x_k = u_k$, which will be verified later. From the quoters' perspective, this means x_k is equally likely to be a buy or a sell.

Consider the fast quoter's continuation value J_k, right after the k-th trade. There are two state variables, the fundamental v_k and the inventory y_k. Because of competition, the fast quoter is indifferent between trading or not. Hence, her value function should satisfy the following Bellman equation:

$$J(v_k, y_k) = (1 - \rho)\Big(y_k v_k - \frac{\gamma}{2} y_k^2\Big) + \rho \cdot (1 - \varphi) y_k \mathbb{E}[v_{k+1} | v_k] + \rho \mathbb{E}[J(v_{k+1}, \varphi y_k) | v_k],$$

which is a sum of three components in the sequence of: (i) the terminal payoff, if the game ends by this period (probability $1 - \rho$); (ii) the proceeds from inventory management, if continues to the next period (probability ρ); (iii) the expected value function next period, when she does nothing this period. To solve for $J(\cdot)$, guess that $J(v, y) = A + Bvy - Cy^2$ for some coefficients $\{A, B, C\}$. Plug these into the above Bellman equation, equate the coefficients on the y polynomial, and jointly

solve for $\{A, B, C\}$ to get

$$J(v, y) = vy - \frac{\delta\gamma}{2}y^2, \text{ with } \delta = \frac{1-\rho}{1-\rho\varphi^2} \in (0, 1).$$

The competitive quotes can be pinned down by the indifference condition. For example, the ask α_k must satisfy

$$\alpha_k + \mathbb{E}[J(v_{k+1}, \varphi y_k - 1)|\, v_k, x_{k+1} = 1\,] = \mathbb{E}[J(v_{k+1}, \varphi y_k)|\, v_k, x_{k+1} = 1\,].$$

A similar expression holds for the bid β_k and is omitted. Using the above solved value function $J(\cdot)$, it can be solved that

$$\alpha_k = v_k + \lambda + \frac{\delta\gamma}{2} - \delta\gamma\varphi y_k \text{ and } \beta_k = v_k - \lambda - \frac{\delta\gamma}{2} - \delta\gamma\varphi y_k,$$

as stated in the proposition. (Note that under the conjecture of $x_{k+1} = u_{k+1}$, $\mathbb{E}[v_{k+1}|\, v_k, x_{k+1}\,] = v_k + \lambda x_{k+1}$, following Equation 5 and 12.)

Next consider the slow quoter. The analysis is almost the same, except that in period k the slow quoter submits $\{\alpha'_{k+1}, \beta'_{k+1}\}$ to be executed with x_{k+2}. Her Bellman equation is simply

$$J'(v_k, y'_k) = (1-\rho)\left(y'_k v_k - \frac{\gamma'}{2}{y'_k}^2\right) + \rho \cdot (1-\varphi')y'_k\mathbb{E}[v_{k+1}|\, v_k\,] + \rho\mathbb{E}\Big[J'(v_{k+1}, \varphi y'_k)|\, v_k\Big].$$

or, solved by conjecturing that $J'(\cdot)$ is quadratic in y'_k,

$$J'(v, y) = vy - \frac{\delta'\gamma'}{2}y^2, \text{ with } \delta' = \frac{1-\rho}{1-\rho{\varphi'}^2} \in (0, 1).$$

The indifference condition implies

$$\alpha'_{k+1} + \mathbb{E}\Big[J'\Big(v_{k+2}, \varphi' y_{k+1} - 1\Big)|\, v_k, x_{k+2} = 1\Big] = \mathbb{E}\Big[J'\Big(v_{k+2}, \varphi' y_{k+1}\Big)|\, v_k, x_{k+2} = 1\Big].$$

Note that $y_{k+1} = \varphi' y_k - (1 - F_{k+1})x_{k+1}$ is also stochastic. A similar condition applies to β'_k. Using the above value function $J'(\cdot)$, it follows that

$$\alpha'_{k+1} = v_k + \lambda + \frac{\delta'\gamma'}{2} - \delta'\gamma'\varphi'^2 y'_k \text{ and } \beta'_{k+1} = v_k - \lambda - \frac{\delta'\gamma'}{2} - \delta'\gamma'\varphi'^2 y'_k,$$

again as stated in the proposition. Note that under assumption C.1, $\alpha'_k > \alpha_k$ and $\beta'_k < \beta_k$, i.e., the fast quotes, if exist, are tighter.

Finally, the conjecture of $x_{k+1} = u_{k+1}$ needs to be verified. It suffices to check that the market order trader's valuation $v_k + q_{k+1}$ (see Equation 4-6) is higher than α'_k when $u_{k+1} = 1$ and lower than β'_k when $u_{k+1} = -1$. This is indeed the case under assumption C.1. The proof is thus complete. □

References

Amihud, Yakov and Haim Mendelson. 1980. "Dealership markets: market making with inventory." *Journal of Financial Economics* 8 (1):31–53.

Barber, Brad M. and Douglas Loeffler. 1993. "The "Dashboard" Column: Second-Hand Information and Price Pressure." *Journal of Financial and Quantitative Analysis* 28 (2):273–284.

Ben-Rephael, Azi, Shmuel Kandel, and Avi Wohl. 2011. "The Price Pressure of Aggregate Mutual Fund Flows." *Journal of Financial and Quantitative Analysis* 46 (2):585–603.

Berman, Gregg E. 2014. "What Drives the Complexity and Speed of our Markets?" Speech.

Brennan, Michael J. and Avanidhar Subrahmanyam. 1996. "Market Microstructure and Asset Pricing: On the Compensation for Illiquidity in Stock Returns." *Journal of Financial Economics* 41:441–464.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan. 2014. "High Frequency Trading and Price Discovery." *The Review of Financial Studies* 27 (8):2267–2306.

Bruche, Max and John C.F. Kuong. 2019. "Dealer Funding and Market Liquidity." Working paper.

Budish, Eric, Peter Cramton, and John Shim. 2015. "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response." *Quarterly Journal of Economics* 14 (3):1547–1621.

Chao, Yong, Chen Yao, and Mao Ye. 2017. "Discrete Pricing and Market Fragmentation: a Tale of Two-Sided Markets." *American Economic Review: Papers and Proceedings* 107 (5):196–199.

Chordia, Tarun, T. Clifton Green, and Badrinath Kottimukkalur. 2018. "Rent Seeking by Low-Latency Traders: Evidence from Trading on Macroeconomic Announcements." *The Review of Financial Studies* 31 (12):4650–4587.

Corwin, Shane A. and Jay F. Coughenour. 2008. "Limited Attention and the Allocation of Effort in Securities Trading." *The Journal of Finance* 63 (6):3031–3067.

Coval, Joshua and Erik Stafford. 2007. "Asset Fire Sales (and Purchases) in Equity Markets." *Journal of Financial Economics* 86 (2):479–512.

Duffie, Darrell. 2010. "Presidential Address: Asset Price Dynamics with Slow-Moving Capital." *The Journal of Finance* 65 (4):1237–1267.

Foucault, Thierry. 1999. "Order Flow Composition and Trading Costs in a Dynamic Limit Order Market." *Journal of Financial Markets* 2 (2):99–134.

Gibson, Scott, Assem Safieddine, and Sheridan Titman. 2000. "Tax-Motivated Trading and Price Pressure: An Analysis of Mutual Fund Holdings." *Journal of Financial and Quantitative Analysis* 35 (3):369–386.

Glosten, Lawrence R. 1994. "Is the Electronic Limit Order Book Inevitable?" *The Journal of Finance* 49 (4):1127–1161.

Glosten, Lawrence R. and Lawrence E. Harris. 1988. "Estimating the Components of the Bid-Ask Spread." *Journal of Financial Economics* 21:123–142.

Glosten, Lawrence R. and Paul R. Milgrom. 1985. "Bid, Ask, and Transaction Prices in a Specialist Market with Heterogeneously Informed Agents." *Journal of Financial Economics* 42 (1):71–100.

Greenwood, Robin and Dimitri Vayanos. 2010. "Price Pressure in the Government Bond Market." *American Economic Review: Papers & Proceedings* 100:585–590.

———. 2014. "Bond Supply and Excess Bond Returns." *The Review of Financial Studies* 27 (3):663–713.

Grossman, Sanford J. and Merton H. Miller. 1988. "Liquidity and Market Structure." *The Journal of Finance* 43 (3):617–633.

Grossman, Sanford J. and Joseph E. Stiglitz. 1980. "On the Impossibility of Informationally Efficient Markets." *American Economic Review* 70 (3):393–408.

Harris, Lawrence and Eitan Gurel. 1986. "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures." *The Journal of Finance* 41 (4):815–829.

Hasbrouck, Joel. 2007. *Empirical Market Microstructure: the Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, New York.

Hasbrouck, Joel and George Sofianos. 1993. "The Trades of Market Makers: An Empirical Analysis of NYSE Specialists." *The Journal of Finance* 48:1565–1593.

Hellwig, Martin F. 1980. "On the Aggregation of Information in Competitive Markets." *Journal of Economic Theory* 22 (3):477–498.

Hendershott, Terrence and Albert J. Menkveld. 2014. "Price pressures." *Journal of Financial Economics* 114 (3):405–423.

Hendershott, Terrence and Mark S. Seasholes. 2007. "Market Maker Inventories and Stock Prices." *American Economic Review: Papers and Proceedings* 97:210–214.

Ho, Thomas and Hans R. Stoll. 1981. "Optimal Dealer Pricing under Transaction Cost and Return Uncertainty." *Journal of Financial Economics* 9 (1):47–73.

———. 1983. "The Dynamics of Dealer Markets Under Competition." *The Journal of Finance* 38:1053–1074.

Jin, Li. 2006. "Capital Gains Tax Overhang and Price Pressure." *The Journal of Finance* 61 (3):1399–1431.

Jovanovic, Boyan and Albert J. Menkveld. 2015. "Dispersion and Skewness of Bid Prices." Working paper.

———. 2019. "Equilibrium Bid-Price Dispersion." Working paper.

Kraus, Alan and Hans Stoll. 1972. "Price Impacts of Block Trading on the New York Stock Exchange." *The Journal of Finance* 27 (3):569–588.

Kyle, Albert S. 1989. "Informed Speculation with Imperfect Competition." *Review of Economic Studies* 56 (3):317–356.

Lee, Charles M. and Mark J. Ready. 1991. "Inferring Trade Direction from Intraday Data." *The*

Journal of Finance 46 (2):733–746.

Liu, Lily Y., Andrew J. Patton, and Kevin Sheppard. 2015. "Does anything beat 5-minute RV? A comparison of realized measures across multiple asset classes." *Journal of Econometrics* 187 (1):293–311.

Madhavan, Ananth, Matthew Richardson, and Mark Roomans. 1997. "Why Do Security Prices Change? A Transaction-Level Analysis of NYSE Stocks." *The Review of Financial Studies* 10 (4):1035–1064.

Madhavan, Ananth and Seymour Smidt. 1991. "A Bayesian model of intraday specialist pricing." *Journal of Financial Economics* 30:99–134.

———. 1993. "An Analysis of Changes in Specialist Inventories and Quotations." *The Journal of Finance* 48 (5):1595–1628.

Madhavan, Ananth and G. Sofianos. 1998. "An Empirical Analysis of NYSE Specialist Trading." *Journal of Financial Economics* 48 (2):189–210.

Menkveld, Albert J. 2013. "High Frequency Trading and the *New Market* Makers." *Journal of Financial Markets* 16 (4):712–740.

Menkveld, Albert J. and Marius A. Zoican. 2017. "Need for Speed? Exchange Latency and Market Quality." *The Review of Financial Studies* 30 (4):1188–1228.

Mitchell, Mark, Todd Pulvino, and Erik Stafford. 2004. "Price Pressure around Mergers." *The Journal of Finance* 59 (1):31–63.

O'Hara, Maureen. 2015. "High Frequency Market Microstructure." *Journal of Financial Economics* 116 (2):257–270.

Parlour, Christine A. 1998. "Price Dynamics in Limit Order Markets." *The Review of Financial Studies* 11 (4):789–816.

Roşu, Ioanid. 2009. "A Dynamic Model of the Limit Order Book." *The Review of Financial Studies* 22 (11):4601–4641.

Sadka, Ronnie. 2006. "Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk." *Journal of Financial Economics* 80 (2):309–349.

Sandås, Patrick. 2001. "Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market." *Review of Financial Studies* 14 (3):705–734.

Scholes, Myron S. 1972. "The Market for Securities: Substitution versus Price Pressure and the Effects of Information on Share Prices." *Journal of Business* 45 (2):179–211.

Seppi, Duane J. 1997. "Liquidity Provision with Limit Orders and a Strategic Specialist." *The Review of Financial Studies* 10 (1):103–150.

Shleifer, Andrei. 1986. "Do Demand Curves for Stocks Slope Down?" *The Journal of Finance* 41 (4):579–590.

Vayanos, Dimitri and Jiang Wang. 2012. "Liquidity and Asset Returns Under Asymmetric Information and Imperfect Competition." *The Review of Financial Studies* 25 (5):1339–1365.

Yao, Chen and Mao Ye. 2018. "Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls." *The Review of Financial Studies* 31 (6):2158–2183.

List of Figures

List of Tables