Trades, Quotes and the Cost of Capital*

Ioanid Roşu†, Elvira Sojli‡, Wing-Wah Tham§

March 23, 2017

Abstract

This paper studies the quote-to-trade (QT) ratio and its relation with liquidity, price discovery, and expected returns. Empirically, we find larger QT ratios in small, illiquid or neglected firms, yet large QT ratios are associated with low expected returns. The results are driven by quotes, not by trades. We propose a model of the QT ratio consistent with these facts. In equilibrium, market makers monitor the market faster (and thus increase the QT ratio) in neglected, difficult-to-understand stocks. They also monitor faster when their clients are less risk averse, which reduces mispricing and lowers expected returns.

KEYWORDS: News, inventory, monitoring costs, volatility, liquidity, risk aversion, high frequency trading.

*We thank Jean-Edouard Colliard, Thierry Foucault, and Daniel Schmidt for their suggestions. We are also grateful to finance seminar participants at HEC Paris, University of Technology of Sydney, and the 2016 Gerzensee Symposium for valuable comments.

†HEC Paris, Email: rosu@hec.fr.

‡University of New South Wales, e.sojli@unsw.edu.au.

§University of New South Wales, w.tham@unsw.edu.au.

1 Introduction

In recent times, the ratio of the number of quotes and trades, or the quote-to-trade ratio (henceforth "QT ratio"), has become an important variable among regulators, practitioners and academics, especially in connection with high-frequency trading (henceforth "HFT").[1] In particular, the QT ratio has been at the center of many policy discussions regarding limits on trading speed, trading fees, or trade surveillance.[2,3]

Despite the widespread interest in the QT ratio, the academic literature has been relatively slow in analyzing this variable. In fact, to our knowledge this paper is the first to directly analyze the QT ratio and its connections with liquidity, price discovery, and the cost of capital.[4] An important difficulty in any study of the QT ratio is that trades and quotes are generated endogenously along with prices. To address this difficulty, in this paper we begin by documenting several new empirical stylized facts about the QT ratio. Then, we propose a theoretical model which is consistent with the stylized facts and provides a framework to interpret our empirical results.

Empirically, we first find that the QT ratio is relatively large in stocks that are small,

[1]For instance, the QT ratio is often connected to HFT by regulators and governmental institutions such as the U.S. Securities and Exchange Commission, U.S. Congressional Research Services, U.K. Government Office of Science, and the European Securities and Market Authorities. Moreover, exchanges such as NASDAQ classify HFT based on the QT ratio (see Brogaard, Hendershott, and Riordan, 2014). Among academics, the QT ratio is associated to the level of algorithmic trading (see Hendershott, Jones, and Menkveld, 2011; Boehmer, Fong, and Wu, 2015) and high-frequency trading (see e.g., Malinova, Park, and Riordan, 2016; Hoffmann, 2014; Conrad, Wahal, and Xiang, 2015; Brogaard, Hendershott, and Riordan, 2016; Subrahmanyam and Zheng, 2016).

[2]The London Stock Exchange was the first to introduce an "order management surcharge" in 2005 based on the number of trades per orders submitted. Euronext, which comprises the Paris, Amsterdam, Brussels, and Lisbon stock exchanges, has operated one since 2007. In 2012 DirectEdge introduced the "Message Efficiency Incentive Program," where the exchange pays full rebates only to traders that have an average monthly messages-to-trade ratio less than 100 to 1. In May 2012 the Oslo Stock Exchange introduced an order-to-execute fee, where traders that exceed a ratio of 70 for a month incur a charge of NOK 0.05 (USD 0.0008) per order. Deutsche Börse and Borsa Italiana announced similar measures in 2012. These fees have been revised across exchanges on a regular basis since their introduction.

[3]More recently, MIFID-II/R requires trading venues to establish a maximum unexecuted order-to-transaction ratio as one of its controls to prevent disorderly trading conditions. It stipulates that "*Trading venues shall calculate the ratio of unexecuted orders to transactions for each of their members or participants at least at the end of every trading session in both of the following ways: (a) in volume terms: (total volume of orders/total volume of transactions); (b) in number terms: (total number of orders/total number of transactions).*" See `http://ec.europa.eu/finance/securities/docs/isd/mifid/rts/160518-rts-9_en.pdf` .

[4]In general, trades and quotes are key in understanding price discovery, and therefore should affect the liquidity of an asset and its cost of capital. O'Hara (2003) highlights the importance of price discovery and liquidity to asset pricing.

illiquid or neglected, i.e., stocks with low market capitalization, institutional ownership, analyst coverage, trading volume, and volatility. Because investors usually demand a return premium for illiquid assets (see Amihud, Mendelson, and Pedersen, 2005, and the references therein) one may think that larger QT ratios are associated with larger expected returns. Surprisingly, the opposite is true: large QT ratios are associated with *low* expected returns. This relation holds both in the first part of our sample (1994–2002) and in the second part (2003–2012). We call this relation the *QT effect*.

One may be tempted to attribute the QT effect to HFT activity. Hendershott et al. (2011) find that algorithmic and high-frequency trading have a positive effect on stock liquidity. Therefore, it is plausible that stocks with higher HFT activity (and therefore higher QT ratio) are more liquid, and thus have a lower cost of capital. This argument, however, does not explain our empirical finding that the QT effect also holds during 1994–2002, when HFT is not known to have a significant impact on trading activity. Thus, we find the HFT explanation of the QT effect less likely.

Further empirical analysis shows that the QT effect is driven by the number of quotes, and not by the number of trades. This suggests that we consider a theoretical model in which trades occurs with exogenous frequency, while quotes arise from the the endogenous decision of market makers to monitor the market and change their quotes when new information arrives. To avoid the complexity of a model with multiple market makers, we assume the existence of a representative market maker.[5]

We thus propose a discrete time, infinite horizon model in which a market maker (called the "dealer" or "she") monitors a risky asset. The fundamental value of the asset follows a random walk. The dealer sets ask and bid quotes to maximize her expected profit subject to a quadratic penalty on her inventory, with coefficient called *inventory aversion*. Given the dealer's quotes, traders submit buy and sell quantities which are linear in the dealer's pricing error, that is, in the difference between the fundamental value and the price. We call the corresponding coefficient the *investor elasticity*.

The model follows Hendershott and Menkveld (2014), with two modifications. First,

[5]It is not obvious ex ante that a representative market maker exists: it is possible that the existence of multiple market makers generates a surge in quotes that cannot be attributed to a single market maker. Nevertheless, we present below evidence that the assumption is reasonable: empirically, a larger number of market makers does not lead to a surge in quotes (relative to trades), but rather to a *decrease*.

we explicitly model the dealer's choice of monitoring frequency: by paying an upfront cost increasing in monitoring frequency, she later receives a stream of signals about the fundamental value. Second, we assume that even when the dealer's pricing error is zero, traders' buy quantity is less than their sell quantity by twice an *imbalance parameter*. To justify this imbalance in trader order flow, we provide micro-foundations for trader behavior.[6] Specifically, we assume that buy and sell quantities arise endogenously in each trading round from risk averse informed investors who receive a random initial asset endowment, and from noise traders who submit inelastic quantities. In equilibrium, the trader order flow is clearly unbalanced: when the dealer's pricing error is zero, because of their risk aversion investors prefer to sell the asset rather than buy it. Our micro-foundations show that investors' risk aversion also affects investor elasticity: low risk aversion causes investors to trade with large elasticity.

Because the trading frequency is normalized to one in our model, the dealer's monitoring frequency can be interpreted as the quote-to-trade ratio. In equilibrium, the QT ratio depends of several parameters: the investor elasticity, the dealer's inventory aversion, her monitoring precision, and her monitoring cost. First, the QT ratio is increasing in the investor elasticity. Indeed, when the investor elasticity is large, the dealer's quotes must stay close to the fundamental value: otherwise, they would attract an unbalanced order flow and the dealer would pay a large inventory penalty. But to keep quotes close to the fundamental value, the dealer must monitor the market frequently, which generates a large QT ratio.

Second, the QT ratio is decreasing in the monitoring precision: a small monitoring precision makes the dealer monitor the market frequently. This result justifies our puzzling empirical finding that the QT ratio is higher in neglected, difficult-to-understand stocks: in these stocks the dealer expects to get less precise signals, and must therefore increase the frequency of monitoring, which is equivalent to increasing the QT ratio.

Third, the QT ratio is increasing in the inventory aversion: when the inventory aversion is larger, the dealer needs to keep quotes closer to the fundamental value, and hence must monitor the market more frequently. This result provides an additional

[6]Order flow imbalance is important in our model, since the cost of capital turns out to be proportional to the imbalance parameter. In Hendershott and Menkveld (2014), the trader order flow is assumed to be exogenous and with an imbalance parameter of zero.

prediction of the model: the QT ratio is smaller in stocks in which the dealer has a lower inventory aversion. The inventory aversion of the representative dealer in a stock is not observable, but in practice we can proxy its inverse with the number of market makers in that stock.[7] We thus obtain the following suprising prediction: stocks with a larger number of market makers have a *lower* QT ratio. This prediction is confirmed in the data. Intuitively, competition among market makers does not lead to a surge in the number of quotes, but rather to a larger aggregate risk bearing capacity and hence to a smaller need to monitor the market and to change quotes.[8]

Fourth, the QT ratio is decreasing in monitoring costs: a smaller monitoring cost increases the dealer's frequency of monitoring. This finding may explain the recent dramatic increase in the QT ratio observed in Figure 1. Indeed, it is plausible that the recent increase in trade automation has translated into a sharp decrease in dealer monitoring costs, which according to our results predicts a large increase in the equilibrium QT ratio.

The equilibrium quotes are governed by an *intermediation irrelevance* result: compared to her value forecast, the dealer's mid-quote is on average set at a discount that is independent of the dealer's characteristics: inventory aversion, cost of monitoring, and signal precision.[9] Intuitively, the average pricing discount must be such that the dealer does not expect her inventory to either increase or decrease. Therefore, the dealer's discount depends only on parameters of the order flow: the imbalance parameter and investor elasticity.[10]

[7]Hendershott and Menkveld (2014) prove the equivalence, up to a quadratic approximation, between a dealer's inventory aversion and her (absolute) risk aversion. A standard risk sharing result implies that the risk tolerance (inverse risk aversion) of a representative trader is the sum of the individual traders' risk tolerance. Thus, assuming a constant risk tolerance of individual marker makers, a larger number of them translates into a higher risk tolerance for the representative dealer.

[8]This result also justifies our assumption of a representative dealer: see Footnote 5.

[9]The intermediation irrelevance result extends also to the quotes themselves: the equilibrium bid-ask spread is the ratio of two parameters that describe the trader order flow.

[10]We stress that the intermediation irrelevance result refers to the *average* discount. This value coincides with the equilibrium discount in a particular state of the system, the *neutral state*, when the dealer's inventory is such that her quotes are not expected to either increase or decrease. In the language of Hendershott and Menkveld (2014), in the neutral state there are no price pressures. In other states, when the inventory deviates from its neutral value, the speed of mean reversion of the pricing discount to its neutral value does in fact depend on the dealer's characteristics, and there is no longer an intermediation irrelevance. Instead, in these states there are price pressures in the sense of Hendershott and Menkveld (2014).

We next discuss the cost of capital, which in our model is in one-to-one relation with the dealer's pricing discount. The intermediation irrelevance implies that the cost of capital should not be affected by the dealer's characteristics, but only by the properties of the trader order flow. In particular, the cost of capital does depend on investor elasticity. Consider an increase in investor elasticity, which means that investors trade more aggressively on the dealer's pricing error. Therefore, the dealer must (i) monitor the market more often to reduce the pricing error; and (ii) reduce the pricing discount by keeping the mid-quote closer to her forecast. The first fact translates into an increase in monitoring frequency, hence an increase of the QT ratio. The second fact translates into a decrease of the pricing discount, hence a decrease of the cost of capital.

Putting these facts together, we obtain the QT effect: an inverse relation between the QT ratio and the cost of capital. This aligns with our main empirical finding. Note that this relation is driven by properties of the order flow, and at a more fundamental level (if we include our micro-foundations) by the investors' risk aversion.

Following the logic of our intermediation irrelevance result, we predict that the number of market makers in a particular stock should not affect its cost of capital. This additional empirical prediction is confirmed by the data: in our sample, the expected return of NASDAQ-listed stocks does not depend on the number of dealers.

Our paper contributes to a large literature on market microstructure and asset pricing (see Amihud and Mendelson, 1986; Amihud, 2002; Brennan and Subrahmanyam, 1996; Chordia, Roll, and Subrahmanyam, 2002, 2000; Chordia, Subrahmanyam, and Anshuman, 2001; Easley, Hvidkjaer, and O'Hara, 2002; Duarte and Young, 2009; Amihud et al., 2005, among many others). While the relation between the quote-to-trade ratio and the cost of capital has not (to our knowledge) been investigated before, our empirical analysis follows the example of many papers that find stock characteristics that matter for average returns (see Harvey, Liu, and Zhu, 2016).

Our theoretical model is closest in spirit to the price pressures model of Hendershott and Menkveld (2014). However, our focus is very different, as we study the quote-to-trade ratio and the cost of capital. We thus depart from their model and endogenize the dealer's monitoring frequency, which allows us to define the quote-to-trade ratio. A second departure is that we introduce imbalances in the order flow (justified by investor

risk aversion), which allows us to obtain a nonzero cost of capital. To avoid the time-varying price pressures that are the focus of Hendershott and Menkveld (2014), we define the cost of capital in the neutral state where the price pressure is zero. The dealer has a positive inventory in this state, and the cost of capital is positive. By contrast, in their paper the cost of capital is zero.

Our paper has implications for the burgeoning literature on High-Frequency Trading (see for example Menkveld, 2016, and the references therein). The recent dramatic increase in the QT ratio (see Figure 1) has been widely attributed to the emergence of algorithmic trading and HFT (see e.g. Hendershott et al., 2011). In our theoretical framework, this is consistent with a sharp decrease in dealer monitoring costs caused by trade automation. Our main focus, however, is on the relation between the QT ratio and the cost of capital. We document a new empirical regularity called the QT effect: large QT ratios are associated with low expected returns.

Our theoretical results provide a possible interpretation of the QT effect: The intermediation irrelevance result implies that the cost of capital does not depend on dealer characteristics, but rather on properties of the order flow, and at a more fundamental level on investors' risk aversion. In particular, a decrease in investor risk aversion increases the QT ratio, while it decreases the cost of capital. Viewed through the lens of our model, other explanations of the QT effect must account for why investor behavior is altered. For instance, if HFT activity changes dealer characteristics but not investors' preferences, it may affect the QT ratio, but not the cost of capital. One piece of evidence that the QT effect is unlikely to be driven by HFT activity is that the QT effect works also in the first part of our sample (1994–2002), before the emergence of HFT.

2 Data and Summary Statistics

2.1 Data

To construct the quote-to-trade ratio, we use the trades and quotes reported in TAQ for the period June 1994 to October 2012.[11] Using TAQ data allows us to construct a long

[11] Our sample starts in June 1994, as TAQ reports opening and closing quotes but not intraday quotes for NASDAQ-listed stocks prior to this date.

time series of the variable QT at the stock level, which is best suited to conduct asset pricing tests. We retain stocks listed on the NYSE, AMEX, and NASDAQ for which information is available in TAQ, Center for Research in Security Prices (CRSP), and Compustat.

Our sample includes only common stocks (Common Stock Indicator Type = 0), common shares (Share Code 10 and 11), and stocks not trading on a "when issued" basis. Stocks that change primary exchange, ticker symbol, or CUSIP are removed from the sample (Hasbrouck, 2009; Goyenko, Holden, and Trzcinka, 2009; Chordia, Roll, and Subrahmanyam, 2000). To avoid extremely illiquid stocks, we also remove stocks that have a price lower than \$2 and higher than \$1,000 at the end of a month.[12] To avoid look-ahead biases, all filters are applied on a monthly basis and not on the whole sample. There are 10,345 individual stocks in the final sample.

Throughout the paper, we follow Shumway (1997) in using returns of -30% for the delisting month (delisting codes 500 and 520–584).[13] All returns are calculated using bid-ask midpoint prices, adjusted for splits and cash distributions, to reduce market microstructure noise effects on observed returns (Asparouhova, Bessembinder, and Kalcheva, 2010, 2013). Risk factors are from Kenneth French's website for the period 1926 to 2017. The PIN factor is from Sören Hvidkjaer's website and is available from 1984 to 2002. Table D.1 in the Appendix reports the definitions and the construction details for all variables and Table D.2 in the Appendix provides the summary statistics.

Consistent with the literature (see Angel, Harris, and Spatt, 2011; Brogaard, Hagströmer, Nordén, and Riordan, 2015), we define QT as the monthly ratio of the number of quote updates at the best national price (National Best Bid Offer) to the number of trades. By quote updates we refer only to changes either in the ask or bid prices, and not to depth updates at the current quotes.[14] Specifically, we calculate the QT variable for

[12]Results are quantitatively similar when removing stocks with price $<$ \$5 and are available from the authors upon demand.

[13]Shumway (1997) reports that the CRSP database has a systematic upward bias on returns of certain delisted stocks. This is because negative delisting returns are coded as missing when the delisting is due to performance reasons.

[14]The results are qualitatively similar if we define QT using the number of both quote and depth updates in the numerator. However, using quotes only is more consistent with our theoretical model in Section 4.

stock i in month t as the ratio:

$$QT_{i,t} = \frac{N(quotes)_{i,t}}{N(trades)_{i,t}}, \tag{1}$$

where $N(quotes)_{i,t}$ is the number of quote updates in stock i during month t, and $N(trades)_{i,t}$ is the number of trades in stock i during month t.

2.2 Determinants of the Quote-to-Trade Ratio

In this section, we examine the summary statistics, time series and the cross-sectional determinants of the QT ratio. Table 1 reports the average firm-level characteristics of ten portfolios sorted on the QT ratio. Specifically, for each month t we divide all stocks into decile portfolios based on their QT during that month. The QT portfolio 1 has the lowest QT, and the QT portfolio 10 has the highest QT. For each QT decile, we compute the cross-sectional mean characteristic for month $t+1$ and report the time-series mean of the cross-sectional average characteristic.[15]

Column (5) in Table 1 shows that the average firm size, as measured by market capitalization, is decreasing in QT. The lowest QT stocks (stocks in QT decile 1) have an average market capitalization of \$8.9 billion, while the highest QT stocks (stocks in QT decile 10) have an average capitalization of \$0.8 billion. Column (6) shows that the average monthly trading volume decreases from \$1.7 billion for the lowest QT stocks to \$0.06 billion for the highest QT stocks. The average monthly trading volume in column (6) decreases from \$1.7 billion for low QT stocks to \$0.06 billion for high QT stocks. Columns (8)–(10) show the averages of three illiquidity measures: the quoted spread, the relative spread, and the Amihud (2002) illiquidity ratio (ILR). The highest QT stocks are roughly three times more illiquid that the lowest QT stocks. The lowest QT stocks are almost twice as volatile as the highest QT stocks, in column (11).

Table 2 formally examines the relation of the above variables as determinants of QT in a regression setting. The dependent variable is the monthly QT measure. We present the results from a panel regression with various specifications for fixed effects and with

[15]The order of the different characteristics across QT portfolios remains unchanged, when we compute the cross-sectional characteristics in month t.

standard errors clustered at the stock and month level. Columns (1)-(4) include variables known to affect expected returns. We find that QT is higher for stocks that have low market capitalization, low institutional ownership, low or no analyst coverage, low trading volume, and low volatility. Generally these are stocks that are neglected by analysts or investors, and are difficult to understand/evaluate (see Hong, Lim, and Stein, 2000; KUMAR, 2009).[16]

Stylized fact 1 (SF1): Neglected stocks (with low market capitalization, institutional ownership, analyst coverage, trading volume, and volatility) have higher quote-to-trade ratios.

This result is puzzling, because in neglected stocks one may expect a lower QT ratio (after controlling for trading volume), as market makers have less precise information based on which to change their quotes. But in our theoretical model, in Section 4.3, a market maker with less precise information actually monitors more often and therefore generates a higher QT ratio.

It is common practice among academics, practitioners and regulators to associate QT with HFT activity (several examples are given in Footnote 1). Results in Tables 1 and 2 suggest that using QT as a proxy for HFT activity must be done with caution. For instance, HFTs are known to trade in larger and more liquid stocks (Hagströmer and Nordén, 2013; Brogaard et al., 2015). In addition, HFTs are more likely to trade in stocks with high institutional ownership, if indeed HFT activity stems from their anticipation of agency and proprietary algorithms of institutional investors such as mutual and hedge funds (O'Hara, 2015). But stylized fact SF1 above shows that QT is actually *lower* in stocks that are large, liquid, or with high institutional ownership. Thus, simply associating HFT activity with QT can be misleading.

[16]In column (5) of Table 2, we include also the number of registered market makers in a particular stock. This is discussed in Section 3.2, as part of the stylized empirical fact SF4.

2.3 Time Series of Quote-to-Trade Ratios

Figure 1 Panel A shows the time series of the equally weighted natural logarithm of monthly QT over the sample period. We note the substantial increase in QT during this time. Panel B is similar to Panel A, but displays separately the evolution of quotes and trades. It shows that the increase in QT is driven by the explosion in quote updates. For instance, in June 1994 the total number of quotes and the total number of trades are roughly equal to each other, at about 1.1 million each. In August 2011, the peak month for both quotes and trades, the number of quotes reached 1,445 million, while trades reached 104 million, an increase ten times larger for quotes than for trades.

Stylized fact 2 (SF2): Quote-to-trade ratios have increased over time.

This stylized fact can be explained theoretically by a decrease in market maker monitoring costs: when these costs are smaller, market makers monitor more often, hence the QT ratio increases (see Section 4.3). Both SF2 and its explanation are consistent with previous literature.[17] Hendershott et al. (2011) study a change of NYSE market structure in 2003 called "Autoquote" and argue that this change resulted in a decrease in monitoring costs among market participants, and especially among algorithmic traders. At the same time, they document an increase in their proxy for algorithmic trading, which is close in spirit to our QT ratio.[18] Angel et al. (2011) argue that the proliferation since 2003 of algorithmic and high-frequency trading has lead to substantial increases in both the number of quotes and trades.

3 Quote-to-Trade Ratio and Stock Returns

In this section, we study the cross-sectional relation between the quote-to-trade ratio and stock returns. We start with an investigation of abnormal expected returns to

[17] In untabulated results, we find that the introduction of Autoquote substantially increases the QT ratio, but it does not affect the relation of QT and the other variables presented in Table 2. These results are available from the authors upon demand.

[18] See Figure 1 in Hendershott et al. (2011). Their proxy for algorithmic trading is defined as the negative of dollar trading volume divided by the number of electronic messages (incuding electronic order submissions, cancellations and trade reports, but excluding specialist quoting or floor orders).

account for various risk factors through portfolio sorts, and then examine other known cross-sectional return predictors through Fama-MacBeth regressions.

3.1 Univariate Analysis

First, we test whether the return differential between the low and high QT stocks can be explained by the market, size, value, momentum, and liquidity factors. Each month, all stocks are divided into portfolios sorted on QT at time t. Portfolio returns are the equally weighted average realized returns of the constituent stocks in each portfolio in month $t+1$.[19] We estimate individual portfolio loadings from a 24-month rolling window regression:

$$r_{p,t+1} = \alpha_p + \sum_{j=1}^{J} \beta_{p,j} X_{j,t} + \varepsilon_{p,t+1}, \tag{2}$$

where $r_{p,t+1}$ is the return in excess of the risk free rate for month $t+1$ of portfolio p constructed in month t based on the QT level, and $X_{j,t}$ is the set of J risk factors: excess market return (r_m), value HML (r_{hml}), size SMB (r_{smb}), Pástor and Stambaugh (2003) liquidity (r_{liq}), momentum UMD (r_{umd}), and PIN (r_{PIN}). Table 3 reports time series averages of alphas obtained from 24-month rolling window regressions.[20] We present results from several asset pricing models that include several risk factors: CAPM (market), FF3 (market, size, value), FF3+PS (with the Pástor and Stambaugh (2003) traded liquidity factor), FF4+PS (with momentum), and FF4+PS+PIN (probability of informed trading, PIN).[21]

Table 3 reports alphas for 10, 25, and 50 QT-sorted portfolios. The low-QT portfolio (QT1) has a statistically significant monthly alpha (α_1) that ranges between 0.60% and 1.88% across various portfolio splits and asset pricing models. The high-QT portfolio alphas range from -0.34% to 0.37%, but are statistically not different from zero in all specifications. This suggests that the high-QT portfolios are priced well by the factor

[19] We also conduct the analysis using value weighted portfolio returns and the results do not change quantitatively.

[20] Since we are using portfolios conditional on QT, we only have portfolio returns from July 1994. We use a 24-month estimation window to increase the sample period. For the Fama-MacBeth individual stock regressions in the next section, we use a 48-month rolling window to estimate factor loadings.

[21] The PIN factor from Sören Hvidkjaer's website is available only until 2002, therefore we restrict our analysis in the last column of Table 3 to the period 1994–2002. This result is discussed in Section 3.3, as part of the sub-sample robustness analysis.

models. However, the risk-adjusted return difference between the low-QT and high-QT portfolios is statistically significant and varies between 0.52% to 1.91% per month across different portfolio splits. Note that the profitability of the long-short strategy derives mainly from the long position (the performance of the low-QT portfolio QT1) rather than from the short position (the performance of the high-QT portfolio QT10). Therefore, short-selling constraints should not impede the implementation of a strategy that exploits the main regularity in Table 3.

3.2 Fama-MacBeth Regressions

To control for other predictive variables in the cross-section of returns, we estimate Fama and MacBeth (1973) cross-sectional regressions of monthly individual stock risk-adjusted returns on different firm characteristics including the QT variable. We use individual stocks as test assets to avoid the possibility that tests may be sensitive to the portfolio grouping procedure. First, we estimate monthly rolling regressions to obtain individual stocks' risk-adjusted returns using a 48-month estimation window. We use a similar procedure as in Brennan, Chordia, and Subrahmanyam (1998) and Chordia, Subrahmanyam, and Tong (2011), to obtain risk-adjusted returns:

$$r^a_{i,t} \;=\; r_{i,t} - \sum_{j=1}^{J} \hat{\beta}_{i,j,t-1} F_{j,t}, \tag{3}$$

where $r_{i,t}$ is the monthly return of stock i in excess of the risk free rate, $\hat{\beta}_{i,j,t-1}$ is the conditional beta estimated by a first-pass time-series regression of risk factor j estimated for stock i by a rolling time series regression up to $t-1$, and $F_{j,t}$ is the realized value of risk factor j at t. Then, we regress the risk-adjusted returns from equation (3) on lagged stock characteristics:

$$r^a_{i,t} \;=\; c_{0,t} + \sum_{m=1}^{M} c_{m,t} Z_{m,i,t-k} + e_{i,t}, \tag{4}$$

where $Z_{m,i,t-k}$ is the characteristic m for stock i at time $t-k$, and M is the total number of characteristics. We use $k = 1$ months for all characteristics.[22] The procedure ensures unbiased estimates of the coefficients $c_{m,t}$, without the need to form portfolios, because errors in the estimation of the factor loadings are included in the dependent variable. The t-statistics are obtained using the Fama-MacBeth standard errors with Newey-West correction with 12 lags.

Table 4 reports the Fama and MacBeth (1973) coefficients for cross-sectional regressions of individual stock risk-adjusted returns on stock characteristics. We consider the risk factors from a four-factor Fama–French model (market, size, value, and momentum), with an additional Pástor and Stambaugh (2003) traded liquidity factor. Column (1) includes only the QT ratio. QT has a highly significant and negative coefficient implying that stocks with higher QT have lower next month risk-adjusted returns. We call this the *QT effect*.

Because the QT effect might be driven by the correlation of QT with liquidity, we include two illiquidity proxies in the regression: the bid-ask spread (SPREAD) and the Amihud (2002) illiquidity ratio (ILR). Column (2) of Table 4 includes QT and SPREAD, column (3) includes QT and ILR, and column (4) includes QT and both SPREAD and ILR. The coefficients on both illiquidity proxies are positive and significant, consistent with higher illiquidity causing higher returns (see Amihud, 2002). However, the inclusion of these known illiquidity proxies does not reduce the effect of QT, which remains negative and significant in all specifications (2)–(4). In column (5), we introduce other firm characteristics that affect expected returns. With these additional control variables, the coefficient on QT remains negative and highly significant, while the illiquidity proxies SPREAD and ILR become both insignificant.

Table 5 explores the question whether the QT effect is driven by the number of quotes or by the number of trades. Column (1) shows that when conditioning on quotes and trades as separate explanatory variables, it is the number of quotes that matters most for risk-adjusted returns. This effect is economically and statistically large. Introducing other liquidity-based control variables in columns (2)–(4) takes away the statistical sig-

[22] Panel A of Table D.3 in the Appendix shows the estimation results where $k = 2$ (excluding the past return variables $R1$ and $R212$).

nificance of the number of trades, but does not affect the number of quotes. Using all firm characteristics as well as liquidity measures as control variables, column (6) shows that the predictive power derives from quotes and not from trades.

Stylized fact 3 (SF3): Higher quote-to-trade ratios predict lower stock returns in the cross-section (the QT effect). The predictability is driven by the number of quotes rather than the number of trades.

This result is puzzling if we compare it with the stylized fact SF1, which shows that the QT ratio is higher in neglected stocks, and in particular in smaller or more illiquid stocks. But, as Table 4 shows, these stocks also tend to have *higher* expected returns, which appears to contradict SF3. Our results then suggest that there is substantial QT variation that is negatively correlated with expected returns even after conditioning on size and illiquidity. In other words, the QT effect remains even across stocks in a portfolio with similar size and illiquidity. The QT effect therefore is distinct from the known effects of other variables: spread, ILR, trading volume, volatility. We thus add to the literature that explores how trading activity and market structure are connected with asset returns (see Amihud and Mendelson, 1986; Amihud, 2002; Brennan and Subrahmanyam, 1996; Chordia et al., 2002, 2000, 2001; Easley et al., 2002; Duarte and Young, 2009, among many others).

One concern is that the QT effect might be driven by the number of market makers that are registered in a stock: it is plausible that a larger number of active market makers drives up the QT ratio because of increased competition, but also decreases the required expected return. We find that neither of these two stories are supported in the data. First, column (5) of Table 2 includes the number of registered market makers in a particular stock (MM) as a control variable. This results in a smaller sample, because the number of market makers in only available for NASDAQ-traded stocks. Nevertheless, we find that the number of market makers has a significant effect on the QT ratio, except that the coefficient is negative: a larger number of market makers in a stock corresponds to a *lower* QT ratio. Second, column (6) of Table 4 shows that the number of market makers in a particular stock has no effect on its cost of capital. We collect

these empirical resuslts in the following stylized fact.

Stylized fact 4 (SF4): The number of market makers in a NASDAQ stock has an inverse relation with the quote-to-trade ratio and no relation to the stock's expected return.

To interpret SF4, recall that the number of market makers can be regarded a proxy for the unobserved risk tolerance of a representative dealer, and also that a dealer's risk aversion is in one-to-one correspondence with her inventory aversion (see Footnote 7). With this interpretation, SF4 implies that a smaller inventory aversion of the representative dealer is associated to a smaller QT ratio, but is unrelated to the stock's expected return.

One prediction of our model is that a dealer with low inventory aversion monitors infrequently (because she is not too concerned with her inventory), and therefore generates a low QT ratio (see Section 4.3). The first part of SF4 shows that this model prediction holds in the data. The result is surprising: one may think that if the representative dealer's inventory aversion is low, or equivalently (if we accept the proxy) if the number of market makers is high, then their competition drives up the quoting activity. But in the data the opposite result is true. Thus, our results suggest that high competition among market makers leads to a high aggregate risk bearing capacity, and hence to a small need for the market makers to monitor, which implies a low QT ratio.

A second prediction of our model arises from the *intermediation irrelevance* result: the expected return of a stock is not affected by the characteristics of the dealer in that stock (the intermediary), but only on the properties of the traders' order flow (see Sections 4.4 and 4.5). In particular, the stock's expected return should not be affected by the dealer's inventory aversion. But according to the second part of SF4 this prediction is true in the data, as long as we consider the number of market makers as a proxy for the representative dealer's (inverse) inventory aversion.

3.3 Robustness

In this section, we verify the robustness of our main empirical result: the QT effect. In Section 3.1 we have considered only one-month holding (portfolio rebalancing) periods. One could therefore raise the concern that the QT effect is caused by temporary price effects. For example, suppose stocks with high or low realized returns attract HFT activity and get a temporary spike in the QT ratio. This type of explanation implies that the QT effect is only a short-term phenomenon. If that were the case, we would expect stocks to switch across QT portfolios, and the alphas of a QT long-short strategy to decrease over longer holding periods.

To test the reversal hypothesis, we examine the average monthly risk-adjusted returns (alphas) of the QT long-short strategies for different holding and formation periods. We use the calendar-time overlapping portfolio approach of Jegadeesh and Titman (1993) to calculate post-performance returns. We assign stocks into portfolios based on QT levels at four different formation periods and examine the average QT level for these portfolios in month $t + k$ keeping the portfolio constituents fixed for k months, where k ranges from 1 to 12 months. We use four formation periods, i.e., we condition on different sets of information about QT: time t, and the 3, 6, and 12-month moving average QT level.

Figure 2 shows the long-short alphas from a five-factor model (Fama-French three-factor model plus momentum and liquidity) for strategies that long the low-QT portfolio and short the high-QT portfolio, at different holding horizons and formation periods. The holding horizons reflect the number of months for which the portfolio constituents are kept fixed after the formation month, i.e., portfolios are rebalanced every k months. We construct the long-short strategies for 25 portfolios and examine 4 different formation periods.[23] The figure shows that the QT effect is very persistent. The one month formation and holding period portfolio has the highest alpha of 1.25%. Overall, the long/short alphas after a year of both formation and holding are 0.60% per month and highly statistically significant.

Another robustness check is to verify whether the QT effect holds during both parts of our sample: 1994–2002 and 2003–2012. Indeed, since QT is often used as a proxy for

[23]The results are robust to other factor model specifications and to the creation of more portfolios. These results are available from the authors upon request.

HFT (see Footnote 1), we would like to study the information content of QT beyond that of HFT. To omit the potential influence of HFT in our study, we conduct both the portfolio analysis and Fama-MacBeth regressions for the two subsamples June 1994 to December 2002 and January 2003 to October 2012. The first subsample is unaffected by changes in technology and algorithmic trading, as Hendershott et al. (2011) document the proliferation of algorithmic and electronic trading only after 2003.

Column (5) in Table 3 (where we include PIN) only covers the first part of the sample June 1994 to December 2002, due to the availability of the PIN factor returns. The effect of QT on risk-adjusted returns using long-short portfolios are strong and even larger for this subsample, in the pre-algorithmic trading period. The long-short alpha in column (5) is the highest in all risk specifications. Table D.4 in the Appendix presents the subsample analysis for the Fama-MacBeth regressions, equivalent to column (5) in Table 4. The effect of QT on risk-adjusted returns is large and statistically significant in the pre- and post-2002 period, despite the reduction in power due to the lower number of time-series observations.

Stylized fact 3′ (SF3′): The relation between quote-to-trade ratio and cross-sectional stock returns holds at longer predictability horizons and is persistent throughout the sample.

3.4 Summary of the Empirical Findings

Our empirical results fall under two large categories: the determinants of the QT ratio of stocks, and the relation of the QT ratio with stock expected returns. We find that high QT is prevalent among neglected stocks, i.e., stocks that have low market capitalization, institutional ownership, analyst coverage, trading volume, and volatility (SF1). In the time series, the QT ratio has increased significantly over time (SF2). Yet, the relation between the QT ratio and expected returns is stable over time (SF3′). This relation, called the QT effect, is that stocks with high QT ratio tend to have low expected returns (SF3). The QT effect appears to be distinct from other known effects on expected returns of spread, ILR, trading volume, volatility, etc. Including the number of market makers

among explanatory variables displays an inverse relation with the QT ratio, but does not affect expected returns (SF4). In the next section we propose a theoretical model that is consistent with all these stylized facts and provides an interpretation for them.

4 Model of the Quote-to-Trade Ratio

This section builds a model of the quote-to-trade ratio, and relates it to the cost of capital and other variables of interest. The model is close in spirit to the the price pressures model of Hendershott and Menkveld (2014, henceforth HM2014) and to the dynamic inventory control model of Ho and Stoll (1981). As in HM2014, we consider a representative intermediary who faces stochastically arriving traders with elastic liquidity demands. At first we consider the liquidity demand in reduced form, but in the Appendix C we add micro-foundations.

Because the focus of our paper is on the quote-to-trade ratio, we depart from HM2014 and endogenize the intermediary's monitoring frequency. As a second departure, we introduce an imbalance in the liquidity demand (justified by investor risk aversion), which allows us to obtain a nonzero cost of capital. To avoid the time-varying price pressures that are the focus of HM2014, we define the cost of capital for a neutral state in which the price pressure is zero. In equilibrium, the intermediary has a positive inventory in this state, and the cost of capital is positive.

4.1 Environment

The market is composed of one risk-free asset and one risky asset. Trading in the risky asset takes place in a market exchange, at discrete dates $t = 0, 1, 2, \ldots$ such that the trading frequency is normalized to one. There are two types of market participants: (a) one monopolistic market maker called the *dealer* ("she") who monitors the market and sets the quotes at which others trade, and (b) traders, who submit market orders.

Assets. The risk-free asset is used as a numeraire and has a return of zero. The risky asset has a net supply of $M > 0$. It pays a dividend D before each trading date. The ex-dividend fundamental value v_t follows a continuous random walk process for

which the increments have variance per unit of time equal to $\Sigma_v = \sigma_v^2$, where σ_v is the *fundamental volatility*. We interpret v_t for t large as the "long-run" value of the asset; in the high frequency world, this can be taken to be the asset value at the end of the trading day, and the increments are then the short term changes in value due to the arrival of new information. Alternatively, v_t can be considered as the cash value that shareholders receive at liquidation, an event which can occur in each period with a fixed probability.[24]

Dealer Monitoring. The dealer monitors the market by periodically obtaining signals about the fundamental value. Monitoring occurs at times $\frac{0}{q}, \frac{1}{q}, \frac{2}{q}, \ldots$, where q is a positive number called the *monitoring rate*.[25] If the monitoring time coincides with the trading time (that is, if the monitoring time is an integer), we assume that monitoring occurs before trading. Per unit of time, the cost of monitoring at the rate q is $C(q)$, which is an increasing function of q.

Monitoring consists in the dealer receiving a set of signals about the fundamental value at each monitoring time. Denote by w_t the dealer's forecast, which is the expected fundamental value of the asset, conditional on all the signals received until time t. We define the *precision function* F_t as the inverse variance of the forecast error. We take a reduced form approach, and assume that the precision function does not depend on t, and is a decreasing function of the monitoring rate q:[26]

$$F(q) \;=\; \frac{1}{\mathsf{Var}(v_t - w_t)}. \tag{5}$$

The intuition is that an increase in the monitoring rate produces more precise forecasts for the dealer.

[24]Suppose there exists $\pi \in (0,1)$ such that the asset liquidates in each period with probability π, in which case the shareholders receive v_t per share. Then it can be showed that the expected profits of a trader with quantities bought and sold at t equal to $-Q_t^b$ and $-Q_t^s$, respectively, has the form described in equation (8) with $\beta = 1 - \pi$, and $\gamma = C(q) = 0$.

[25]With this interpretation of monitoring, q should take only integer values. However, we allow q to be any positive real number because we take a reduced form approach and specify directly the signal precision that the dealer derives from monitoring (micro-foundations for the signal structure are provided in Appendix A.) Ideally, we would like to solve a model in which trading and monitoring follow independent Poisson processes with intensities 1 and q, respectively. That model is much more difficult to solve, although we conjecture that the equilibrium is qualitatively the same. Thus, in the rest of the paper we say, with a slight abuse of terminology, that monitoring takes place at a rate $q > 0$.

[26]In Appendix A we show how to generate $F(q)$ using a specific signal structure.

To simplify the equilibrium formulas, we assume that the monitoring cost $C(q)$ and the precision function $F(q)$ are linear increasing functions,

$$C(q) \;=\; c\,q, \qquad F(q) \;=\; f\,q, \tag{6}$$

where c and f are positive constants.[27]

Dealer's Quotes and Objective. After monitoring at time τ, the dealer sets the quotes: the ask quote a_τ and the bid quote b_τ. We therefore interpret the monitoring rate q as the quote rate.[28] Let $\mathcal{I}_\tau$ be the dealer's information set after monitoring at τ, and $w_\tau = \mathsf{E}_\tau(v_\tau) = \mathsf{E}(v_\tau|\mathcal{I}_\tau)$ her forecast of the asset value.

In general, a quoting strategy for the dealer is a set of processes a_t (the ask quote) and b_t (the bid quote) which are measurable with respect to the dealer's information set. Let x_t be the dealer's inventory in the risky asset just before trading at t.[29] If Q^b_t is the aggregate buy market order at t, and Q^s is the aggregate sell market order at t, the dealer's inventory evolves according to

$$x_{t+1} \;=\; x_t - Q^b_t + Q^s_t. \tag{7}$$

Then, for a given quoting strategy, the dealer's expected utility at τ is equal to the expected profit from date τ onwards, minus the quadratic penalty in the inventory, and minus the monitoring costs:

$$\mathsf{E}_\tau \sum_{t=\tau}^{\infty} \beta^{t-\tau}\Big(x_t\,D \;+\; \big((v_t - b_t)Q^s_t + (a_t - v_t)Q^b_t\big) \;-\; \gamma\,x_t^2 \;-\; C(q)\Big), \tag{8}$$

where $\beta \in (0,1)$ and $\gamma > 0$. Thus, the dealer maximizes expected profit, but at each t faces a utility loss that is quadratic in the inventory. Note that except for the dividend

[27] In the proof of Proposition 2, we describe the equilibrium conditions for more general F and C.

[28] Technically, there is no need for the dealer to change her quotes when no trading is expected (at noninteger monitoring times k/q). But, since the dealer incurs no cost from modifying quotes, it makes sense intuitively to allow her to adjust the quotes to new information, especially if she is not certain that no trading takes place at that time. Thus, in such a "trembling hand" equilibrium the dealer's quote rate is indeed equal to q. This is consistent also with the alternative model described in Footnote 25.

[29] We let the initial inventory x_0 as a free parameter, although later (in Section 4.5) we set it equal to the parameter $\bar{x}$ from equation (15), which is the long-term mean of the dealer's equilibrium inventory.

payment this utility function is essentially the same as the one specified in HM2014.[30]

Traders' Order Flow. Upon observing the quotes a_t (the ask quote) and b_t (the bid quote), traders submit at t the following aggregate market orders:

$$\begin{aligned} Q_t^b &= \frac{k}{2}(v_t - a_t) + \ell - m + \varepsilon_t^b, \quad \text{with} \quad \varepsilon_t^b \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \\ Q_t^s &= \frac{k}{2}(b_t - v_t) + \ell + m + \varepsilon_t^s, \quad \text{with} \quad \varepsilon_t^s \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \end{aligned} \tag{9}$$

where Q_t^b is the *buy demand* and Q_t^s is the *sell demand*. The numbers k, ℓ, m and Σ_N are exogenous constants. Together, Q_t^b and Q_t^s are called the *liquidity demand*, or the traders' *order flow*. The parameter k is the *investor elasticity*, ℓ is the *inelasticity parameter*, and m is the *imbalance parameter*.[31] In Appendix C, we provide micro-foundations for the liquidity demand.

Equilibrium Concept. Because the dealer is a monopolist market maker in our model, the structure of the game is simple. First, before trading begins (before $t = 0$), the dealer chooses a constant monitoring rate q. Second, in the trading game the dealer continuously chooses the quotes (the ask quote a_t and the bid quote b_t) such that objective function (8) is maximized.

4.2 Optimal Quotes

We solve for the equilibrium in two steps. In the first step (Section 4.2), we take the dealer's monitoring rate q as given and describe the optimal quoting behavior. In the second step (Section 4.3), we determine the optimal monitoring rate q as the rate which maximizes the dealer's expected utility.

We thus start by fixing the monitoring rate q. The optimal behavior of the dealer in the trading game is described in Proposition 1. This result is obtained by applying

[30] This penalty can be justified either by the dealer facing external funding constraints, or by her being risk averse. The latter explanation is present in HM2014 (Section 3). There, the dealer maximizes quadratic utility over non-storable consumption. To solve the dynamic optimization problem, HM2014 consider an approximation of the resulting objective function (see their equation (16)). This approximation coincides with our dealer's expected utility in (8) when $C(q) = 0$.

[31] HM2014 use a similar reduced form approach, except that they set $m = 0$. By providing micro-foundations for traders' order flow, we find that $m > 0$ when investors are risk averse and the asset is in positive net supply.

standard methods in linear-quadratic dynamic programming.[32] The solution depends on a few parameters that describe traders' order flow in (9).

Consider the game described in Section 4.1, with positive parameters D, k, ℓ, m, Σ_N. Define the following constants:

$$\begin{aligned} h &= \frac{\ell}{k}, \quad \omega = \frac{1-\beta}{\beta k}, \quad \alpha = \beta\,\frac{(\gamma-\omega)+\sqrt{(\gamma-\omega)^2+\frac{4\gamma}{\beta k}}}{2}, \\ \lambda &= \frac{\alpha}{1+k\alpha} = \frac{-(\gamma+\omega)+\sqrt{(\gamma-\omega)^2+\frac{4\gamma}{\beta k}}}{2}, \\ \Delta &= \frac{1-\beta+2k\alpha}{k(1-\beta+k\alpha)}\,m - \frac{\beta}{2(1-\beta+k\alpha)}\,D. \end{aligned} \tag{10}$$

The next result describes the optimal quotes set by the dealer.

Proposition 1. *The dealer's optimal quotes at $t = 0, 1, \ldots$ are*

$$a_t = w_t - \lambda x_t + h - \Delta, \qquad b_t = w_t - \lambda x_t - h - \Delta, \tag{11}$$

where w_t is the dealer's value forecast, and x_t is her inventory. The mid-quote price $p_t = (a_t + b_t)/2$ satisfies

$$p_t = w_t - \lambda x_t - \Delta = w_t - \lambda x_t - \frac{1-\beta+2k\alpha}{k(1-\beta+k\alpha)}\,m + \frac{\beta}{2(1-\beta+k\alpha)}\,D. \tag{12}$$

To get intuition for this result, suppose the imbalance parameter m and the dividend D are both zero (hence $\Delta = 0$). Consider first the particular case when the dealer is risk-neutral: $\gamma = 0$. In that case, both α and λ are equal to zero, and the dealer's inventory x_t does not affect her strategy. Equation (11) implies that the dealer sets her quotes at equal distance around her forecast w_t. Hence, the ask quote at t is $a_t = w_t + h$, and the bid quote is $b_t = w_t - h$, where h is the constant half-spread. The equilibrium value $h = \ell/k$ corresponds to two opposite concerns for the dealer. If she sets too large a half-spread, then investors (whose price sensitivity is increasing in k) submit a smaller

[32] For a general treatment of such problems, see Sargent and Ljungqvist (2000). See also HM2014 for an application to their price pressures model.

expected quantity at the quotes.[33] If she sets too small a half-spread, this decreases the part of the profit that comes from the inelastic part ℓ of traders' order flow.

When the dealer has inventory concerns ($\gamma > 0$), her inventory affects the optimal quotes: according to equation (11), the quotes are equally spaced around an inventory-adjusted forecast ($w_t - \lambda x_t$). The effect of the dealer's inventory on the mid-quote price is in fact the *price pressure* mechanism identified by HM2014. To understand this phenomenon, suppose that before trading at t the dealer has zero inventory, and at t traders submit a net demand Q. The dealer's inventory then becomes negative ($-Q$). To avoid the inventory penalty, the dealer must bring back the inventory to zero. For that, the dealer must raise the quotes to convince more sellers to arrive. Quantitatively, according to (11) the dealer must increase both quotes by λQ, with the coefficient λ as in equation (10). This makes the corresponding slope coefficient λ essentially a price impact coefficient, in the spirit of Kyle (1985).[34]

According to (12), the mid-quote price is decreasing in the imbalance parameter m, and increasing in the dividend D. To understand why, suppose the imbalance parameter m is large, yet the dealer sets the mid-quote price equal to her forecast (that is, $p_t = w_t$). The dealer then expects the sell demand to be much larger than the buy demand. Thus, in order to avoid inventory buildup and to attract more buyers, she must lower her price well below her forecast. A similar intuition works when the dividend D is large, but the above argument reverses: because investors prefer getting a large dividend, to attract more sellers the dealer must set a price higher than the forecast.

[33] For instance, equation (9) implies that the expected quantity traded at the ask is $\mathsf{E}_t(Q_t^b) = \frac{k}{2}(w_t - a_t) + \ell$, which is decreasing in a_t.

[34] We stress that in our model price impact is caused by inventory considerations and not by adverse selection between the dealer and the traders. Nevertheless, adverse selection occurs as long as the dealer's signal precision f is not infinite. The interested reader can separate the effect of inventory and information by analyzing more carefully the dealer's signal structure described in Appendix A.2. There we see that the informativeness of trading depends on the noise parameter Σ_N. The signal structure, however, is chosen there to justify the reduced-form assumption in (5). Under that structure, the dealer is only concerned about her forecast just before trading, and not on what effect trading has on this forecast. But under a different signal structure this fact is no longer true, e.g., if we set $\tilde{V}_\eta = V_\eta$ and $\tilde{V}_\psi = V_\psi$ (see the discussion before equation (A15)).

4.3 Optimal Monitoring and the QT Ratio

We now discuss the dealer's optimal monitoring rate q. Because the trading rate is normalized to one, we identify the *quote-to-trade ratio* as the monitoring rate q:

$$q \;=\; \text{Quote-to-Trade Ratio.} \tag{13}$$

Thus far, the description of the equilibrium does not depend on a particular specification for the precision function $F(q)$ or the monitoring function $C(q)$. To provide explicit formulas, however, we now assume that both functions are linear: $F(q) = fq$ and $C(q) = cq$. In the proof of Proposition 2, we describe the equilibrium conditions for more general F and C. Proposition 2 shows how to compute the dealer's optimal monitoring rate, which as we discussed above is the equilibrium QT ratio.

Proposition 2. *The dealer's optimal monitoring rate q satisfies*

$$q^2 \;=\; \frac{k(k\alpha+1)}{fc} \;=\; \frac{k\beta}{fc}\,\frac{(\gamma-\omega)+\sqrt{(\gamma-\omega)^2+\frac{4\gamma}{\beta k}}}{-(\gamma+\omega)+\sqrt{(\gamma-\omega)^2+\frac{4\gamma}{\beta k}}}. \tag{14}$$

Using the formula in (14), we provide some comparative statics for q.

Corollary 1. *The QT ratio q is increasing in investor elasticity k and inventory aversion γ, and is decreasing in signal precision f and in monitoring cost c.*

If the investor elasticity k is larger, investors are more sensitive to the quotes, and the dealer increases her monitoring rate to prevent large fluctuations in inventory. If the inventory aversion γ is larger, the dealer is relatively more concerned about her inventory than about her profit. She then increases her monitoring rate to stay closer to the fundamental value, such that her inventory does not fluctuate too much.

Empirically, the parameter γ is not directly observable. One possible proxy for γ is the number of market makers that provide liquidity in the asset: arguably, a larger number of intermediaries is correlated with a smaller γ for the representative dealer. With this interpretation, a larger number of market makers should correspond to a smaller dealer monitoring rate, hence to a smaller QT ratio. But this is exactly the first part of the stylized empirical fact SF4 in Section 3.2.

If the signal precision parameter f is smaller, the dealer gets noisier signals each time she monitors, hence she must monitor the market more often in order to avoid getting a large inventory. As a result, in neglected stocks where we expect dealer's signals to be noisier, the QT ratio q should be larger. This is counter-intuitive, since one could think that the QT ratio is actually smaller in neglected stocks. This theoretical result is, however, consistent with our stylized empirical fact SF1 that the QT ratio is larger in neglected stocks (with low market capitalization, institutional ownership, analyst coverage, trading volume, and volatility).

Similarly, if the monitoring cost parameter c is smaller, the dealer can afford to monitor more often in order to maintain the same precision, which increases the QT ratio. There is much evidence that the costs of monitoring have decreased dramatically in recent times (see Hendershott et al., 2011). Accordingly, our stylized empirical fact SF2 documents a sharp rise in the QT ratio, especially in the second part of our sample (2003–2012).

4.4 Intermediation Irrelevance

In this section, we study the equilibrium evolution of the dealer's inventory. As we see in Proposition 1, the dealer's inventory is an important state variable. Corollary 2 computes its long-term mean and describes the equilibrium quotes by considering deviations of the dealer's inventory from its long-term mean.

Corollary 2. *The dealer's inventory is an AR*(1) *process:*

$$x_{t+1} - \bar{x} \;=\; \frac{1}{1+k\alpha}\big(x_t - \bar{x}\big) + \varepsilon_t, \qquad \bar{x} \;=\; \frac{1+k\alpha}{k\alpha}\,\frac{(1-\beta)m + \beta k D/2}{1-\beta+k\alpha}. \tag{15}$$

where ε_t *is IID with mean zero and variance* $\frac{k^2}{fq} + \Sigma_N$*. The mid-quote price satisfies*

$$p_t \;=\; w_t - \lambda\big(x_t - \bar{x}\big) - \bar{\delta}, \qquad \bar{\delta} \;=\; \frac{2m}{k}. \tag{16}$$

The mean inventory $\bar{x}$ represents the dealer's bias in holding the risky asset. In HM2014 both m and D are zero, and therefore the mean inventory $\bar{x}$ is also zero. In our case both m and D are positive, hence $\bar{x}$ is also positive. Intuitively, the case when

m is positive corresponds to the case when investors are risk averse and the risky asset is in positive net supply (see the micro-foundations in Appendix C). But the dealer also behaves approximately as a risk averse investor because of the quadratic penalty in inventory (see Footnote 30). Therefore, our model becomes essentially a risk sharing problem, in which the dealer holds a positive inventory on average.[35]

If we write the mid-quote equation (16) at both t and $t+1$, we compute

$$p_{t+1} - p_t \;=\; w_{t+1} - w_t + \psi\left(x_t - \bar{x}\right) - \lambda\varepsilon_{t+1}, \qquad \psi \;=\; \frac{\lambda k\alpha}{1+k\alpha}. \tag{17}$$

We define the *neutral state* the situation in which the dealer's inventory is at its long-term mean ($x_t = \bar{x}$). In this state, equation (17) implies that the expected change in price is zero, which in the language of HM2014 means that there is no price pressure.

We define the *pricing discount* as the difference between the dealer's forecast and the mid-quote price,

$$\delta_t \;=\; w_t - p_t. \tag{18}$$

From (16) it follows that the pricing discount in the neutral state is the same as its long-term average, and is equal to $\bar{\delta} = 2m/k$. Note that this value is independent on the characteristics of the dealer, that is, on the inventory aversion γ, the signal precision f, or the monitoring cost c. We have thus proved the main result of this section.

Corollary 3 (Intermediation Irrelevance)**.** *The average pricing discount is $\bar{\delta} = 2m/k$, and does not depend on dealer characteristics.*

In particular, the average pricing discount does not depend on the dealer's inventory aversion γ. This is because in the neutral state there is no price pressure and the dealer just needs to balance the order flow such that the inventory does not accumulate in either direction. This result is surprising, because one may expect the discount to be larger if the dealer has a larger inventory aversion γ. But while a larger coefficient γ just increases the speed of convergence of the pricing discount to its mean, it does not change the mean itself, which depends only on the properties of the order flow.[36]

[35] Even if $m = 0$, the dealer tends to hold inventory when the dividend D is positive. Indeed, in that case the dealer must increase her quotes to attract sellers (see equation (12)), which tends to raise her inventory and thus increase the dividend collected.

[36] According to (16), the equilibrium discount satisfies $\delta_t - \bar{\delta} = \lambda(x_t - \bar{x})$, and thus δ_t and x_t are

The average pricing discount $\bar{\delta}$ does depend on the properties of the order flow: the imbalance parameter m and the investor elasticity k. If the imbalance parameter m is larger, the dealer expects the difference between the sell and buy demands to be larger. To compensate, the dealer must lower price to encourage demand, and therefore increase the discount. If the investor elasticity k is larger, investors are more sensitive to mispricing and therefore trade more intensely when the price is different from the fundamental value. To prevent an expected accumulation of inventory, the dealer must then set the price closer to her forecast, which implies a lower discount.

Empirically, it is difficult to find evidence for intermediation irrelevance, since parameters such as the inventory aversion γ are not easily observable. One proxy for γ that was suggested above is the number of market makers that provide liquidity in the asset (see the discussion after Corollary 1). The intuition is that a larger number of intermediaries is correlated with a smaller γ for the representative dealer. In the next section we see that the cost of capital (expected return) of a stock is in one-to-one correspondence with the pricing discount. Therefore, according to our intermediation irrelevance result the cost of capital of a stock should be unrelated to the number of market makers active in that stock. But this is exactly the second part of the stylized empirical fact SF4 in Section 3.2.

4.5 Cost of Capital

In this section, we define and analyze the cost of capital in the context of our model. We consider the point of view of an econometrician that has access to the quote and trade information, but not necessarily to the dealer's inventory and forecast (in practice, dealers' inventories and forecasts are not public information). The expected return (including dividends) at date t is then

$$r_t \;=\; \frac{\mathsf{E}_t(p_{t+1}) + D - p_t}{p_t}, \tag{19}$$

both $AR(1)$ processes with the same autoregressive coefficient: $1/(1+k\alpha)$. From (10), α is increasing in γ, therefore the speed of mean reversion of both processes is also increasing in γ.

where E_t be the expectation operator conditional on the past information, p_t is the mid-quote price, and D is the dividend per share.

To simplify the presentation, we assume that the dealer's inventory starts at its long-term mean, that is, we set $x_0 = \bar{x}$. In this neutral state the price does not change in expectation (see Section 4.4). We define the *cost of capital* to be the expected return in the initial state.[37] Denote the initial dealer forecast by $w_0 = \overline{w}$. Then, the cost of capital is

$$r \;=\; \frac{D}{\overline{w} - \bar{\delta}} \;=\; \frac{D}{\overline{w} - \frac{2m}{k}}. \tag{20}$$

Note that the cost of capital is in one-to-one correspondence with the pricing discount $\bar{\delta} = 2m/k$. Thus, the intermediary irrelevance result (Corollary 3) applies equally to the cost of capital, which should therefore not depend on dealer characteristics. The cost of capital should depend on the characteristics of the order flow: the imbalance parameter m and the investor elasticity k. The intuition for this dependence is the same as in the discussion after Corollary 3.

The next result connects the cost of capital to the equilibrium QT ratio.

Corollary 4 (QT Effect)**.** *Holding all parameters constant except for the investor elasticity k, there is an inverse relation between the cost of capital and the QT ratio.*

Thus, the key driver of the QT effect in our model is investor elasticity. When k is larger, Corollary 1 shows that the QT ratio is also larger: because traders are more sensitive to the quotes, in order to prevent large fluctuations in inventory the dealer must monitor more often. At the same time, when k is larger, the cost of capital is smaller: because investors trade more intensely when the price differs from the fundamental value, in order to prevent an expected accumulation of inventory the dealer must set the price closer to her forecast, which implies a lower discount and hence a lower cost of capital.

In Appendix C we provide micro-foundations for the order flow, and we show that

[37] We define the cost of capital only in the initial (neutral) state, because we want to avoid price pressures that appear later in other states. Another reason is that in general it is difficult to analyze risk premia in dynamic microstructure models. Indeed, if the expected return is constant, return compounding implies that the price process grows exponentially on average, and to keep up the fundamental value should also follow a geometric Brownian motion. But to maintain a tractable model we need the fundamental value to follow an arithmetic Brownian motion.

the investor elasticity k is larger when traders are less risk averse. Therefore, trader risk aversion drives the QT effect: less risk averse traders cause both a larger QT ratio and a smaller cost of capital.

The QT effect is documented empirically in the cross-section of stock returns by the stylized empirical fact SF3 in Section 3.2. The inverse relation between the cost of capital and the QT ratio hold empirically in both parts of our sample: 1994–2002 and 2003–2012. This is the stylized empirical fact SF3′ in Section 3.3.

5 Conclusion

This paper studies the quote-to-trade (QT) ratio and its relation with liquidity, price discovery, and expected returns. Empirically, we find that the QT ratio is larger in stocks that are small, illiquid or neglected. Our main finding (the QT effect) is that stocks with higher QT ratio tend to have lower average returns. Despite the fact that the QT ratio has increased significantly over time, especially in the second part of our sample (2003–2012), the QT effect is almost equally strong in both parts of the sample. Overall, our results are driven by quotes, not by trades.

In the theoretical part of our paper, we propose a model of the QT ratio that is consistent with our empirical findings. In equilibrium, market makers receive less precise signals in neglected stocks, and therefore monitor the market faster in those stocks, thus increasing their QT ratio. A QT effect arises in our model: market makers also monitor faster when investors have a higher elasticity (or, more fundamentally, when investors are less risk averse), which increases the QT ratio, but at the same time reduces mispricing and lowers expected returns. If we interpret the representative dealer's risk tolerance (inverse risk aversion) as a proxy for the number of market makers, then our model provides two additional empirical predictions: a larger number of market makers lowers the QT ratio, but has no effect on expected return. We find that indeed these results hold in the data.

Our results help understand the determinants of the QT ratio, and which determinants are related to liquidity and the cost of capital. Many proposals to regulate automated trading in financial markets are based on the QT ratio, under the supposi-

tion that this variable reflects unnecessary or even destabilizing high-frequency trading activity. Our results, however, suggest that HFT activity may have little effect on important variables such as the cost of capital. Indeed, our theoretical results suggest that in general the QT ratio is determined by the activity of various market participants, such as market makers and investors, and we find empirical evidence for this connection even before the proliferation of HFT in the late 2000s. At the same time, our intermediation irrelevance result implies that the cost of capital is affected only by the characteristics of investors (such as risk aversion) and not by the characteristics of intermediaries such as market makers. Thus, we obtain a useful negative theoretical result: a certain regulation may affect the QT ratio, but as long as it does not affect the investors' liquidity demand, the cost of capital should not change.

REFERENCES

Amihud, Y. (2002). "Illiquidity and stock returns: Cross-section and time-series effects." *Journal of Financial Markets*, 5, 31–56.

Amihud, Y. and H. Mendelson (1986). "Asset pricing and the bid-ask spread." *Journal of Financial Economics*, 17, 223–249.

Amihud, Y., H. Mendelson, and L. H. Pedersen (2005). "Liquidity and asset prices." *Foundations and Trends in Finance*, 1(4), 1–96.

Ang, A., R. J. Hodrick, Y. Xing, and X. Zhang (2009). "High idiosyncratic volatility and low returns: International and further U.S. evidence." *Journal of Financial Economics*, 91, 1–23.

Angel, J., L. Harris, and C. Spatt (2011). "Equity trading in the 21st century." *Quarterly Journal of Finance*, 1, 1–53.

Asparouhova, E., H. Bessembinder, and I. Kalcheva (2010). "Liquidity biases in asset pricing tests." *Journal of Financial Economics*, 96, 215–237.

Asparouhova, E., H. Bessembinder, and I. Kalcheva (2013). "Noisy prices and inference regarding returns." *Journal of Finance*, 68, 665–714.

Boehmer, E., K. Y. L. Fong, and J. J. Wu (2015). "International evidence on algorithmic trading." Working paper.

Brennan, M. and A. Subrahmanyam (1996). "Market microstructure and asset pricing: On the compensation for illiquidity in stock returns." *Journal of Financial Economics*, 41, 441–464.

Brennan, M. J., T. Chordia, and A. Subrahmanyam (1998). "Alternative factor specifications, security characteristics and the cross-section of expected stock returns." *Journal of Financial Economics*, 49, 345–373.

Brogaard, J., B. Hagströmer, L. Nordén, and R. Riordan (2015). "Trading fast and slow: Colocation and liquidity." *The Review of Financial Studies*, 28(12), 3407–3443.

Brogaard, J., T. Hendershott, and R. Riordan (2014). "High frequency trading and price discovery." *Review of Financial Studies*, 28, 3407–3443.

Brogaard, J., T. Hendershott, and R. Riordan (2016). "High frequency trading and the 2008 short sale ban." *Journal of Financial Economics*, Forthcoming.

Chordia, T., R. Roll, and A. Subrahmanyam (2000). "Commonality in liquidity." *Journal of Financial Economics*, 56, 3–28.

Chordia, T., R. Roll, and A. Subrahmanyam (2002). "Order imbalance, liquidity, and market returns." *Journal of Financial Economics*, 65, 111–130.

Chordia, T., A. Subrahmanyam, and V. Anshuman (2001). "Trading activity and expected stock returns." *Journal of Financial Economics*, 59(1), 3–32.

Chordia, T., A. Subrahmanyam, and Q. Tong (2011). "Trends in the cross-section of expected stock returns." Working paper.

Conrad, J., S. Wahal, and J. Xiang (2015). "High-frequency quoting, trading, and the efficiency of prices." *Journal of Financial Economics*, 116(2), 271–291.

Duarte, J. and L. Young (2009). "Why is PIN priced?" *Journal of Financial Economics*, 91, 119–138.

Easley, D., S. Hvidkjaer, and M. O'Hara (2002). "Is information risk a determinant of asset returns?" *Journal of Finance*, 57, 2185–2221.

Fama, E. and J. MacBeth (1973). "Risk, return, and equilibrium: Empirical tests." *Journal of Political Economy*, 81, 607–636.

Fama, E. F. and K. R. French (1992). "The cross-section of expected stock returns." *Journal of Finance*, 47, 427–465.

Fama, E. F. and K. R. French (1993). "Common risk factors in the returns on stocks and bonds." *Journal of Financial Economics*, 33, 3–56.

Goyenko, R. Y., C. W. Holden, and C. A. Trzcinka (2009). "Do liquidity measures measure liquidity?" *Journal of Financial Economics*, 92, 153–181.

Hagströmer, B. and L. Nordén (2013). "The diversity of high-frequency traders." *Journal of Financial Markets*, 16(4), 741–770.

Harvey, C. R., Y. Liu, and H. Zhu (2016). " and the cross-section of expected returns." *The Review of Financial Studies*, 29(1), 5–68.

Hasbrouck, J. (2009). "Trading costs and returns for U.S. equities: Estimating effective costs from daily data." *Journal of Finance*, 64, 1445–1477.

Hendershott, T., C. M. Jones, and A. J. Menkveld (2011). "Does algorithmic trading improve liquidity." *Journal of Finance*, 66(1), 1–33.

Hendershott, T. and A. Menkveld (2014). "Price pressures." *Journal of Financial Economics*, 114(3), 405–423.

Ho, T. and H. R. Stoll (1981). "Optimal dealer pricing under transactions and return uncertainty." *Journal of Financial Economics*, 9, 47–73.

Hoffmann, P. (2014). "A dynamic limit order market with fast and slow traders." *Journal of Financial Economics*, 113, 159–169.

Hong, H., T. Lim, and J. C. Stein (2000). "Bad news travels slowly: Size, analyst coverage, and the profitability of momentum strategies." *The Journal of Finance*, 55(1), 265–295.

Jegadeesh, N. and S. Titman (1993). "Returns to buying winners and selling losers: Implications for stock market efficiency." *Journal of Finance*, 48(1), 65–91.

KUMAR, A. (2009). "Who gambles in the stock market?" *The Journal of Finance*, 64(4), 1889–1933.

Kyle, A. S. (1985). "Continuous auctions and insider trading." *Econometrica*, 53(6), 1315–1335.

Malinova, K., A. Park, and R. Riordan (2016). "Taxing high frequency market making: Who pays the bill?" Working paper.

Menkveld, A. (2016). "The economics of high-frequency trading: Taking stock." *Annual Review of Financial Economics*, 8, 1–24.

O'Hara, M. (2003). "Liquidity and price discovery." *Journal of Finance*, 58, 1335–1354.

O'Hara, M. (2015). "High frequency market microstructure." *Journal of Financial Economics*, 116(2), 257–270.

Pástor, L. and R. F. Stambaugh (2003). "Liquidity risk and expected stock returns." *Journal of Political Economy*, 111, 642–685.

Sargent, T. and L. Ljungqvist (2000). *Recursive Macroeconomic Theory, Second edition.* MIT.

Shumway, T. (1997). "The delisting bias in CRSP data." *Journal of Finance*, 52, 327–340.

Subrahmanyam, A. and H. Zheng (2016). "Limit order placement by high-frequency traders." Working paper.

Table 1: Characteristics of quote-to-trade ratio portfolios

The table presents the monthly average characteristics for 10 quote-to-trade ratio (QT) portfolios constructed in month t. Portfolio 1 consists of stocks with the lowest QT and portfolio 10 consists of stocks with the highest QT in month t. Each portfolio contains on average 309 stocks. Stocks priced below \$2 or above \$1000 at the end of month t are removed. The sample period is June 1994 to October 2012. For each QT decile, we compute the cross-sectional mean characteristic for month $t+1$. The reported characteristics are computed as the time-series mean of the mean cross-sectional characteristic. Column (2) is the QT level, columns (3) and (4) are the number of trades and quote updates in thousands, column (5) shows market capitalization (in million USD), columns (6) and (7) show the share volume (in million shares) and USD volume traded (in million USD), columns (8) and (9) show the quoted spread and relative spread (in % of the mid-quote), column (10) shows the Amihud illiquidity ratio (ILR) in %, column (11) shows volatility (calculated as the absolute monthly return in %), column (12) shows price, column (13) shows the average Book-to-Market value measured at the end of the previous calendar year, and column (14) shows the average monthly portfolio return in excess of the risk free rate (r_{t+1}) for each portfolio. Individual stock returns are mid-quote returns corrected for delisting bias in CRSP by adding a -30% return for delisting codes 500 and 520-584.

Average portfolio characteristics at $t+1$													
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
QT		N(trades)	N(quotes)	MCAP	VOLUME (mill.)		SPREAD						
portf (t)	QT	(x 1000)	(x 1000)	(mill.)	Shares	USD	Quoted	Relative (%)	ILR (%)	VOLA (%)	PRC	BM	$r_{p,t+1}$ (%)
1	1.4	131	187	8912	75.1	1725	0.140	1.41	2.75	3.71	15.7	0.63	1.52
2	2.8	47	215	3629	20.0	692	0.160	1.61	3.75	3.58	18.0	0.63	1.30
3	3.9	33	224	2863	13.5	533	0.178	1.71	3.96	3.34	20.3	0.64	1.10
4	5.2	26	229	2497	10.5	444	0.201	1.82	4.41	3.22	22.2	0.64	1.04
5	6.7	20	216	2091	8.1	352	0.233	1.98	5.25	3.13	23.7	0.65	0.95
6	9.0	15	201	2315	7.2	321	0.275	2.14	6.79	2.84	24.6	0.70	0.81
7	13.9	11	166	3302	7.7	335	0.259	1.98	5.69	2.32	24.9	0.76	0.94
8	20.8	7	131	2034	4.7	207	0.278	1.86	4.55	1.98	25.7	0.76	0.84
9	43.6	3	97	1431	2.8	126	0.323	1.95	5.13	1.89	25.9	0.78	0.84
10	154.4	1	85	828	1.2	58	0.441	2.38	7.91	1.73	27.9	1.01	0.65

Table 2: Determinants of the quote-to-trade ratio

The table shows the two-way fixed effects panel regression on the determinants of the quote-to-trade ratio (QT). The dependent variable is the monthly QT. The independent variables are: annual number of analysts covering a stock (*AN.COVER.*), dummy equal to one when a company has no analyst coverage and zero otherwise (*NO ANALYST*), quarterly institutional ownership (*INST.OWN.*), log-book-to-market as of the previous year end (*BM*); previous month return (r_{t-1}); as well as contemporaneous (monthly) variables: log-market capitalization (*MCAP*), log-price (*PRC*), share trading volume (*VOLUME*), Amihud illiquidity ratio (*ILR*), bid-ask spread (*SPREAD*), volatility (*VOLA*), and number of NASDAQ market makers (*MM*). Standard errors are double-clustered at the stock and month level.

	(1)	(2)	(3)
AN.COVER.	-0.87***	-0.62***	-0.72***
	(-6.83)	(-5.80)	(-5.74)
NO ANALYST	20.54***	-16.69***	-11.51***
	(3.92)	(-4.51)	(-3.01)
INST.OWN.	-22.91***	-47.94***	-59.65***
	(-4.77)	(-6.46)	(-9.25)
BM	16.52***	-3.68	-5.27
	(3.46)	(-1.41)	(-1.55)
MCAP	-1.66	-2.78*	-4.42**
	(-1.32)	(-1.71)	(-2.04)
*R*1	-14.19***	0.18	-0.03
	(-3.64)	(0.10)	(-0.01)
PRC	0.56***	0.36***	0.55***
	(5.98)	(3.91)	(3.12)
VOLUME	1.60e-08**	-4.30e-08***	-4.99e-08***
	(2.21)	(-6.81)	(-6.85)
ILR	3.49	-3.52	-2.86
	(1.21)	(-1.44)	(-0.89)
SPREAD	-8.79***	0.63	-2.67
	(-4.16)	(0.21)	(-0.49)
VOLA	-57.01***	-19.75***	-15.53***
	(-5.24)	(-4.28)	(-3.37)
MM			-0.46***
			(-3.78)
Stock FE	NO	YES	YES
Time FE	NO	YES	YES
N	672,952	672,888	453,736
Adj. R^2	0.03	0.19	0.20

Table 3: Risk-adjusted returns for quote-to-trade ratio portfolios

The table shows risk-adjusted monthly returns for various portfolios sorted on the quote-to-trade ratio (QT). The α's reported in the table are time series averages of intercepts (risk-adjusted returns) obtained from 24-month rolling window regressions. The monthly returns of the QT portfolios are risk-adjusted using several asset pricing models: CAPM, Fama and French (1993) model (FF3), a model that adds the Pástor and Stambaugh (2003) traded liquidity factor (FF3+PS), a five factor model that adds a momentum factor (FF4+PS), and a model that adds the PIN factor for the period June 1994 to December 2002 (FF4+PS+PIN). We show the alpha for the lowest and highest QT portfolios and the alpha for the difference in returns between the low and high portfolios. In Panel A, stocks are assigned to ten portfolios based on their QT level in month t. Then returns are calculated for each portfolio for month $t+1$. Panels B and C show stocks assigned to 25 and 50 portfolios. ***, **, and * indicate rejection of the null hypothesis that the risk-adjusted portfolio returns are significantly different from zero at the 1%, 5%, and 10% level, respectively.

	Risk-adjusted returns (%)				
	CAPM	FF3	FF3+PS	FF4+PS	FF4+PS+PIN
	Panel A: 10 QT portfolios				
α_1	0.92*	1.05***	1.03***	1.67***	1.69***
α_{10}	0.37	-0.08	-0.09	0.09	0.08
α_{1-10}	0.55	1.14***	1.11***	1.58***	1.61***
	Panel B: 25 QT portfolios				
α_1	0.89	1.10**	1.10**	1.88***	1.91***
α_{25}	0.22	-0.22	-0.21	-0.03	-0.04
α_{1-25}	0.67	1.31***	1.31**	1.91***	1.95***
	Panel C: 50 QT portfolios				
α_1	0.60	0.82*	0.81*	1.56***	1.57***
α_{50}	0.08	-0.33	-0.34	-0.18	-0.19
α_{1-50}	0.52	1.15**	1.15**	1.74***	1.76***

Table 4: Stock risk-adjusted returns and quote-to-trade ratio

The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted monthly returns on firm characteristics. The dependent variable is the risk-adjusted return $r^a_{i,t} = r_{i,t} - \sum_{j=1}^{J} \beta_{i,j,t-1} F_{j,t}$, where the risk factors $F_{j,t}$ come from the FF4+PS model (market, size, value, momentum and the Pástor and Stambaugh (2003) traded liquidity factor). The firm characteristics are measured in month $t-1$. The characteristics included are: quote-to-trade ratio (QT), relative bid/ask spread ($SPREAD$), Amihud illiquidity ratio (ILR), log-market capitalization ($MCAP$), log-book-to-market ratio (BM) calculated as the natural logarithm of the book value of equity divided by the market value of equity from the previous fiscal year, previous month return ($R1$), cumulative return from month $t-2$ to $t-12$ ($R212$), idiosyncratic volatility ($IDIOVOL$) measured as the standard deviation of the residuals from a FF3 regression of daily raw returns within each month as in Ang, Hodrick, Xing, and Zhang (2009), log-dollar-volume ($USDVOL$), log-price (PRC), and number of NASDAQ market makers (MM). All coefficients are multiplied by 100. The standard errors are corrected by using the Newey-West method with 12 lags. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Const.	0.006***	0.004***	0.013***	0.011***	0.036***	0.041***
$QT_{i,t-1}$	-0.222***	-0.244***	-0.286***	-0.297***	-0.119***	-0.088*
$SPREAD_{i,t-1}$		0.141***		0.067**	0.035	0.041
$ILR_{i,t-1}$			0.097***	0.075***	-0.004	0.035
$MCAP_{i,t-1}$					-0.224***	-0.321***
$BM_{i,t-1}$					0.073	0.074
$R1_{i,t-1}$					-4.413***	-4.651***
$R212_{i,t-1}$					0.061	0.086
$IDIOVOL_{i,t-1}$					-12.544***	-16.133***
$USDVOL_{i,t-1}$					0.163***	0.279***
$PRC_{i,t-1}$					-0.433***	-0.556***
$MM_{i,t-1}$						0.000
R^2	0.00	0.01	0.01	0.01	0.04	0.04
Time series (months)	216	216	216	216	216	216

Table 5: Quotes versus Trades

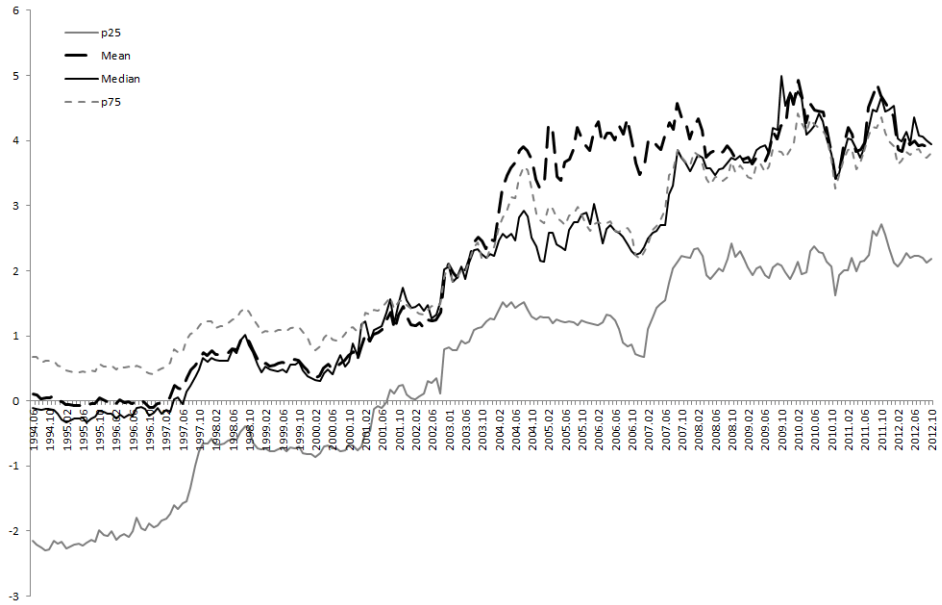
The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted monthly returns on firm characteristics including the number of quotes and trades. The dependent variable is the risk-adjusted return $r^a_{i,t} = r_{i,t} - \sum_{j=1}^{J} \beta_{i,j,t-1} F_{j,t}$, where the risk factors $F_{j,t}$ come from the FF4+PS model (market, size, value, momentum and the Pástor and Stambaugh (2003) traded liquidity factor). The firm characteristics are measured in month $t-1$. The characteristics included are: number of quotes ($QUOTE$), number of trades ($TRADE$), relative bid/ask spread ($SPREAD$), Amihud illiquidity ratio (ILR), log-market capitalization ($MCAP$), log-book-to-market ratio (BM) calculated as the natural logarithm of the book value of equity divided by the market value of equity from the previous fiscal year, previous month return ($R1$), cumulative return from month $t-2$ to $t-12$ ($R212$), idiosyncratic volatility ($IDIOVOL$) measured as the standard deviation of the residuals from a FF3 regression of daily raw returns within each month as in Ang et al. (2009), log-dollar-volume ($USDVOL$), and log-price (PRC). All coefficients are multiplied by 100. The standard errors are corrected by using the Newey-West method with 12 lags. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Const.	0.018***	0.016***	0.015***	0.014***	0.025***	0.025***
$QUOTE_{i,t-1}$	-0.326***	-0.342***	-0.285***	-0.307***	-0.097**	-0.118**
$TRADE_{i,t-1}$	0.206*	0.245**	0.286**	0.298**	-0.121	-0.104
$SPREAD_{i,t-1}$		0.052		0.035		0.010
$ILR_{i,t-1}$			0.107**	0.087**		0.001
$MCAP_{i,t-1}$					-0.245***	-0.228***
$BM_{i,t-1}$					0.060	0.065
$R1_{i,t-1}$					-4.562***	-4.464***
$R212_{i,t-1}$					0.050	0.057
$IDIOVOL_{i,t-1}$					-9.315***	-11.320***
$USDVOL_{i,t-1}$					0.376***	0.057***
$PRC_{i,t-1}$					-0.604***	-11.320***
R^2	0.01	0.01	0.01	0.01	0.04	0.04
Time series (months)	216	216	216	216	216	216

Figure 1: Time series evolution in the quote-to-trade ratio

The graphs show the time series of the natural logarithm of the quote-to-trade ratio $QT_{i,t} = \frac{N(quotes)_{i,t}}{N(trades)_{i,t}}$. Panel A shows the monthly time series of the cross-sectional mean, median, 25th, and 75th percentile of the QT variable. Panel B shows the monthly average number of quote updates and number of trades.

(a) *Quote-to-Trade Ratio*



(b) *Quotes and Trades*

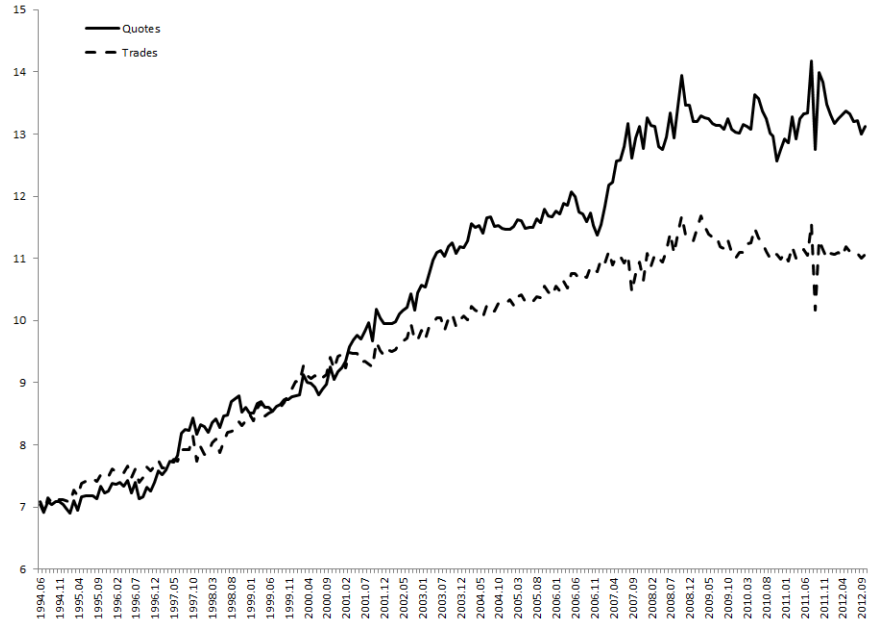
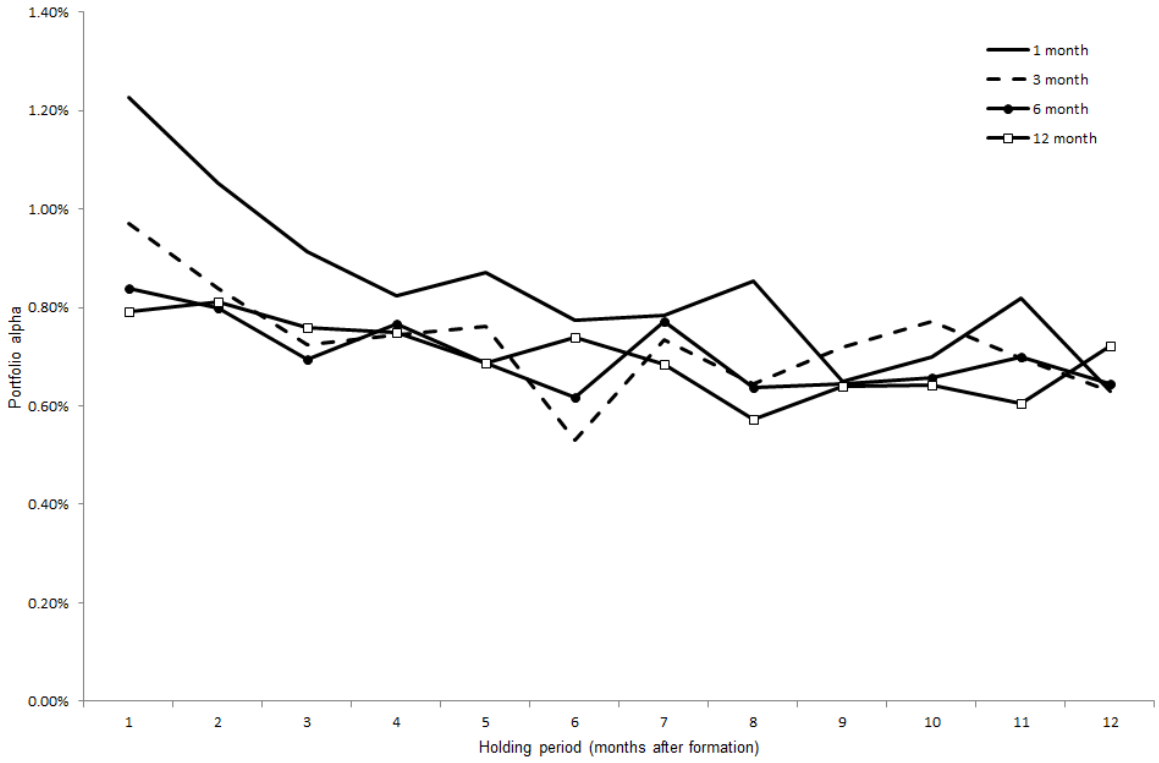


Figure 2: Portfolio alphas for different holding horizons and formation periods

The figure shows the long-short alpha for the difference between risk-adjusted returns for low-quote-to-trade ratio (QT1) and high-quote-to-trade ratio (QT25) portfolios for 25 QT-sorted portfolios across different holding and formation periods. The alphas are estimated using the FF4+PS model (market, size, value, momentum and the Pástor and Stambaugh (2003) traded liquidity factor). Stocks are assigned into portfolios based on their quote-to-trade ratio level over the past 1, 3, 6, and 12 months (formation period), and holding horizons range from 1 to 12 months.



Appendix A. Monitoring

The purpose of this section is to provide micro-foundations for the dealer's precision function in (5). If w_t is the dealer's forecast of v_t before trading at t, equation (5) implies that her forecast precision $1/\mathsf{Var}(v_t - w_t)$ is independent of the trading time t, and has a linear expression, $F(q) = fq$, in the monitoring rate q. In this section, we show that this specification arises from an actual set of signals observed by the dealer.

Recall that in our model in Section 4.1, trading takes place at integer times $t = 0, 1, 2, \ldots$, while monitoring takes place at fractional times $\frac{0}{q}, \frac{1}{q}, \frac{2}{q}, \ldots$, where q is the monitoring rate. In this Appendix, we consider the monitoring rate q to be a positive integer, while in the rest of the paper we use the results derived here in reduced form, and consider q to be any positive real number.

To simplify notation, we index monitoring times by $\tau = 0, 1, 2, \ldots$ rather than by the corresponding fractional times. With this notation, trading takes place at $\tau = 0, q, 2q, \ldots$, which are integer multiples of the monitoring rate; by convention, we assume that on these dates monitoring occurs before trading. Equations (9) imply that traders' order flow satisfies

$$\begin{aligned} Q^b_\tau &= \frac{k}{2}(v_\tau - a_\tau) + \ell - m + \varepsilon^b_\tau, \quad \text{with} \quad \varepsilon^b_\tau \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \\ Q^s_\tau &= \frac{k}{2}(b_\tau - v_\tau) + \ell + m + \varepsilon^s_\tau, \quad \text{with} \quad \varepsilon^s_\tau \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \end{aligned} \tag{A1}$$

A.1. Uninformative Trading

We first analyze the simpler case when the trading process is uninformative to the dealer. Formally, this occurs when the trading noise measured by Σ_N is sufficiently large (see equation (A17) below). In this case, we ignore the trading process altogether and focus instead on the monitoring process. Denote by $\mathcal{I}_\tau$ the dealer's information set after monitoring at τ, and by $w_\tau = \mathsf{E}(v_\tau|\mathcal{I}_\tau)$ the dealer's forecast at τ.

We now show that any positive function $F(q)$, not necessarily linear, can arise as the dealer's precision function for a certain set of signals. Define

$$G \;=\; G(q) \;=\; \frac{1}{F(q)}. \tag{A2}$$

Fix $q > 0$, and define $V_\eta = V_\eta(q) > 0$ as follows: if $F(q) \le q/\Sigma_v$, choose any $V_\eta > 0$; and if $F(q) > q/\Sigma_v$, choose any $V_\eta \in \left(0, \frac{1}{F(q) - q/\Sigma_v}\right)$. Also, define $V_v = V_v(q)$ and $V_\psi = V_\psi(q)$ by

$$V_v \;=\; \frac{\Sigma_v}{q}, \qquad V_\psi \;=\; G^2\,\frac{V_\eta + V_v}{V_\eta V_v} - G. \tag{A3}$$

Clearly, $V_v > 0$. We show that $V_\psi > 0$ as well. Indeed, from the definition of V_η, we see that $\left(F(q) - q/\Sigma_v\right)V_\eta < 1$ for all $q > 0$. Using the notation above, this is the same as $\left(\frac{1}{G} - \frac{1}{V_v}\right)V_\eta < 1$, which is equivalent to $\frac{1}{G} < \frac{1}{V_v} + \frac{1}{V_\eta}$. Thus, $G\frac{V_\eta + V_v}{V_\eta V_v} > 1$ or equivalently

$V_\psi = G\big(G\frac{V_\eta+V_v}{V_\eta V_v} - 1\big) > 0$. Note that equation (A3) implies

$$\frac{G^2}{G+V_\psi} \;=\; \frac{V_v V_\eta}{V_v+V_\eta}. \tag{A4}$$

We define the signal observed by the dealer at $\tau = 0$. Since we can choose freely the initial variance $\mathsf{Var}(v_0) = \Sigma_{v_0}$, consider $\Sigma_{v_0} > G$, and suppose that at $\tau = 0$ the dealer observes $s_0 = v_0 + \nu$, with $\nu \sim \mathcal{N}\big(0, \frac{G\Sigma_{v_0}}{\Sigma_{v_0}-G}\big)$. Then, the dealer's forecast is $w_0 = \mathsf{E}(v_0|s_0) = \beta_0 s_0$, where $\beta_0 = G/\Sigma_{v_0}$. A direct computation shows that indeed $\mathsf{Var}(v_0 - w_0) = G$. Thus, if we define

$$G_\tau \;=\; \mathsf{Var}(v_\tau - w_\tau), \quad \tau \;\geq\; 0, \tag{A5}$$

we have $G_0 = G$.

At each $\tau = 1, 2, \ldots$, the dealer observes two signals:

$$\begin{cases} r_\tau = (v_{\tau-1} - w_{\tau-1}) + \psi_\tau, \text{ with } \psi_\tau \overset{IID}{\sim} \mathcal{N}(0, V_\psi), \text{ and} \\ s_\tau = (v_\tau - v_{\tau-1}) + \eta_\tau \text{ with } \eta_\tau \overset{IID}{\sim} \mathcal{N}(0, V_\eta). \end{cases} \tag{A6}$$

Since the forecast is $w_\tau = \mathsf{E}(v_\tau|r_\tau, s_\tau, r_{\tau-1}, s_{\tau-1}, \ldots)$, its increment is $\Delta w_\tau = w_\tau - w_{\tau-1} = \mathsf{E}(v_\tau - w_{\tau-1}|r_\tau, s_\tau) = \mathsf{E}(v_\tau - v_{\tau-1}|s_\tau) + \mathsf{E}(v_{\tau-1} - w_{\tau-1}|r_\tau)$. Then,

$$\Delta w_\tau \;=\; \frac{V_v}{V_v+V_\eta}\, s_\tau + \frac{G_{\tau-1}}{G_{\tau-1}+V_\psi}\, r_\tau. \tag{A7}$$

We compute $v_\tau - w_\tau = v_{\tau-1} - w_{\tau-1} + \Delta v_\tau - \Delta w_\tau = \frac{V_\psi}{G_{\tau-1}+V_\psi}(v_{\tau-1} - w_{\tau-1}) - \frac{G_{\tau-1}}{G_{\tau-1}+V_\psi}\psi_\tau + \frac{V_\eta}{V_v+V_\eta}\Delta v_\tau - \frac{V_v}{V_v+V_\eta}\eta_\tau$. Taking variance on both sides, we obtain the recursive equation

$$G_\tau \;=\; \frac{G_{\tau-1}V_\psi}{G_{\tau-1}+V_\psi} + \frac{V_v V_\eta}{V_v+V_\eta}. \tag{A8}$$

From (A4), we substitute $\frac{V_v V_\eta}{V_v+V_\eta}$ by $\frac{G^2}{G+V_\psi}$, and the recursive equation (A8) becomes

$$G_\tau - G_{\tau-1} \;=\; \left(1 - \frac{V_\psi^2}{(G+V_\psi)(G_{\tau-1}+V_\psi)}\right)(G - G_{\tau-1}). \tag{A9}$$

Because $G_0 = G$, equation (A9) implies that G_τ is constant and equal to G for all τ.[38] Since $G = \frac{1}{F(q)}$, this finishes the proof.

For future reference, we use equation (A7) to compute $\mathsf{Var}(\Delta w_\tau) = \frac{V_v^2}{V_v+V_\eta} + \frac{G^2}{G+V_\psi}$. Equation (A4) then implies that $\mathsf{Var}(\Delta w_\tau) = \frac{V_v^2}{V_v+V_\eta} + \frac{V_v V_\eta}{V_v+V_\eta} = V_v$. Thus, we have proved that

$$\mathsf{Var}(\Delta w_\tau) \;=\; \mathsf{Var}(\Delta v_\tau) \;=\; V_v \;=\; \frac{\Sigma_v}{q}. \tag{A10}$$

[38] Note that the coefficient in front of $G - G_{\tau-1}$ in equation (A9) is a number in the interval $(0, 1)$. It is then straightforward to show that G_τ converges monotonically to the constant G regardless of the initial value G_0.

A.2. Informative Trading

We now analyze the general case when the trading process is informative, meaning that the noise parameter Σ_N can be any positive real number. Thus, beside the monitoring times, we also need to analyze the dealer's inference at the trading times $\tau = 0, q, 2q, \ldots$, where q is the monitoring rate and is a positive integer. (Recall that on the these dates monitoring occurs before trading.)

We now show that any linear function $F(q) = fq$ that satisfies a mild condition (see equation (A14) below) can arise as the dealer's precision function for a set of signals. As before, given $F(q)$ we define $G = G(q) = \frac{1}{F(q)} = \frac{1}{fq}$. Denote by $\mathcal{I}_\tau$ the dealer's information set after monitoring at τ, by $w_\tau = \mathsf{E}(v_\tau|\mathcal{I}_\tau)$ the dealer's forecast at τ, and by $e_\tau = v_\tau - w_\tau$ her forecast error. Then, equations (A1) become

$$\begin{aligned} Q^b_\tau &= \frac{k}{2}e_\tau - (a_\tau - w_\tau) + \ell + \varepsilon^b_\tau, \quad \text{with} \quad \varepsilon^b_\tau \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \\ Q^s_\tau &= -\frac{k}{2}e_\tau - (w_\tau - b_\tau) + \ell + \varepsilon^s_\tau, \quad \text{with} \quad \varepsilon^s_\tau \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_N/2\right), \end{aligned} \tag{A11}$$

At trading time $t = 0, q, 2q, \ldots$, define also

$$w_{\tau_+} = \mathsf{E}(v_\tau \mid \mathcal{I}_\tau, Q^b_\tau, Q^s_\tau), \qquad G_{\tau_+} = \mathsf{Var}(v_\tau - w_{\tau_+}). \tag{A12}$$

As in the informative case, we look for a stationary equilibrium, which here means that we want the dealer to have a periodic signal precision with periodicity equal to the monitoring rate q. Thus, the signal precision follows a periodic sequence of the form

$$G_0\ ,\ G_{0_+}\ ,\ G_1\ ,\ \cdots\ ,\ G_q\ =\ G_0\ ,\ G_{q_+}\ ,\ G_{q+1}\ ,\ \cdots \tag{A13}$$

We show that there is a simple solution for which G_τ are equal to $G = \frac{1}{fq}$, as long as the following condition is satisfied:

$$f\ >\ \frac{1}{\Sigma_v} \qquad \text{or} \qquad \frac{\Sigma_N f^2}{k^2}\ >\ \frac{1}{\Sigma_v} - f. \tag{A14}$$

To understand intuitively the role played by this condition, suppose (A14) fails to hold. This means that the noise component of trading, measured by Σ_N, is small. Then, the increase in precision $(1/G_0 - 1/G_{0_+})$ that comes from the information content of trading is also small. By contrast, the decrease in precision $(1/G_{0_+} - 1/G_1)$ that comes from the diffusion in fundamental value during the interval $[0, 1]$ is large, and thus the equation $G_0 = G_1$ cannot hold when (A14) fails. Note that the condition (A14) also translates into the requirement that the dealer's monitoring precision f is sufficiently high.

Suppose now condition (A14) is satisfied. We then assume that the dealer receives the same signals r_τ and s_τ as in the uninformative case, except for the monitoring times that come just after trading: $\tau = 1, q+1, 2q+1, \ldots$. At those times, we modify the variance of r_τ and s_τ, by defining new values for V_ψ and V_η. To see how this is done, consider the following cases:

- If $f > 1/\Sigma_v$, we multiply by q to obtain $fq = F = 1/G > 1/V_v$, where $V_v = \Sigma_v/q$. In this case, we choose $\frac{1}{V_\eta}$ in the positive interval $\left(\frac{1}{G} - \frac{1}{V_v}, \frac{\Sigma_N}{k^2G^2} + \frac{1}{G} - \frac{1}{V_v}\right)$.

- If $f \leq 1/\Sigma_v$, we have $1/G \leq 1/V_v$. Because q is a positive integer, condition (A14) implies $\frac{\Sigma_N f^2}{k^2} q^2 > \big(\frac{1}{\Sigma_v} - f\big)q$, which is equivalent to $\frac{\Sigma_N}{k^2 G^2} > \frac{1}{V_v} - \frac{1}{G}$. In this case, we choose $\frac{1}{\tilde{V}_\eta}$ in the interval $\big(0, \frac{\Sigma_N}{k^2G^2} + \frac{1}{G} - \frac{1}{V_v}\big)$. Since $1/G - 1/V_v \leq 0$, it follows that $\frac{1}{\tilde{V}_\eta}$ also belongs to the larger interval $\big(\frac{1}{G} - \frac{1}{V_v}, \frac{\Sigma_N}{k^2G^2} + \frac{1}{G} - \frac{1}{V_v}\big)$.

Thus, in both cases $\frac{1}{\tilde{V}_\eta}$ lies in the interval $\big(\frac{1}{G} - \frac{1}{V_v}, \frac{\Sigma_N}{k^2G^2} + \frac{1}{G} - \frac{1}{V_v}\big)$, or equivalently $\frac{1}{\tilde{V}_\eta} + \frac{1}{V_v} - \frac{1}{G}$ lies in the interval $\big(0, \frac{\Sigma_N}{k^2G^2}\big)$. Now define

$$\tilde{V}_\psi \;=\; \frac{\Sigma_N}{k^2}\,\frac{\big(\frac{\Sigma_N}{k^2G^2} + \frac{1}{G} - \frac{1}{V_v}\big) - \frac{1}{\tilde{V}_\eta}}{\tilde{V}_\eta - \big(\frac{1}{G} - \frac{1}{V_v}\big)}. \tag{A15}$$

From the above discussion, it follows that both $\tilde{V}_\eta$ and $\tilde{V}_\psi$ are positive, and hence when $\tau = 1, q+1, 2q+1, \ldots$, the modified signals r_τ and s_τ are well defined.

We show $G_\tau = G$ for all $\tau \geq 0$. Because the only difference between the informative and the uninformative case occurs at $\tau = 1, q+1, 2q+1, \ldots$, without loss of generality we only need to prove that $G_1 = G$. Since trading at $\tau = 0$ is informative for the dealer, her forecast after trading is $w_{0_+} = \mathsf{E}(v_0|\mathcal{I}_0, Q_0^b, Q_0^s) = w_0 + \mathsf{E}(e_0|Q_0^b, Q_0^s)$, where $e_0 = v_0 - w_0$ and

$$\mathsf{E}(e_0|Q_0^b, Q_0^s) \;=\; \frac{kG}{k^2G + \Sigma_N}\big(Q_0^b - Q_0^s\big), \qquad \mathsf{Var}(e_0|Q_0^b, Q_0^s) \;=\; \frac{G\Sigma_N}{k^2G + \Sigma_N}. \tag{A16}$$

We apply the recursive formula (A8) for $\tau = 1$, by replacing (i) V_η with $\tilde{V}_\eta$, (ii) V_ψ with $\tilde{V}_\psi$, and (iii) G_0 with $G_{0_+} = \frac{G\Sigma_N}{k^2G+\Sigma_N}$. Then, a direct computation shows that $G_1 = G$. Since all G_τ are equal to G, it follows that $F(q) = fq$.

We can now determine when trading is uninformative for the dealer. From the above analysis, this translates into the update $w_{0_+} - w_0$ being much smaller than a generic increment $w_\tau - w_{\tau-1}$ (for τ not of the form $1, q+1, 2q+1, \ldots$). This translates into the condition that the variance Σ_N is sufficiently large:[39]

$$\Sigma_N \;\gg\; \frac{k^2}{f^2\Sigma_v}. \tag{A17}$$

Appendix B. Proofs of Results

Proof of Proposition 1. Fix the monitoring rate $q > 0$. Let $\mathcal{I}_t$ be the dealer's information set just before trading at t, and by E_t the expectation operator conditional on $\mathcal{I}_t$. Let $w_t = \mathsf{E}_t(v_t)$ be the dealer's forecast of the fundamental value, and G the variance of the forecast error. From (5), we have

$$G \;=\; \mathsf{Var}(v_t - w_t) \;=\; \frac{1}{fq}. \tag{B1}$$

[39] Using equations (A10) and (A16), the condition $\mathsf{Var}(w_{0_+} - w_0) \ll \mathsf{Var}(\Delta w_\tau)$ becomes $\frac{k^2G^2}{k^2G+\Sigma_N} \ll \frac{\Sigma_v}{q}$, which translates to $\frac{\Sigma_N}{k^2G^2} \gg \frac{q}{\Sigma_v}$, or since $G = \frac{1}{fq}$, to $\Sigma_N \gg \frac{k^2\Sigma_v}{qf^2\Sigma_v}$. But the monitoring rate q is a positive integer, hence the condition is equivalent to $\Sigma_N \gg \frac{k^2}{f^2\Sigma_v}$.

We now compute the dealer's expected utility coming from a quoting strategy (a_t, b_t). If we define

$$h_t \;=\; \frac{a_t - b_t}{2}, \qquad \delta_t \;=\; w_t - \frac{a_t + b_t}{2}, \qquad e_t \;=\; v_t - w_t, \tag{B2}$$

then the quoting strategy is equivalent to choosing (h_t, δ_t). Equation (9) implies that traders' buy and sell demands at t are given, respectively, by $Q_t^b = \frac{k}{2}(v_t - a_t) + \ell - m + \varepsilon_t^b$ and $Q_t^s = \frac{k}{2}(b_t - v_t) + \ell + m + \varepsilon_t^s$, with $\varepsilon_t^b, \varepsilon_t^s \sim \mathcal{N}(0, \Sigma_N/2)$. Let $\varepsilon_t = -ke_t + \varepsilon_t^s - \varepsilon_t^b$. This is uncorrelated with the past information and has a normal distribution $\mathcal{N}(0, k^2G + \Sigma_N)$. If x_t is the dealer's inventory before trading at t, equation (7) shows that x_t describes the recursive equation $x_{t+1} = x_t - Q_t^b + Q_t^s$, which translates into

$$x_{t+1} \;=\; x_t - k\delta_t + 2m + \varepsilon_t \quad \text{with} \quad \varepsilon_t \overset{IID}{\sim} \mathcal{N}\big(0, k^2G + \Sigma_N\big). \tag{B3}$$

Substituting Q_t^b and Q_t^s in the dealer's objective function from (8), and ignoring the monitoring costs $C(q)$, we get $\mathsf{E}_\tau \sum_{t=\tau}^\infty \beta^{t-\tau}\, \mathsf{E}_t\Big(Dx_t - \frac{k}{2}(a_t - v_t)^2 - \frac{k}{2}(v_t - b_t)^2 + (\ell - m)(a_t - v_t) + (\ell + m)(v_t - b_t) - \gamma x_t^2\Big)$. We decompose $\mathsf{E}_t(v_t - b_t)^2 = \mathsf{E}_t(v_t - w_t + w_t - b_t)^2 = G + (w_t - b_t)^2$, and similarly $\mathsf{E}_t(a_t - v_t)^2 = G + (a_t - w_t)^2$. Using the notation in (B2), it follows that the dealer's maximization problem at τ is

$$\max_{(h_t, \delta_t)_{t \geq \tau}} \; \mathsf{E}_\tau \sum_{t=\tau}^\infty \beta^{t-\tau}\Big(Dx_t - kG - k\delta_t^2 - kh_t^2 + 2\ell h_t + 2m\delta_t - \gamma x_t^2\Big), \tag{B4}$$

where x_t evolves according to (B3). Using the Bellman principle of optimization, we reduce the dynamic optimization in (B4) to the following static optimization problem:

$$V(x_t) \;=\; \max_{h_t, \delta_t} \Big(2dx_t - kG - k\delta_t^2 - kh_t^2 + 2\ell h_t + 2m\delta_t - \gamma x_t^2 + \beta\, \mathsf{E}_t\, V(x_{t+1})\Big), \tag{B5}$$

where $d = \frac{D}{2}$. We guess that $V(x)$ is a quadratic function of the form

$$V(x) \;=\; W_0 - 2W_1 x - W_2 x^2 \tag{B6}$$

for some constants W_0, W_1, W_2. Substituting x_{t+1} from (B3), the problem becomes

$$\begin{aligned} V(x_t) \;=\; \max_{h_t, \delta_t} \Big(&2dx_t - kG - k\delta_t^2 - kh_t^2 + 2\ell h_t + 2m\delta_t - \gamma x_t^2 \\ &+ \beta W_0 - 2\beta W_1(x_t - k\delta_t + 2m) - \beta W_2(x_t - k\delta_t + 2m)^2 - \beta W_2(k^2G + \Sigma_N)\Big). \end{aligned} \tag{B7}$$

The first order condition in (B7) with respect to h_t implies $h_t = \frac{\ell}{k}$, which shows that the optimal $h_t = h$, the constant defined in (10). The first order condition in (B7) with respect to δ_t implies $\delta_t = \frac{\beta W_2}{1 + k\beta W_2}\, x_t + \frac{m + k\beta W_1 + 2km\beta W_2}{k(1 + k\beta W_2)}$, which shows that the optimal $\delta_t = \lambda x_t + \Delta$, where

$$\lambda \;=\; \frac{\alpha}{1 + k\alpha}, \qquad \Delta \;=\; \frac{m + k\alpha_1 + 2km\alpha}{k(1 + k\alpha)}, \qquad \alpha_1 \;=\; \beta W_1, \qquad \alpha \;=\; \beta W_2. \tag{B8}$$

Because $V(x_t) = W_0 - 2W_1 x_t - W_2 x_t^2$, we solve for W_0, W_1, W_2:

$$\begin{aligned} W_0 &= \frac{1}{1-\beta}\Big(\frac{\ell^2}{k} - k(1+k\alpha)G - \alpha\Sigma_N + \frac{(1+k\alpha)\big((1-\beta)m+\beta kd\big)^2}{k(1-\beta+k\alpha)^2}\Big), \\ W_1 &= \frac{\alpha}{1-\beta+k\alpha}\, m - \frac{1+k\alpha}{1-\beta+k\alpha}\, d, \qquad W_2 = \frac{\beta W_2}{1+k\beta W_2} + \gamma. \end{aligned} \tag{B9}$$

For a maximum, we need to have $W_2 > 0$. The quadratic equation for W_2 in (B9) has a unique positive solution,

$$W_2 = \frac{\gamma - \omega + \sqrt{(\gamma-\omega)^2 + 4\frac{\gamma}{\beta k}}}{2}, \quad \text{with} \quad \omega = \frac{1-\beta}{\beta k}. \tag{B10}$$

This implies that $\alpha = \beta W_2$ indeed satisfies (10).

If the dealer has an inventory of $x_t = x$, from equation (B6) it follows that the maximum expected utility she can achieve at t is $V(x) = W_0 - 2W_1 x - W_2 x^2 = \frac{1}{1-\beta}\big(\frac{\ell^2}{k} - \alpha\Sigma_N - k(1+k\alpha)G + \frac{(1+k\alpha)((1-\beta)m+\beta kd)^2}{k(1-\beta+k\alpha)^2}\big) - 2W_1 x - W_2 x^2$. Since $G = \frac{1}{fq}$, we get

$$U(q) = \frac{1}{1-\beta}\left(\widetilde{W}_0 - \frac{k(1+k\alpha)}{fq}\right) - 2W_1 x - W_2 x^2, \tag{B11}$$

where $\widetilde{W}_0$, W_1 and W_2 do not depend on q. Also, using $\alpha_1 = \beta W_1$, we compute $\Delta = \frac{1-\beta+2k\alpha}{k(1-\beta+k\alpha)}\, m - \frac{\beta}{1-\beta+k\alpha}\, d$. Since $d = \frac{D}{2}$, this proves that the formula for Δ in (10). □

Proof of Proposition 2. Consider a more general function $F(q) = 1/\mathsf{Var}(v_t - w_t)$ that is increasing in the monitoring rate q. If $G(q) = 1/F(q)$, we have showed in the proof of Proposition 1 that the dealer's maximum expected utility is of the form $V(x_t) = W_0 - 2W_1 x_t - W_2 x_t^2$, where W_0, W_1 and W_2 are as in (B9). This formula, however, does not include the monitoring costs per unit of time, $C(q)$. If we include these costs, the dealer's maximum utility is $W_0 - 2W_1 x_t - W_2 x_t^2 - \frac{C(q)}{1-\beta}$. But up to a constant that does not depend on q, this utility is equal to $\frac{-k(1+k\alpha)G(q) - C(q)}{1-\beta}$. The first order condition with respect to q is equivalent to $-k(1+k\alpha)G'(q) - C'(q) = 0$. Thus, the optimal monitoring rate satisfies

$$-\frac{C'(q)}{G'(q)} = \frac{C'(q)F^2(q)}{F'(q)} = k(k\alpha+1). \tag{B12}$$

The second order condition for a maximum is $k(k\alpha+1)G''(q) + C''(q) > 0$, which is satisfied if the functions G and C are convex, with at least one of them strictly convex.

We now use the linear specification $C(q) = cq$ and $F(q) = fq$, and compute the optimal monitoring rate q. Since $G(q) = \frac{1}{fq}$, from (B12) it follows that q satisfies $fcq^2 = k(k\alpha+1)$, which proves the first part of equation (14). Because the function G is strictly convex, note that the second order condition is satisfied.

The second part of (14) follows by using the expression for α in (10). □

Proof of Corollary 1. We first prove that α is decreasing in k and increasing in γ.

Equation (B9) implies that $\alpha = \beta W_2$ satisfies the equation $\frac{\alpha}{\beta} - \gamma = \frac{\alpha}{1+k\alpha}$. Differentiating this equation with respect to k, we get $\frac{\partial \alpha}{\partial k} = -\frac{\beta \alpha^2}{(1+k\alpha)^2 - \beta} < 0$. Similarly, differentiation with respect to γ implies $\frac{\partial \alpha}{\partial \gamma} = \frac{\beta(1+k\alpha)^2}{(1+k\alpha)^2-\beta} > 0$.

Equation (14) implies that q and the term $Q = k(1 + k\alpha)$ have the same dependence on the parameters k and γ. Using the formula above for $\frac{\partial \alpha}{\partial k}$, we compute $\frac{\partial Q}{\partial k} = \frac{(1+k\alpha)^2(1-\beta+2k\alpha)}{(1+k\alpha)^2-\beta} > 0$. Finally, Q is increasing in α, which (as proved above) is increasing in γ, hence Q is also increasing in γ.

By visual inspection of equation (14), it is clear that the quote-to-trade ratio q is decreasing in f and increasing in σ_v. □

Proof of Corollary 2. Using equation (B3) and the fact that in equilibrium $\delta_t = \lambda x_t + \Delta$, it follows that the dealer's inventory evolves according to $x_{t+1} = (1 - k\lambda)x_t - k\Delta + 2m + \varepsilon_t$, with $\varepsilon_t \sim \mathcal{N}(0, k^2 G + \Sigma_N)$ and $G = \frac{1}{F} = \frac{1}{fq}$. From (10), the coefficient $\phi = 1 - k\lambda = \frac{1}{1+k\alpha} \in (0, 1)$, hence $x_{t+1} = \frac{1}{1+k\alpha}\, x_t - k\Delta + 2m + \varepsilon_t$. Thus, x_t follows an $AR(1)$ process with auto-regressive coefficient ϕ, mean $\bar{x} = (2m - k\Delta)/(1 - \phi)$, and variance $\Sigma_x = \big(\frac{k^2}{fq} + \Sigma_N\big)/(1 - \phi^2)$. Using the formula for Δ in (10), it is straightforwad to prove the formula for $\bar{x}$ in (15). One can also show that $\Sigma_x = \frac{k\alpha(2+k\alpha)}{(1+k\alpha)^2}\big(\frac{k^2}{fq} + \Sigma_N\big)$. □

Proof of Corollary 3. This has already been proved in the discussion that precedes the statement of the Corollary. Alternatively, Proposition 1 implies that the pricing discount at t is equal to $w_t - p_t = \lambda x_t + \Delta$, whose average equals $\lambda \bar{x} + \Delta$. Using (10), we compute the average discount to be $2m/k$, which is the same as $\bar{\delta}$. □

Proof of Corollary 4. First, we prove rigorously equation 20. Since the system is initially in the neutral state ($x_0 = \bar{x}$), according to (17) the expected price change $\mathsf{E}_0\, p_1 - p_0$ is zero. But, if $\overline{w}$ is the initial forecast, by definition $\overline{w} - p_0$ is the pricing discount. Since in the neutral state the pricing discount is $\bar{\delta} = 2m/k$, it follows that $p_0 = \overline{w} - \bar{\delta}$, which proves (20). Suppose now we hold all parameters constant except for k. Clearly, the cost of capital is decreasing in k, as the pricing discount $\bar{\delta}$ is decreasing in k. At the same time, Corollary 1 implies that the QT ratio q is increasing in k. This proves the inverse relation between r and q. □

Appendix C. Micro-Foundations of Order Flow

In this section we provide assumptions under which the traders' liquidity demand is approximately of the form described in (9). The proofs are in Appendix C.3.

C.1. Environment

We assume that are two types of traders: noise traders and investors. Noise traders are either buyers or sellers. At each trading date t, noise buyers submit an aggregate buy order for N_t^b shares, and noise sellers submit an aggregate buy order for N_t^s shares. Both N_t^b and N_t^s have IID normal distribution $\mathcal{N}(\ell_N, \Sigma_N/2)$, therefore by subtracting the mean we decompose them as follows:

$$N_t^b \;=\; \ell_N + \varepsilon_t^b, \quad N_t^s \;=\; \ell_N + \varepsilon_t^s, \quad \text{with} \quad \varepsilon_t^b, \varepsilon_t^s \;\sim\; \mathcal{N}(0, \Sigma_N/2). \tag{C1}$$

Investors have CARA utility with coefficient A. A mass one of investors is born in each period t, and starts with an initial endowment in the risky asset that has a normal distribution $\mathcal{N}(M, \sigma_M^2)$, where $M > 0$ is the risky asset supply. Investors born at t do the following: (i) observe the fundamental value v_t before trading (ii) trade at the quotes set by the dealer (a_t and b_t) in period t on the exchange, (iii) collect the dividend before $t+1$, and (iv) liquidate the asset at $t+1$ for a liquidation value equal to $v_t + u$, where u has a normal distribution $\mathcal{N}(0, \sigma_u^2)$.[40]

C.2. Equilibrium

To simplify notation, rather than treating the dividend D separately, we include it in the liquidation value. We thus define $V_t = v_t + D$ to be the expected part of the total liquidation value of an investor born at t. Before we analyze the equilibrium, we describe the behavior of a CARA investor in the presence of ask and bid quotes. Define the *lower target* $\underline{X}_t$ and the *higher target* $\overline{X}_t$ by:

$$\underline{X}_t \;=\; \frac{V_t - a_t}{A\sigma_u^2}, \qquad \overline{X}_t \;=\; \frac{V_t - b_t}{A\sigma_u^2}. \tag{C2}$$

The next standard lemma shows that a CARA investor born at t trades only when his initial endowment in the risky asset is outside of the target interval $[\underline{X}_t, \overline{X}_t]$. In that case, he trades exactly so that his final inventory is equal to the closest target.

Lemma C.1. *Consider a risky asset with liquidation value $V + u$, with $u \sim \mathcal{N}(0, \sigma_u^2)$, and a CARA investor with coefficient A who observes the value V and has endowment x_0 in the risky asset. The investor can buy any positive quantity at the ask quote a, or sell any positive quantity at the price b, where $a > b$. Suppose the risk-free rate is zero. Let $\underline{X} = \frac{V-a}{A\sigma_u^2}$ and $\overline{X} = \frac{V-b}{A\sigma_u^2}$. Then, the investor's optimal trade makes his final inventory equal to either (i) $\underline{X}$, if $x_0 < \underline{X}$, (ii) x_0, if $x_0 \in [\underline{X}, \overline{X}]$, or (iii) $\overline{X}$, if $x_0 > \overline{X}$.*

Define the following numeric constants:

$$\rho_0 \;=\; \frac{1}{\sqrt{8\pi}} \approx 0.1995, \qquad \rho_1 = \frac{1}{2\pi} + \frac{1}{4} \approx 0.4092. \tag{C3}$$

By aggregating the orders of all traders, we obtain the main result of this section.

Proposition C.1. *The investors born at t submit aggregate orders Q_t^b and Q_t^s of the form*

$$\begin{aligned} Q_t^b \;&\approx\; \frac{k_0}{2}(v_t - a_t) + \ell_0 - m_0 + \varepsilon_t^b, \qquad Q_t^s \;\approx\; \frac{k_0}{2}(b_t - v_t) + \ell_0 + m_0 + \varepsilon_t^s, \\ \text{with} \quad k_0 \;&=\; \frac{2\rho_1}{A\sigma_u^2}, \qquad \ell_0 \;=\; \ell_N + \rho_0\sigma_M, \qquad m_0 \;=\; \frac{\rho_1}{A\sigma_u^2}D + \rho_1 M, \end{aligned} \tag{C4}$$

[40] A particular case occurs if investors' liquidation value is v_{t+1}. In that case, $u = v_{t+1} - v_t$, which has a normal distribution with standard deviation $\sigma_u = \sigma_v$. But in the paper we do not make this restriction, and instead we regard σ_u as an independent parameter.

and the error terms ε_t^b and ε_t^s are IID with normal distribution $\mathcal{N}(0, \Sigma_N/2)$. Both approximations in (C4) represent equality up to terms of the order of $1/\sigma_M$.

Proposition C.1 provides micro-foundations for the equations (9). For instance, the imbalance parameter m_0 arises from the fact that investors are risk averse and therefore are more likely to be sellers than buyers when the asset is in positive net supply ($M > 0$). The investor risk aversion A is therefore a key determinant of the order flow characteristics. The next result shows how risk aversion affects the order flow parameters (k_0 and m_0), as well as the average pricing discount which in equilibrium is the ratio $\bar{\delta}_0 = 2m_0/k_0$.

Corollary 5. *The investor elasticity k_0 and the imbalance parameter m_0 are decreasing in risk aversion A, while the average pricing discount $\bar{\delta}_0 = 2m_0/k_0$ is increasing in A.*

The intuition for this result is straightforward. If investors are more risk averse (A is larger), they trade less aggressively and therefore their demands are less sensitive in the mispricing (k is smaller). For the same reason as above, there is a smaller imbalance between the sell and buy demands (m is smaller). Among the two, the effect of risk aversion on investor elasticity dominates, and therefore when investors are more risk averse, the average pricing discount is larger ($\bar{\delta}$ is larger). Since the cost of capital is in one-to-one correspondence with the pricing discount (see Section 4.5), Corollary 5 implies that the cost of capital is also increasing in investors' risk aversion.

C.3. Proofs of Results

Proof of Lemma C.1. This is a standard result in asset pricing, and therefore we only provide the intuition. First, suppose there is only one trading price p (the buy and sell prices are equal). Then, an investor with constant absolute risk aversion has an optimal target inventory of the form $X = \frac{V-p}{A\sigma_u^2}$. Therefore, regardless of his initial endowment x_0, the investor submits a market order such that his final inventory equals X. When the buy and sell prices are different, there are two targets corresponding to each price: $\underline{X} < \overline{X}$. A key fact is that the investor optimally must either buy at the ask, or sell at the bid, but not both.[41] In the first case, when the investor only buys, he behaves like a CARA agent that faces the ask quote a, hence optimally trades up to the lower target $\underline{X}$. For this trade to be a buy, however, his initial endowment x_0 must be below $\underline{X}$. Similarly, when x_0 is above the higher target $\overline{X}$, he sells down to $\overline{X}$. Finally, when x_0 is in between the two targets, there is no incentive to trade and the CARA agent's target inventory in this case remains equal to x_0. □

Proof of Proposition C.1. We first introduce some notation. Define

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{x^2}{2}\right), \quad \Phi(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t, \\ \psi(x) &= \Phi(-x)\Big(\phi(x) - x\Phi(-x)\Big), \end{aligned} \tag{C5}$$

[41] Because of the positive bid-ask spread, any quantity simultaneously bought and sold represents a deadweight loss.

where $\phi(x)$ is the standard normal density, and $\Phi(x)$ is the standard cumulative density. One can check that the function $\psi(x)$ defined in (C5) is positive and decreasing.

By assumption, there is a mass one of investors whose endowments are independent and distributed according to the normal distribution $\mathcal{N}(M, \sigma_M^2)$, with density $g(x) = \phi\big(\frac{x-M}{\sigma_M}\big)/\sigma_M$. Then, investors' endowments integrate to $\int_{-\infty}^{\infty} xg(x)\mathrm{d}x = M$, which, since the dealer has zero endowment, is indeed equal to the net supply of the risky asset.

To compute investor i's optimal demand at t, note that by assumption his liquidation value is $V_t + u_i$, where $V_t = v_t + D$ is known by the investor, and u_i is unknown and distributed according to $\mathcal{N}(0, \sigma_u^2)$. Thus, investor i computes $\mathsf{E}(V_t + u_i) = V_t$ and $\mathsf{Var}(V_t + u_i) = \sigma_u^2$. Thus, the targets $\underline{X}_t = \frac{V_t - a_t}{A\sigma_u^2}$ and $\overline{X}_t = \frac{V_t - b_t}{A\sigma_u^2}$ are common to all investors.

According to Lemma C.1, the optimal demand of an investor depends on his initial endowment. By assumption, traders' endowments are IID with normal distribution $\mathcal{N}(M, \sigma_M^2)$ and corresponding probability density function $g(x) = \frac{1}{\sigma_M}\phi\big(\frac{x-M}{\sigma_M}\big)$. Therefore, investors' aggregate buy market order at t is equal to $I_t^b = \underline{P} \int_{-\infty}^{\underline{X}_t} (\underline{X}_t - x)g(x)\mathrm{d}x$, where $\underline{P} = \int_{-\infty}^{\underline{X}_t} g(x)\mathrm{d}x$ is the mass of investors with endowments below $\underline{X}_t$. Similarly, investors' aggregate sell market order at t is equal to $I_t^s = \overline{P} \int_{\overline{X}_t}^{\infty} (x - \overline{X}_t)g(x)\mathrm{d}x$, where $\overline{P} = \int_{\overline{X}_t}^{\infty} g(x)\mathrm{d}x$ is the mass of investors with endowments above $\overline{X}_t$. Finally, investors with endowments between $\underline{X}_t$ and $\overline{X}_t$ do not submit any order. With the definition of ψ in (C5), we compute

$$I_t^b \;=\; \psi\left(\frac{M - \underline{X}_t}{\sigma_M}\right), \qquad I_t^s \;=\; \psi\left(\frac{\overline{X}_t - M}{\sigma_M}\right). \tag{C6}$$

Consider the linear approximation of ψ near $x = 0$:

$$\psi(x) \;=\; \rho_0 - \rho_1 x + O(x^2), \quad \rho_0 \;=\; \psi(0) \;=\; \tfrac{1}{\sqrt{8\pi}}, \quad \rho_1 \;=\; -\psi'(0) \;=\; \tfrac{1}{2\pi} + \tfrac{1}{4}, \tag{C7}$$

where $O(x^2)$ represents the standard "big O" notation.[42] The investors' aggregate buy order is thus $I_t^b = \rho_0\sigma_M + \rho_1(\underline{X}_t - M) + O(1/\sigma_M) = \frac{\rho_1}{A\sigma_u^2}(V_t - a) + \rho_0\sigma_M - \rho_1 M + O(1/\sigma_M)$. Also, from (C1), the noise buyers' aggregate order at t is $N_t^b = \ell_N + \varepsilon_t^b$, with $\varepsilon_t^b \sim \mathcal{N}(0, \Sigma_N/2)$. By adding I_t^b and N_t^b, and using the fact that $V_t = v_t + D$, we obtain that the aggregate traders' buy order at t, $Q_t^b = I_t^b + N_t^b$, satisfies

$$Q_t^b \;=\; \frac{\rho_1}{A\sigma_u^2}(v_t - a) + \big(\ell_N + \rho_0\sigma_M\big) - \Big(\frac{\rho_1}{A\sigma_u^2}D + \rho_1 M\Big) + \varepsilon_b^t + O(1/\sigma_M). \tag{C8}$$

Let $k_0 = \frac{2\rho_1}{A\sigma_u^2}$, $\ell_0 = \ell_N + \rho_0\sigma_M$, $m_0 = \frac{\rho_1}{A\sigma_u^2}D + \rho_1 M$. Thus, we have $Q^b = \frac{k_0}{2}(v_t - a) + \ell_0 - m_0 + \varepsilon_b^t + O(1/\sigma_M)$ and similarly $Q^s = \frac{k_0}{2}(b - v_t) + \ell_0 + m_0 + \varepsilon_s^t + O(1/\sigma_M)$. This proves (C4). □

Proof of Corollary 5. From (C4) we get $k_0 = \frac{2\rho_1}{A\sigma_u^2}$, $m_0 = \frac{k_0 D}{2} + \rho_1 M$, which implies $\frac{2m_0}{k_0} = D + \frac{2\rho_1 M}{k_0}$. Simple inspection shows that k_0 and m_0 are decreasing in A, while $\frac{2m_0}{k_0}$ is increasing in A. □

[42] This means that there is a number $B > 0$ such that $\big|\psi(x) - (\rho_0 - \rho_1 x)\big| < Bx^2$.

Proof of Corollary 5. From (C4) we get $k_0 = \frac{2\rho_1}{A\sigma_u^2}$, $m_0 = \frac{k_0 D}{2} + \rho_1 M$, which implies $\frac{2m_0}{k_0} = D + \frac{2\rho_1 M}{k_0}$. Simple inspection shows that k_0 and m_0 are decreasing in A, while $\frac{2m_0}{k_0}$ is increasing in A. □

Appendix D. Additional Tables

Table D.1: Variable description

$N(quotes)_{i,t}$	Total number of quote updates in stock i over period t. (Source: TAQ)
$N(trades)_{i,t}$	Total number of trade executions in stock i over period t. (Source: TAQ)
$QT_{i,t} = \frac{N(quotes)_{i,t}}{N(trades)_{i,t}}$	Quote to trade ratio for stock i over period t. (Source: TAQ)
$R_{f,t}$	Risk free rate, one month Treasury bill rate. (Source: WRDS/Kenneth French Webpage)
$R_{m,t}$	Value weighted return on the market portfolio. (Source: WRDS/Kenneth French Webpage)
$R_{i,t}$, $R_{p,t}$	return on stock i or portfolio p. (Source: WRDS/CRSP)
$r_{m,t} = R_{m,t} - R_{f,t}$	Excess return on the market. (Source: WRDS/Kenneth French Webpage)
$r_{i,t} = R_{i,t} - R_{f,t}$	Excess return on stock i. (Source: WRDS/TAQ)
$r_{p,t} = R_{p,t} - R_{f,t}$	Excess return on portfolio p. (Source: WRDS/TAQ)
$r^a_{i,t}$	Risk-adjusted return on stock (or portfolio) i. (Source: WRDS/TAQ)
$r_{hml,t}$	Value factor constructed by Kenneth French. (Source: WRDS/Kenneth French Webpage)
$r_{smb,t}$	Size factor constructed by Kenneth French. (Source: WRDS/Kenneth French Webpage)
$r_{umd,t}$	Momentum factor (up-minus-down) constructed by Kenneth French. (Source: WRDS/Kenneth French Webpage)
$r_{liq,t}$	Liquidity factor constructed by Pástor and Stambaugh (2003). (Source: WRDS)
$r_{pin,t}$	Liquidity factor constructed by Easley et al. (2002). (Source: Soren Hvidkjaer Webpage)
$QSPREAD_{i,t}$	Quoted spread. Difference between best ask quote and best bid quote (measured in USD). (Source: TAQ)
$SPREAD_{i,t}$	Relative spread. The quoted spread divided by the mid-quote price (measured in %). (Source: TAQ)
$PRC_{i,t}$	Price in USD. (Source: WRDS/TAQ)
$USDVOL_{i,t}$	Trading volume in USD (measured in mill. USD). (Source: WRDS/TAQ)
$VOLUME_{i,t}$	Share volume (measured in mill.). (Source: WRDS/TAQ)
$ILR_{i,t}$	Amihud (2002) illiquidity ratio for stock i over period t calculated as $ILR_{i,t} = [\sum(USDvol_{i,t})/\lvert r_{i,t}\rvert] \cdot 10^6$. (Source: WRDS/TAQ)
$VOLAT_{i,t}$	Return volatility for stock i calculated as absolute return over period t. (Source: WRDS/TAQ)
$IDIOVOL_{i,t}$	Idiosyncratic volatility for stock i measured as the standard deviation of the residual from a three-factor Fama/French model on daily data as in Ang et al. (2009). (Source: WRDS/TAQ)
$MCAP_{i,t}$	Market Capitalization of a stock, calculated as the number of outstanding shares multiplied by price. (measured in mill. USD)
$BM_{i,t}$	Book-to-Market value for stock i calculated as the log of the book value of equity divided by the market value of equity measured for the previous fiscal year.
Analyst following	Log of one plus the number of analysts following the firm. (Source: IBES)
Institutional ownership	Holdings of institutions at the end of the year constructed from 13F files. (Source: WRDS)

Table D.2: Sample stock descriptives

The table presents the monthly time-series averages of the cross-sectional 25th percetiles, means, medians, 75th percentiles, and standard deviations of the variables for the sample stocks. The sample period is June 1994 through October 2012, and only NYSE/AMEX and NASDAQ listed stocks are included in the sample. Stocks with a price less than USD 5, above USD 1000, or with less than 100 trades in month t-1 are removed. Stocks that change listings exchange, CUSIP or ticker symbol are removed.

	p25	Mean	Median	p75	Std.dev
Number of sample stocks (whole sample=6278)	2854	3126	3048	3368	391
MCAP (in mill. USD)	70	2570	252	1047	13418
PRC (Price in USD)	8	22	17	29	24
USDVOL (in mill. USD)	2	393	19	140	2261
VOLUME (in 1000 shares)	237	12331	1307	6391	69839
N(quotes) (in 1000)	1	166	9	110	504
N(trades) (in 1000)	0	28	2	16	111
QT (quote to trade ratio)	0.80	25.03	3.13	9.88	162.91
RSPREAD (%)	0.26	2.19	1.19	2.93	3.03
SPREAD	0.04	0.28	0.18	0.39	0.48
ILR (%)	0.036	8.331	3.402	2.389	121.071
VOLA	0.006	0.027	0.012	0.029	0.066
BM (log)	0.32	0.74	0.56	0.89	1.03
r_m (value weighted excess market return)	-0.018	0.001	-0.001	0.021	0.035
r_i (indiv. stock mid-quote excess returns, delist adj.)	-0.014	0.003	0.002	0.018	0.032
r_{smb} (SMB factor return)	-0.017	0.005	0.002	0.024	0.039
r_{hml} (HML factor return)	0.198	0.456	0.435	0.692	0.297
r_{umd} (UMD factor return)	-0.058	0.014	0.003	0.071	0.153
r_{liq} (Pastor/Stambaugh liquidity factor return)	-0.129	0.129	0.094	0.334	0.489
Institutional Ownership	0.000	0.000	0.000	0.000	0.000
R1 (lagged 1 month return in month t-1)	0.000	0.000	0.000	0.000	0.000
R212 (cumulative returns month t-12 through t-2)	0.000	0.000	0.000	0.000	0.000

Table D.3: FMB regressions using t-2 information

The table reports the Fama and MacBeth (1973) coefficients from a regression of risk-adjusted returns using lag QT. The firm characteristics are measured in month $t-2$, except $R1$ and $R212$. The variables included are: relative bid/ask spread ($SPREAD$), Amihud illiquidity ratio (ILR), log market value of equity ($MCAP$), log book to market ratio (BM) calculated as the log of the book value of equity divided by the market value of equity measured for the previous fiscal year, previous month return ($R1$), and the cumulative return from month $t-2$ to $t-12$ ($R212$), idiosyncratic volatility ($IDIOVOL$) measured as the standard deviation of the residuals from a Fama and French (1992) three factor model regressed on daily raw returns within each month as in Ang et al. (2009), and log USD volume ($USDVOL$). All coefficients are multiplied by 100. The standard errors are corrected by using the Newey-West method with 12 lags. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively. Panel A presents the results for information delay and Panel B presents the results on liquidity.

	(1)	(2)	(3)	(4)	(5)
Const.	0.004***	0.012***	0.009***	0.027***	0.030***
$QT_{i,t-2}$	-0.200***	-0.240***	-0.248***	-0.145***	-0.148***
$SPREAD_{i,t-2}$	0.132***		0.072*		0.034
$ILR_{i,t-2}$		0.088***	0.057*		-0.047*
$MCAP_{i,t-2}$				-0.060	-0.073
$BM_{i,t-2}$				0.063	0.063
$R1_{i,t-2}$				-5.111***	-5.111***
$R212_{i,t-2}$				0.100	0.129
$IDIOVOL_{i,t-2}$				-9.254***	-11.167***
$USDVOL_{i,t-2}$				0.034	0.004
$PRC_{i,t-2}$				-0.473***	-0.439***
R^2	0.01	0.01	0.01	0.03	0.04
Time series (months)	216	216	216	216	216

Table D.4: Stock risk-adjusted returns and quote-to-trade ratio subsample

The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted returns for single stocks, given by $r^a_{i,t} = r_{i,t} - \sum_{j=1}^{J} \beta_{i,j,t-1} F_{j,t}$ for two subsamples, before and after the introduction of algorithmic trading in 2002. Pre-2002 refer to t he period from June 1994 to December 2002 and Post-2002 refers to the period from January 2003 to October 2013. The firm characteristics are measured in month $t-1$. The variables included are: relative bid/ask spread ($SPREAD$), Amihud illiquidity ratio (ILR), log market value of equity ($MCAP$), log book to market ratio (BM) calculated as the log of the book value of equity divided by the market value of equity measured for the previous fiscal year, previous month return ($R1$), and the cumulative return from month $t-2$ to $t-12$ ($R212$), idiosyncratic volatility ($IDIOVOL$) measured as the standard deviation of the residuals from a Fama and French (1992) three factor model regressed on daily raw returns within each month as in Ang et al. (2009), log USD volume ($USDVOL$), and log price (PRC). All coefficients are multiplied by 100. The standard errors are corrected by using the Newey-West method with 12 lags. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

	Pre-2002	Post-2002
Const.	0.030**	0.041**
$QT_{i,t-1}$	-0.156**	-0.088*
$SPREAD_{i,t-1}$	0.068**	0.007
$ILR_{i,t-1}$	0.031	-0.033
$MCAP_{i,t-1}$	-0.351***	-0.115
$BM_{i,t-1}$	0.302***	-0.123*
$R1_{i,t-1}$	-4.144***	-4.643***
$R212_{i,t-1}$	0.566	-0.369
$IDIOVOL_{i,t-1}$	-17.318***	-8.464**
$USDVOL_{i,t-1}$	0.390***	-0.031
$PRC_{i,t-1}$	-0.516***	-0.361***
R^2	0.04	0.04
Time series (months)	100	116