Quotes, Trades and the Cost of Capital*

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

July 24, 2017

Abstract

We study the quoting activity of market makers in relation with trading, liquidity, and expected returns. Empirically, we find larger quote-to-trade (QT) ratios in small, illiquid or neglected firms, yet large QT ratios are associated with low expected returns. The last result is driven by quotes, not by trades. We propose a model of quoting activity 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: Liquidity, price discovery, volatility, trading volume, monitoring, neglected stocks, risk aversion, inventory, high frequency trading.

*We thank Dion Bongaerts, Jean-Edouard Colliard, David Easley, Thierry Foucault, Amit Goyal, Johan Hombert, Dashan Huang, Maureen O'Hara, Rohit Rahi, and Daniel Schmidt for their suggestions. We are also grateful to finance seminar participants at the University of British Columbia, Pontifical University of Chile, University of Chile, University of Technology Sydney, HEC Paris, as well as conference participants at the 8th Erasmus Liquidity Conference, 11th Risk Management conference in Singapore, CEPR-Imperial-Plato Inaugural Conference in London, 2017 Frontiers of Finance conference, 2nd Sydney Market Microstructure conference, and the 2016 CEPR Gerzensee European Summer Symposium in Financial Markets, 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

Market participants in stock exchanges around the world are usually divided into two categories: market makers who provide liquidity via quotes (or limit orders), and market takers who demand liquidity via marketable orders and thus generate trades. Several natural questions arise: What is the role of market makers in the price discovery process? How do they set their quotes? What effect do market makers have on the liquidity of a stock and its expected return (cost of capital)?[1] Directly answering these questions is difficult, as explicit market maker data is not readily available. Nevertheless, we can still observe the market makers' activity indirectly via the quoting process, and analyze how this process is related to a stock's liquidity and cost of capital.

In many market structure models, such as Glosten and Milgrom (1985), the market makers set their quotes at the expected asset value given the information contained in trades. There are two consequences of this mechanism: first, there is no expected price appreciation in the model, and hence the expected return is zero. Second, suppose we define the *quote-to-trade ratio* (henceforth "QT ratio") as the number of quote updates divided by the number of trades.[2] Then, as market makers set their bid and ask quotes mechanically, in response to trades, the quote-to-trade ratio is always equal to two. Models such as Glosten and Milgrom (1985) are of course stylized, but if we believe they provide a reasonable description of how market makers behave, then in practice we should not expect to find any systematic patterns in the QT ratio, or any connection between the QT ratio and the cost of capital.

In this paper, we find that the QT ratio in fact exhibits clear patterns across stocks, and we summarize these patterns as a list of new empirical stylized facts. Our main stylized fact, called the *QT effect*, is that the QT ratio has an inverse relation with expected returns, even after controlling for variables known to affect asset returns. The

[1] A large literature in asset pricing relates the liquidity of securities to their expected return, see e.g., Amihud, Mendelson, and Pedersen (2005) and the references therein.

[2] Empirically we define the quote-to-trade ratio by using in the numerator only the updates at the best quotes (highest bid and lowest ask). If the numerator is instead the number of all quotes (limit orders), we obtain a variable closely related to the order-to-trade (or message-to-trade) ratio used by various regulators, academics and practitioners in analyzing high-frequency trading or algorithmic trading. We choose our definition both because it is closer to our theoretical measure, and because of data availability issues.

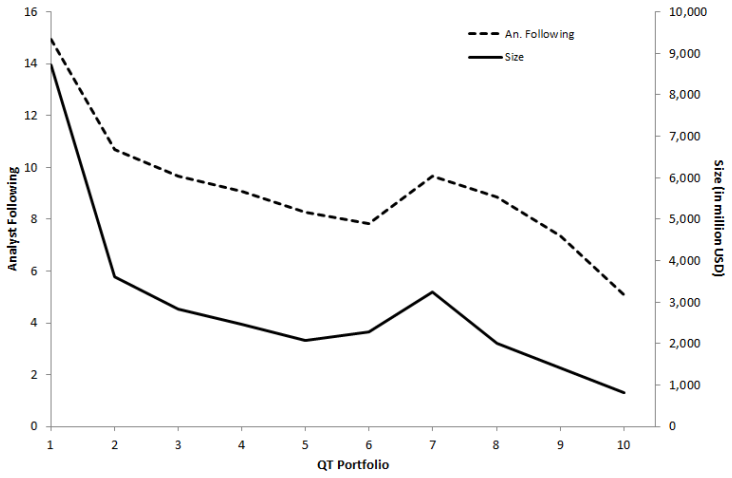
QT effect turns out to be driven by quotes and not by trades. These results suggest a new channel by which market structure affects expected returns: via the quoting activity of market makers.

To explore this channel, we propose a theoretical model of quoting activity, and we verify that it is consistent with our stylized empirical facts. In the model, the QT ratio is affected by many stock characteristics, but only one of them affects the cost of capital. This key characteristic, called *investor elasticity*, measures the extent to which informed investors in a stock respond to mispricing, and we show that the measure is inversely related to the investors' risk aversion. The investor elasticity is however not observable, and thus we are bound to rely on theory to interpret our empirical results.

To our knowledge, this paper is the first to directly analyze market maker quoting activity and its connections with liquidity and asset pricing. Our results suggest that market makers, far from passively reacting to trades, are in fact active producers of information and in doing so affect a stock's liquidity and its expected return.

Figure 1: Size and analyst coverage for 10 quote-to-trade ratio portfolios

The figure shows the average size (market capitalization) and number of analysts following a stock for ten portfolios sorted on the quote-to-trade ratio (QT). Portfolio 1 has the smallest QT ratio, and portfolio 10 has the largest QT ratio.



Our first stylized fact (SF1) connects the quote-to-trade ratio with certain stock characteristics. In particular, Figure 1 displays the average market capitalization and the average number of analysts following a stock for ten portfolios sorted by the QT

ratio. Firms that are small and illiquid (with few analysts following them) appear to have larger QT ratios than firms that are large and liquid. This result also holds for stocks with low institutional ownership, low volume, and low volatility. In general, we call a stock *neglected* if it has low market capitalization, low analyst coverage, low institutional ownership, low volume, or low volatility, and we show using a more rigorous regression analysis that neglected stocks have on average larger QT ratios.

The second stylized empirical fact (SF2) is that the QT ratio has increased significantly over time, especially after the emergence of algorithmic and high-frequency trading in 2003. This fact is documented by Hendershott, Jones, and Menkveld (2011) for their proxy of algorithmic trading, the message-to-trade ratio, but we show that the same pattern works for our quote-to-trade ratio measure.

Returning to SF1, we have seen that neglected stocks have larger QT ratios. As neglected firms tend to be small and illiquid, one may expect that a large QT ratio is associated with a large expected return. The third stylized fact (SF3) shows that the opposite is in fact true: large QT ratios are associated with small expected returns. This is our main empirical result, the QT effect. We verify that this effect holds both in the first part (1994–2002) and in the second part (2003–2012) of our sample.

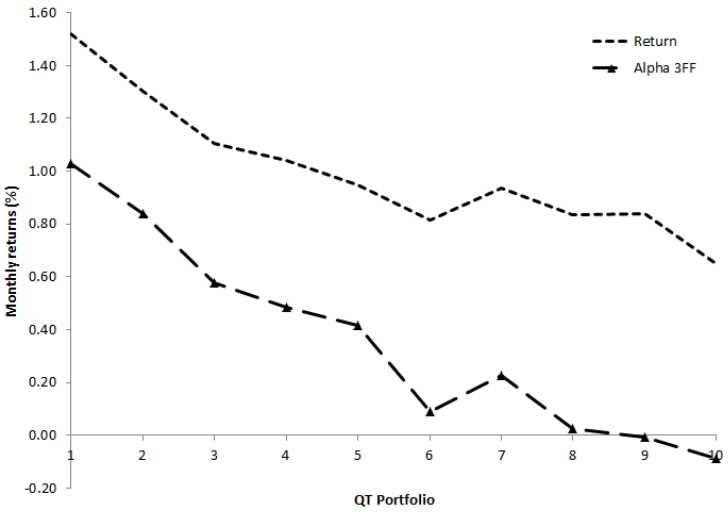
Figure 2 illustrates the QT effect: stocks with large QT ratios have small average returns, whether computed in excess of the risk-free rate, or after risk adjusting with the factors of Fama and French (1993). Figure 2 illustrates also the fourth stylized empirical fact (SF4), which is the asymmetry of the QT effect: stocks with low QT ratios have positive and significant alpha with respect to standard factor models, while stocks with high QT ratios have alphas that are close to zero and insignificant. Finally, the fifth stylized fact (SF5) is that the QT effect is driven by quotes and not by trades.

To interpret these stylized facts, we consider a model of quoting activity, in which a representative market maker (called the "dealer" or "she") sets ask and bid quotes to profit from trading.[3] The dealer maximizes her expected profit subject to a quadratic penalty on her inventory, with a coefficient called *inventory aversion*. After trading, the asset liquidates at a random price called the fundamental value. Trading occurs at the

[3]In the Internet Appendix we present two main extensions of our baseline model: a multi-dealer model (see Internet Appendix Section 3), and a multi-trade version with a single dealer (see Internet Appendix Section 4). We find that our results are robust to these extensions.

Figure 2: Excess return and alpha for 10 quote-to-trade ratio portfolios

The figure plots the average return in excess to the 1-month T-bill rate ("Return") and the alpha with respect to the Fama-French three factor model ("Alpha 3FF") for ten portfolios sorted on the quote-to-trade ratio (QT). Portfolio 1 has the smallest QT ratio, while portfolio 10 has the largest QT ratio. The variables are computed monthly and presented in percentages.



first arrival of a Poisson process with frequency normalized to one. The dealer monitors the market according to a Poisson process: at each monitoring time she observes a signal about the asset's fundamental value. Monitoring is costly and the cost increases in monitoring frequency, which is chosen ex ante.

Given the dealer's quotes, traders submit buy and sell quantities which, except for a noise term, are linear in the dealer's pricing error (the fundamental value minus the mid-quote price). The corresponding coefficient is our key *investor elasticity* parameter. The specification is the same as in Ho and Stoll (1981) or Hendershott and Menkveld (2014), except that we introduce an additional *imbalance parameter* which measures the difference between buy and sell quantities when the dealer's pricing error is zero.

To justify a nonzero imbalance parameter, we provide micro-foundations for trader behavior.[4] 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 liquidity traders who submit inelastic quantities. In equilibrium, the order flow is clearly unbalanced: risk averse investors demand a positive return for

[4]Order flow imbalance is important in our model, since the cost of capital turns out to be proportional to the imbalance parameter.

holding the asset, such that the price that equates buy and sell quantities is below the fundamental value. Our micro-foundations show that investors' risk aversion also affects investor elasticity: low risk aversion causes investors to trade with large elasticity.

As 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 on 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. 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 monitoring precision: a small monitoring precision makes the dealer monitor the market frequently. This result explains our puzzling stylized fact (SF1) 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 inventory aversion is large, the dealer needs to keep quotes closer to the fundamental value, and hence must monitor the market more frequently. This result provides an additional 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.[5] We thus obtain the following surprising 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. Instead, as the quotes are public information, each market maker's monitoring exerts a positive externality on the others and thus leads to under-investment in monitoring in equilibrium.

[5]In Internet Appendix Section 3 we introduce an extension of the model to multiple dealers, and we show that a larger number of market makers is indeed associated with a smaller QT ratio.

Fourth, the QT ratio is decreasing in monitoring costs: a smaller monitoring cost increases the dealer's monitoring frequency. This finding is consistent with the stylized fact SF2, i.e., the recent dramatic increase in the QT ratio (see Figure 3). 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 depend on a state variable: the dealer's initial inventory. The dependence works as in Hendershott and Menkveld (2014): with a large initial inventory, the dealer needs to attract more buying than selling on average, and therefore sets lower quotes. In general, our results are true when the dealer's initial inventory is positive. We define the dealer's *pricing discount* (or simply *discount*) as the difference between the dealer's forecast of the fundamental value and her mid-quote price. As the discount is in one-to-one relation with the expected return, we define the *cost of capital* to be equal to the discount.

A key determinant of the equilibrium discount (or cost of capital) is the 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 into a decrease of the cost of capital. Putting these facts together, we obtain a theoretical version of the QT effect: an inverse relation between the QT ratio and the cost of capital (stylized fact SF3). Note that this relation is driven by properties of the order flow, and at a more fundamental level (if we use our micro-foundations) by the investors' risk aversion.

The theoretical QT effect implies that stocks with large QT ratio have a small pricing discount, or equivalently their price is close to the fundamental value. If we identify the fundamental value with the expected return according to a standard factor model, and the price with the actual expected return, then the discount is equal to alpha with respect to the factor model. This results aligns well with the stylized fact SF4, which is the asymmetry of the QT effect: stocks with low QT ratios have positive and significant

alphas with respect to standard factor models, while stocks with high QT ratios have close to zero and insignificant alphas.

Our paper contributes to a large literature on market microstructure and asset pricing (see Amihud and Mendelson, 1986; Brennan and Subrahmanyam, 1996; Chordia, Roll, and Subrahmanyam, 2000, 2002; Chordia, Subrahmanyam, and Anshuman, 2001; Amihud, 2002; Easley, Hvidkjaer, and O'Hara, 2002; Easley and O'Hara, 2004; Amihud et al., 2005; Duarte and Young, 2009, among many others). While the relation between quoting activity and the cost of capital has not, to our knowledge, been investigated before, our empirical analysis follows the example of many papers, which find stock characteristics that matter for average returns.

The main message of our paper is that market makers produce public information (via quotes) in a way that affects the cost of capital. Another paper that analyzes the role of information in asset pricing is Easley and O'Hara (2004). One of their main findings is that more public information leads to a lower cost of capital.[6] In their rational expectations equilibrium model, however, there are no quotes and thus our stylized facts cannot be accommodated in their paper.

Our paper has also implications for the burgeoning literature on high-frequency trading (HFT).[7] The quote-to-trade ratio is often connected to HFT by regulators, practitioners and academics.[8] The recent dramatic increase in the QT ratio apparent in Figure 3 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. As the QT ratio is frequently used as a proxy for HFT, one may be tempted to attribute

[6]Easley and O'Hara (2004) show that the cost of capital is decreasing in the fraction of the signals that are public (which in their notation is equal to $1 - \alpha$), and the total number of signals (public or private). The intuition is that in both cases the uninformed investors can better learn from prices and therefore view the stock as less risky and demand a lower cost of capital.

[7]See for example Menkveld (2016) and the references therein.

[8]In practice, the QT ratio is typically defined with the numerator including not just the updates at the best quotes, but all orders or messages (see the discussion in Footnote 2). 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 et al., 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).

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, is not consistent with our stylized fact SF1, which shows that a large QT ratio is associated in fact to *illiquid* stocks. Moreover, the argument 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. We thus find the HFT explanation of the QT effect unlikely.

The paper is organized as follows. Section 2 describes the data and analyzes the quote-to-trade ratio in connection to various stock characteristics. Section 3 studies the relation between the quote-to-trade ratio and stock returns. Section 4 provides a theoretical model of the quote-to-trade ratio and compares the equilibrium results with the previous empirical stylized facts. Section 5 concludes. All proofs are in the Appendix or the Internet Appendix. The Internet Appendix provides several extensions of the baseline model in Section 4.

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.[9] Using TAQ data allows us to construct a long time series of the variable QT at the stock level, such that we can 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 (Chordia, Roll, and Subrahmanyam, 2000; Hasbrouck, 2009; Goyenko,

[9]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.

Holden, and Trzcinka, 2009). 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.[10] 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.

All returns are calculated using bid-ask midpoint prices, to reduce market microstructure noise effects on observed returns (Asparouhova, Bessembinder, and Kalcheva, 2010, 2013).[11] All returns are adjusted for splits and cash distributions. We follow Shumway (1997) in using returns of -30% for the delisting month (delisting codes 500 and 520–584).[12] Risk factors are from WRDS and Kenneth French's website for the period 1926 to 2013. The PIN factor is from Sören Hvidkjaer's website and is available from 1984 to 2002. Table IA.1 in the Appendix reports the definitions and the construction details for all variables and Table IA.2 in the Appendix provides the summary statistics.

Consistent with the literature (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.[13] Specifically, we calculate the QT variable for 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.

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

[11] Calculating returns from end of day prices does not change the results qualitatively. These results are available from the authors upon demand.

[12] 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.

[13] The results are qualitatively similar if QT is defined by using in the numerator both quote and depth updates. Using only quotes, however, is more consistent with our theoretical model in Section 4.

2.2 Stock Characteristics and the Quote-to-Trade Ratio

In this section, we analyze the relation of the QT ratio with various stock characteristics. To alleviate concerns about the effect of market-wide events during our sample period, we use time fixed effects in our regressions. We also use stock fixed effects to control for unobservable time-invariant stock characteristics.

To get some perspective about the firms with different QT ratios, we report in Table 1 average values of various firm-level characteristics. Specifically, each month we divide all stocks into decile portfolios based on their QT during at month t. 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 and report the time-series mean of the average cross-sectional characteristic.[14]

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 \$8.7 billion, while the highest QT stocks (stocks in QT decile 10) have an average capitalization of \$0.8 billion. Column (7) shows that the average monthly trading volume decreases from \$1.7 billion for the lowest QT stocks to \$0.05 billion for the highest 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 three times 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 ratio. We present the results from a panel regression with various specifications for fixed effects and with standard errors clustered at the stock and month level. Column (1) presents the results without any fixed effects. To control for unobservable time-invariant stock characteristics, we introduce stock fixed effect in column (2). To alleviate concerns about the effect of market-wide events during our sample period, we use time fixed effects in column (3). Finally, the regression presented in column (4) includes both firm and time fixed effects,

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

as both play an important role in our analysis. We find that QT is higher for stocks that have low analyst coverage, low institutional ownership, low market capitalization, low trading volume, and low volatility.[15] 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).

Stylized fact 1 (SF1): Neglected stocks (with low market capitalization, analyst coverage, institutional ownership, 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, as market makers have less precise information based on which to change their quotes. But in our theoretical model a market maker with less precise information actually monitors more often to prevent getting a large inventory, and therefore generates a higher QT ratio (see Section 4.3).

It is common practice among academics, practitioners and regulators to associate QT with HFT activity (several examples are given in Footnote 8). 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 the 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.

[15] QT also has an inverse relation with the stock price, which we interpret as evidence of the importance of the discrete tick size. Indeed, holding other variables constant (including the tick size), a stock with large price has a smaller *relative* tick size, and hence it is likely to exhibit fewer opportunities for market makers to update their quotes. But because our model does not speak to a discrete tick size, we leave an analysis along these lines to future research.

2.3 Time Series of Quote-to-Trade Ratios

Figure 3 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 monthly number of quotes at the best price 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.[16] 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.[17] Angel et al. (2011) argue that the proliferation of algorithmic and high-frequency trading since 2003 has lead to substantial increases in both the number of quotes and trades.

[16]Table IA.3 shows 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. This is essentially the same as the first stage in the IV regression of Hendershott et al. (2011).

[17]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 (including electronic order submissions, cancellations and trade reports, but excluding specialist quoting or floor orders).

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 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 investigate the raw return differential between the low and high QT stocks. Every time period, we sort stocks into decile portfolios based on their QT for each month t. We then compute the risk-free adjusted return for each of these portfolios for month $t+1$. Column (1) of Table 3 reports the excess returns for the ten portfolios. The QT1 portfolio has a return of 1.52% and QT10 has a return of 0.65% per month. The raw excess return of long-short portfolio based on QT is 0.87% a month.

This raw excess return differential might be driven by compensation for known risk factors. Therefore, we test whether the return differential between the low and high QT stocks can be explained by the market, size, value, momentum, liquidity, profitability and investment 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 $t+1$.[18] We estimate individual portfolio loadings from the 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}), the extra Fama and French (2015) factors: profitability (r_{RWM}) and investment (r_{CMA}), Pástor and Stambaugh (2003) liquidity (r_{liq}), momentum UMD (r_{umd}), and PIN (r_{PIN}). Table 3 reports

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

alphas obtained from regression.[19] 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), FF3+PS+MOM (with momentum), FF5 (with profitability and investment), and FF4+PS+PIN (probability of informed trading, PIN).[20]

Columns (2)-(8) in Table 3 report alphas for the ten QT-sorted portfolios. We first focus on the full sample analysis in columns (2)-(7). The low-QT portfolio (QT1) has a statistically significant monthly alpha (α_1) that ranges between 0.92% and 1.67% across various asset pricing models. The high-QT portfolio alphas range from 0.10% 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 models. However, the risk-adjusted return difference between the low-QT and high-QT portfolios is statistically significant and varies between 0.55% to 1.58% per month across the different asset pricing models. Table IA.4 shows that the differences between the low and high QT portfolio are not sensitive to the number of formed portfolios.

Stylized fact 3 (SF3): Higher quote-to-trade ratios are associated to lower stock returns in the cross-section (the QT effect).

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. In the literature, smaller or illiquid stocks also tend to have *higher* expected returns, which appears to contradict SF3. To address these issues, in the next section we perform a multivariate analysis and control for other variables that are potentially important in the cross-section of stock returns.

Table 3 also reveals an asymmetry in the QT effect. Thus, the profitability of the

[19] One can also estimate the individual portfolio loadings from rolling window regressions, to account for time-varying factor loadings. We construct time series averages of alphas obtained from 24-month rolling window regressions and obtain quantitatively similar results. These results are available from the authors upon demand.

[20] 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.

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.

Stylized fact 4 (SF4): Stocks with low quote-to-trade ratios have positive and significant alphas with respect to various asset pricing models, while stocks with high quote-to-trade ratios have insignificant alphas.

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. In addition, the Fama-MacBeth procedure accounts for time fixed effects that could arise from market-wide events during our sample period.

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.[21] 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 three-factor Fama-French model (market, size, and value), with the addition of the momentum and of the 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 thus confirm again the QT effect (the stylized fact SF3 in the previous section).

As 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 for 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 for QT remains negative and highly significant, while the illiquidity proxies SPREAD and ILR become both insignificant. The QT effect therefore is distinct from the known effects of other variables: spread, ILR, trading volume, volatility. The coefficients of control variables are quantitatively similar to papers using a similar sample period, e.g. Hou and Loh (2016) We thus add to the literature that explores how trading activity and market structure are

[21] Panel A of Table IA.5 in the Appendix shows the estimation results where $k = 2$ (excluding the past return variables $R1$ and $R212$).

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).

An important question is whether the QT effect is driven by the number of quotes or by the number of trades. Table 5 explores this question. 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 significance 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 of QT derives from quotes and not from trades.

Stylized fact 5 (SF5): Higher quote-to-trade ratios predict lower stock returns in the cross-section, and the predictability is driven by the number of quotes rather than the number of trades.

This result justifies our later choice to model the trades as exogenous (at a rate normalized to one) and focus instead on the quotes and how they result from the market makers' monitoring decisions.

3.3 Robustness

In this section, we investigate 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 4 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.[22] 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 HFT (see Footnote 8), 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

[22]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.

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 IA.6 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.

The stylized fact SF3′ essentially states that our main result, SF3, is robust under different specifications. In the next section we propose a theoretical model that is consistent with the stylized facts in Section 3 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 Ho and Stoll (1981) and Hendershott and Menkveld (2014). To simplify the presentation, in this section we consider a model with a single trading round. In the Internet Appendix we present an extension of the model to multiple trading rounds, and also an extension with multiple dealers, and we see that the main results of the paper remain robust.

4.1 Environment

The market consists of one risk-free asset and one risky asset. Trading in the risky asset takes place in a market exchange based on the mechanism described below. There are two types of market participants: (a) one monopolistic market maker called the *dealer* ("she") who monitors the market and sets ask and bid 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 $M > 0$. After trading, the risky asset liquidates at a fundamental value equal to v, which has a normal distribution $v \sim \mathcal{N}(v_0, \sigma_v^2)$, where σ_v is the *fundamental volatility*.

Trading. Trading occurs at at the first arrival τ in a Poisson process with frequency parameter normalized to one. Upon observing the ask quote a and the bid quote b, traders submit at τ the following aggregate market orders:

$$\begin{aligned} Q^b &= \frac{k}{2}(v-a) + \ell - m + \varepsilon^b, \quad \text{with} \quad \varepsilon^b \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_L/2\right), \\ Q^s &= \frac{k}{2}(b-v) + \ell + m + \varepsilon^s, \quad \text{with} \quad \varepsilon^s \overset{IID}{\sim} \mathcal{N}\left(0, \Sigma_L/2\right), \end{aligned} \tag{5}$$

where Q^b is the *buy demand* and Q^s is the *sell demand*. The numbers k, ℓ, m and Σ_L are exogenous constants. Together, Q^b and Q^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*.

In Appendix B, we provide micro-foundations for the order flow by showing that a population of (i) liquidity traders and (ii) risk averse investors with random initial inventories generates aggregate orders that approximately satisfy (5).[23] In Internet Appendix Section 2 we show that the equilibrium is qualitatively similar if instead of aggregating the order flow over the whole population, we consider only the optimal order from one individual trader selected at random from the population.

Dealer Monitoring. The dealer monitors the market according to an independent Poisson process with frequency parameter $q > 0$ called the *monitoring frequency* (or *monitoring rate*). Let t_n be the n-th arrival of this process, and let $t_0 = 0$. Monitoring consists in the dealer receiving a signal s_n at each monitoring time t_n for $n \geq 0$:

$$s_n = v + \varepsilon_n, \quad \text{with} \quad \varepsilon_n \overset{IID}{\sim} \mathcal{N}\left(0, \frac{1}{F(q)}\right). \tag{6}$$

In the rest of the paper we consider the initial signal s_0 at $t_0 = 0$ as the dealer's prior,

[23] Hendershott and Menkveld (2014) use a similar reduced form approach, except that they set $m = 0$. By providing micro-foundations for the order flow, we find that $m > 0$ when investors are risk averse and the asset is in positive net supply.

while monitoring refers to the subsequent signals s_n with $n > 0$. Note that we allow the signal precision F to depend on the monitoring rate. Intuitively, if $F(q)$ is increasing in q, monitoring has increasing returns to scale: monitoring more often produces more precise signals each time. The cost of monitoring at the rate q is $C(q)$, and is paid only once before monitoring begins at $t = 0$.

Dealer's Quotes and Objective. A quoting strategy for the dealer is a pair (a_t, b_t) of right-continuous functions in $t \geq 0$, where a_t is the ask quote at t and b_t is the bid quote at t. Let x_0 be the dealer's initial inventory in the risky asset and x_{end} the inventory after trading. If Q^b is the aggregate buy market order and Q^s is the aggregate sell market order, the dealer's inventory after trading is

$$x_{\text{end}} \;=\; x_0 - Q^b + Q^s. \tag{7}$$

Denote by τ the random trading time, which is exponentially distributed with parameter equal to one. Then, for a given quoting strategy (a_t, b_t) the dealer's expected utility is equal to the expected profit minus the quadratic penalty in the inventory and minus the monitoring costs:

$$\mathsf{E}_0\left(x_0\, v \;+\; \big((v - b_\tau)Q^s + (a_\tau - v)Q^b\big) \;-\; \gamma\, x_{\text{end}}^2 \;-\; C(q)\right), \tag{8}$$

where the parameter $\gamma > 0$ is the dealer's *inventory aversion*.[24]

Equilibrium Concept. As the dealer is a monopolist market maker in our model, the structure of the game is simple. First, the dealer chooses a constant monitoring rate q. Second, in the trading game the dealer chooses the quoting strategy (a_t, b_t) such that objective function (8) is maximized.

[24]This utility function is justified if the dealer either faces external funding constraints, or is risk averse. The latter explanation is present in Hendershott and Menkveld (2014, Section 3), where the dealer maximizes quadratic utility over non-storable consumption. To solve for the equilibrium, they consider an approximation of the resulting objective function (see their equation (16)). This approximation coincides with our dealer's expected utility in (8) when $C(q) = 0$.

4.2 Optimal Quotes

We solve for the equilibrium in two steps. In the first step (this section), 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. Consider the game described in Section 4.1, with positive parameters D, k, ℓ, m, Σ_L. Also, let x_0 be the dealer's initial inventory. Define the following constants:

$$h \;=\; \frac{\ell}{k}, \qquad \delta \;=\; \frac{m}{k}\,\frac{1+2\gamma k}{1+\gamma k} + \frac{\gamma}{1+\gamma k}\,x_0. \tag{9}$$

The next result describes the optimal quoting strategy of the dealer. The strategy is conditional on the dealer's value forecast w_t. In Section 4.3 we describe the process w_t, which is exogenous to the dealer once the monitoring decision is made.

Proposition 1. *Suppose the dealer has initial inventory x_0 and her forecast at t is w_t. Then the dealer's optimal quotes at t are*

$$a_t \;=\; w_t + h - \delta, \qquad b_t \;=\; w_t - h - \delta, \tag{10}$$

where h and δ are as in (9). The mid-quote price $p_t = (a_t + b_t)/2$ satisfies

$$p_t \;=\; w_t - \delta \;=\; w_t - \frac{m}{k}\,\frac{1+2\gamma k}{1+\gamma k} - \frac{\gamma}{1+\gamma k}\,x_0. \tag{11}$$

To get intuition for this result, suppose the imbalance parameter m is zero. Consider first the particular case when the dealer is risk-neutral: $\gamma = 0$. In that case, the dealer's inventory x_0 does not affect her strategy. Equation (10) implies that the dealer sets her quotes at equal distance around her forecast w_t. Hence, the ask quote 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 $h = \ell/k$ reflects 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 expected

quantity at the quotes.[25] 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 positive inventory aversion ($\gamma > 0$), her initial inventory affects the optimal quotes. Indeed, according to equation (10), the quotes at t are equally spaced around an inventory-adjusted forecast $(w_t - \frac{\gamma}{1+\gamma k} x_0)$. The effect of the dealer's inventory on the mid-quote price is the *price pressure* mechanism identified by Hendershott and Menkveld (2014). To understand this phenomenon, suppose that the initial inventory is large and positive. To avoid the inventory penalty, the dealer must reduce the inventory. This implies that the dealer must lower the quotes to attract more buyers than sellers.

According to (11), the mid-quote price is also decreasing in the imbalance parameter m. 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.

4.3 Optimal Monitoring and the QT Ratio

Suppose the dealer monitors the market at the rate q, which means that at t_n, the n-th arrival in a Poisson rate with frequency q, she receives a signal s_n with precision $F(q)$. The next standard result describes the evolution of the dealer's forecast w_t that arises from monitoring.

Lemma 1. *Let $n \geq 0$ and $t \in [t_n, t_{n+1})$. Then, the dealer's value forecast is the average current signal, $w_t = (s_0 + \cdots + s_n)/(n+1)$, and its precision is*

$$\frac{1}{\mathsf{Var}(v - w_t)} \;=\; (n+1)F(q). \tag{12}$$

Intuitively, the forecast changes only when there is a new signal, at the monitoring time t_n. The forecast is clearly the average signal. Since each signal has the same precision $f(q)$, the precision increases linearly with the number of monitoring times.

Proposition 1 implies that the dealer's equilibrium quotes change only when her

[25] For instance, equation (5) implies that the expected quantity traded at the ask is $\mathsf{E}_\tau(Q^b) = \frac{k}{2}(w_\tau - a_\tau) + \ell$, which is decreasing in a_τ.

forecast changes. Therefore, the monitoring rate is the same as the quote rate, and as the trading rate is normalized to one, the quote rate is the same as the *quote-to-trade ratio*. We thus define

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

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 the following expressions:

$$F(q) \;=\; f\ln(q+1), \qquad C(q) \;=\; cq, \tag{14}$$

where $f > 0$ and $c > 0$ are constant parameters.[26] We call the parameter f the *signal precision* and c the *monitoring cost*.

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

$$q^2 \;=\; \frac{k(1+k\gamma)}{fc}. \tag{15}$$

Using the formula in (15), we obtain the following straightforward result.

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 both adverse selection and large variation in inventory. To better understand the reasons behind this increase, we write equation (15) as a sum: $q^2 = \frac{k}{fc} + \frac{k^2\gamma}{fc}$. The first term (which does not depend on the dealer's inventory aversion γ) simply reflects that by increasing her monitoring rate, the dealer reduces the adverse selection that comes from trading with investors with superior information. The second term depends on the inventory aversion γ. If this parameter 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 is not expected to vary too much.

[26] The results are qualitatively the same if we take $F(q) = f$ or $F(q) = fq$, but the formulas are less explicit. In the proof of Proposition 2, we describe the equilibrium conditions for general F and C.

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).

We finish this section by showing that the equilibrium described above remains essentially the same if we replace the monitoring process by a unique signal with the appropriate precision.

Corollary 2. *Suppose instead of monitoring at the rate q and receiving signals with precision $F(q)$ the dealer receives a unique signal with precision*

$$\tilde{F}(q) \;=\; \frac{qF(q)}{\ln(q+1)}. \tag{16}$$

Then, in the new equilibrium the dealer chooses the same half spread h, pricing discount δ, and monitoring rate q.

From the previous section it is clear that the equilibrium half spread and pricing discount do not depend on the dealer's signal structure. Thus, the main statement of Corollary 2 is the equivalence of monitoring rates under the two different signal structures. In particular, if we choose the monitoring precision $F(q) = f\ln(q+1)$ as in (14), the equivalent signal precision becomes linear: $\tilde{F}(q) = fq$. In the Internet Appendix we use this equivalent formulation to simplify the presentation of the various extensions of our model.

4.4 Pricing Discount and the Cost of Capital

In this section we analyze the equilibrium cost of capital. We first define the *pricing discount*, or simply the *discount*, at t to be the difference between the dealer's forecast w_t and the mid-quote price p_t. According to Proposition 1, the equilibrium discount is always equal to the constant δ from equation (9). We compute the expected return at t (using the mid-quote price):

$$\frac{\mathsf{E}_t(v) - p_t}{p_t} \;=\; \frac{w_t - p_t}{p_t} \;=\; \frac{\delta}{w_t - \delta}, \tag{17}$$

and we see that the expected return is in one-to-one correspondence with the discount. We then define the *cost of capital* r to be equal to the discount:[27]

$$r \;=\; \delta \;=\; \frac{m}{k}\,\frac{1+2\gamma k}{1+\gamma k} + \frac{\gamma}{1+\gamma k}\,x_0. \tag{18}$$

Thus, the cost of capital depends on a state variable: the dealer's initial inventory x_0. In the rest of the paper, we assume that $x_0 \geq 0$. We obtain the following result.

Corollary 3. *If $x_0 \geq 0$, then the cost of capital is increasing in the imbalance parameter m and decreasing in the investor elasticity k.*

Intuitively, if the imbalance parameter m increases, the dealer expects the difference between the sell and buy demands to increase as well. To attract buyers, the dealer must lower the price and thus increase the discount. If the investor elasticity k increases, investors trade more aggressively when the price deviates from the fundamental value. To stop the inventory from accumulating too much in either direction, the dealer must raise the price closer to her forecast, which translates into a lower discount.

The next result connects the cost of capital to the equilibrium quote-to-trade ratio.

Corollary 4 (QT Effect)**.** *If $x_0 \geq 0$, then holding all parameters constant except for the investor elasticity k, there is an inverse relation between the discount (or cost of capital) and the QT ratio.*

[27] This is standard in one-period models, e.g., Easley and O'Hara (2004).

The key driver of the QT effect in our model is investor elasticity. When k is larger, Corollary 1 shows that the QT ratio q 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 discount δ 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 B we provide micro-foundations for the order flow, and we show that the investor elasticity k is larger when traders are less risk averse. Therefore the QT effect is driven, at a more fundamental level, by traders' risk aversion: 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 (see the stylized empirical fact SF3′ in Section 3.3).

The stylized empirical fact SF4 shows that the QT effect is asymmetric: stocks with low QT ratios have positive and significant alpha with respect to various factor models, while stocks with high QT ratios have alpha close to zero and insignificant. Now, recall that the QT ratio is increasing in the investor elasticity k. Also, if we identify the fundamental value with the expected return according to the factor model, and the price with the actual expected return, then the discount is just the alpha with respect to the factor model. We thus obtain the following theoretical translation of SF4: stocks with low k have positive discount, while stocks with high investor elasticity k have discount close to zero. But this result is the consequence of equation (18) for the equilibrium discount when k becomes very large.

4.5 Additional Predictions

In this section, we provide several additional predictions of our model. Some of these predictions involve the dealer's inventory aversion γ. Since this parameter is not directly observable, as empirical proxy we use instead the number of market makers that provide

liquidity in the asset: arguably, a larger number of intermediaries implies a smaller γ for the representative dealer.

Corollary 1 implies that the dealer's optimal monitoring rate q is increasing in her inventory aversion γ. But the QT ratio is an empirical proxy for the monitoring rate q, and the number of market makers is an empirical proxy for the (inverse) inventory aversion γ. We obtain the following empirical prediction.

Prediction 1: The number of market makers in a stock has an inverse relation to the stock's quote-to-trade ratio.

Based on the intuition of Corollary 1, a larger number of market makers can be interpreted as a smaller inventory aversion γ of the aggregate market maker. But a less averse dealer monitors less often the stock, as she is less concerned about accumulating inventory. Therefore, the resulting QT ratio is also smaller.

In Internet Appendix Section 3 we provide an extension of the model to N dealers, and we show that the inventory aversion of a representative dealer is $1/N$ of the individual inventory aversion. In that extension, we also prove directly that the QT ratio is smaller in the N-dealer case (see Corollary IA.3). This result provides additional intuition to Prediction 1: because the quotes are public information, each market maker's monitoring exerts a positive externality on the others and thus leads to under-investment in monitoring in equilibrium.

We test this prediction in column (1) of Table IA.3 in the Internet Appendix. This is essentially the same as column (4) of Table 2 in this paper, except that we add as explanatory variable the number of registered market makers in a particular stock (MM). 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 indeed has a negative effect on the QT ratio.

We now consider the effect of the dealer's inventory aversion on the cost of capital (or discount). Equation (18) shows that the equilibrium discount depends on the dealer's initial inventory x_0. We now describe the equilibrium corresponding to a particular

value:

$$x_{0,\text{neutral}} \;=\; \frac{m}{\gamma k}. \tag{19}$$

We call this value the dealer's *neutral* (or *preferred*) inventory.[28]

Corollary 5. *When the dealer's inventory has the neutral value, the expected buy and sell quantities from equation (5) are equal. The equilibrium cost of capital (discount) is*

$$\delta_{\text{neutral}} \;=\; \frac{2m}{k}. \tag{20}$$

The first statement of Corollary 5, that the traders' order flow is balanced in the neutral state, is in fact the reason behind our definition of neutral inventory in (19). The neutral inventory represents the dealer's bias in holding the risky asset, and mathematically it is positive because the imbalance parameter m is positive. Intuitively, the neutral inventory is positive because the investors are risk averse and the risky asset is in positive net supply (see the micro-foundations in Appendix B). But the dealer also behaves approximately as a risk averse investor because of the quadratic penalty in inventory (see Footnote 24). Therefore, our model becomes essentially a risk sharing problem, in which the dealer prefers to hold a positive inventory. Based of the formula (19) and Corollary 5, we obtain the following empirical prediction.

Prediction 2: Market makers have a preferred inventory in a stock that is positive and increasing in the stock's expected return and the total number of market makers active in that stock.

Formally, equations (19) and (20) imply that the neutral (or preferred) inventory $x_{0,\text{neutral}} = \frac{m}{\gamma k}$ is positive, and is equal to the product of the half discount $\delta_{\text{neutral}}/2 = m/k$ and the inverse inventory aversion $1/\gamma$. But the discount is a proxy for the expected return, and the inverse inventory aversion is a proxy for the number of market makers.

Intuitively, to see that the dealer's preferred inventory is increasing in the number of market makers, we show that it is decreasing in her aversion parameter γ: when the

[28] In the multi-trade extension in the Internet Appendix Section 4, we show that the neutral inventory is equal to the long-term average inventory *regardless* of the value of the initial inventory..

dealer is more inventory averse, she prefers to hold less of the risky asset. To see that the dealer's preferred inventory is increasing in the expected return (or cost of capital), this is equivalent to showing that the preferred inventory is increasing in the imbalance parameter m and decreasing in the investor elasticity k.[29] First, an increase in m should increase the dealer's preferred inventory: note that according to the micro-foundations in Appendix B, m is proportional to the supply parameter M; but when the risky supply is higher, there is more of it to share and the dealer's preferred inventory is also higher. Second, a decrease in k should also increase the dealer's preferred inventory: when investors are more aggressive (or less risk averse, according to the micro-foundations in Appendix B), they hold relatively more of the risky asset, which leaves a smaller preferred inventory to the dealer.

We do not test Prediction 2, because we do not have data on market maker inventories. There is, however, evidence that market makers tend to have positive inventories even overnight, and even when they are high-frequency market makers (see for example Comerton-Forde, Hendershott, Jones, Moulton, and Seasholes, 2010, Figure 2).

Returning to equation (20), we see that the discount (or cost of capital) in the neutral state no longer depends on the dealer's inventory aversion γ. But the number of market makers is an empirical proxy for the (inverse) inventory aversion γ. We obtain the following empirical prediction.

Prediction 3: The number of market makers in a stock has no relation to the stock's expected return.

This result is surprising, because one may expect the discount to be larger if the dealer has a larger inventory aversion γ. But in the neutral state this is not the case, because the neutral discount reflects the dealer's desire to balance the order flow, and therefore only the coefficients of the order flow may affect the discount, and not the dealer's characteristics, including the aversion parameter γ.[30]

[29] The discount (cost of capital) $\delta_{\text{neutral}} = 2m/k$ is higher either when m is higher or when k is lower (investors in the stock are more risk averse), because in both cases the dealer must set a higher discount to accommodate the investors.

[30] In the dynamic extension of the model in Internet Appendix Section 4, we see that the dealer's

We test this prediction in Table IA.7 in the Internet Appendix. This is essentially the same as column (5) of Table 4 in this paper, except that we add as explanatory variable the number of registered market makers in a particular stock (MM). We find that the coefficient on the number of market makers is insignificant.

Note that Prediction 3 depends crucially on the dealer's initial inventory being equal to the neutral value in (19). Suppose instead the dealer starts with zero inventory.[31] The equilibrium discount is then

$$\delta_{\text{zero}} = \frac{m}{k}\frac{1+2\gamma k}{1+\gamma k}, \tag{21}$$

and we can see that the discount is now increasing in the dealer's inventory aversion γ. This suggests that Prediction 3 is not robust unless we have strong reasons to believe that the dealer starts with the neutral inventory. (Prediction 2 may also help to test that statement.) Empirically, the fact that the number of market makers is insignificant in Table IA.7 could be simply due to the error in variables, as the number of market is undoubtedly a noisy proxy for the dealer's inventory aversion.

5 Conclusion

This paper studies the quoting activity of market makers, and how the resulting quote-to-trade (QT) ratio is related to liquidity, price discovery, and expected returns. Empirically, we find that the QT ratio is larger in neglected stocks, that is, in stocks with low market capitalization, analyst coverage, institutional ownership, trading volume, and volatility. Our main finding, called the QT effect, is that stocks with higher QT ratio 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. Further analysis shows that the QT effect is driven by quotes and not by trades, and is robust after controlling for other variables known to affect returns.

desire to balance the order flow (on average) arises as an equilibrium result, as an imbalanced order flow would result in a permanent expected accumulation of inventory, which cannot be optimal.

[31] In the context of the micro-foundations in Appendix B, the zero inventory choice corresponds to the particular case when the liquidity traders have a zero average initial endowment (see Footnote 32).

In the theoretical part of our paper, we propose a model of quoting activity 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 theoretical counterpart of the 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. Our model provides several additional empirical predictions, e.g., a larger number of market makers lowers the QT ratio, but has no effect on expected return. These predictions are borne out in the data.

Future research using more detailed data should clarify how market makers set their quotes and explore further the channel by which quoting activity is related to liquidity and expected returns. Nevertheless, our results already suggest that market makers are not simply reacting to trades as in some standard market structure models, but they actively produce information, and by doing so affect a stock's liquidity and cost of capital. If this information channel extends to how market makers set their quotes beyond the best bid and ask, then we would expect our results to extend from the QT ratio (which we define using only updates at the best quotes) to the more general order-to-trade ratio that is targeted by exchanges in their attempt to regulate algorithmic and high-frequency trading. If that is the case, then regulators should be aware that restricting or taxing the quoting activity of market makers can have unintended consequences on their information production, and thus may adversely affect a stock's liquidity and its cost of capital.

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.

Comerton-Forde, C., T. Hendershott, C. M. Jones, P. C. Moulton, and M. S. Seasholes (2010). "Time variation in liquidity: The role of market-maker inventories and revenues." *The Journal of Finance*, 65(1), 295–331.

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.

Easley, D. and M. O'Hara (2004). "Information and the cost of capital." *Journal of Finance*, 59, 1553–1583.

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

Fama, E. F. and K. R. French (2015). "A five-factor asset pricing model." *Journal of Financial Economics*, 116(1), 1 – 22.

Glosten, L. and P. Milgrom (1985). "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders." *Journal of Financial Economics*, 14, 71–100.

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.

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.

Hou, K. and R. K. Loh (2016). "Have we solved the idiosyncratic volatility puzzle?" *Journal of Financial Economics*, 121(1), 167 – 194.

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.

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. (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.

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. 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 %) (VOLAT), column (12) shows price, column (13) shows the average Book-to-Market value measured at the end of the previous calendar year (BM), column (14) shows the average number of analysts following the stock (ANF), and column (15) shows the average institutional ownership (INST).

QT		N(trades)	N(quotes)	MCAP	VOLUME (mill.)		SPREAD		ILR	VOLAT				
portf	QT	(x 1000)	(x 1000)	(mill.)	Shares	USD	Quoted	Rel.(%)	(%)	(%)	PRC	BM	ANF	INST
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	1.0	130	178	8,720	75	1,718	0.138	1.41	2.50	4.33	15.6	0.63	15	0.50
2	2.2	47	205	3,607	20	690	0.162	1.63	3.37	3.57	17.8	0.63	11	0.51
3	3.2	33	216	2,841	13	530	0.180	1.74	3.96	3.36	20.1	0.64	10	0.51
4	4.3	25	223	2,474	10	434	0.204	1.85	4.38	3.18	22.1	0.64	9	0.51
5	5.6	19	213	2,085	8	341	0.237	2.02	5.38	3.06	23.7	0.65	8	0.50
6	7.5	15	199	2,292	7	309	0.282	2.20	7.12	2.90	24.6	0.70	8	0.48
7	10.8	11	168	3,239	7	326	0.264	2.01	6.08	2.36	24.9	0.77	10	0.50
8	16.9	6	133	2,012	4	199	0.279	1.88	4.56	1.93	25.6	0.76	9	0.50
9	32.9	3	98	1,414	3	119	0.324	1.96	5.18	1.74	25.9	0.79	7	0.46
10	175.7	1	96	823	1	54	0.442	2.37	8.18	1.51	27.9	1.01	5	0.39

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

The table shows panel regressions of the quote-to-trade ratio (QT) on different stock characteristics. The dependent variable is the monthly QT. The independent variables are: annual number of analysts following the stock (*ANF*), quarterly institutional ownership (*INST*), log-book-to-market as of the previous year end (*BM*); previous month return (*R1*); as well as contemporaneous (monthly) variables: log-market capitalization (*MCAP*), price (*PRC*), U.S. dollar trading volume (*USDVOL*), Amihud illiquidity ratio (*ILR*), relative bid-ask spread (*SPREAD*), and volatility (*VOLAT*). The coefficient on *USDVOL* is multiplied by 10^9. Standard errors are double-clustered at the stock and month level.

	(1)	(2)	(3)	(4)
ANF	-0.69***	-0.16**	-0.33***	-0.53***
	(-5.82)	(-2.44)	(-2.71)	(-5.24)
INST	-34.84***	12.23**	-67.03***	-49.18***
	(-5.33)	(2.45)	(-8.51)	(-6.41)
BM	16.90***	-0.14	14.94***	-3.45
	(3.50)	(-0.05)	(3.19)	(-1.33)
*R*1	-16.04***	-8.34***	-6.51***	-0.62
	(-4.22)	(-3.43)	(-3.23)	(-0.36)
MCAP	-5.23***	3.70**	-4.33***	-3.77**
	(-4.31)	(2.31)	(-3.53)	(-2.32)
PRC	0.52***	0.24***	0.63***	0.43***
	(5.20)	(2.83)	(5.50)	(4.48)
USDVOL	0.35	-0.68***	-1.54***	-2.28***
	(1.17)	(-3.15)	(-4.16)	(-3.77)
ILR	14.41***	0.79	1.98	-1.79
	(3.60)	(0.33)	(0.74)	(-0.77)
SPREAD	-491.47***	-339.00***	3.68	-146.77***
	(-9.60)	(-7.35)	(0.12)	(-3.69)
VOLAT	-37.87***	-13.08***	-46.60***	-16.02***
	(-4.29)	(-3.31)	(-5.38)	(-4.30)
Stock FE	NO	YES	NO	YES
Time FE	NO	NO	YES	YES
N	672,952	672,888	672,952	672,888
Adj. R^2	0.028	0.173	0.070	0.192

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

The table shows monthly returns for various portfolios sorted on the quote-to-trade ratio (QT). We form ten portfolios based on the QT level in month t, and returns are calculated for each portfolio for month $t+1$. Column (1) shows the average monthly portfolio raw return in excess of the risk free rate (r^e_{t+1}) for each portfolio at time $t+1$. Columns (2)-(7) report the risk-adjusted returns, α's. The α's reported in the table are the intercepts (risk-adjusted returns) obtained from regressions of returns on the risk factors. 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 (FF3+PS+MOM), the Fama and French (2015) five factor model (FF5), 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. ***, **, 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 (%)					
					FF3+PS		FF3+PS
	r^e_{t+1}	CAPM	FF3	FF3+PS	+MOM	FF5	+MOM+PIN
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
α_1	1.52***	0.92*	1.05***	1.03***	1.67***	1.08***	1.69***
α_2	1.30***	0.91*	0.85***	0.84***	1.15***	0.68***	1.16***
α_3	1.10***	0.72*	0.62***	0.58***	0.92***	0.50***	0.92***
α_4	1.04***	0.61	0.48***	0.49***	0.76***	0.40***	0.75***
α_5	0.95***	0.58*	0.42**	0.42**	0.60***	0.30***	0.60***
α_6	0.81***	0.30	0.09	0.09	0.33*	0.12	0.32*
α_7	0.94***	0.54*	0.24	0.23	0.57***	0.19*	0.58***
α_8	0.84***	0.43	0.05	0.03	0.32**	0.07	0.31**
α_9	0.84***	0.41	-0.01	-0.01	0.22	0.15	0.21
α_{10}	0.65***	0.37	-0.08	-0.09	0.09	-0.10	0.08
α_{1-10}	0.87***	0.55	1.14***	1.11***	1.58***	1.17***	1.61***

Table 4: Stock 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 FF3+PS+MOM 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 ($MCAP$), 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), dollar volume ($USDVOL$), and price (PRC). All characteristics apart from returns are logged and all coefficients are multiplied by 100. The standard errors are corrected by using the Newey-West method with 12 lags. T-statistics for the QT variable are presented in brackets. ***, **, and * indicate significance at the 1%, 5%, and 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)
Const.	0.006***	0.004***	0.013***	0.011***	0.036***
$QT_{i,t-1}$	-0.222**	-0.244***	-0.286***	-0.297***	-0.119**
	(-2.49)	(-3.03)	(-3.51)	(-3.47)	(-2.39)
$SPREAD_{i,t-1}$		0.141***		0.067	0.035
$ILR_{i,t-1}$			0.097***	0.075**	-0.004
$MCAP_{i,t-1}$					-0.224***
$BM_{i,t-1}$					0.073
$R1_{i,t-1}$					-4.413***
$R212_{i,t-1}$					0.061
$IDIOVOL_{i,t-1}$					-12.544***
$USDVOL_{i,t-1}$					0.163*
$PRC_{i,t-1}$					-0.433***
R^2	0.00	0.01	0.01	0.01	0.04
Time series (months)	216	216	216	216	216

Table 5: Quotes versus Trades

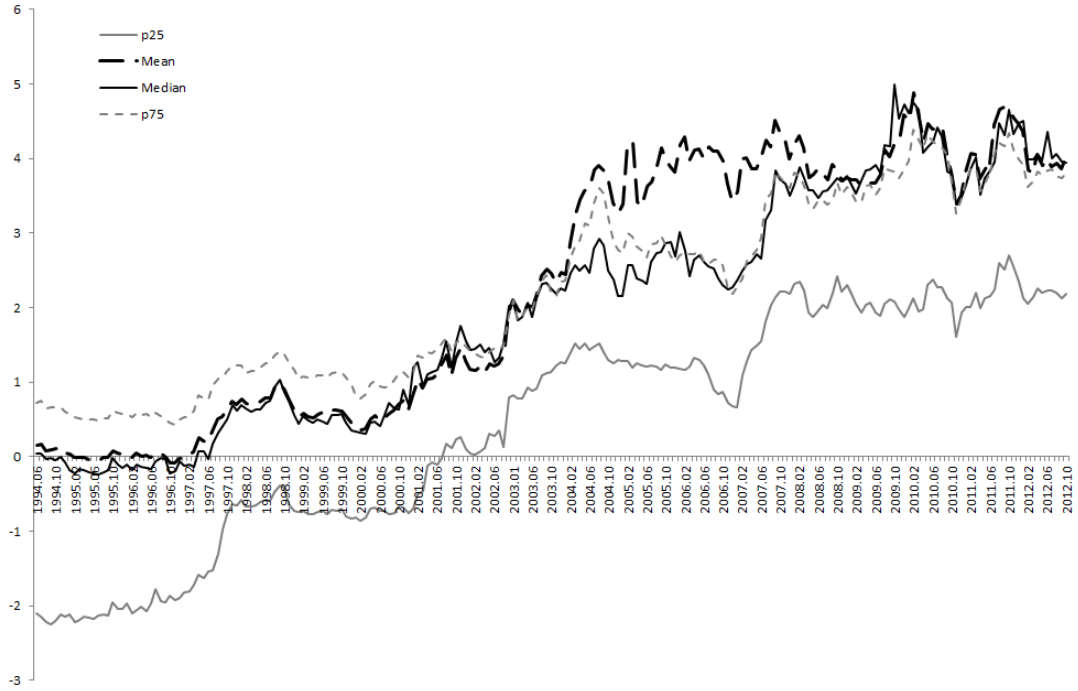
The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted monthly returns on firm characteristics including the number of quotes and trades. The dependent variable is the risk-adjusted return $r^a_{i,t} = r_{i,t} - \sum_{j=1}^{J} \beta_{i,j,t-1} F_{j,t}$, where the risk factors $F_{j,t}$ come from the FF3+PS+MOM 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), market capitalization ($MCAP$), 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), dollar volume ($USDVOL$), and price (PRC). All characteristics apart from returns are logged and 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***	0.365***
$PRC_{i,t-1}$					-0.604***	-0.572***
R^2	0.01	0.01	0.01	0.01	0.04	0.04
Time series (months)	216	216	216	216	216	216

Figure 3: 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*

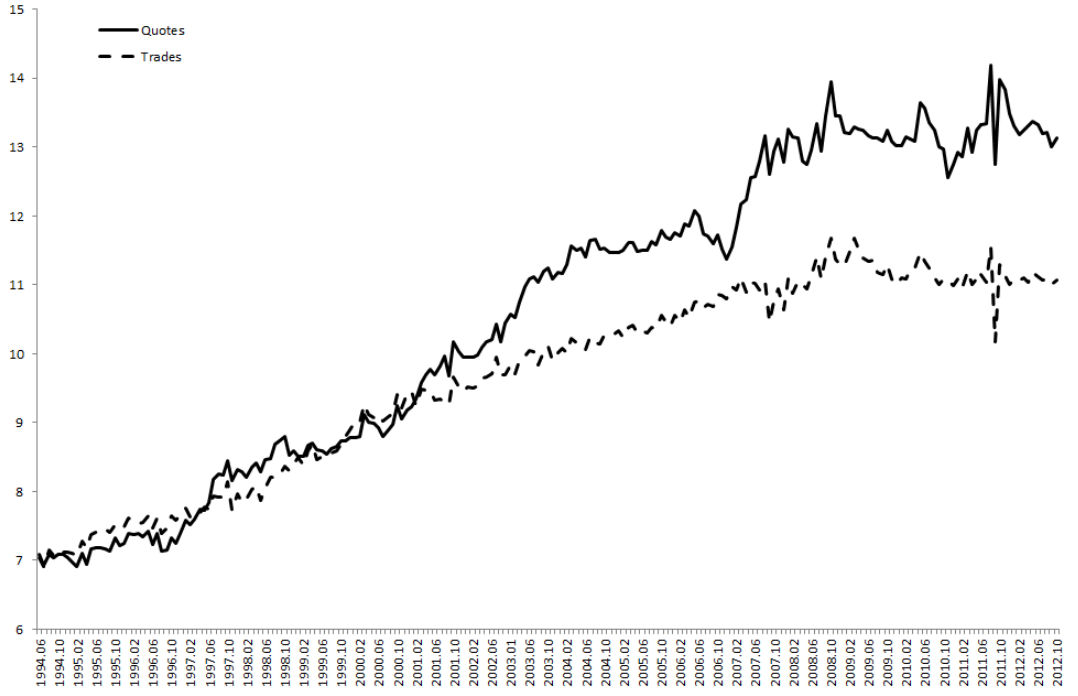
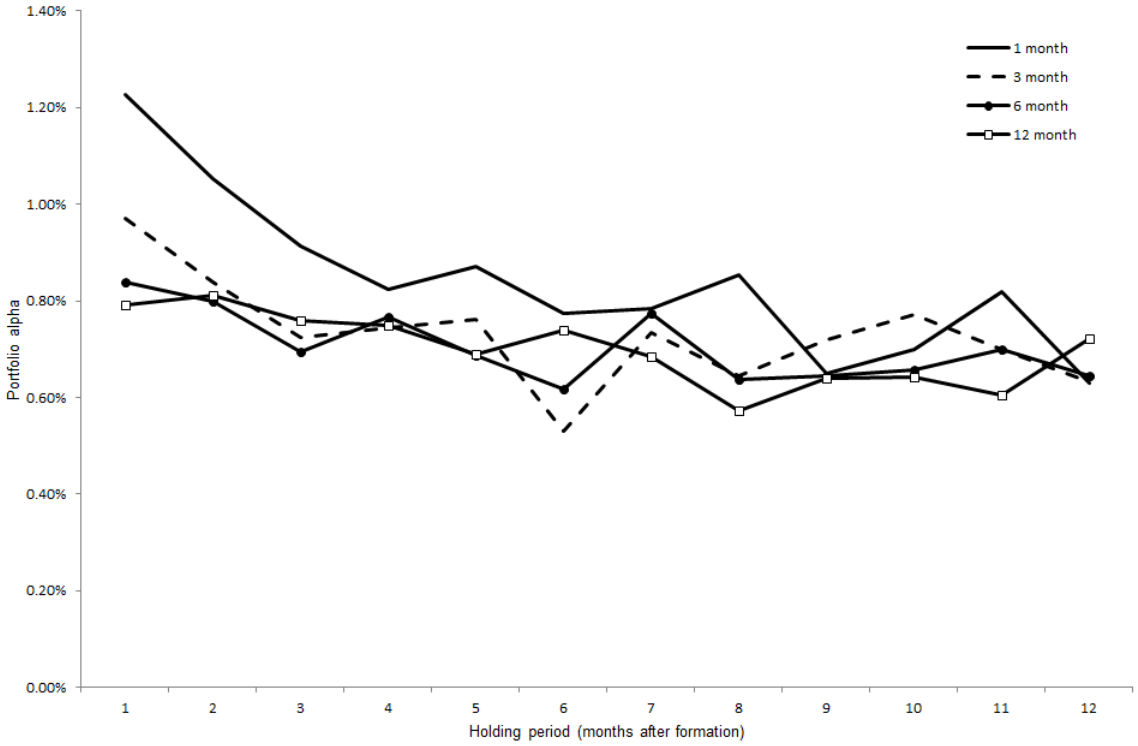


Figure 4: Portfolio alphas for different holding horizons and formation periods

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. Proofs of Results

Proof of Proposition 1. Fix the monitoring rate $q > 0$. Let $\mathcal{I}_\tau$ be the dealer's information set just before trading at τ, and by E_τ the expectation operator conditional on $\mathcal{I}_\tau$. Let $w_\tau = \mathsf{E}_\tau(v)$ be the current dealer's forecast of the fundamental value, and G_τ the variance of the forecast error:

$$G_\tau \;=\; \mathsf{Var}(v - w_\tau). \tag{A1}$$

For simplicity of notation, we omit the subscript τ in the remainder of this proof.

We now compute the dealer's expected utility coming from quoting (a, b) at τ. If we define

$$h \;=\; \frac{a-b}{2}, \qquad \delta \;=\; w - \frac{a+b}{2}, \qquad e \;=\; v - w, \tag{A2}$$

then the quoting strategy is equivalent to choosing (h, δ). Equation (5) implies that traders' buy and sell demands at t are given, respectively, by $Q^b = \frac{k}{2}(v-a) + \ell - m + \varepsilon^b$ and $Q^s = \frac{k}{2}(b-v) + \ell + m + \varepsilon^s$, with $\varepsilon^b, \varepsilon^s \sim \mathcal{N}(0, \Sigma_L/2)$. If x_0 is the dealer's initial inventory, the final inventory x_{end} satisfies $x_{\text{end}} = x_0 - Q^b + Q^s$, which translates into

$$x_{\text{end}} \;=\; x_0 - k\delta + 2m + \varepsilon, \qquad \varepsilon \;=\; -ke + \varepsilon^s - \varepsilon^b \;\overset{IID}{\sim}\; \mathcal{N}\big(0, k^2 G + \Sigma_L\big). \tag{A3}$$

Substituting Q^b and Q^s in the dealer's objective (8), and ignoring monitoring costs, we get $\mathsf{E}_\tau\Big(x_0 v + \frac{k}{2}(a-v)^2 - \frac{k}{2}(v-b)^2 + (\ell - m)(a-v) + (\ell+m)(v-b) - \gamma x_{\text{end}}^2\Big)$. We decompose $\mathsf{E}_\tau(v-b)^2 = \mathsf{E}_\tau(v - w + w - b)^2 = G + (w-b)^2$, and similarly $\mathsf{E}_\tau(a-v)^2 = G + (a-w)^2$. Also, $\mathsf{E}_\tau(x_{\text{end}}^2) = (x_0 - k\delta + 2m)^2 + (k^2 G + \Sigma_L)$. Using the notation in (A1) and (A2), the dealer's maximization problem is equivalent to

$$\max_{h,\delta} \Big(x_0 w - kG - k\delta^2 - kh^2 + 2\ell h + 2m\delta - \gamma\big(x_0 - k\delta + 2m\big)^2 - \gamma\big(k^2 G + \Sigma_L\big)\Big). \tag{A4}$$

The first order condition in (A4) with respect to h implies $h = \frac{\ell}{k}$, which shows that the optimal half spread satisfies (9). The first order condition in (A4) with respect to δ implies $\delta = \frac{\gamma}{1+k\gamma}\, x_0 + \frac{m}{k}\, \frac{1+2k\gamma}{1+k\gamma}$, which shows that the optimal discount satisfies (9). The second order conditions are satisfied for both h and δ. The maximum expected utility

the dealer can achieve (not accounting for monitoring costs) is

$$U_{\max} \;=\; x_0 w + \frac{\ell^2}{k} - k(1+k\gamma)G - \gamma\Sigma_L + \frac{m^2 - 2\gamma k m x_0 - \gamma k x_0^2}{k(1+k\gamma)}. \tag{A5}$$

Note that this formula is linear in the forecast w, hence by the law of iterated expectations it is time-consistent and well defined as a value function. This also implies that the optimal quotes change along with the forecast w, as described above. □

Proof of Lemma 1. In general, the forecast is the average signal with weights given by the precision of each signal. But the precision of each signal is the same: $\frac{1}{\mathsf{Var}(\varepsilon_0)} = F(q)$. Hence, the forecast is the equal-weighted average signal: $w_t = v + \frac{\varepsilon_0 + \cdots + \varepsilon_n}{n+1}$. The variance of the forecast error is $\mathsf{Var}(v - w_t) = \mathsf{Var}\big(\frac{\varepsilon_0 + \cdots + \varepsilon_n}{n+1}\big) = \frac{\mathsf{Var}(\varepsilon_0)}{n+1}$, hence the forecast precision is $\frac{1}{\mathsf{Var}(v-w_t)} = (n+1)F(q)$. □

Proof of Proposition 2. Recall that we consider the initial signal s_0 as the dealer's prior, while the other signals s_n with $n > 0$ as resulting from monitoring. Trading has frequency 1 while monitoring has frequency q. Hence, at each time before trading occurs, the probability that monitoring occurs before trading is $q/(q+1)$, while the probability that trading occurs before monitoring is $1/(q+1)$. Denote by n the event in which exactly n monitoring times occur before trading. The ex ante probability (before monitoring starts at $t = 0$) of event n is $\big(\frac{q}{q+1}\big)^n \frac{1}{q+1} = \frac{q^n}{(q+1)^{n+1}}$. In that case, Lemma 1 implies that the forecast variance is $G_n = \frac{1}{(n+1)F(q)}$. Thus, the ex ante expected forecast variance is

$$G(q) \;=\; \mathsf{Var}\big(v - w_n\big) \;=\; \sum_{n=0}^{\infty} \frac{q^n}{(q+1)^{n+1}} \, \frac{1}{(n+1)F(q)} \;=\; \frac{\ln(q+1)}{qF(q)}, \tag{A6}$$

where the last equality comes from the Taylor series: $\ln(1-\alpha) = -\sum_{n=0}^{\infty} \frac{\alpha^{n+1}}{n+1}$, with $\alpha = \frac{q}{q+1}$. When $F(q) = f\ln(q+1)$, we get $G(q) = \frac{1}{fq}$.

Consider general functions $G(q)$ and $C(q)$. Then, equation (A5) from the proof of Proposition 1 implies that the dealer's maximum expected utility (accounting for the monitoring costs $C(q)$ is of the form $U_{\max} = D - k(1+k\gamma)G(q) - C(q)$, where D is a constant that does not depend on q. The first order condition with respect to q is

equivalent to $-k(1+k\gamma)G'(q) - C'(q) = 0$. Thus, the optimal monitoring rate satisfies

$$-\frac{C'(q)}{G'(q)} \;=\; k(k\gamma+1). \tag{A7}$$

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

We now use the specification $F(q) = f\ln(q+1)$ and $C(q) = cq$, and compute the optimal monitoring rate q. From above, $G(q) = \frac{1}{fq}$, hence (A7) implies that q satisfies $fcq^2 = k(k\gamma+1)$, which proves the first part of equation (15). As G is strictly convex, the second order condition is satisfied. One verifies that $F(q) = fq$ and $F(q) = f$ correspond respectively to $G(q) = \frac{\ln(q+1)}{fq^2}$ and $G(q) = \frac{\ln(q+1)}{fq}$, which are strictly convex functions as well. □

Proof of Corollary 1. By visual inspection of equation (15), it is clear that q is increasing in k and γ, and decreasing in f and c. □

Proof of Corollary 2. The proof of Proposition 1 implies that the dealer's equilibrium choice of h and δ does not depend on the forecast variance. It remains to show under what conditions the dealer chooses the same monitoring rate. Equation (A5) in the proof of Proposition 1 shows that it is enough to have the same ex ante variance $G(q)$. Equation (A6) in the proof of Proposition 2 implies that the ex ante forecast variance $G(q)$ satisfies $G(q) = \frac{\ln(q+1)}{qF(q)}$. With a unique signal, the ex ante forecast variance is $\tilde{G}(q) = \frac{1}{\tilde{F}(q)}$. If we want the two variances to be equal, we need $\frac{\ln(q+1)}{qF(q)} = \frac{1}{\tilde{F}(q)}$, which is equivalent to (16). □

Proof of Corollary 3. Equation (18) implies that the cost of capital (discount) is equal to $\frac{m}{k}\,\frac{1+2\gamma k}{1+\gamma k} + \frac{\gamma}{1+\gamma k}\,x_0$. This is clearly increasing in m. The derivative with respect to k is $-\frac{m(2\gamma^2k^2+2\gamma k+1)}{k^2(1+\gamma k)^2} - \frac{\gamma^2}{(1+\gamma k)^2}x_0$, which is negative if $x_0 \geq 0$. □

Proof of Corollary 4. Suppose we hold all parameters constant except for k. According to Corollary 3, the discount δ is decreasing in k. At the same time, the QT ratio q is increasing in k (see Corollary 1). This proves the inverse relation between δ and q. □

Proof of Corollary 5. Equation (20) follows by simply substituting (19) in (18) and applying Proposition 1 to show that these values correspond to the equilibrium. It remains only to show that the neutral inventory $x_0 = \frac{m}{\gamma k}$ indeed balances the expected order flow. We use the notation from the proof of Proposition 1. From (5) it follows that in equilibrium $Q^b - Q^s = k\big(v - \frac{a+b}{2}\big) - 2m + \varepsilon^b - \varepsilon^s$. Since $\mathsf{E}_\tau(v) = w$ and $w - \frac{a+b}{2} = \delta$, we have $\mathsf{E}_\tau(Q^b - Q^s) = k\delta - 2m$. Thus, when δ is equal to its neutral value, $\delta_{\text{neutral}} = \frac{2m}{k}$, the order flow is balanced, i.e., $\mathsf{E}_\tau(Q^b) = \mathsf{E}_\tau(Q^s)$. But equation (9) shows that x_0 and δ are in one-to-one correspondence. Thus, if δ is equal to its neutral value, x_0 is also equal to its neutral value. Hence, when $x_{0,\text{neutral}} = \frac{m}{\gamma k}$, the expected order flow is balanced, and this completes the proof. □

Appendix B. Micro-Foundations of Order Flow

In this section we provide assumptions under which the traders' liquidity demand is approximately of the form described in (5). The proofs are in Appendix B.3.

B.1. Environment

There are two types of traders: investors and liquidity (or noise) traders. Liquidity traders are either buyers or sellers. When trading occurs, liquidity buyers submit to the exchange an aggregate buy order for L^b shares, and liquidity sellers submit an aggregate buy order for L^s shares. Both L^b and L^s have IID normal distribution $\mathcal{N}(\ell_L, \Sigma_L/2)$, therefore by subtracting the mean we decompose them as follows:

$$L^b \;=\; \ell_L + \varepsilon^b, \quad L^s \;=\; \ell_L + \varepsilon^s, \quad \text{with} \quad \varepsilon^b, \varepsilon^s \;\sim\; \mathcal{N}(0, \Sigma_L/2). \tag{B1}$$

Investors have CARA utility with coefficient A. A mass one of investors starts with an initial endowment in the risky asset that is normally distributed as $\mathcal{N}(M, \sigma_M^2)$.[32] Investors observe the asset value v before trading, and then trade on the exchange at the quotes set by the dealer: the ask quote a and the bid quote b. The asset liquidates

[32]In addition, suppose the liquidity traders' initial average endowment is $x_\ell = 0$. As the investors' average endowment is the asset supply M, market clearing implies that the dealer's initial inventory must be zero. If instead we allow other values of x_ℓ, this section applies to any initial dealer inventory.

at $v+u$, where u has a normal distribution $\mathcal{N}(0,\sigma_u^2)$.[33]

B.2. Equilibrium

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* $\underline{X}$ and the *upper target* $\overline{X}$ by:

$$\underline{X} \;=\; \frac{v-a}{A\sigma_u^2}, \qquad \overline{X} \;=\; \frac{v-b}{A\sigma_u^2}. \tag{B2}$$

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

Lemma B.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 $\underline{X}$ and $\overline{X}$ be defined as in (B2). 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 $x_0>\overline{X}$.*

Next, define the following numeric constants:

$$\rho_0 \;=\; \frac{1}{\sqrt{8\pi}} \approx 0.1995, \qquad \rho_1 = \frac{1}{2\pi}+\frac{1}{4} \approx 0.4092. \tag{B3}$$

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

Proposition B.1. *Investors submit aggregate orders Q^b and Q^s of the form*

$$\begin{aligned} &Q^b \;\approx\; \frac{k}{2}(v-a)+\ell-m+\varepsilon^b, \qquad Q^s \;\approx\; \frac{k}{2}(b-v)+\ell+m+\varepsilon^s, \\ &\textit{with}\quad k \;=\; \frac{2\rho_1}{A\sigma_u^2}, \qquad \ell \;=\; \ell_L+\rho_0\sigma_M, \qquad m \;=\; \rho_1 M, \end{aligned} \tag{B4}$$

[33]The notation in Appendix B differs slightly from the rest of the paper, where v denotes the liquidation value. Here the liquidation value here is $v+u$, while v is the forecast of the informed investors. This notation, however, is compatible with the rest of the paper as long as the dealer learns about v rather than $v+u$.

and the error terms ε^b and ε^s are IID with normal distribution $\mathcal{N}(0, \Sigma_L/2)$. Both approximations in (B4) represent equality up to terms of the order of $1/\sigma_M$.

Proposition B.1 provides micro-foundations for the order flow equations (5). For instance, the imbalance parameter m 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 a key determinant of the investor elasticity k. Intuitively, 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).

B.3. Proofs of Results

Proof of Lemma B.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.[34] 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 upper 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 B.1. We first introduce some notation. Define

$$\begin{aligned} \phi(x) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{x^2}{2}\right), \quad \Phi(x) = \int_{-\infty}^{x} \phi(t)\mathrm{d}t, \\ \psi(x) &= \Phi(-x)\Big(\phi(x) - x\Phi(-x)\Big), \end{aligned} \tag{B5}$$

[34] 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 (B5) 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 function

$$\phi_M(x) \;=\; \frac{1}{\sigma_M}\,\phi\Big(\frac{x-M}{\sigma_M}\Big). \tag{B6}$$

Then, investors' endowments integrate to $\int_{-\infty}^{\infty} x\phi_M(x)\mathrm{d}x = M$, 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, note that by assumption his liquidation value is $v+u$, where v is known by the investor, and $u \sim \mathcal{N}(0,\sigma_u^2)$ is unknown. Thus, investor i computes $\mathsf{E}(v+u) = v$ and $\mathsf{Var}(v+u) = \sigma_u^2$. Thus, the targets $\underline{X} = \frac{v-a}{A\sigma_u^2}$ and $\overline{X} = \frac{v-b}{A\sigma_u^2}$ are common to all investors.

According to Lemma B.1, the optimal demand of an investor depends on his initial endowment. By assumption, traders' endowments are IID with density $\phi_M(x)$ as in (B6). Therefore, investors' aggregate buy market order is equal to $I^b = \underline{P}\int_{-\infty}^{\underline{X}}(\underline{X}-x)\phi_M(x)\mathrm{d}x$, where $\underline{P} = \int_{-\infty}^{\underline{X}}\phi_M(x)\mathrm{d}x$ is the mass of investors with endowments below $\underline{X}$. Similarly, investors' aggregate sell market order is equal to $I^s = \overline{P}\int_{\overline{X}}^{\infty}(x-\overline{X})\phi_M(x)\mathrm{d}x$, where $\overline{P} = \int_{\overline{X}}^{\infty}\phi_M(x)\mathrm{d}x$ is the mass of investors with endowments above $\overline{X}$. Finally, investors with endowments between $\underline{X}$ and $\overline{X}$ do not submit any order. We compute

$$I^b \;=\; \psi\left(\frac{M-\underline{X}}{\sigma_M}\right), \qquad I^s \;=\; \psi\left(\frac{\overline{X}-M}{\sigma_M}\right), \tag{B7}$$

where ψ is as in (B5. Consider the linear approximation of ψ near $x=0$:

$$\psi(x) \;=\; \rho_0 - \rho_1 x + O(x^2), \quad \rho_0 \;=\; \psi(0) \;=\; \tfrac{1}{\sqrt{8\pi}}, \quad \rho_1 \;=\; -\psi'(0) \;=\; \tfrac{1}{2\pi}+\tfrac{1}{4}, \tag{B8}$$

where $O(x^2)$ represents the standard "big O" notation.[35] The investors' aggregate buy order is thus $I^b = \rho_0\sigma_M + \rho_1(\underline{X}-M) + O(1/\sigma_M) = \frac{\rho_1}{A\sigma_u^2}(v-a) + \rho_0\sigma_M - \rho_1 M + O(1/\sigma_M)$. Also, from (B1), the liquidity buyers' aggregate order is $L^b = \ell_L + \varepsilon^b$, with $\varepsilon^b \sim \mathcal{N}(0, \Sigma_L/2)$. By adding I^b and L^b, we obtain that the aggregate traders' buy order,

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

$Q^b = I^b + L^b$, satisfies

$$Q^b \;=\; \frac{\rho_1}{A\sigma_u^2}(v-a) + \big(\ell_L + \rho_0\sigma_M\big) - \rho_1 M + \varepsilon^b + O(1/\sigma_M). \tag{B9}$$

Let $k = \frac{2\rho_1}{A\sigma_u^2}$, $\ell = \ell_L + \rho_0\sigma_M$, $m = \rho_1 M$. Thus, we have $Q^b = \frac{k}{2}(v-a) + \ell - m + \varepsilon^b + O(1/\sigma_M)$ and similarly $Q^s = \frac{k}{2}(b-v) + \ell + m + \varepsilon^s + O(1/\sigma_M)$. This proves (B4). $\square$