Quoting Activity and the Cost of Capital

Ioanid Roşu, Elvira Sojli, and Wing Wah Tham*

August 7, 2020

Abstract

We study the quoting activity of market makers in relation to 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 more precisely informed, which reduces mispricing and lowers expected returns.

JEL: G12, G14, D82

KEYWORDS: Quote-to-trade ratio, market making, liquidity, price discovery, monitoring, information acquisition, neglected stocks, inventory, high frequency trading.

---

*Roşu (corresponding author), rosu@hec.fr, HEC Paris; Sojli, e.sojli@unsw.edu.au, University of New South Wales; and Tham, w.tham@unsw.edu.au, University of New South Wales. We thank two anonymous referees, Hendrik Bessembinder (the editor), Dion Bongaerts, Jean-Edouard Colliard, David Easley, Thierry Foucault, Michael Goldstein, Amit Goyal, Johan Hombert, Dashan Huang, Maureen O'Hara, Rohit Rahi, Daniel Schmidt, and Yajun Wang for their suggestions. We are also grateful to finance seminar participants at the Stockholm Business School, Academy of Economic Studies in Bucharest, Cass Business School, University of British Columbia, Pontifical University of Chile, University of Chile, University of Technology Sydney, HEC Paris, as well as conference participants at the 2018 Australasian Finance and Banking Conference, 2018 Financial Management Association, 2018 Behavioural Finance and Capital Markets Conference in Melbourne, 2018 European Finance Association, 2018 Monash Workshop on Financial Markets, 2017 AFFI/Eurofidai, 2017 Hong Kong Conference on Market Design and Regulation in the Presence of High-Frequency Trading, 2017 FIRN, 2017 Northern Finance Association, 2017 SAFE Market Microstructure Conference in Frankfurt, 2017 Erasmus Liquidity Conference, 2017 RMI Risk Management Conference in Singapore, 2017 CEPR-Imperial-Plato Inaugural Conference in London, 2017 Frontiers of Finance Conference in London, 2017 FIRN Sydney Market Microstructure Conference, and the 2016 CEPR Gerzensee European Summer Symposium in Financial Markets, for valuable comments.

I. Introduction

Information is related to a firm's cost of capital, which is important for firm shareholders and market participants alike.[1] In financial markets, information is incorporated into prices by market makers, who provide liquidity via quotes (or limit orders), and market takers, who demand liquidity via marketable orders and generate trades. A natural approach then is to study the relation between information and cost of capital through the lens of quoting and trading. Quoting activity in particular has attracted the attention of exchanges and regulators due to its rapid increase in recent years.[2]

Quoting activity, however, has played a limited role in the academic literature on price discovery. The reason is that in many market structure models, such as Glosten and Milgrom (1985), the market makers mechanically set their quotes at the expected asset value given the information contained in trades. In this setup, there is no expected price appreciation, and hence the expected return (cost of capital) is zero. Moreover, suppose we define the quote-to-trade ratio, henceforth "QT ratio," as the number of quote updates divided by the number of trades. Then, as market makers set their bid and ask quotes passively in response to trades, the QT ratio is always equal to two (or a higher constant, if one adds exogenous public news to the model). In contrast, we show empirically that the QT ratio exhibits various patterns across stocks, and we summarize these patterns as a list of new empirical stylized facts.

Our first stylized fact (SF1) is that the QT ratio is larger in stocks that are neglected or difficult to understand (with low analyst coverage, institutional ownership, trading volume, or volatility). To illustrate this result, Figure 1 shows the average number of analysts following a stock and the average trading volume for ten portfolios sorted by the QT ratio. Firms with lower analyst coverage or lower trading volume have larger QT ratios than firms with higher analyst coverage or higher trading volume.

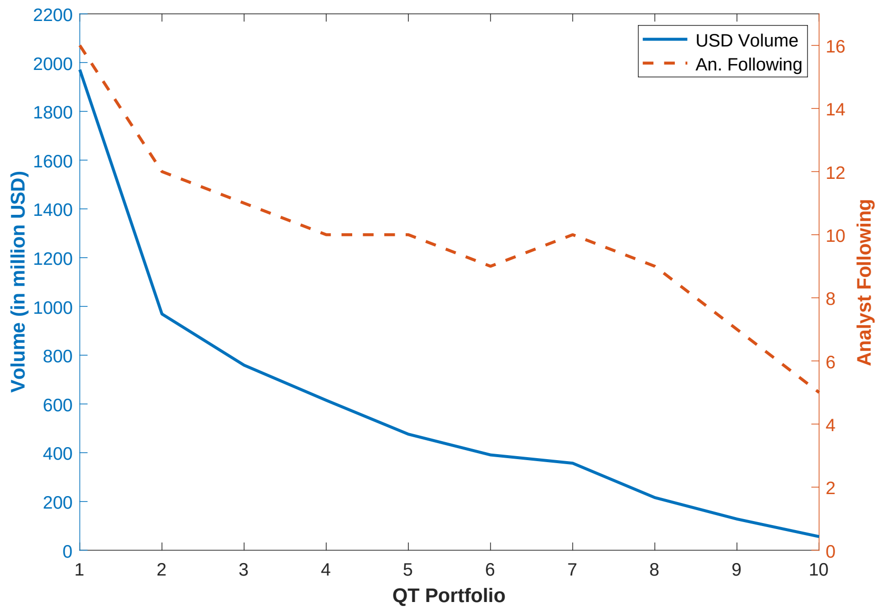
[1] See Diamond and Verrecchia (1991), Easley and O'Hara (2004), Amihud, Mendelson and Pedersen (2005) and references therein.

[2] Many exchanges have implemented limits on quoting activity. London Stock Exchange was the first to introduce, in 2005, an "order management surcharge" based on the number of orders per trades submitted. Several exchanges followed: Euronext in 2007, DirectEdge, Oslo Stock Exchange, Deutsche Börse and Borsa Italiana in 2012. In 2014, MIFID-II/R required trading venues to establish a maximum unexecuted order-to-transaction ratio as one of its controls to prevent disorderly trading conditions (see http://ec.europa.eu/finance/securities/docs/isd/mifid/rts/160518-rts-9_en.pdf).

FIGURE 1

Volume and Analyst Coverage for 10 Quote-to-Trade Ratio Portfolios

The figure shows the average U.S. dollar trading volume and the average 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.



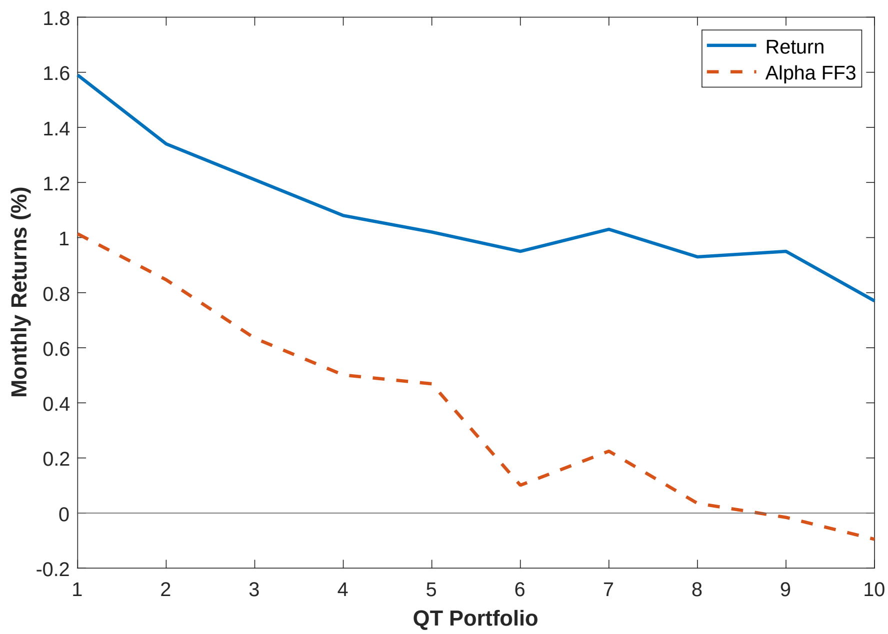
The second stylized empirical fact (SF2) is that the QT ratio has increased significantly over time, especially after the emergence of algorithmic and high-frequency trading in 2003. This fact is documented by Hendershott, Jones and Menkveld (2011) for their proxy of algorithmic trading, the message-to-trade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

The third stylized empirical fact (SF3) is that stocks with higher QT ratios tend to have lower expected returns (cost of capital). We regard this as our main result, and we call it the QT effect. This result is surprising, because stocks with higher QT ratios are usually smaller and more illiquid (SF1), and thus one might expect them to have a higher cost of capital. Figure 2 illustrates the QT effect: stocks with large QT ratios have low average returns, whether computed in excess of the risk-free rate, or after risk adjusting with the Fama and French (1993) factors. Further empirical analysis using Fama-MacBeth regressions confirms the QT effect, and verifies that it holds in different

FIGURE 2

Excess Return and Alpha for 10 Quote-to-Trade Ratio Portfolios

The figure shows 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 FF3") 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. Returns are computed monthly and presented in percentages.



subsamples. The fourth stylized empirical fact (SF4) is that the QT effect appears to be driven by quotes and not by trades.

To interpret our empirical findings, we construct a model that focuses on the quoting activity of market makers and its relation to the cost of capital. Our model extends the monopolistic dealer models of Ho and Stoll (1981) and Hendershott and Menkveld (2014), with two changes: First, the dealer learns about the value of the risky asset via costly monitoring (which generates quote changes). Second, the order flow is driven by risk averse investors (that demand a positive risk premium). The dealer ("she") sets ask and bid quotes to maximize the expected profit, subject to a quadratic penalty in her inventory. She monitors the market by getting signals about the fundamental value according to a Poisson process with frequency called the monitoring rate. There is one round of trading, after which the asset liquidates at a random fundamental value. Trading occurs at the first arrival of a Poisson process with frequency normalized to one.[3] As the dealer optimally changes her quotes every time she receives a signal, her monitoring frequency is the same as her quoting frequency, or the "quote rate."

In equilibrium, the dealer's quote rate is decreasing in the monitoring precision. Indeed, a small precision of the signals obtained from monitoring makes the dealer monitor more often (and increase her quote rate) in order not to stay far away from the fundamental value and incur a large expected inventory penalty. This is consistent with our stylized fact SF1: the quote-to-trade ratio (which is an empirical proxy for the quote rate) is higher in neglected, difficult-to-understand stocks, in which monitoring is expected to produce imprecise signals.

The dealer's quote rate is increasing in the monitoring cost. Indeed, when this cost is small, the dealer can afford to monitor more often in order to maintain the same precision. There is evidence that monitoring costs have decreased dramatically in recent times (see Hendershott et al., 2011). This is consistent with our stylized fact SF2: QT ratios have increased significantly over time, especially after the emergence of algorithmic and high-frequency trading.

We define the cost of capital in the model simply as the price discount, which is the difference between the dealer's value forecast and her mid-quote price (the midpoint between the bid and ask). A key determinant of the equilibrium price discount is the investor elasticity, which measures how aggressively investors trade in response to the dealer's pricing error. Consider an increase in investor elasticity, which happens if investors have more precise private signals or if they are less risk averse. Then, first, the dealer must monitor the market more often (hence increase the quote rate) to reduce the pricing error, as large errors would make aggressive investors cause large dealer inventories. Second, the dealer must reduce the pricing discount (hence reduce the cost of capital) by keeping the mid-quote closer to her forecast. Intuitively, this reduction occurs because more informed or less risk averse investors need a smaller compensation in the form of a smaller price discount. These facts together imply an inverse relation between the quote to trade ratio and the cost of capital, which is consistent with our stylized fact SF3.

[3]Our results are robust if we extend our baseline model to multiple dealers (see Section 4 in the Internet Appendix), or to multiple trading rounds (see Section 5 in the Internet Appendix).

The model generates two additional empirical predictions, which are borne out in the data. Our first prediction is that the number of market makers in a stock has an inverse relation to the stock's QT ratio. This is surprising, because one might think that a larger number of market makers generates more quoting activity. However, in our model we interpret a larger number of market makers as a smaller inventory aversion for the representative dealer, and a less inventory-averse dealer can afford to monitor the stock less often and set a lower quote rate. An extension of the model to multiple dealers provides additional intuition for our first prediction: because dealer quotes are public information, each dealer's monitoring exerts a positive externality on the others and thus leads to under-investment in monitoring in equilibrium.[4]

Our second additional prediction is that the number of market makers in a stock has no relation to the stock's cost of capital. This prediction depends on the dealer being in the "neutral state," meaning that her initial inventory is such that there is a zero expected imbalance between the traders' buy and sell quantities.[5] Intuitively, in the neutral state, the price that balances the incoming order flow is affected only by the properties of the order flow and not by the characteristics of the dealer, including her inventory aversion (or the number of dealers if we consider the case of multiple dealers).

We also examine several alternative explanations for the QT effect. In particular, we study the role of frictions (e.g., tick size and impediments to arbitrage), institutional investors and governance, return reversals, and market structure events (e.g., Regulation NMS). While our analysis supports some of these alternative explanations, they explain only part of the QT effect. Overall, our model provides a consistent explanation for several stylized empirical facts (SF1–SF4 and two additional predictions) that describe the market makers' quoting behavior and its relation with the cost of capital. While alternative explanations for each individual stylized fact may exist, we note that our model provides a coherent explanation for all of these findings, based on the production of public information by market makers.

[4]In Section 4 in the Internet Appendix, we extend the model to N dealers and show directly that the aggregate quote rate is decreasing in N (see Corollary IA.4).

[5]In Section 5 in the Internet Appendix, we extend the model to multiple trading rounds and show that the neutral state is the long-term average state, regardless of the initial state.

Related Literature

This paper contributes to the literature on market making.[6] Our work is closest to Hendershott and Menkveld (2014). In their setup, the order flow is exogenously specified, and the equilibrium QT ratio is constant and equal to two. Another related paper is Easley and O'Hara (2004), which analyzes the relation between information and the cost of capital. One of their main findings is that more public information leads to a lower cost of capital.[7] In their rational expectations equilibrium model, however, there are no quotes and thus our results cannot be accommodated in their paper. A related paper is Foucault, Röell and Sandas (2003), which in its analysis of NASDAQ's professional day traders (the "SOES bandits") shares a finding similar to ours: news monitoring by one dealer generates a positive externality on the other dealers. In their model, there is also a negative externality, because the bandits may discover that some dealer quotes are stale after one dealer updates her quotes.

Our paper has implications for the burgeoning literature on high-frequency trading (HFT).[8] The quote-to-trade ratio is often connected to HFT by regulators, practitioners and academics.[9] 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. 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 proxy for HFT, one may be tempted to attribute the QT effect to HFT activity. Hendershott et al. (2011) find that algorithmic trading has a positive effect on stock liquidity. Therefore, it is plausible that stocks with higher HFT activity (and therefore higher QT ratio)

[6]See O'Hara (1995), Hendershott and Menkveld (2014), and references therein.

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

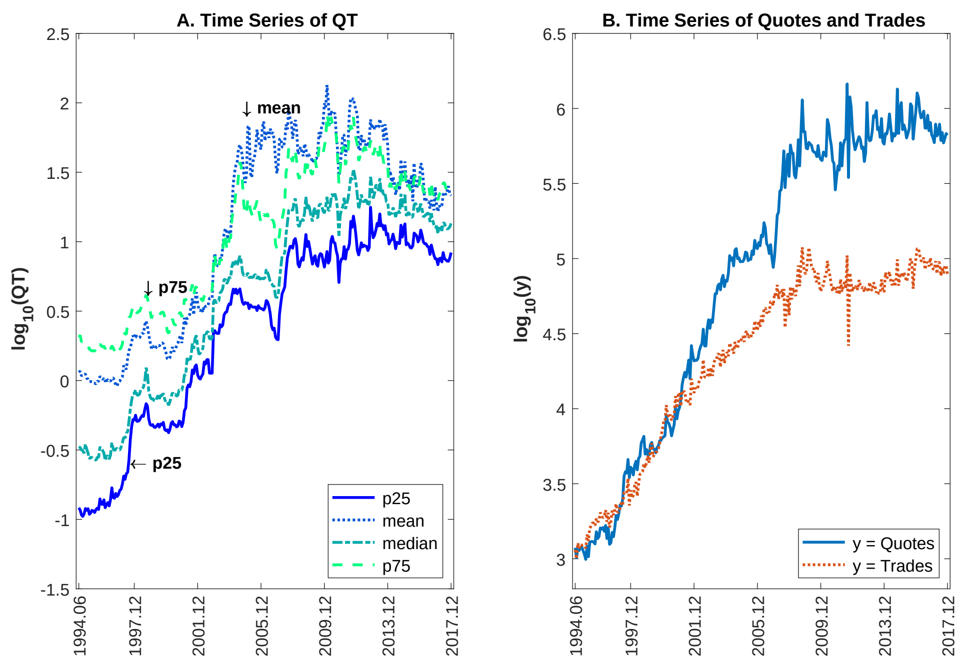
[8]See for example Menkveld (2016) and references therein.

[9]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. 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, 2018), high-frequency trading (see, e.g., Conrad, Wahal and Xiang, 2015; Brogaard, Hendershott and Riordan, 2017), and arbitrage activity (see Foucault, Kozhan and Tham, 2016).

FIGURE 3

Time Series Evolution of Quote-to-Trade Ratio

The figure shows the base-10 logaritm of the time series of quotes, trades, and the quote-to-trade ratio $QT_{i,t} = \frac{N(quotes)_{i,t}}{N(trades)_{i,t}}$. Graph A shows the monthly time series of the cross-sectional mean, median, 25th, and 75th percentile of the QT variable. Graph B shows the monthly average of the number of quote updates and the number of trades.



are more liquid, and thus have a lower cost of capital. This argument, however, is not consistent with our empirical stylized fact SF1 which shows that a large QT ratio is typically to be found in illiquid, neglected stocks. Moreover, the argument does not explain why the QT effect also holds during 1994–2002, when HFT is not known to have a significant impact in financial markets (see Section IV.D). We thus find the HFT explanation of the QT effect unlikely.

The paper is organized as follows. Section II describes the data and provides our main empirical results. Section III describes the model, solves for the equilibrium quote rate and cost of capital, and provides additional predictions. Section IV investigates alternative explanations of the QT effect. Section V concludes. All proofs are in the Appendix.

II. Quotes, Trades, and Returns

In this section, we construct our quote-to-trade ratio measure (also called "QT ratio" or "QT"), and provide stylized facts on quotes, trades, and stock returns.

A. Data

To construct our QT ratio variable, we use the trades and quotes reported in TAQ for the period June 1994 to December 2017.[10] Using TAQ data allows us to generate a long time series of the variable QT at the stock level, in order to conduct asset pricing tests. We retain stocks listed on the NYSE, AMEX, and NASDAQ for which information is available in TAQ, Center for Research in Security Prices (CRSP), and Compustat.

Our sample includes only common stocks (Common Stock Indicator Type = 0), common shares (Share Code 10 and 11), and stocks not trading on a "when issued" basis. Stocks that change primary exchange, ticker symbol, or CUSIP are removed from the sample (Chordia, Roll and Subrahmanyam, 2000; Hasbrouck, 2009). To avoid illiquidity issues related to the price level, we also remove stocks that have a price lower

[10]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.

than \$2 and higher than \$1,000 at the end of a month.[11] To avoid look-ahead bias, all filters are applied on a monthly basis and not on the whole sample. There are 10,670 individual stocks in the final sample.

All returns are calculated using bid-ask midpoint prices, following our definition of the price p_t (see Proposition 1), and to reduce market microstructure noise effects on observed returns (see Asparouhova, Bessembinder and Kalcheva, 2010, 2013).[12] 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). Risk factors are from WRDS and Kenneth French's website for the period 1926 to 2017. The PIN factor is from Sören Hvidkjaer's website and is available from 1984 to 2002. Table IA.1 in the Internet Appendix reports the definitions and the construction details for all variables, and Table IA.2 in the Internet Appendix provides the summary statistics.

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 across all exchanges, and $N(trades)_{i,t}$ is the number of trades in stock i during month t.

B. Stock Characteristics and the Quote-to-Trade Ratio

In this section, we analyze the relation of the QT ratio with various firm-level 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

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

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

[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 III.A.

control for unobservable time-invariant stock characteristics.

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 290 stocks. Stocks priced below \$2 or above \$1000 at the end of month t are removed. The sample period is June 1994 to December 2017. 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 portf	QT	N(trades) (x 1000)	N(quotes) (x 1000)	MCAP (mill.)	VOLUME (mill.) Shares	VOLUME (mill.) USD	SPREAD Quoted	SPREAD Rel.(%)	ILR (%)	VOLAT (%)	PRC	BM	ANF	INST
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
1	1.4	156	280	10,764	84	1,971	0.118	1.21	2.09	4.01	16.1	0.64	16	0.51
2	3.1	65	356	5,555	24	969	0.142	1.41	2.90	3.29	19.6	0.63	12	0.51
3	4.5	47	377	4,571	16	759	0.158	1.50	3.41	3.02	23.1	0.63	11	0.51
4	5.9	36	376	3,812	12	615	0.178	1.59	3.76	2.84	25.7	0.62	10	0.52
5	7.5	27	349	3,045	9	476	0.206	1.73	4.58	2.73	27.8	0.63	10	0.51
6	9.6	20	313	2,826	7	391	0.244	1.88	6.04	2.58	28.9	0.67	9	0.49
7	13.1	14	259	3,308	7	357	0.232	1.74	5.30	2.12	28.9	0.73	10	0.50
8	19.2	8	207	2,069	4	216	0.245	1.64	4.01	1.75	29.0	0.73	9	0.50
9	34.4	5	160	1,425	3	128	0.285	1.71	4.53	1.59	28.7	0.76	7	0.46
10	177.3	2	137	826	1	56	0.403	2.11	7.34	1.43	30.3	0.99	5	0.39

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 month t. The QT portfolio 1 has the lowest QT, and the QT portfolio 10 has the highest QT. For each QT decile at time t, 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 \$10.8 billion, while the highest QT stocks (stocks in

[14]The order of the different characteristics across QT portfolios remains unchanged, when we compute the cross-sectional characteristics for two equal subsamples, see Table IA.3.

QT decile 10) have an average capitalization of $0.8 billion. Column (7) shows that the average monthly trading volume decreases from $1.97 billion for the lowest QT stocks to $0.06 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 than 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 QT ratio with various firm characteristics in a multivariate 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 (e.g., Autoquote, Reg NMS, changes in the tick size), we use time fixed effects in column (3). Finally, the regression presented in column (4) includes both firm and time fixed effects, as both play an important role in our analysis. All non-binary variables have been standardized, demeaned and divided by the standard deviation.

We find that the QT ratio is higher for stocks that have low analyst coverage, low institutional ownership, 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 analyst coverage, institutional ownership, trading volume, and volatility) have higher quote-to-trade ratios.

This result may appear puzzling, because in neglected stocks one may expect a lower QT ratio, as market makers have less precise information based on which to change

[15] Table IA.4 in the Internet Appendix shows that the results in Table 2 remain robust to the addition of two control variables: (i) the number of registered NASDAQ market makers and (ii) several time dummy variables related to market wide events, previously subsumed by the time fixed effects.

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*), trading volume in mill. U.S. dollars (*USDVOL*), Amihud illiquidity ratio (*ILR*), relative bid-ask spread (*SPREAD*), and idiosyncratic volatility (*IVOLAT*) 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). All non-binary variables have been standardized, demeaned and divided by the standard deviation. Time fixed-effects are at the month-year level. Standard errors are double-clustered at the stock and month-year level.

	(1)	(2)	(3)	(4)
ANF	-0.05***	-0.04***	-0.01**	-0.02***
	(-5.07)	(-3.79)	(-2.55)	(-3.04)
INST	-0.06***	-0.11***	0.01**	-0.08***
	(-4.94)	(-7.84)	(2.06)	(-7.05)
BM	0.18***	0.17***	0.01	-0.01
	(2.93)	(2.73)	(0.36)	(-0.29)
*R*1	-0.01***	-0.01***	-0.01***	0.00
	(-4.44)	(-3.35)	(-2.66)	(-1.19)
MCAP	-0.02	0.00	0.03	0.00
	(-0.84)	(-0.12)	(1.45)	(-0.09)
PRC	0.07***	0.08***	0.02***	0.04***
	(5.83)	(5.99)	(3.01)	(4.69)
USDVOL	0.002	-0.02***	-0.01**	-0.02***
	(-0.63)	(-3.42)	(-2.37)	(-3.10)
ILR	0.02**	0.00	0.00	0.00
	(2.16)	(0.10)	(0.83)	(-0.21)
SPREAD	-0.06***	0.00	-0.04***	-0.02***
	(-8.97)	(-0.39)	(-6.34)	(-2.67)
IVOLAT	-0.04***	-0.02***	-0.01***	-0.01***
	(-6.02)	(-3.86)	(-5.81)	(-3.55)
Stock FE	NO	YES	NO	YES
Time FE	NO	NO	YES	YES
N	805,763	805,763	805,655	805,655
Adj. R^2	0.044	0.068	0.292	0.305

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 III.B.2).[16]

[16] One alternative view is that the QT ratio is driven by the exogenous arrival of public news. If a neglected stock is expected to have a low number of trades but a relatively large flow of news (e.g., coming from stock index changes), then we should also expect the stock to have a relatively high quote-to-trade ratio, which is consistent with SF1. As we do not incorporate exogenous news in our

It is common practice among academics, practitioners and regulators to associate QT with HFT activity (several examples are given in Footnote 9). Results in Tables 1 and 2 suggest that using QT as 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, Hagströmer, Nordén and Riordan, 2015). In addition, HFTs are more likely to trade in stocks with high institutional ownership, if HFT activity stems from their anticipation of agency and proprietary algorithms of institutional investors such as mutual and hedge funds (O'Hara, 2015). But SF1 shows that the QT ratio is actually lower in stocks with high institutional ownership. Thus, simply associating HFT activity with the QT ratio can be misleading.

C. Time Series of Quote-to-Trade Ratios

Figure 3 Graph A shows the time series of the equally weighted base-10 logarithm of the monthly QT ratio over the sample period. We note the substantial increase in QT during this time. Graph B is similar to Graph A, but displays separately the evolution of quotes and trades. Graph B shows that the increase in QT is driven by the explosion in quote updates. For exmaple, 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.[17]

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 III.B.2). Both the empirical fact and its explanation are consistent with previous literature. Hendershott et al. (2011) study a change of NYSE market structure in 2003 called "Autoquote" and argue that this change

theoretical model, we leave an exploration of this alternative view to future research.

[17] The positive relation between quotes and trades is established in Skjeltorp, Sojli and Tham (2018).

resulted in a decrease in monitoring costs among market participants, and especially among algorithmic traders. At the same time, they document an increase in their proxy for algorithmic trading, which is close in spirit to our QT ratio.[18] Angel, Harris and Spatt (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.

D. Quote-to-Trade Ratio and Stock Returns

In this section, we study the relation between QT ratios and average 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.

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 return in excess of the risk free rate 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 1.59% and QT10 has a return of 0.77% per month. The raw excess return of long-short portfolio based on QT is 0.83% 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 $t+1$.[19] We estimate individual portfolio

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

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

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_{Mkt}), value (r_{HML}), size (r_{SMB}), the additional Fama and French (2015) factors: profitability (r_{RMW}) and investment (r_{CMA}), Pástor and Stambaugh (2003) liquidity (r_{Liq}), momentum (r_{UMD}), and probability of informed trading (r_{PIN}). Table 3 reports alphas obtained from regression (2).[20] We present results from several asset pricing models that include several risk factors: CAPM (market), FF3 (market, size, value), FF3+PS (with the Pástor and Stambaugh (2003) traded liquidity factor), FF3+PS+MOM (with momentum), FF5 (with profitability and investment), and FF4+PS+PIN (with probability of informed trading).[21]

Columns (2)-(7) in Table 3 report alphas for the ten QT-sorted portfolios. We first focus on the full sample analysis in columns (2)-(6). The low-QT portfolio (QT1) has a monthly alpha (α_1) that ranges between 0.90% and 1.64% across various asset pricing models, which is statistically different from zero. The high-QT portfolio alphas range from -0.10% to 0.40%, but are statistically not different from zero in all specifications, except the CAPM. 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 different from zero and varies between 0.50% to 1.56% per month across the different asset pricing models. Table IA.5 in the Internet Appendix shows that the differences between the low and high QT portfolio alphas are not sensitive to the number of formed portfolios.

[20]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.

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

TABLE 3

Risk-Adjusted Returns for Quote-to-Trade Ratio Portfolios

The table shows monthly returns (in percentage points) 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, alphas. The alphas 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 (%)					
	r^e_{t+1}	CAPM	FF3	FF3+PS	FF3+PS +MOM	FF5	FF3+PS +MOM+PIN
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
α_1	1.59***	0.90*	1.05***	1.01***	1.64***	0.95***	1.66***
α_2	1.34***	0.91**	0.86***	0.85***	1.17***	0.59***	1.17***
α_3	1.21***	0.76*	0.67***	0.63***	0.97***	0.52***	0.98***
α_4	1.08***	0.62*	0.49***	0.50***	0.76***	0.34***	0.76***
α_5	1.02***	0.63*	0.47**	0.47**	0.67***	0.25***	0.67***
α_6	0.95***	0.30	0.08	0.10	0.32*	0.19**	0.31*
α_7	1.03***	0.58**	0.24	0.22	0.59***	0.21**	0.60***
α_8	0.93***	0.48***	0.06	0.04	0.32**	0.08	0.32**
α_9	0.95***	0.44	-0.01	-0.02	0.20	0.18**	0.18
α_{10}	0.77***	0.40***	-0.08	-0.10	0.08	0.01	0.06
α(QT1-QT10)	0.83***	0.50	1.13***	1.11***	1.56***	0.94***	1.60***

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

This result is puzzling when compared with the stylized fact SF1, which shows that the QT ratio is higher in neglected stocks, and in particular in less traded or more illiquid stocks. In the literature, less traded or illiquid stocks also tend to have higher expected returns, which appears to contradict the QT effect. To address these issues, in the next section we provide 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. The profitability of the long-short strategy derives mainly from the long position (the performance of the low-QT portfolio QT1) rather than from the short position (the performance of the high-QT portfolio QT10). Therefore, short-selling constraints are not likely to impede the implementation of a strategy that exploits the main regularity in Table 3.

2. Fama-MacBeth Regressions

To control for other predictive variables of 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) 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 time t. Then, we regress the risk-adjusted returns from equation (3) on lagged stock characteristics:

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

where $Z_{m,i,t-k}$ is the characteristic m for stock i at time $t-k$, and M is the total number of characteristics. We use $k = 1$ months for all characteristics.[22] The procedure ensures

[22]Table IA.6 in the Internet Appendix shows the estimation results where $k = 2$ for all conditioning

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

Stock Returns and Quote-to-Trade ratio

The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted monthly returns on firm characteristics. The dependent variable is the risk-adjusted return $r^a_{i,t} = r_{i,t} - \sum_{j=1}^{J} \beta_{i,j,t-1} F_{j,t}$, where the risk factors $F_{j,t}$ come from the 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), number of quotes ($QUOTE$), number of trades ($TRADE$), 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 ($IVOLAT$) 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), trading volume in mill. U.S. dollars ($USDVOL$), 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.	0.005***	0.004***	0.014***	0.012***	0.035***
$QT_{i,t-1}$	-0.172**	-0.199***	-0.251***	-0.261***	-0.156***
	(-2.23)	(-2.91)	(-3.61)	(-3.59)	(-3.85)
$SPREAD_{i,t-1}$		0.166***		0.074*	0.021
$ILR_{i,t-1}$			0.102***	0.081***	-0.008
$MCAP_{i,t-1}$					-0.136*
$BM_{i,t-1}$					0.032
$R1_{i,t-1}$					-3.350***
$R212_{i,t-1}$					0.084
$IVOLAT_{i,t-1}$					-9.525***
$USDVOL_{i,t-1}$					0.052
$PRC_{i,t-1}$					-0.345***
R^2	0.00	0.01	0.01	0.01	0.03
Time series (months)	278	278	278	278	278

Table 4 reports the Fama and MacBeth (1973) coefficients for cross-sectional regressions of individual stock risk-adjusted returns on stock characteristics. We con-

variables with the exception of the past return variables $R1$ and $R212$.

struct risk-adjusted returns using the Fama-French three-factor model (market, size, and value), with the momentum and 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.

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 with a t-statistic of -3.85, 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). The results are quantitatively unchanged when introducing other control variables, like short interest, institutional ownership or analyst following in Table IA.7 in the Internet Appendix. Furthermore, focusing on the NASDAQ-listed only subset or using excess returns, rather than risk-adjusted returns, does not alter the QT effect. Our results add to the literature that explores how trading activity and market structure are connected with asset returns (see Amihud and Mendelson, 1986; Amihud, 2002; Brennan and Subrahmanyam, 1996; Chordia, Roll and Subrahmanyam, 2002; Chordia et al., 2000; Easley, Hvidkjaer and O'Hara, 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. We explore this question in Table 5. Column (1) shows that when conditioning on quotes and trades as separate explanatory variables, it is

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 ($IVOLAT$) 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.	0.022***	0.021***	0.019***	0.020***	0.020***	0.021***
$QUOTE_{i,t-1}$	-0.320***	-0.337***	-0.290***	-0.313***	-0.130***	-0.154***
$TRADE_{i,t-1}$	0.175*	0.204**	0.235**	0.244**	-0.160	-0.134
$SPREAD_{i,t-1}$		0.023		0.011		-0.006
$ILR_{i,t-1}$			0.081**	0.064*		-0.003
$MCAP_{i,t-1}$					-0.159**	-0.143**
$BM_{i,t-1}$					0.023	0.026
$R1_{i,t-1}$					-3.496***	-3.418***
$R212_{i,t-1}$					0.069	0.073
$IVOLAT_{i,t-1}$					-7.348**	-8.466***
$USDVOL_{i,t-1}$					0.346***	0.322***
$PRC_{i,t-1}$					-0.573***	-0.534***
R^2	0.01	0.01	0.01	0.01	0.03	0.03
Time series (months)	278	278	278	278	278	278

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) does not affect the statistical significance of the number of trades and of the number of quotes. Column (6) includes all firm characteristics as well as liquidity measures as control variables and shows that the predictive power of QT derives from quotes and not from trades.

Stylized fact 4 (SF4): The QT predictability is driven by the number of quotes rather than the number of trades.

This result justifies our modeling choice in Section III to consider the trades as exogenous and focus instead on the quotes and how they result from the market makers' monitoring decisions.

III. Model of Quoting Activity

Our model is close in spirit to the dealer models of Ho and Stoll (1981) and Hendershott and Menkveld (2014). As we are interested in the relation between quoting activity and the cost of capital, we make two key modifications. First, the dealer learns about the value of the risky asset via costly monitoring, which generates endogenous quote changes. Second, the order flow is generated by risk averse investors, which cause a pricing discount in our model and thus generates a positive cost of capital.

A. 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 numeraire and has a zero return. After trading takes place, the risky asset liquidates at a value v per share called the fundamental value or asset value. The random variable v has a normal distribution $v \sim \mathcal{N}(v_0, \sigma_v^2)$, where σ_v is the fundamental volatility.

Trading. Trading occurs 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:

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

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 traders' order flow. The parameter k is the investor elasticity, ℓ is the inelasticity parameter, and m is the imbalance parameter. Hendershott and Menkveld (2014) use a similar reduced form approach, except that they exogenously set the imbalance parameter m to zero. We endogenize the value of m and other parameters by providing micro-foundations for the order flow, and we find that $m > 0$, when investors are risk averse and the asset is in positive net supply.

Order Flow Micro-foundations. To get more intuition for the equations in (5), we provide micro-foundations for the order flow. (For more details, see Section 2 in the Internet Appendix.) First, we assume that the risky asset has a positive net supply $M > 0$. There are two types of traders: (i) liquidity traders, who submit inelastic aggregate buy order L^b and aggregate sell order L^s, where both L^b and L^s have IID normal distributions $\mathcal{N}(\ell_L, \Sigma_L/2)$; and (ii) informed investors with CARA utility and coefficient of risk aversion $A > 0$. A mass one of informed investors starts with an initial endowment in the risky asset that is normally distributed as $\mathcal{N}(M, \sigma_M^2)$. To simplify notation, we redefine the asset value to be $v+u$, with u normally distributed as $\mathcal{N}(0, \Sigma_u)$. Investors observe the same signal v before trading, and then trade on the exchange at the existing quotes. As a result, we show in Section 2 in the Internet Appendix that the aggregate order flow approaches the form in (5) when the endowment volatility σ_M is large.[23] Moreover, the investor elasticity k is proportional to the informed investors' signal precision $1/\Sigma_u$ and their risk tolerance $1/A$, while the imbalance parameter m is proportional to the net supply M.

Dealer Monitoring. The dealer monitors the market according to an independent

[23]In Section 3 in the Internet Appendix, we show that the equilibrium is qualitatively similar if instead of aggregating the order flow over the whole population, we consider only the optimal orders from one individual trader selected at random from the population.

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, 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]

[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 $C(q) = 0$.

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.

B. Equilibrium Quoting

We solve for the equilibrium in two steps. In the first step (Section III.B.1), we take the dealer's monitoring rate q as given and describe the optimal quoting behavior. In the second step (Section III.B.2), we determine the optimal monitoring rate q as the rate which maximizes the dealer's expected utility.

1. Optimal Quotes

We start by fixing the monitoring rate q. Consider the game described in Section III.A, with positive parameters $k, \ell, m, \Sigma_L, f, c, \gamma$. Also, let x_0 be the dealer's initial inventory. Define the following constants:

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

Proposition 1 describes the optimal quoting strategy of the dealer, which is conditional on the dealer's value forecast w_t. In Section III.B.2, we describe the process for w_t, which is exogenous to the dealer once the initial 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:*

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

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. Furthermore, 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 spread, then investors (whose price elasticity is increasing in k) submit a smaller expected quantity at the quotes.[25] If she sets too small a 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.

2. Optimal Monitoring and the Quote Rate

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 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 $1/\mathsf{Var}(v - w_t) = (n+1)F(q)$.*

[25] For example, 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.

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 forecast changes. Therefore, we interpret the monitoring rate q as the dealer's quote rate:

$$q = \text{Quote Rate}. \tag{12}$$

Thus far, the description of the equilibrium does not depend on a particular specification for the precision function $F(q)$ or the monitoring function $C(q)$.[26] Proposition 2, however, provides explicit formulas by assuming that:

$$F(q) = f\ln(q+1), \qquad C(q) = cq, \tag{13}$$

where $f > 0$ is a signal precision parameter and $c > 0$ is a monitoring cost parameter.[27]

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

$$q^2 = \frac{k(1+k\gamma)}{fc}. \tag{14}$$

Corollary 1 uses the formula in (14) to generate some comparative statics for the quote rate.

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

If investor elasticity k is larger, investors trade more aggressively on the pricing error, 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 (14) as a sum: $q^2 = \frac{k}{fc} + \frac{k^2\gamma}{fc}$. The first term (which does not depend on the

[26] In Section 4.1 in the Internet Appendix, we show that the equilibrium remains essentially the same if we replace the monitoring process at rate q by a unique signal with precision $\tilde{F}(q) = \frac{qF(q)}{\ln(q+1)}$.

[27] 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.

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.

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

If the monitoring cost c is smaller, the dealer can afford to monitor more often in order to maintain the same precision, which increases the quote rate. There is much evidence that monitoring costs have decreased dramatically in recent times (see Hendershott et al., 2011). Therefore, according to Corollary 1, we should also expect a large increase in the QT ratio. This is consistent with our stylized fact SF2 in Section II.C that documents a sharp rise in the QT ratio.

C. 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: $\frac{\mathsf{E}_t(v)-p_t}{p_t} = \frac{w_t-p_t}{p_t} = \frac{\delta}{w_t-\delta}$. Therefore, the expected return is in one-to-one correspondence with the discount. We thus define the "cost of capital" r to be equal to the discount:[28]

$$r = \delta = \frac{m}{k}\,\frac{1+2\gamma k}{1+\gamma k} + \frac{\gamma}{1+\gamma k}\,x_0. \tag{15}$$

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 $x_0 \geq 0$. Corollary 2 provides some comparative statics for the cost of capital.

Corollary 2. *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.

Corollary 3 connects the cost of capital to the equilibrium quote rate.

Corollary 3 (Quote 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 quote rate.*

The quote effect in our model is driven by investor elasticity. When k is larger, Corollary 1 shows that the quote rate 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.

If we consider also the micro-foundations for the order flow (see Section III.A), the investor elasticity k is larger when the investors have more precise information. Therefore, at a more fundamental level the quote effect is driven by traders' information

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

precision: more precise investors cause both a larger quote rate and a smaller cost of capital.

The quote effect is documented empirically in the cross-section of stock returns by the stylized fact SF3 (see Section II.D).[29]

D. Neutral State

In this section, we describe the equilibrium when the dealer's initial inventory x_0 has a particular value:

$$x_{0,\text{neutral}} = \frac{m}{\gamma k}. \tag{16}$$

We call this value the dealer's neutral inventory (or preferred inventory), and we say that in this case the sytem is in its "neutral state."[30]

Corollary 4 shows that in the neutral state the dealer expects her inventory to stay the same, that is, the expected change in her inventory is zero.

Corollary 4. *When the dealer's inventory is equal to its neutral value, the expected buy and sell quantities from equation (5) are equal. The equilibrium cost of capital (discount) is:*

$$\delta_{\text{neutral}} = \frac{2m}{k}. \tag{17}$$

The first statement of Corollary 4, that the traders' order flow is balanced in the neutral state, is in fact the reason behind our definition of neutral inventory in (16). 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

[29]One concern remains whether our theoretical explanation is systematic enough to justify the QT effect: it is possible that the pricing discount in any particular stock is just an idiosyncratic error that should vanish in the large cross section of stocks. We argue, however, that the price discount is likely to be driven by systematic variables. Indeed, Corollary 2 shows that a key determinant of the price discount is the investor elasticity k, which our microfoundations show to be determined by the precision of the informed investors' signals. If investors analyze multiple stocks, then their signal precision is likely to comove across stocks, which makes the investor elasticity systematic.

[30]In Section 5 in the Internet Appendix, we extend our model to multiple trading rounds, and we show that the neutral inventory is equal to the long-term average of the dealer's inventory, regardless of its initial value.

inventory is positive because the investors are risk averse and the risky asset is in positive net supply (see the order flow micro-foundations described in Section III.A). 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.

Formally, equations (16) and (17) 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.

The dealer's preferred inventory is decreasing in γ, as the dealer prefers to hold less of the risky asset when she is more inventory averse. The preferred inventory is increasing in the imbalance parameter m, as the dealer can engage in more risk sharing when the risky asset is in higher supply. The preferred inventory is decreasing in the investor elasticity k, as more aggressive investors hold relatively more of the risky asset and decrease the share left to the dealer.

A surprising consequence of Corollary 4 is that the discount (or cost of capital) in the neutral state is independent of the dealer's inventory aversion γ. One may indeed 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. In the multi-trade model in Section 5 of the Internet Appendix, we see that the dealer's 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 that would not be optimal.

E. Additional Predictions

In this section, we provide two additional predictions of our model. Corollary 1 implies that the dealer's optimal monitoring rate q is increasing in her inventory aversion γ. As proxy for the inventory aversion γ of a dealer in a stock we use $1/N$, where N

is the number of market makers that provide liquidity in that stock. We expect that a larger number of intermediaries implies a smaller γ for the representative dealer. 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.

Intuitively, 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 the stock less often, as she is less concerned about accumulating inventory. Therefore, the resulting QT ratio is also smaller.

In Section 4 in the Internet Appendix, we provide an extension of the model to N dealers and 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.4). 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.4 in the Internet Appendix. This augments column (4) of Table 2 with the number of registered market makers in a particular stock (*MM*) as an explanatory variable. This results in a smaller sample, because the number of market makers is only available for NASDAQ-traded stocks. Nevertheless, we find that the number of market makers has a negative effect on the QT ratio. This is surprising, because one may think that competition among market makers results in an increase of the QT ratio.

Corollary 4 implies that in the neutral state, when the traders' order flow is balanced, the dealer's discount (cost of capital) 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 2: The number of market makers in a stock has no relation to

the stock's expected return.

Intuitively, when the dealer's initial inventory is in the neutral state (where the expected imbalance between buy and sell quantities is zero) the dealer wants only to balance the incoming order flow, and hence the pricing discount (or cost of capital) is affected only by the properties of the order flow and not by the characteristics of the dealer, including her inventory aversion (or the number of dealers if we consider the case of multiple dealers).[31]

We test this prediction in Table IA.8 in the Internet Appendix. The table presents the results of Fama-MacBeth regressions similar to those in Table 4, but we include the number of registered market makers (MM) in a stock as an explanatory variable. Indeed, the introduction of the MM variable does not affect the QT effect. Moreover, all other explanatory variables have qualitatively similar magnitudes and levels of significance as in Table 4.

IV. Alternative Explanations of the QT Effect

In this section, we discuss several alternative explanations of the QT effect that involve various frictions, e.g., the tick size and impediments to arbitrage, institutional investors and governance, temporary price effects, and market structure changes.

A. Tick Size and Impediments to Arbitrage

Yao and Ye (2018) and Albuquerque, Song and Yao (2019) provide a connection between an impediments-to-arbitrage hypothesis and the QT effect: First, Yao and Ye (2018) find that stocks with a larger tick size relative to price have higher QT ratios. Second, Albuquerque et al. (2019) find that an exogenous increase in tick size negatively affects stock prices and is associated with an increase in the cost of trading. Therefore, the QT effect might be driven in part by the effect of illiquidity related to the tick size.

[31] This result depends on the system being initially in the neutral state, which we argue is a plausible assumption. Indeed, in an extension with multiple trading rounds, we show that the neutral inventory is equal to the long-term average of the dealer's inventory, regardless of its initial value (see Section 5 in the Internet Appendix).

To investigate this alternative explanation, we follow two approaches: (i) long-short portfolios double sorted by QT and transaction cost proxy variables, and (ii) Fama and MacBeth (1973) regressions for high and low transaction cost levels.

To form the long-short portfolio alphas, we first sort stocks into terciles based on transaction cost proxies (relative spread, Amihud illiquidity ratio, and turnover), and then create either three or five QT portfolios within each liquidity tercile. We examine the alpha from a strategy that, within each liquidity tercile, goes long in low-QT stocks and short in high-QT stocks. Then, if the illiquidity level explains the QT effect according to the impediments-to-arbitrage hypothesis, abnormal profits should concentrate in the most illiquid group and the size of the risk-adjusted returns should be of the same magnitude as the cost of transacting in the U.S. equity market. Moreover, abnormal returns for the other liquidity groups should not be statistically different from zero.

The results in Panels A and B of Table 6 provide support for the impediments-to-arbitrage hypothesis. The strongest QT effect occurs among the most illiquid stocks for relative spread and ILR. The abnormal return for the most illiquid group is as high as over 100 basis points and as low as 17 basis points for the most liquid group across the illiquidity proxies. The difference in abnormal returns across these illiquidity groups implies a difference in average transaction costs (impediment to trade) of 60 to 90 basis points between the illiquid and liquid group of stocks. However, the results that the abnormal returns for other liquidity groups remain statistically and economically different from zero suggest that impediments-to-arbitrage explain a proportion of the estimated economic effect of the inventory mechanism, but not all. The results on turnover (see Panel C of Table 6) show that the QT effect is quite strong across all turnover levels.[32]

In Table 7, we extend the univariate analysis to a multivariate analysis for stocks divided into two groups, depending on the median liquidity level in each month: one group with low liquidity and one group with high liquidity. Then, we re-estimate the Fama-MacBeth regressions as in Table 4. Consistent with the impediments-to-arbitrage hypothesis, Table 7 shows that the QT effect diminishes when liquidity is high, but

[32]Results remain quantitatively similar when using other proxies for transactions costs, like quoted spread, number of trades, or dollar traded volume. We do not report these results for brevity, but they are available from the authors upon request.

TABLE 6

Risk-Adjusted Returns for Quote-to-Trade Ratio Double Sorted Portfolios

The table shows monthly alphas of a long-short strategy for various portfolios double sorted on a variable of interest and the quote-to-trade ratio (QT). The strategy longs low QT ratio stocks and shorts high QT ratio stocks within three levels of liquidity and institutional investors at the end of month t. We first assign all stocks to three portfolios based on their liquidity and institutional ownership levels. Then, we construct three or five portfolios based on the level of QT ratio within each liquidity and institutional ownership portfolio, and long the low QT ratio portfolio and short the high QT ratio portfolio. The alphas 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 a five factor asset pricing model including the Fama and French (1993) model (FF3) with the added Pástor and Stambaugh (2003) traded liquidity factor and momentum factor. All portfolio returns are equally weighted. Panels A to C report the long-short α for the liquidity portfolios with low, medium and high relative bid/ask spread, ILR, and turnover, respectively. Panel D reports the long-short α for the institutional ownership level. ***, **, 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.

	3x3 Portfolios	3x5 Portfolios
Panel A: Relative Spread		
Low	0.27*	0.37**
	1.68	1.92
Medium	0.48***	0.63***
	2.77	2.96
High	0.89***	1.19***
	6.02	6.24
Panel B: ILR		
Low	0.17	0.23
	1.08	1.31
Medium	0.57***	0.78***
	3.50	3.79
High	0.83***	1.15***
	6.41	6.92
Panel C: Turnover		
Low	0.64***	0.82***
	6.02	6.23
Medium	0.55***	0.67***
	4.09	4.17
High	0.59***	0.86***
	3.53	4.02
Panel D: Institutional Investors		
Low	0.50***	0.72***
	3.50	3.99
Med	0.51***	0.65***
	3.52	3.67
High	0.36***	0.45***
	2.57	2.79

TABLE 7

Stock Returns and Quote-to-Trade Ratio Across Subgroups

The table reports the Fama and MacBeth (1973) coefficients from regressions of risk-adjusted monthly returns on firm characteristics for two levels of explanatory variables. 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 sample is divided into two groups of high and low relative spread, ILR, turnover, and institutional ownership based on sample median values for every time-period. Panel A presents the results for low (below median) levels, and Panel B presents the results for high (above median) levels of the variables of interest. We exclude *SPREAD* from the regressions in columns (1) and (2) and *ILR* for columns (3) and (4). 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 ($IVOLAT$) 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), trading volume in mill. U.S. dollars ($USDVOL$), 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.

	Relative Spread		ILR		Turnover		IO	
Panel A. Low								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Const.	0.033***	0.035***	0.032***	0.031***	0.040***	0.037***	0.043***	0.036***
$QT_{i,t-1}$	-0.194***	-0.191***	-0.179***	-0.189***	-0.159***	-0.161***	-0.184***	-0.193***
	(-2.65)	(-2.68)	(-2.72)	(-2.67)	(-4.17)	(-3.88)	(-4.40)	(-4.17)
$ILR_{i,t-1}$		-0.027				0.000		0.046*
$SPREAD_{i,t-1}$				-0.077		0.084***		0.025
$MCAP_{i,t-1}$	0.016	0.014	-0.005	0.001	-0.250***	-0.275***	-0.212***	-0.203***
$BM_{i,t-1}$	-0.012	-0.010	0.016	0.021	0.083	0.073	0.138	0.137
$R1_{i,t-1}$	-2.960***	-2.888***	-2.944**	-2.923**	-5.661***	-5.465***	-3.684***	-3.528***
$R212_{i,t-1}$	0.011	0.012	-0.040	-0.030	0.453**	0.465**	0.137	0.149
$IVOLAT_{i,t-1}$	-3.641	-2.937	-8.901*	-10.302*	-8.312*	-15.080***	-10.027**	-13.567***
$USDVOL_{i,t-1}$	-0.149	-0.173	-0.106*	-0.106	0.149**	0.193***	0.125***	0.172*
$PRC_{i,t-1}$	0.004	-0.003	-0.088*	-0.073	-0.262***	-0.250***	-0.461***	-0.429***
R^2	0.04	0.04	0.04	0.04	0.03	0.04	0.04	0.04
Time series (months)	278	278	278	278	278	278	278	278
Panel B. High								
Const.	0.066***	0.064***	0.060***	0.061***	0.038***	0.030***	0.029***	0.036***
$QT_{i,t-1}$	-0.198***	-0.202***	-0.218***	-0.227***	-0.192**	-0.220***	-0.160**	-0.176**
	-4.841	-4.839	-4.153	-4.056	-2.213	-2.873	-2.333	-2.525
$ILR_{i,t-1}$		0.023				-0.013		-0.062**
$SPREAD_{i,t-1}$				-0.014		0.313		-0.132
$MCAP_{i,t-1}$	-0.356***	-0.352***	-0.349***	-0.350***	0.178	0.161	-0.059	-0.044
$BM_{i,t-1}$	0.077	0.076	0.043	0.041	-0.017	-0.022	-0.053	-0.047
$R1_{i,t-1}$	-3.810***	-3.680***	-3.759***	-3.751***	-2.622***	-2.621***	-3.176***	-3.129***
$R212_{i,t-1}$	0.162	0.164*	0.185	0.195*	-0.060	-0.057	-0.029	-0.028
$IVOLAT_{i,t-1}$	-10.604***	-11.289***	-8.036**	-8.978*	-0.460*	-3.954	-2.342	-0.839
$USDVOL_{i,t-1}$	0.143***	0.160*	0.167***	0.156*	-0.297**	-0.254	-0.027	-0.109
$PRC_{i,t-1}$	-0.456***	-0.452***	-0.418***	-0.386***	-0.336*	-0.299	-0.178	-0.167
R^2	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04
Time series (months)	278	278	278	278	278	278	278	278

nevertheless it remains statistically and economically significant for both subgroups.

Next, we perform a back-of-the-envelope calculation to estimate how large a transaction cost is needed to eliminate the QT alpha from Table 6. Consider the Amihud illiquidity ratio, which is a price impact measure computed as a stock's absolute daily return divided by its daily trading volume. For stocks that exhibit the strongest QT effect, the average value of this ratio is 0.17, i.e., a \$1M trade triggers a 0.17% price impact. Given a five-factor alpha of 0.83% in this group, it would take a trade of 0.83%/0.17% = \$4.88 million to eliminate the profit over a single day. This is a substantial amount for the equity markets, where the average trade size is about \$1 million. Moreover, Panel C of Table 6 shows that the QT effect is strong across stocks with different levels of trading turnover. Our results, therefore, suggest that the impediments-to-arbitrage hypothesis provides only a partial explanation for the QT effect.

B. Institutional Investors and Governance

Institutional investors as equity holders have become more active participants in the governance of modern corporations, whether by "voice" or by exit threat (see, e.g., Admati, Pfleiderer and Zechner, 1994; Maug, 1998; Gillan and Starks, 2000). Good corporate governance in general increases firm value and reduces agency costs, leading to lower costs of capital (see, e.g., Becht, Franks, Mayer and Rossi, 2008; Brav, Jiang, Partnoy and Thomas, 2008).

Albuquerque et al. (2019) find that long-term institutional investors tend to hold stocks with larger relative tick size and higher QT ratio. This suggests an alternative explanation of the QT effect where the QT ratio simply captures the proportion of institutional investors and level of corporate governance, which has an inverse relation with the cost of capital.

Thus, the QT effect should be stronger among stocks with lower institutional holdings and insignificant for stocks with higher institutional ownership. We test this hypothesis using both univariate and multivariate approaches. We construct institutional holdings of equity in firm i at the end of the year using 13F files. For the univariate analysis, we first sort stocks by institutional holdings and examine the alphas from a strategy that, within each institutional ownership tercile, goes long in low-QT stocks and short in high-QT stocks.

In Panel D of Table 6, we report the results for portfolios double sorted on QT and institutional ownership. The results provide partial support for the institutional investors hypothesis. We find the strongest QT effect among stocks with lowest institutional ownership. However, alphas for all institutional ownership terciles are statistically significant. The abnormal return for stocks with the smallest institutional ownership is between 47 and 75 basis points, and between 36 and 45 basis points for the group with the most institutional investors. The difference in abnormal returns across these groups is consistent with the hypothesis that institutional investors play a role in reducing the cost of equity. Nevertheless, the abnormal returns across all institutional ownership groups remain statistically and economically significant. Thus, we find that institutional ownership dilutes but does not subsume the QT effect.

In columns (7) and (8) of Table 7, we show the estimated QT effect using Fama-MacBeth regressions across sub-samples with high and low institutional ownership. Consistent with the univariate analysis, we find that the QT effect is partly mitigated by institutional ownership, decreasing by 20pbs between the low and high institutional ownership samples. However, the QT effect remains statistically and economically significant. In additional analysis, Column (2) of Table IA.7 in the Internet Appendix analyzes the QT effect, while directly controlling for institutional ownership. Column (2) of Table IA.7 shows that the magnitude of the QT effect (in column (5) of Table 4) decreases with the inclusion of the institutional ownership variable by 0.013%, but QT remains still highly statistically and economically significant.

C. Return Continuations and Reversals

In Section II.D.1 we consider only one-month holding (portfolio rebalancing) periods. One could 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 IA.1 in the Internet Appendix 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.[33] The figure shows that the QT effect is very persistent. The one month formation and holding period portfolio has the highest alpha of 1.00%. Overall, the long/short alphas after a year of both formation and holding periods are 0.50% per month and highly statistically significant.

D. Algorithmic Trading and Reg NMS

The emergence of algorithmic trading and the introduction of Reg NMS are two major events during the sample period that are likely to have important effects on the U.S. equity market structure.[34] Thus, to investigate whether our QT effect is robust to these market structure changes, we verify the QT effect for the different subsamples defined by these two events.

We first consider the emergence of algorithmic trading. As Hendershott et al. (2011)

[33]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.

[34]Table IA.4 in the Internet Appendix shows that the QT ratio increases significantly after both of these events.

document the proliferation of algorithmic and electronic trading only after 2003, one might think that the QT effect should hold differently in the subsample June 1994 to December 2002, compared to the subsample January 2003 to December 2017. Indeed, as the QT ratio is often used as a proxy for algorithmic trading and HFT (see Footnote 9), one may argue that the QT effect is driven by the effect of algorithmic trading on the cost of capital. However, in the earlier subsample the QT ratio is less likely to be related to algorithmic trading. Thus, if we find that the QT effect holds similarly in both subsamples, algorithmic trading is less likely to be an explanation for the QT effect.

Table IA.9 in the Internet Appendix shows the results of Fama-MacBeth regressions similar to those in Table 4 but performed over the two subsamples (June 1994 to December 2002 and January 2003 to December 2017). The effect of the QT ratio on risk-adjusted returns is large and statistically significant in the pre- and post-2003 period, despite the reduction in power due to the lower number of time-series observations.

Additional evidence is shown in column (7) in Table 3. In this column, we report the alphas of ten portfolios sorted on QT, where alpha is computed according to the standard risk factors plus a PIN factor. As the PIN factor is available only until 2002, we are in effect computing the alphas during only the first part of our sample. We find that during this pre-algorithmic period, the effect of the QT ratio on risk-adjusted returns is strong and even larger than for the other columns in Table 3, which are computed using the whole sample.

Next, we investigate the introduction of Reg NMS, which transformed the market landscape by introducing more competition and led to unprecedented market fragmentation. Table IA.9 in the Internet Appendix shows the results of Fama-MacBeth regressions similar to those in Table 4 but performed over two different subsamples: June 1994 to December 2006 and January 2007 to December 2017. The effect of QT on risk-adjusted returns is large and statistically significant in the pre- and post-2007 period.

In conclusion, Sections IV.C and IV.D show that the QT effect holds at longer predictability horizons and is persistent throughout the sample.

V. Conclusion

This paper studies the quoting activity of market makers, and how the resulting quote-to-trade 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 analyst coverage, institutional ownership, trading volume, and volatility. Our main finding, 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 since 2003), the QT effect is qualitatively unchanged across sample periods. 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.

As the quoting activity of market makers is clearly an important determinant of the QT ratio, we propose a model that incorporates (i) the quoting activity that comes from the market makers' monitoring of the market, and (ii) the cost of capital that comes from risk averse investors. The model is consistent with our stylized empirical findings and produces additional predictions that are borne out in the data: a larger number of market makers lowers the QT ratio, but has no effect on expected returns. In our model the QT effect is driven by investors' aggressiveness: e.g., when investors are more precisely informed, market makers monitor faster and thus increase the QT ratio, but at the same time reduce mispricing and lower expected returns. Although we rule out several likely alternative interpretations, we acknowledge that there could be other, non-mutually exclusive explanations for the surprising association between the quote-to-trade ratio and cost of capital.

Appendix A. Proofs

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 $G_\tau = \mathsf{Var}(v - w_\tau)$ the variance of the forecast error. To simplify notation, in the remainder of this proof we omit the subscript τ for the forecast w_τ, etc.

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

$$\text{(A-1)} \qquad h \;=\; \frac{a-b}{2}, \qquad \delta \;=\; w - \frac{a+b}{2}, \qquad e \;=\; v - w,$$

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:

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

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 (A-1), the dealer's maximization problem is equivalent to:

$$\text{(A-3)} \;\; \max_{h,\delta} \Big(x_0 w - kG - k\delta^2 - kh^2 + 2\ell h + 2m\delta - \gamma\big(x_0 - k\delta + 2m\big)^2 - \gamma\big(k^2 G + \Sigma_L\big)\Big).$$

The first order condition in (A-3) with respect to h implies $h = \frac{\ell}{k}$, which shows that the optimal half spread satisfies (9). The first order condition in (A-3) 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 (ignoring monitoring costs) is:

$$\text{(A-4)} \qquad U_{\max} \;=\; x_0 w + \frac{\ell^2}{k} - k(1+k\gamma)G - \gamma\Sigma_L + \frac{m^2 - 2\gamma k m x_0 - \gamma k x_0^2}{k(1+k\gamma)}.$$

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. □

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:

$$G(q) \;=\; \mathsf{Var}\big(v - w_n\big) \;=\; \sum_{n=0}^{\infty} \frac{q^n}{(q+1)^{n+1}} \, \frac{1}{(n+1)F(q)} \;=\; \frac{\ln(q+1)}{qF(q)}, \tag{A-5}$$

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 (A-4) 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 first order condition for q is:

$$-\frac{C'(q)}{G'(q)} \;=\; k(k\gamma+1). \tag{A-6}$$

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 q. Since $G(q) = \frac{1}{fq}$, equation (A-6) implies that q satisfies $fcq^2 = k(k\gamma+1)$, which proves the first part of equation (14). 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 (14), it is clear that q is increasing in k and γ, and decreasing in f and c. $\square$

Proof of Corollary 2. Equation (15) 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 $x_0 \geq 0$. $\square$

Proof of Corollary 3. Suppose we hold all parameters constant except for k. According to Corollary 2, the discount δ is decreasing in k. At the same time, the quote rate q is increasing in k (see Corollary 1). This proves the inverse relation between δ and q. $\square$

Proof of Corollary 4. Equation (17) follows by simply substituting (16) in (15) 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 $Q^b - Q^s = k\left(v - \frac{a+b}{2}\right) - 2m + \varepsilon^b - \varepsilon^s$. 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. $\square$

References

Admati, A. R.; P. Pfleiderer; and J. Zechner. "Large Shareholder Activism, Risk Sharing, and Financial Market Equilibrium." *Journal of Political Economy*, 102 (1994), 1097–1130.

Albuquerque, R.; S. Song; and C. Yao. "The Price Effects of Liquidity Shocks: A Study of SEC's Tick-Size Experiment." *Journal of Financial Economics*, forthcoming (2019).

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

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

Amihud, Y.; H. Mendelson; and L. H. Pedersen. "Liquidity and Asset Prices." *Foundations and Trends in Finance*, 1 (2005), 1–96.

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

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

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

Asparouhova, E.; H. Bessembinder; and I. Kalcheva. "Noisy Prices and Inference Regarding Returns." *Journal of Finance*, 68 (2013), 665–714.

Becht, M.; J. Franks; C. Mayer; and S. Rossi. "Returns to Shareholder Activism: Evidence from a Clinical Study of the Hermes U.K. Focus Fund." *Review of Financial Studies*, 22 (2008), 3093–3129.

Boehmer, E.; K. Y. L. Fong; and J. J. Wu. "Algorithmic Trading and Market Quality: International Evidence."Working paper, Singapore Management University (2018).

Brav, A.; W. Jiang; F. Partnoy; and R. Thomas. "Hedge Fund Activism, Corporate Governance, and Firm Performance." *Journal of Finance*, 63 (2008), 1729–1775.

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

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

Brogaard, J.; B. Hagströmer; L. Nordén; and R. Riordan. "Trading Fast and Slow: Colocation and Liquidity." *Review of Financial Studies*, 28 (2015), 3407–3443.

Brogaard, J.; T. Hendershott; and R. Riordan. "High Frequency Trading and Price Discovery." *Review of Financial Studies*, 28 (2014), 3407–3443.

Brogaard, J.; T. Hendershott; and R. Riordan. "High Frequency Trading and the 2008 Short Sale Ban." *Journal of Financial Economics*, 124 (2017), 22–42.

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

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

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

Diamond, D. W., and R. E. Verrecchia. "Disclosure, Liquidity, and the Cost of Capital." *Journal of Finance*, 46 (1991), 1325–1359.

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

Easley, D.; S. Hvidkjaer; and M. O'Hara. "Is Information Risk a Determinant of Asset Returns?" *Journal of Finance*, 57 (2002), 2185–2221.

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

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

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

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

Foucault, T.; R. Kozhan; and W. W. Tham. "Toxic Arbitrage." *Review of Financial Studies*, 30 (2016), 1053–1094.

Foucault, T.; A. Röell; and P. Sandas. "Market Making with Costly Monitoring: An Analysis of the SOES Controversy." *Review of Financial Studies*, 16 (2003), 345–384.

Gillan, S. L., and L. T. Starks. "Corporate governance proposals and shareholder activism: The role of institutional investors." *Journal of Financial Economics*, 57 (2000), 275–305.

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

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

Hasbrouck, J. "Trading Costs and Returns for U.S. Equities: Estimating Effective Costs from Daily Data." *Journal of Finance*, 64 (2009), 1445–1477.

Hendershott, T.; C. M. Jones; and A. J. Menkveld. "Does Algorithmic Trading Improve Liquidity." *Journal of Finance*, 66 (2011), 1–33.

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

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

Hong, H.; T. Lim; and J. C. Stein. "Bad News Travels Slowly: Size, Analyst Coverage, and the Profitability of Momentum Strategies." *Journal of Finance*, 55 (2000), 265–295.

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

Jegadeesh, N., and S. Titman. "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency." *Journal of Finance*, 48 (1993), 65–91.

Kumar, A. "Who Gambles in the Stock Market?" *Journal of Finance*, 64 (2009), 1889–1933.

Maug, E. "Large Shareholders as Monitors: Is There a Trade-Off between Liquidity and Control?" *Journal of Finance*, 53 (1998), 65–98.

Menkveld, A. "The Economics of High-Frequency Trading: Taking Stock." *Annual Review of Financial Economics*, 8 (2016), 1–24.

O'Hara, M. *Market Microstructure Theory*, Malden, MA: Blackwell Publishers (1995).

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

Pástor, L., and R. F. Stambaugh. "Liquidity Risk and Expected Stock Returns." *Journal of Political Economy*, 111 (2003), 642–685.

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

Skjeltorp, J. A.; E. Sojli; and W. W. Tham. "Cross-Sided Liquidity Externalities." *Management Science*, 64 (2018), 2901–2929.

Yao, C., and M. Ye. "Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls." *Review of Financial Studies*, 31 (2018), 2157–2183.