Journal of Financial Economics 91 (2009) 119–138

Contents lists available at ScienceDirect

Journal of Financial Economics

journal homepage: www.elsevier.com/locate/jfec

# Why is *PIN* priced?☆

Jefferson Duarte, Lance Young*

*Foster School of Business, University of Washington, Box 353200, Seattle, WA 98195, USA*

ARTICLE INFO

*Article history:*
Received 18 May 2007
Received in revised form
31 August 2007
Accepted 15 October 2007
Available online 11 November 2008

*JEL classification:*
G12
G14

*Keywords:*
Liquidity
Information asymmetry

ABSTRACT

Recent empirical work suggests that a proxy for the probability of informed trading (*PIN*) is an important determinant of the cross-section of average returns. This paper examines whether *PIN* is priced because of information asymmetry or because of other liquidity effects that are unrelated to information asymmetry. Our starting point is a model that decomposes *PIN* into two components, one related to asymmetric information and one related to illiquidity. In a two-pass Fama-MacBeth [1973. Risk, return, and equilibrium: empirical tests. Journal of Political Economy 81, 607–636] regression, we show that the *PIN* component related to asymmetric information is not priced, while the *PIN* component related to illiquidity is priced. We conclude, therefore, that liquidity effects unrelated to information asymmetry explain the relation between *PIN* and the cross-section of expected returns.



## 1. Introduction

Market microstructure's impact on asset prices has captured substantial attention in the finance literature in recent years. Easley and O'Hara (2004) argue that stocks with more information asymmetry have higher expected returns. They construct a rational expectations asset pricing model with asymmetric information and find that everything else held constant, uninformed investors demand a premium to hold shares in firms with higher information asymmetry. In this model, the effects of information asymmetry are undiversifiable since the uninformed expect to lose to the informed and therefore demand to be compensated for this expected loss. On the other hand, Hughes, Liu, and Liu (2007) and Lambert, Leuz, and Verrecchia (2005) show that in a large economy the effect of asymmetric information on expected returns is diversifiable. They argue that asymmetric information is priced in Easley and O'Hara (2004) because the number of assets in their model is finite and hence asymmetric information risk cannot be diversified away. In spite of the fact that private information should be diversifiable in a large economy, empirically a proxy for information asymmetry, *PIN* is positively and significantly related to average stock returns. Specifically, Easley, Hvidkjaer, and O'Hara (2002) show that a 10% difference in the *PINs* of two stocks results in a 250 basis point difference in their annual expected returns. Therefore, the empirical evidence does not agree with theories that predict that asymmetric information is diversifiable and thus raises the question: why is *PIN* priced? Our results indicate that *PIN* is priced because it is a proxy for illiquidity unrelated to asymmetric information. We arrive at this conclusion in three steps.

First, we examine the Easley, Kiefer, O'Hara, and Paperman (1996) structural microstructure model (the

☆ The authors thank Yakov Amihud, Hank Bessembinder, Kathy Dewenter, Jaehoon Hahn, Xi Han, Tyler Henry, Avi Kamara, Jon Karpoff, Jennifer Koski, Jun Liu, Paul Malatesta, Ed Rice, Jay Shanken, Masahiro Watanabe, Fan Yu, and seminar participants at the University of Washington, 2007 Pacific Northwest Finance Conference, Brigham Young University, Rice University, University of Illinois at Urbana-Champaign, University of Miami, University of Virginia, and the University of Wisconsin-Madison for their helpful comments. Special thanks go to Lew Thorson and Tim Yao for their invaluable computing expertise and assistance. Duarte acknowledges financial support from the 2006 CFO Forum Summer Fellowship. Any remaining errors are our own.

* Corresponding author. Fax: + 1 206 543 7472.
*E-mail addresses:* jduarte@u.washington.edu (J. Duarte), youngla@u.washington.edu (L. Young).

0304-405X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jfineco.2007.10.008

*PIN* model hereafter). Our empirical examination shows that the original *PIN* model cannot match the pervasive positive correlation between buyer and seller initiated order flow (hereafter buy order flow and sell order flow) or the variances of buy and sell order flow. Our results suggest that the *PIN* model cannot match these moments because the *PIN* model specifies only two possible motives for trades, information and exogenous liquidity needs. All of the trades that are not initiated by informed traders are considered liquidity trades. Information-related trading happens on days in which informed traders receive a private signal, which induces a larger number of buy orders if the private signal is positive and a larger number of sell orders on days with negative private signals. As a result, large numbers of buys and sells arrive on different days, creating a negative correlation between buys and sells. However, we find that for more than 95% of stocks, buys and sells are positively correlated. Furthermore, with only two motives for trade, the *PIN* model cannot match the relatively large variances of buys and sells.

Our second step is an extension of the *PIN* model which accommodates the positive correlation between buys and sells and generates variances closer to those observed in the data. Our extension of the *PIN* model accomplishes this by allowing for simultaneous positive shocks to both buy and sell order flow. The extended model also allows us to compute a measure of asymmetric information, *AdjPIN*, which like the *PIN* measure, is identified by periods of abnormal order flow imbalance as motivated by sequential trade models such as Glosten and Milgrom (1985) and as assumed in Easley, Kiefer, O'Hara, and Paperman (1996). In contrast to the *PIN* measure, however, *AdjPIN* is consistent with the high variances of buys and sells and the positive correlation between buys and sells. Consequently, a Fama-MacBeth (1973) regression that includes *AdjPIN* instead of *PIN* as a proxy for asymmetric information allows us to examine whether the relation between expected returns and *PIN* obtains because *PIN* is a proxy for information asymmetry or because of the original *PIN* model's inability to match some of the characteristics of the order flow data. Our results indicate that *AdjPIN* is orthogonal to expected returns. Thus, the evidence suggests that *PIN* is not priced because it is a proxy for information asymmetry.

Our third step is to explore the relation between *PIN* and expected returns using our extended model to develop a measure of illiquidity unrelated to information asymmetry. We call this measure *PSOS* (probability of symmetric order-flow shock). The *PSOS* is the probability that any given trade happens during a shock to both the number of buyer initiated and seller initiated trades. To connect *PSOS* to illiquidity, we show that stocks with a high *PSOS* match the usual trading activity patterns of illiquid stocks, namely, very low trading activity on most days with sudden spikes in volume associated with the release of public information. We also show that firms in the highest *PSOS* decile have Amihud (2002) illiquidity measures nearly 30 times that of the median *PSOS* firm. In addition to being related to illiquidity, *PSOS* is strongly correlated with *PIN* while the correlation between *PSOS* and *AdjPIN* is relatively low, indicating that *PSOS* is the component of *PIN* that proxies for illiquidity unrelated to asymmetric information. Consequently, the relation between expected returns and *PSOS* reveals the extent to which *PIN* is priced because it is a proxy for illiquidity effects unrelated to information asymmetry. The estimated relation between expected returns and *PSOS* is strong, indicating that the relation between *PIN* and expected returns is due to the fact that *PIN* is also a proxy of illiquidity not related to private information. Naturally, our evidence does not address the open question in the literature of why illiquidity would matter for the cross-section of expected returns. (See, for instance, Constantinides, 1986; Amihud and Mendelson, 1986; Gârleanu and Pedersen, 2004.) Our evidence does suggest, however, that the relation between expected returns and illiquidity cannot be explained by information asymmetry effects.

It is important to note that our interpretation of the results is conditional on the assumption that periods of private information can be identified by abnormal order flow imbalance. Both *PIN* and *AdjPIN* identify the arrival of private information as periods of abnormal order flow imbalance. Despite the evidence that the original *PIN* model captures information asymmetry,[1] it is possible that the abnormal order flow imbalances used to identify *PIN* and *AdjPIN* are not the result of informed trade, but instead reflect liquidity shocks or the effect of the changes in the demand for immediacy in the sense of Grossman and Miller (1988). In this case, the separation of illiquidity into asymmetry of information and inventory concerns implied by the original *PIN* model is not correct. Even though we recognize that this is an important caveat, this paper's objective is not to question the notion that informed trade can be identified by abnormal order flow imbalance. Instead, adopting the same identification methodology as Easley, Hvidkjaer, and O'Hara (2002), our objective is to explore the reasons why *PIN* is priced.

This paper adds to the growing literature on the analysis of order flow data and market microstructure's effect on asset prices. Studies closely related to this paper include Easley, Kiefer, O'Hara, and Paperman (1996), Easley, Kiefer, and O'Hara (1997), Easley, Hvidkjaer, and O'Hara (2002), Vega (2006), and Duarte, Han, Harford, and Young (2008). All of these studies use order flow data and the *PIN* model to extract information from the trade process that is important to asset prices. Our paper complements these papers by allowing for simultaneous positive shocks to both buy and sell order flow. Our paper also relates to the recent papers by Venter and de Jongh (2004) and Boehmer, Grammig, and Theissen (2007) that examine statistical issues related to *PIN*. Our focus, however, is different because we explore a large cross-section of stocks to focus on the economic effect of asymmetric information on expected returns.

The remainder of the paper is outlined as follows. In Section 2, we outline the data we use for our empirical results. In Section 3, we outline our extension of the Easley, Kiefer, O'Hara, and Paperman (1996) structural

[1] See Easley, Kiefer, and O'Hara (1997), Easley, Kiefer, O'Hara, and Paperman (1996), and Vega (2006).

**Table 1**
Percentiles of summary statistics on the number of buyer and seller initiated trades.

This table presents the median and the percentiles of the cross-sectional distribution of a series of statistics on the daily number of buys and sells for each stock in the sample. The number of buys and sells in a day for each stock is estimated with the Lee and Ready (1991) algorithm applied to intraday transactions and quotes data. The mean, variance, and correlation between buyer and seller initiated trade are estimated for each of 48,512 firm-years between 1983 and 2004.

| | 95th percentile | 75th percentile | Median | 25th percentile | 5th percentile |
|---|---|---|---|---|---|
| Mean buys | 322 | 37 | 10 | 3 | 0.4 |
| Mean sells | 266 | 35 | 10 | 3 | 0.3 |
| Variance buys | 14,263 | 513 | 71 | 10 | 0.9 |
| Variance sells | 9,805 | 379 | 54 | 8 | 0.7 |
| Correlation between buys and sells | 0.84 | 0.66 | 0.50 | 0.34 | 0.15 |

microstructure model and present empirical evidence that the model matches important characteristics of the data. In Section 4, we present our empirical results relating *AdjPIN* to the cross-section of average returns. Section 5 concludes.

## 2. Data

Estimation of the structural microstructure models requires data on the number of buyer and seller initiated trades for each firm-day. To compute the number of buyer and seller initiated trades each day, we use the intra day data from the Institute for the Study of Securities Markets (ISSM) (1983–1992) and the NYSE Trade and Quote (TAQ) (1993–2005) databases. We exclude all trades and quotes that occur before the open and at the open, as well as those at the close and after the close, to avoid including trades that occurred during the opening and closing auctions. Furthermore, we exclude all trades with non-typical settlement conditions because these trades may have been made under special arrangements from which the models abstract. We exclude all quotes with zero bid or ask prices, quotes for which the bid–ask spread is greater than 50% of the price, and trades with zero prices to eliminate possible data errors. We employ the Lee and Ready (1991) algorithm to sign the trades. That is, trades above the mid-point between the bid and ask prices are considered buyer initiated and trades below the mid-point of the bid and ask prices are considered seller initiated. Trades that occur at the mid-point of the bid and ask prices are classified as buyer or seller initiated according to a tick test.[2] In addition, if there are no quotes posted during the trading day, we use the tick test to sign any trades made during the day. Each firm-day, we compute the number of trades classified as buyer initiated and the number of trades classified as seller initiated. As in Easley, Kiefer, O'Hara, and Paperman (1996), we do not consider information on the sizes of the trades. For a model that considers such information see Bernhardt and Hughson (2002). The full sample includes all NYSE and Amex stocks. As in Easley, Hvidkjaer, and O'Hara (2002), we exclude American Depositary Receipts (ADRs), closed-end funds and firm-years with less than 60 days of trading data. Appendix A provides more detail on the intra day data.

Table 1 presents the median, first, and third quartile as well as the fifth and 95th percentiles of the mean, variance of both buys and sells, and correlation between buys and sells. Interestingly, buys appear more volatile than sells. Furthermore, the median correlation between buys and sells is in excess of 0.50. In fact, even the fifth percentile is positive for every year in the sample. Indeed, on average, 50% of firms had correlations of buys and sells between 0.34 and 0.66, and more than 95% of firms had buy and sell correlations above 0.15.[3] Fig. 1 presents graphs of the number of buys and sells for the firm at the 25th percentile, the median, and the 75th percentile. The positive correlation is easily visible in the graphs. As we will see in Section 3, the *PIN* model cannot reproduce the positive correlations between buys and sells or the high variances observed in the data.

To understand the relation between symmetric order-flow shocks and illiquidity, we also use a comprehensive database of news events.[4] The database contains a random sample of firms, which represents between 20% and 25% of the firms in the Center for Research in Security Prices (CRSP) database. The database records each day each firm in the sample is mentioned in publications that are covered in the Dow Jones Interactive Publications Library. The database is described in detail in Chan (2003).

The asset pricing tests require daily and monthly return data which we gather from CRSP. The tests also require book-value data which we gather from COMPUSTAT, specifically data item 60. We compute monthly turnover (*TURNOVER*) as the logarithm of the monthly volume divided by shares outstanding. Following Amihud (2002), we compute the Amihud measure (*ILLIQ*) as the average ratio of daily absolute price change and daily (dollar) volume for the year $t-1$. Our asset pricing regressions also require us to compute the market $\beta$'s. We estimate market $\beta$'s as follows: first, for each firm-month, we estimate the market loadings using 60 months of past data. We then form portfolios based on these pre-ranking

[2] The tick test classifies a trade as buyer initiated if the price was above that of the previous trade, and a sell if the price was below that of the previous trade.

[3] We have also computed the correlation between buyer and seller initiated order flow at the hourly level and find that the correlations are similar to daily correlations.

[4] We thank Wesley Chan for graciously providing us with his database.

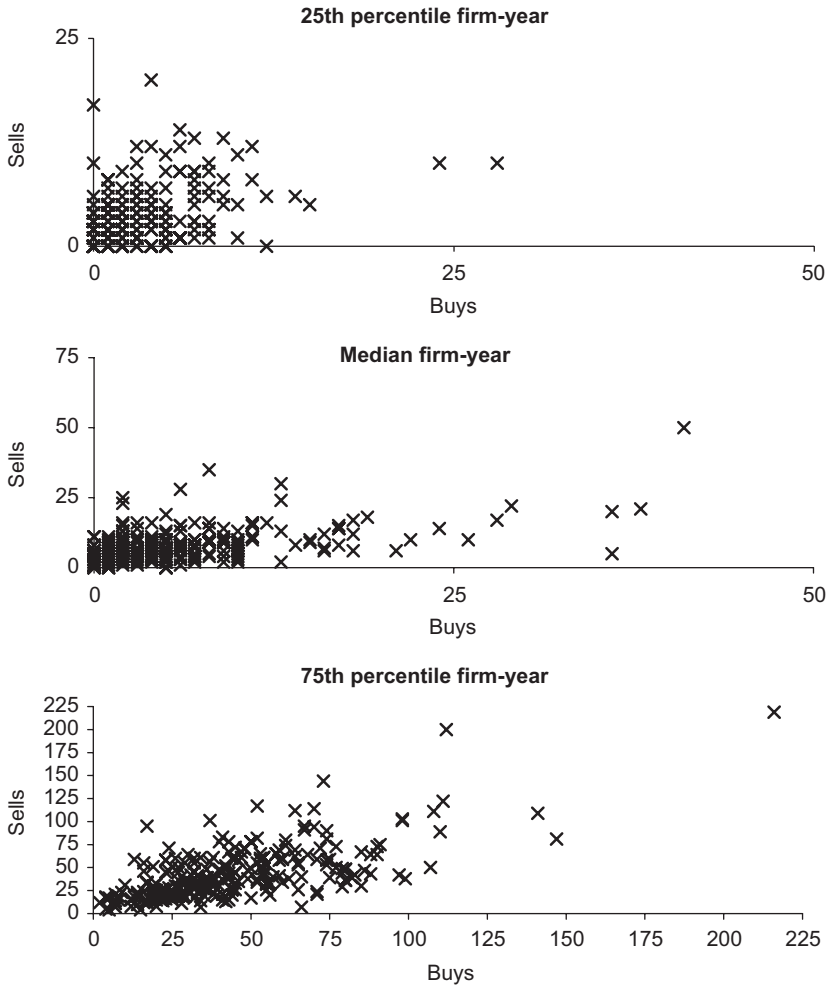


**Fig. 1.** Example of number of buyer and seller initiated trades in the sample. This figure displays the plots of the daily number of buyer and seller initiated trades for three different stock-years in the sample. The number of buyer and seller initiated trades are calculated by Lee and Ready (1991) algorithm. The stock-year in the top panel has an estimated correlation between buys and sells equal to the 25th percentile of the correlations in the sample. The stock-year in the second panel has an estimated correlation between buys and sells equal to median, correlation and the stock-year in the third panel has a correlation between buys and sells equal to the 75th percentile.

factor loadings. Using the returns from these portfolios, we estimate the full period $\beta$ for each portfolio and assign this $\beta$ to each firm in the portfolio.

## 3. The *PIN* model and its extension

This section briefly discusses the *PIN* model, then extends the model and presents estimates of various parameterizations of the extension as well as tests of these parameterizations.

### 3.1. The PIN model

Before considering our extension, we will briefly review the Easley, Kiefer, O'Hara, and Paperman (1996) model. This model is based on the Glosten and Milgrom (1985) and Easley and O'Hara (1987) sequential trade models. The model contains both informed traders who trade for speculative purposes based on private information, and noise traders whose reasons for trading are exogenous. It also posits the existence of an uninformed liquidity provider who sets the bid and ask quotes by observing the flow of buy and sell orders, and assessing the probability that the orders come from informed traders. The bid–ask spread compensates the liquidity provider for the possibility of trading with the informed traders. At the beginning of each day, nature decides whether a private information event will occur. The probability that a private information event will occur on a given day is $a$. If a private information event occurs on a particular day, informed traders receive a private signal which is positive with probability $d$. If the signal is positive, buy order flow for that day arrives according to a Poisson distribution with intensity parameter $u+\varepsilon_b$ and sell order flow arrives according to a Poisson distribution with intensity parameter $\varepsilon_s$. The intuition is that on days with positive private information, both informed traders

and noise traders arrive in the market as buyers. The total buy order flow for the day therefore consists of arrivals of both noise traders, who arrive at rate $\varepsilon_b$, and informed traders who arrive at rate $u$. On the other hand, only noise traders arrive to sell, so the arrival rate of sell order flow is $\varepsilon_s$. If the signal is negative, buy orders consist only of noise traders with intensity parameter $\varepsilon_b$, and sell order flow arrives according to a Poisson distribution with intensity parameter $\varepsilon_s + u$ to reflect both the arrivals of noise sellers and of informed sellers. If there is no private signal, only noise traders will arrive in the market, so buy and sell order flow arrives by Poisson distributions with intensity parameters $\varepsilon_b$ and $\varepsilon_s$, respectively. A tree outlining the information structure in the model can be found in Fig. 2. The *PIN* is computed (suppressing the firm subscript) as

$$PIN = \frac{a \times u}{a \times u + \varepsilon_s + \varepsilon_b}. \tag{1}$$

The intuition behind the formula for *PIN* is that the probability of an informed trade is the ratio of expected informed order flow to expected total order flow.

### 3.1.1. Estimation of the PIN model

We estimate the *PIN* model numerically via maximum likelihood for each firm-year from 1983 to 2004. The likelihood function of the Easley, Kiefer, O'Hara, and Paperman (1996) model is

$$L(\theta|B,S) = (1-a)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} + ade^{-(u+\varepsilon_b)}\frac{(u+\varepsilon_b)^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} + a(1-d)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(u+\varepsilon_s)}\frac{(u+\varepsilon_s)^S}{S!}, \tag{2}$$

where $B$ and $S$ are the number of buys and sells for a given day and $\theta = (a, u, \varepsilon_b, \varepsilon_s, d)$ is the parameter vector. The likelihood equation shows that at each node of the tree in Fig. 2, buys and sells arrive according to independent Poisson distributions, with the intensity parameters differing according to the node of the tree. For instance, conditional on positive information, buys arrive at a rate $\varepsilon_b + u$ and sells arrive at a rate of $\varepsilon_s$. The direct computation of this likelihood function when $B$ is large may result in numerical overflow since $\varepsilon_b^B$ or $B!$ become very large numbers. To avoid this problem we compute $e^{-\varepsilon_b}\varepsilon_b^B/B!$ as

$$e^{[-\varepsilon_b + B\ln(\varepsilon_b) - \sum_{i=1}^{B}\ln(i)]}. \tag{3}$$

We compute the ratios $e^{-\varepsilon_s}\varepsilon_s^S/S!$, $e^{-(u+\varepsilon_b)}(u+\varepsilon_b)^B/B!$ and $e^{-(u+\varepsilon_s)}(u+\varepsilon_s)^S/S!$ similarly. To account for the fact that a particular numerical optimization may arrive at a local maximum, we run the likelihood optimization 10 times for each firm-year, each with randomly selected starting points, and we then select the maximum of these 10 optimizations. Because the optimization procedure is

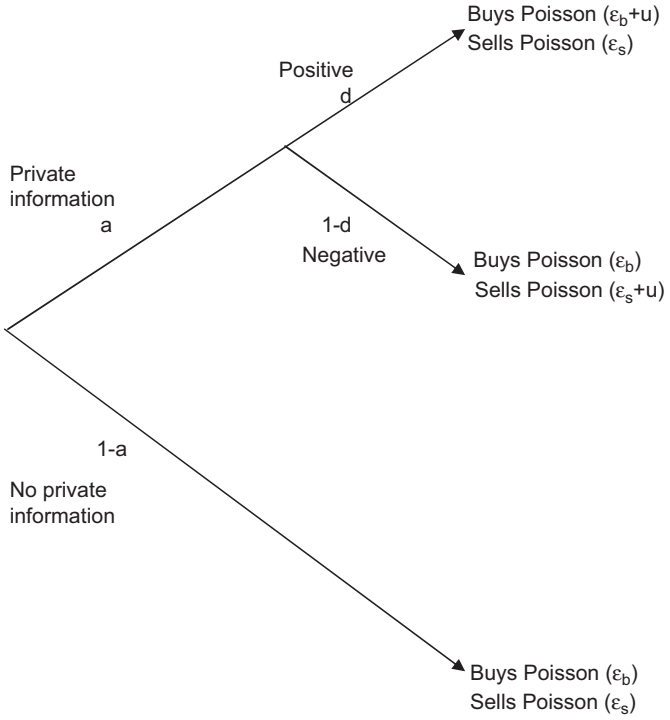


**Fig. 2.** The *PIN* model. This tree represents the Easley, Hvidkjaer, and O'Hara (2002) trading model. In any given trading day, private information arrives with probability $a$. When there is no private information, the number of buyer initiated trades is Poisson distributed with intensity $e_b$ and the number of seller initiated trades is Poisson distributed with intensity $e_s$. Positive private information induces an increase in the number of buyer initiated trades, as a result the number of buyer initiated trades is Poisson distributed with intensity $e_b + u$ when there is positive private information. Analogously, the number of seller initiated trades is Poisson distributed with intensity $e_s + u$ when there is negative private information.

**Table 2**
Summary of estimation results.

Panel A presents the cross-sectional distribution of the estimated parameters of the original *PIN* model along with the cross-sectional distribution of the estimated probability of informed trading (*PIN*). The parameter $a$ represents the probability that an information event will occur on a particular day, $u$ represents the arrival rate of informed traders on information days, $d$ is the probability that an information event will be positive, and $\varepsilon_b/(\varepsilon_s)$ is the rate at which liquidity traders arrive to buy/(sell). *PIN* is the probability of informed trade. The model is estimated on 48,512 firm-years between 1983 and 2004. The $t$-statistics of the *PIN*s are calculated with delta method based on the asymptotic covariance matrix of the estimated model parameters. Panel B displays the cross-sectional distribution of the moments of the number of buyer initiated and seller initiated trades implied by the original *PIN* model.

| | 95th percentile | 75th percentile | Median | 25th percentile | 5th percentile |
|---|---|---|---|---|---|
| *Panel A: Estimated parameters of the original PIN model* | | | | | |
| $a$ | 0.46 | 0.35 | 0.26 | 0.18 | 0.08 |
| $u$ | 193.84 | 44.78 | 17.53 | 7.04 | 2.13 |
| $d$ | 0.98 | 0.82 | 0.69 | 0.54 | 0.26 |
| $\varepsilon_b$ | 270.13 | 28.58 | 7.47 | 1.79 | 0.13 |
| $\varepsilon_s$ | 245.15 | 31.07 | 8.96 | 2.49 | 0.19 |
| *PIN* | 0.51 | 0.28 | 0.20 | 0.15 | 0.10 |
| $t$-Statistic | 15.05 | 11.65 | 9.14 | 6.96 | 4.44 |
| *Panel B: Moments implied by the original PIN model* | | | | | |
| Mean buys | 321 | 37 | 10 | 3 | 0.4 |
| Mean sells | 267 | 35 | 10 | 3 | 0.4 |
| Variance buys | 6,581 | 297 | 48 | 8 | 0.8 |
| Variance sells | 3,303 | 162 | 29 | 6 | 0.7 |
| Correlation between buys and sells | 0 | −0.04 | −0.08 | −0.13 | −0.22 |

computationally intensive,[5] we divide the optimization across more than 100 personal computers. The estimation procedure converges for virtually all firm-years.

Panel A of Table 2 presents quartiles, median, fifth, and 95th percentiles of the parameters of the *PIN* model across all firm-years. The median firm-year has a *PIN* of 0.2, with a first quartile of 0.15 and a third quartile of 0.28. Panel A also shows the median as well as the first and third quartiles of the $t$-statistics for the hypothesis that $PIN = 0$. The $t$-statistics range between 4.44 and 15.05 indicating that the *PIN* is relatively precisely estimated. In general, our *PIN* estimates are consistent with the ones in Easley, Hvidkjaer, and O'Hara (2002).

### 3.1.2. Empirical and PIN model implied moments

Fig. 3 presents graphs of data simulated under the *PIN* model for the same firm-years as in Fig. 1. The parameters used in these simulations are the same as those estimated from the samples displayed in Fig. 1.[6] In Fig. 3, sell order flow appears on the vertical axis and buy order flow appears on the horizontal axis. In each of the graphs, days with negative information appear in the upper left hand corner of the graph and days with positive private information appear in the lower right hand corner of the graph. The non-information days are represented by the cluster of points in the left bottom corner.

The graphs in Fig. 3 show how *PIN* is identified in the data. Days with negative or positive private information produce large order flow imbalances relative to the normal level of trade. Another interesting feature of the graphs is that buy and sell order flow are negatively correlated under the *PIN* model. In fact, this is not peculiar to these firm-years, it is a general feature of the *PIN* model, in which contemporaneous covariance between the number of buys and sells in the *PIN* model is

$$cov[B,S] = (au)^2(d-1)d \leqslant 0. \quad (4)$$

This result is central to the intuition behind the model. Informed trade can inflate the arrival rate of buy or sell order flow, but not on the same day. Thus, when buy order flow is inflated, sell order flow is expected to be below its mean and vice versa. Panel B of Table 2 shows the correlations between buys and sells computed using the estimated parameters of the model. As is clear from Eq. (4), all of the correlations are below or equal to zero. Note that the negative correlation between buys and sells in the original *PIN* model is not consistent with the data. Recall from Table 1 that even the fifth percentile of firms has a positive correlation. Indeed, Table 1 indicates that more than 95% of the firm-years in the sample have correlation between buys and sells above 0.15.

Panel B in Table 2 also presents the implied means and variances of buys and sells under the *PIN* model, calculated for each firm using the estimated parameters. (See Appendix B for the formulas of these means and variances.) Comparison with the corresponding sample means and variances in Table 1 shows that while the model closely matches the buys' and sells' means, the model produces buy and sell variances that are smaller than the actual variances in the data. For about 50% of the sample, the actual variances of buys and sells are nearly twice as large as the implied variance from the model. This problem is most pronounced for firms with large buy and sell variances.

Thus, it appears that some aspect of the data-generating process that creates a large positive correlation between buys and sells and high levels of buy and sell volatility is

[5] The estimation of the original and of the extended models for all 48,611 firm-years entails 2,916,660 optimizations.

[6] The parameter values are $a = 0.225$, $\varepsilon_b = 2.568$, $\varepsilon_s = 2.749$, $u = 7.395$, and $d = 0.452$ in the first panel; $a = 0.175$, $\varepsilon_b = 4.004$, $\varepsilon_s = 7.179$, $u = 15.236$, and $d = 0.723$ in the second panel; and $a = 0.410$, $\varepsilon_b = 37.827$, $\varepsilon_s = 32.206$, $u = 45.870$, and $d = 0.369$ in the third panel.

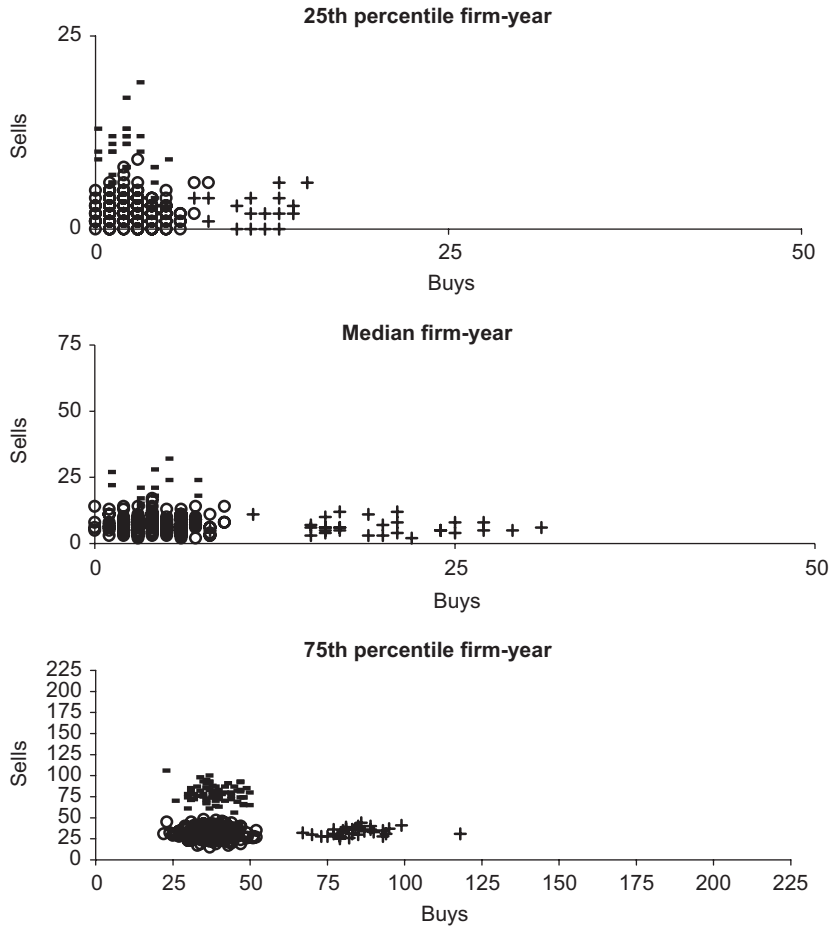


**Fig. 3.** Simulation of the number of buyer and seller initiated traders under the original *PIN* model. The plots in this figure are constructed by simulating daily buyer and seller initiated trades under the *PIN* model. Days with positive private information are represented by a positive sign (see the cloud of observations on the right bottom corner of each plot). Days with negative private information are represented by negative signs (see the cloud of observations on the left top corner of the each plot). Days with only noise trading are represented by circles (see the cloud of observations on the left bottom corner of each plot). To conduct the simulation, the *PIN* model is estimated using data from the firm-year with a correlation between buyer and seller initiated trades equal to the 25th percentile, the firm-year with the median correlation between buyer and seller initiated trades, and the firm-year with the 75th percentile correlation between buyer and seller initiated trades. The estimated parameters from these firm-years are then used to simulate the number of buyer and seller initiated trades under the *PIN* model for 240 hypothetical trading days. The top panel contains the 240 simulated trading days for the 25th percentile firm-year. The middle panel contains the simulated trading days for the median firm-year. The bottom panel contains the simulated trading days for the 75th percentile firm-year. The actual data for these firm-years are graphed in Fig. 1.

missing from the original *PIN* model. In the next section, we extend the *PIN* model to better match the positive correlations between buys and sells, as well as the high volatility in buy and sell order flow.

### 3.2. Model extension

We extend the *PIN* model to account for the large buy and sell volatility and pervasive positive correlation between buys and sells by allowing both buy and sell order flow to increase on certain days. Naturally, there are many possible different extensions of the *PIN* model that match the data well. However, we consider only extensions that are mixtures of a finite number of Poisson random variables. We focus on these extensions because they are parsimonious and imply measures, such as *PIN*, which have economic interpretations. In this section we present the unrestricted extended model, while in Section 3.2.1, we search for a more parsimonious version of the extended model.

The unrestricted extended model's information structure is in Fig. 4. The first small difference between the extension and the *PIN* model is that the extension allows for the arrival rate of informed buyers, $u_b$, to be different from the arrival rate of informed sellers, $u_s$. The motivation for this is to better allow the model to account for the fact that in the data, buy order flow has a larger variance than sell order flow for almost all firms. More importantly, the model allows for increased buy and sell variation and a positive correlation between buys and sells because each day an event can occur that causes both buy and sell order flow to increase. We call this event a symmetric order-flow shock. The probability of such an event in the

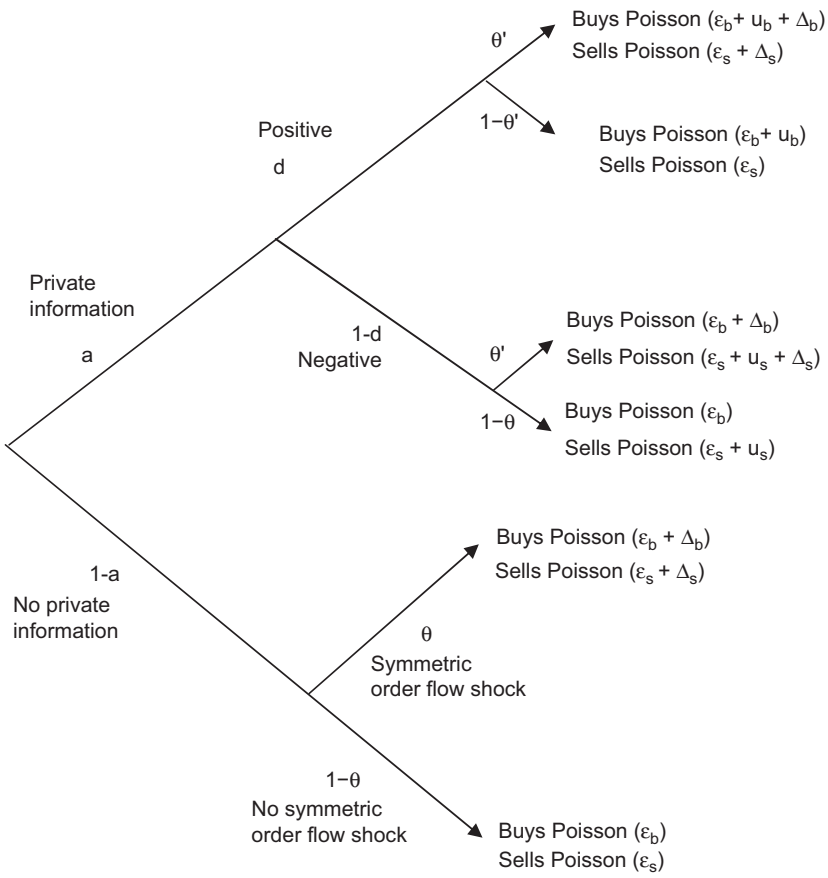


**Fig. 4.** Unrestricted extended trading model. The tree represents the extended trading model. The difference between this model and the Easley, Hvidkjaer, and O'Hara (2002) trading model is twofold. First, the extended model has additional branches for days in which both the number of buys and sells increase (symmetric order-flow shocks). These days happen with probability $q'$ on the days with private information and with probability $q$ on the days without private information. Second, the number of buyer initiated informed trades has a different distribution from the number of seller initiated trades.

model, conditional on the absence of private information, is represented by $\theta$. The probability of such an event, conditional on the arrival of private information, is $\theta'$. In the event of symmetric order-flow shock, the additional arrival rate of buys is $\Delta_b$ and sells is $\Delta_s$.

There are at least two possible explanations for symmetric order-flow shocks. One possible cause of a symmetric order-flow shock is the occurrence of a public news event about whose implications traders disagree. This disagreement causes both buy and sell trades to arrive at higher rates. The notion that disagreement about public news events is an important source of volume has a long history in the literature dating back to Bachelier (1900). As Kandel and Pearson (1995) note, agents observing identical signals can disagree about how to interpret these signals if they disagree about the model that should be used to interpret them. A number of studies have found empirical support for this hypothesis including, Harris and Raviv (1993), Kandel and Pearson (1995), Bessembinder, Chan, and Seguin (1996), Bamber, Barron, and Stober (1999), and Sarkar and Schwartz (2007). Another potential cause for symmetric order-flow shocks is that traders simply coordinate on trading on certain days to reduce trading costs, see Admati and Pfleiderer (1988). Even though there is more than one possible cause for symmetric order-flow shocks, our purpose here is not to distinguish between the possible causes of such shocks. Instead, our objective is to present evidence that their inclusion in the model adds to its ability to explain the data, and has important consequences for the relation between information asymmetry and average returns.

The addition of symmetric order-flow shocks to the model allows the correlation of buys and sells to be positive. Under the extended model, the covariance of buys and sells is

$$\begin{aligned} cov[B,S] = {} & a \times d \times u_b \times (a \times (d-1) \times u_s \\ & - (1-a) \times \Delta_s \times (\theta - \theta')) \\ & - \Delta_b \times (\Delta_s \times (\theta \times (\theta - 1) + a \\ & \times (2 \times \theta' \times \theta - 2 \times \theta^2 + (\theta - \theta')) \\ & + a^2(\theta'^2 - \theta \times \theta')) + (-1 + a) \\ & \times a \times (-1 + d) \times (\theta - \theta') \times u_s). \end{aligned} \tag{5}$$

Unlike the *PIN* model, this expression can be positive or negative. To see this, note that when $\theta = \theta'$, the formula above becomes

$$cov[B,S] = a^2 \times u_b \times u_s \times (d-1) \times d - \Delta_b \times \Delta_s \times (\theta - 1) \times \theta. \tag{6}$$

As with the original *PIN* model, the probability of informed trade in the adjusted model is the ratio of the expected informed order to the total expected order flow:

$$AdjPIN = \frac{a \times (d \times u_b + (1-d) \times u_s)}{a \times (d \times u_b + (1-d) \times u_s) + (\Delta_b + \Delta_s) \times (a \times \theta' + (1-a) \times \theta) + \varepsilon_s + \varepsilon_b}. \tag{7}$$

The extended model also allows us to introduce a related probability, the (*PSOS*). The *PSOS* is the unconditional probability that a given trade will come from a shock to both buy and the sell order flows. The *PSOS* is given by

$$PSOS = \frac{(\Delta_b + \Delta_s) \times (a \times \theta' + (1-a) \times \theta)}{a \times (d \times u_b + (1-d) \times u_s) + (\Delta_b + \Delta_s) \times (a \times \theta' + (1-a) \times \theta) + \varepsilon_s + \varepsilon_b}. \tag{8}$$

As we show in Section 3.2.3, high *PSOS* firms tend to be firms with low volume, relative to other firms, on most days but who experience large increases in both buy and sell order flow on days associated with public news events. For instance, a high *PSOS* firm has about one-fifth the volume of the median firm on days without news and nearly twice as much volume as the median firm on days with news. As we will see later, this tendency of high *PSOS* firms to have very low volume on most days and very high volume on a few days means that these firms tend to be very illiquid relative to other firms. To see this, note that an investor who wishes to trade shares in a high *PSOS* firm faces a trade-off between waiting, perhaps for some time, for a period of high volume when the stock will be more liquid than on a typical day and when his trade can presumably be crossed with another trade, or trading immediately and involving a market maker or liquidity provider at additional cost. Thus, these firms' shares tend to be among the least liquid firms in the market, as measured by their Amihud measures. Based on this we argue that *PSOS* is effectively a proxy for illiquidity.

### 3.2.1. Model selection and estimation of extended model

As with the *PIN* model, we estimate the extended model numerically via maximum likelihood for each available firm-year. The likelihood function of the extended model is

$$\begin{aligned} L(\theta|B,S) = {} & (1-a)(1-\theta)\mathrm{e}^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}\mathrm{e}^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ & + (1-a)\theta\mathrm{e}^{-(\varepsilon_b+\Delta_b)}\frac{(\varepsilon_b+\Delta_b)^B}{B!}\mathrm{e}^{-(\varepsilon_s+\Delta_s)}\frac{(\varepsilon_s+\Delta_s)^S}{S!} \\ & + a(1-\theta')(1-d)\mathrm{e}^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}\mathrm{e}^{-(u_s+\varepsilon_s)}\frac{(u_s+\varepsilon_s)^S}{S!} \\ & + a\theta'(1-d)\mathrm{e}^{-(\varepsilon_b+\Delta_b)}\frac{(\varepsilon_b+\Delta_b)^B}{B!} \\ & \times \mathrm{e}^{-(u_s+\varepsilon_s+\Delta_s)}\frac{(u_s+\varepsilon_s+\Delta_s)^S}{S!} \\ & + a(1-\theta')d\mathrm{e}^{-(u_b+\varepsilon_b)}\frac{(u_b+\varepsilon_b)^B}{B!}\mathrm{e}^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\ & + a\theta d\mathrm{e}^{-(u_b+\varepsilon_b+\Delta_b)}\frac{(u_b+\varepsilon_b+\Delta_b)^B}{B!} \\ & \times \mathrm{e}^{-(\varepsilon_s+\Delta_s)}\frac{(\varepsilon_s+\Delta_s)^S}{S!}, \end{aligned} \tag{9}$$

where $B$ and $S$ are the number of buys and sells for a given day and $\theta = (a, u_b, u_s, \varepsilon_b, \varepsilon_s, d, \theta, \theta', \Delta_b, \Delta_s)$ is the parameter vector.

The estimation procedure is similar to the one described in Section 3.1.1, and it converges in virtually all cases. Analogous to the original *PIN* model, each term in the likelihood function above corresponds to a branch in the trading tree in Fig. 4.

Because the general model contains twice as many parameters as the *PIN* model, before continuing, we seek a parsimonious parameterization that matches the characteristics of the data. Table 3 presents implied means, and variances of buys and sells as well as the correlation between buys and sells for four restricted versions of the general model, as well as the unrestricted version. These implied means, variances, and covariances are calculated with the formulas in Appendix B. These models are meant to represent a variety of potential parameter restrictions. Specifically, in the first three, we allow symmetric order-flow shocks only on days with no private information ($\theta' = 0$). Moreover, in Model 1 we restrict the buyer and seller initiated information-related trades to have the same arrival intensities ($u_b = u_s$). Furthermore, we restrict buyer and seller symmetric order-flow shocks to have the arrival intensities ($\Delta_b = \Delta_s$). In Model 2, we keep the restriction $u_s = u_b$ and relax the restriction $\Delta_b = \Delta_s$. Model 3 relaxes the restrictions $u_s = u_b$ and $\Delta_b = \Delta_s$. Model 4 allows symmetric order-flow shocks on all days, but restricts the probability of symmetric order flow arrival to be the same on days with and without private information ($\theta = \theta'$). Model 5 is the unrestricted general model.

Each of the models produces positive correlations between buys and sells because, as opposed to the original *PIN* model, all the estimated extended models do not restrict the probability of asymmetric order-flow shocks to be zero. Furthermore, the buy and sell correlations are close to those found in the data. The median correlation between buys and sells ranges between 0.38 and 0.45, while in the data the median correlation is 0.50. Furthermore, each extended model is able to produce higher buy and sell variances than the *PIN* model, while closely matching the means of buys and sells. The first three models, which assume that symmetric order-flow shocks arrive only on days without private information, produce somewhat higher correlations than the last two

**Table 3**
The percentiles of the moments of buys and sells implied by the extended *PIN* model.

This table presents the median and the percentiles of the mean, variance, and correlation of buys and sells implied by the extended trading model and calculated using estimated parameters for each stock-year in the sample. The parameter $a$ represents the probability that an information event will occur on a particular day, $u_b/(u_s)$ represents the arrival rate of informed buyers/(sellers) on information days, $d$ is the probability that an information event will be positive, and $\varepsilon_b/(\varepsilon_s)$ is the rate at which liquidity traders arrive to buy/(sell). The parameter $q$ represents the probability of a symmetric order-flow shock conditional on a private information event, $q'$ represents the probability of a symmetric order-flow shock conditional on there being no private information event, and $\Delta_b/(\Delta_s)$ is the increase in the arrival rate of buys/(sells) that occurs in a symmetric order-flow shock. Each panel displays the results based on different versions of the extended model. The models are estimated on 48,512 firm-years between 1983 and 2004.

| | 95th percentile | 75th percentile | Median | 25th percentile | 5th percentile |
|---|---|---|---|---|---|
| *Panel A*: *Model* 1 ($u_b = u_s$, $\Delta_b = \Delta_s$, $\theta' = 0$) | | | | | |
| Mean buys | 322 | 37 | 10 | 3 | 0.4 |
| Mean sells | 270 | 35 | 11 | 3 | 0.4 |
| Variance buys | 8,123 | 339 | 53 | 9 | 0.9 |
| Variance sells | 7,740 | 320 | 49 | 8 | 0.9 |
| Correlation between buys and sells | 0.79 | 0.61 | 0.46 | 0.31 | 0.10 |
| *Panel B*: *Model* 2 ($u_b = u_s$, $\theta' = 0$) | | | | | |
| Mean buys | 320 | 37 | 10 | 3 | 0.4 |
| Mean sells | 266 | 35 | 10 | 3 | 0.3 |
| Variance buys | 9,636 | 381 | 57 | 9 | 0.8 |
| Variance sells | 6,401 | 268 | 41 | 7 | 0.7 |
| Correlation between buys and sells | 0.71 | 0.53 | 0.39 | 0.26 | 0.12 |
| *Panel C*: *Model* 3 ($\theta' = 0$) | | | | | |
| Mean buys | 321 | 37 | 10 | 3 | 0.4 |
| Mean sells | 266 | 35 | 10 | 3 | 0.3 |
| Variance buys | 9,684 | 382 | 57 | 9 | 0.8 |
| Variance sells | 6,337 | 269 | 42 | 7 | 0.7 |
| Correlation between buys and sells | 0.80 | 0.61 | 0.45 | 0.30 | 0.10 |
| *Panel D*: *Model* 4 ($\theta' = \theta$) | | | | | |
| Mean buys | 321 | 37 | 10 | 3 | 0.4 |
| Mean sells | 266 | 35 | 10 | 3 | 0.3 |
| Variance buys | 10,973 | 393 | 57 | 9 | 0.8 |
| Variance sells | 7,005 | 283 | 43 | 7 | 0.7 |
| Correlation between buys and sells | 0.67 | 0.50 | 0.37 | 0.25 | 0.11 |
| *Panel E*: *Model* 5 (*unrestricted extended model*) | | | | | |
| Mean buys | 321 | 37 | 10 | 3 | 0.4 |
| Mean sells | 266 | 35 | 10 | 3 | 0.3 |
| Variance buys | 10,350 | 391 | 58 | 9 | 0.9 |
| Variance sells | 6,758 | 281 | 43 | 7 | 0.7 |
| Correlation between buys and sells | 0.66 | 0.50 | 0.37 | 0.25 | 0.10 |

models, which allow for symmetric order-flow shocks on private information days. The final two models, however, produce somewhat higher buy and sell variances, particularly for the high variance firms.

In order to choose the model which best fits the data, we run a series of likelihood ratio tests. We start by testing the original *PIN* model as null against Model 1 in Panel A of Table 3, which restricts $u_b = u_s$, $\Delta_b = \Delta_s$, and $\theta' = 0$. This amounts to a test of the null that $\theta' = \theta = 0$ or $\Delta_b = \Delta_s = 0$. Note that the standard regularity conditions that assure that the likelihood ratio statistic is asymptotically $\chi^2$ do not apply in this case because in the *PIN* model $\theta' = \theta = 0$ and thus, the parameters lie on the frontier of the parameter space under the null hypothesis. In order to obtain the distribution of the test statistics, we conduct Monte Carlo simulations. For 100 randomly selected firm-years, we estimate the parameters of both models. We then generate one firm-year of trading days under the null using the estimated parameters and estimate both models using the simulated data to compute the likelihood ratio statistic. We repeat this process 500 times for each firm-year. We then compare the likelihood ratio from the data to critical values of the simulated distribution. Table 4 shows that the *PIN* model is rejected at the 5% level in favor of Model 1 in 99% of the firm-years in the randomly generated sample.

Having rejected the *PIN* model in favor of Model 1, we then test Model 1 as null against Model 2. This amounts to a test of the null hypothesis that $\Delta_s = \Delta_b$. As the parameters do not lie on the frontier of the parameter space under the null, the standard asymptotic results apply and the likelihood ratio is distributed asymptotically $\chi^2$ with one degree of freedom. This being the case, we can apply this test to the entire sample. The results in Table 4 show that the null is rejected at the 5% level in 53% of the firm-years. We interpret this as evidence that Model 2 is preferable to Model 1 and, therefore, that the restriction that $\Delta_s = \Delta_b$ should not be enforced.

**Table 4**
Likelihood ratio tests.

This table presents the 5% rejection frequency of the null hypothesis for a series of likelihood ratio tests. The parameter $u_b/(u_s)$ represents the arrival rate of informed buyers/(sellers) on information days, $q$ represents the probability of a symmetric order-flow shock conditional on a private information event, $q'$ represents the probability of a symmetric order-flow shock conditional on there being no private information event, and $\Delta_b/(\Delta_s)$ is the increase in the arrival rate of buys/(sells) that occurs in a symmetric order-flow shock. The test statistics in Tests 2, 3, and 5 are distributed chi-square with one degree of freedom. Tests 2, 3, and 5 were conducted using the likelihood ratios computed using separate estimates of the models for each of the 48,512 firm-years between 1983 and 2004. The test statistics in Tests 1 and 4 have non-standard asymptotic distributions and hence, the reported rejection frequencies are estimated using Monte Carlo simulations. Tests 1 and 4 are performed on a sample of 100 randomly selected firm-years.

| Test number | Null hypothesis | Alternative hypothesis | Frequency of rejection of null |
|---|---|---|---|
| 1 | Original *PIN* model | Extended model with the restrictions $u_b = u_s$, $\Delta_b = \Delta_s$, $\theta' = 0$ | 0.99 |
| 2 | Extended model with the restrictions $u_b = u_s$, $\Delta_b = \Delta_s$, $\theta' = 0$ | Extended model with the restrictions $u_b = u_s$, $\theta' = 0$ | 0.53 |
| 3 | Extended model with the restrictions $u_b = u_s$, $\theta' = 0$ | Extended model with the restriction $\theta' = 0$ | 0.23 |
| 4 | Extended model with the restriction $\theta' = 0$ | Unrestricted extended model | 0.79 |
| 5 | Extended model with the restriction $\theta' = \theta$ (preferred model) | Unrestricted extended model | 0.09 |

We then conduct a similar experiment and test Model 2 as null against Model 3. This amounts to a test of $u_b = u_s$, and, therefore, the likelihood ratio is distributed asymptotically $\chi^2$ with one degree of freedom. The results indicate that the null is rejected in favor of the alternative 23% of the time. While this is not as high a rejection rate as we see above, it is higher than the 5–10% rejection rate we would expect by pure chance. Therefore, we conclude that the restriction $u_b = u_s$ should not be enforced and reject Model 2 in favor of Model 3.

Finally, we test the last two models, which involve restrictions on the parameter $\theta'$. First, we test Model 3 against Model 5, the unrestricted model. This amounts to a test of the null that $\theta' = 0$. As in the first likelihood ratio test, the likelihood ratio statistic is no longer distributed $\chi^2$. Therefore, we conduct Monte Carlo simulations, as described above, to obtain the likelihood ratio statistics' critical values under the null. Table 4 shows that the null is rejected in 79% of the randomly selected firm-years. We therefore conclude that the restriction that $\theta' = 0$ should not be imposed. Lastly, we test Model 4 against the unrestricted alternative, Model 5. This is a test of $\theta' = \theta$ and therefore, the likelihood ratio is distributed $\chi^2$ with one degree of freedom. The null is rejected in 9% of the firm-years. As this is close to what we would expect by chance, we conclude that the restriction $\theta' = \theta$ is innocuous, and choose Model 4 as our preferred model. Henceforth, we refer to this model as the *AdjPIN* model. In what follows, all results use estimates from *AdjPIN* model. However, all of the results that follow are robust to any of the other four alternative specifications. The preferred extended model has nine free parameters, which is four more parameters than the original *PIN* model.

### 3.2.2. The preferred extended model

Fig. 5 presents graphs of buys and sells simulated under the preferred model, for the firm-years featured in Figs. 1 and 3.[7] Here, there are six groups of points, one for each branch of the tree in Fig. 4. The group of points in the lower left show the normal level of noise trade in the stock. The cluster of points in the upper right reflect days with symmetric order-flow shocks. The graphs show that the preferred model better matches the data in Fig. 1 than the *PIN* model. It is clear from the graphs that symmetric order-flow shocks allow the preferred model to generate positive correlations between buys and sells. Furthermore, the graphs indicate that buys and sells have higher variances under the preferred model than under the *PIN* model, particularly for the firm-year representing the third quartile. This is because the preferred model has a larger number of possible reasons for trade than the original *PIN* model, which is reflected in a tree in Fig. 4 with more branches and consequently more variation on the number of buys and sells.

Table 5 presents the fifth percentile, first quartile, median, 75th percentile, and 95th percentile of the parameters of the *AdjPIN* model for all firm-years. The median and both quartiles of *AdjPIN* are lower than *PIN*, ranging between a fifth percentile of 0.08 and a 95th percentile of 0.37. The *t*-statistics indicate that the *AdjPIN* estimates are large relative to their standard errors, with even the first quartile being over three. The median and both quartiles of *PSOS* are also displayed in Table 5. They range between a fifth percentile of 0.12 and a 95th percentile of 0.66. The *t*-statistics indicate that the *PSOS* estimates are also large relative to their standard errors, with even the first quartile being over two. Fig. 6 shows the time series of the first quartile, median, and third

[7] The parameter values used in these simulations are $\alpha = 0.585$, $\varepsilon_b = 1.553$, $\varepsilon_s = 1.624$, $u_b = 4.075$, $u_s = 2.631$, $\Delta_b = 5.151$, $\Delta_s = 6.112$, $\delta = 0.377$, and $\theta = 0.173$ in the first panel; $\alpha = 0.426$, $\varepsilon_b = 2.067$, $\varepsilon_s = 5.810$, $u_b = 4.665$, $u_s = 14.143$, $\Delta_b = 12.711$, $\Delta_s = 7.492$, $\delta = 0.867$, and $\theta = 0.171$ in the second panel; and $\alpha = 0.498$, $\varepsilon_b = 19.749$, $\varepsilon_s = 23.274$, $u_b = 24.261$, $u_s = 39.520$, $\Delta_b = 41.558$, $\Delta_s = 34.536$, $\delta = 0.659$, and $\theta = 0.410$ in the third panel.

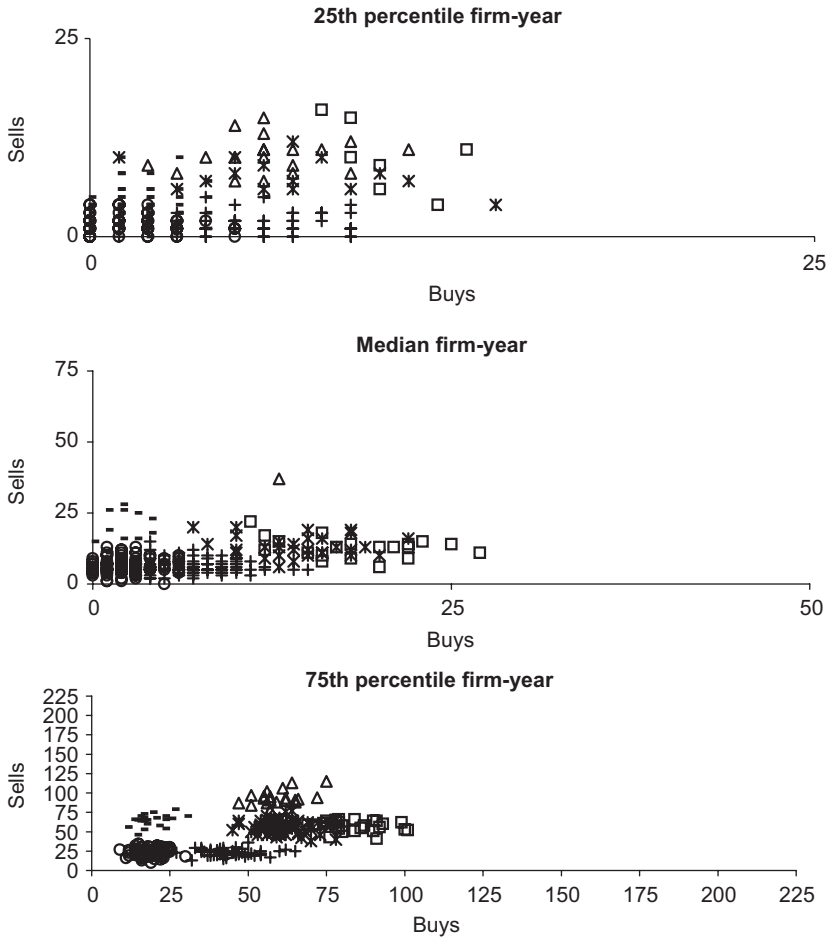


**Fig. 5.** Simulation of the preferred extended trading model. The plots in this figure are constructed by simulating daily buyer and seller initiated trades under the preferred extended model. Days with positive private information are represented by a positive sign (see the cloud of observations on the right bottom corner of each plot). Days with positive private information and symmetric order-flow shocks are represented by squares. Days with negative private information are represented by negative signs (see the cloud of observations on the left top corner of the each plot). Days with negative private information and symmetric order-flow shocks are represented by triangles. Days with symmetric order-flow shocks and no private information are represented by asterisks. Days with only noise trading are represented by circles (see the cloud of observations on the left bottom corner of each plot). To conduct the simulation, the preferred extended model is estimated using data from the firm-year with a correlation between buyer and seller initiated trades equal to the 25th percentile, the firm-year with the median correlation between buyer and seller initiated trades, and the firm-year with the 75th percentile correlation between buyer and seller initiated trades. The estimated parameters from these firm-years are then used to simulate the number of buyer and seller initiated trades under the preferred extended model for 240 hypothetical trading days. The top panel contains the 240 simulated trading days for the 25th percentile firm-year. The middle panel contains the simulated trading days for the median firm-year. The bottom panel contains the simulated trading days for the 75th percentile firm-year. The actual data for these firms years are graphed in Fig. 1.

quartile of *PIN*, *AdjPIN*, and *PSOS*. The graph indicates that *PIN* and *AdjPIN* are stable over time. Furthermore, the graph indicates that *AdjPIN*'s quartiles and its median are consistently lower than *PIN*'s quartiles and median. The median *PSOS* is also stable over time at about 0.2, though it appears that *PSOS* is somewhat more variable than *PIN* and *AdjPIN*, particularly in the third quartile.

The results in Table 5 show that the probability of trade due to symmetric order-flow shocks is non-trivial. The median estimated $\theta$ is 0.25, indicating that one in four days includes simultaneously elevated buy and sell order flow. In fact, the probability of a symmetric order-flow shock day is similar in magnitude to the probability of an information day ($\alpha$) for the median firm as well as the first and third quartiles. The median, first, and third quartiles of *PSOS* are also larger than the median, first, and third quartiles of *AdjPIN*. Thus, for most firms, trade related to positive order-flow shocks is more likely than trade related to private information on any given day.

### 3.2.3. *PIN, PSOS and Illiquidity*

Having examined the models' relative ability to match the data, we now contrast *PIN*, *AdjPIN*, and *PSOS*. Table 6 presents the correlations between *PIN*, *AdjPIN*, and *PSOS* and a number of other variables employed in this study. The correlation between *PIN* and *AdjPIN* is high, at around 0.71. The correlation between *PIN* and *PSOS* is also around 0.71. The correlation between *AdjPIN* and *PSOS* is not as high, at around 0.34. This reflects the fact that the *AdjPIN* model explicitly accounts for the possibility of symmetric

**Table 5**
Summary of preferred extended model estimation results.

This table presents the cross-sectional distribution of the estimated parameters of the preferred extended model along with the cross-sectional distribution of the estimated probability of information trading in the extended model (*AdjPIN*), and the cross-sectional distribution of the estimated probability of symmetric order-flow shock (*PSOS*). The preferred extended model is the extended model with the constraint $\theta' = \theta$. The sample consists of 48,512 firm-years between 1983 and 2004. The *t*-statistics for *PIN* and *PSOS* are calculated using the delta method and the asymptotic covariance matrix of the estimated model parameters.

| | 95th percentile | 75th percentile | Median | 25th percentile | 5th percentile |
|---|---|---|---|---|---|
| $a$ | 0.62 | 0.48 | 0.39 | 0.28 | 0.08 |
| $u_b$ | 153 | 32 | 13 | 5 | 1.0 |
| $u_s$ | 148 | 33 | 13 | 4 | 0.8 |
| $d$ | 0.98 | 0.78 | 0.56 | 0.32 | 0.03 |
| $\varepsilon_b$ | 212 | 21 | 5 | 1 | 0 |
| $\varepsilon_s$ | 185 | 23 | 6 | 1 | 0 |
| $\theta$ | 0.55 | 0.36 | 0.25 | 0.14 | 0.04 |
| $\Delta_b$ | 188 | 39 | 13 | 4 | 1.2 |
| $\Delta_s$ | 156 | 31 | 11 | 4 | 1.1 |
| *AdjPIN* | 0.37 | 0.23 | 0.17 | 0.13 | 0.08 |
| *t*-Statistic | 15.86 | 11.85 | 8.47 | 5.70 | 3.03 |
| *PSOS* | 0.66 | 0.38 | 0.25 | 0.18 | 0.12 |
| *t*-Statistic | 16.24 | 10.41 | 7.51 | 5.24 | 2.50 |

order-flow shocks, so the model is better able to discriminate between days with symmetric order-flow shocks and days that involve only abnormal buy or sell volume, which we identify as informed trade. The correlations between *AdjPIN* and size, and opening spreads (*SPREAD*) are consistent with the idea that *AdjPIN* is a proxy for the asymmetry of information. The correlation between *PSOS* and the Amihud illiquidity measure, *ILLIQ,* is also consistent with the interpretation that *PSOS* is a measure of illiquidity.

We explore the relation between *PSOS* and illiquidity further in Table 7. First, in each year $t$, we sort firms into deciles based on the firms' *PSOS* estimated in year $t-1$. Using the previously mentioned database of news events, we then compute the average number of buys, sells, the average product of buys and sells, the average volume (number of trades), the average net order flow imbalance, and the average value of the Amihud measure separately on news and non-news days. The results indicate that for the low *PSOS* deciles, the average number of buys and sells on news days is approximately two times larger than non-news days. By contrast, for the large *PSOS* deciles, the average number of buys and sells is around 20 times larger on news days than on non-news days. The average product of buys and sells for large *PSOS* firms is 74 times larger on news days versus non-news days. Furthermore, relative to the other *PSOS* deciles, the high *PSOS* firms experience much smaller volumes on non-news days than the lower *PSOS* deciles; only around six buys and five sells per day as opposed to upwards of 40 of each for lower *PSOS* deciles. This evidence indicates that the expected number of both buys and sells increases on news days for most firms, but the increase is particularly pronounced for high *PSOS* firms. The evidence in Table 7 also indicates that high *PSOS* firms have much higher Amihud measures than low *PSOS* firms. This is particularly true on non-news days. However, despite the fact that high *PSOS* firms have nearly twice the volumes on news days than the median firm, their Amihud measures on news days are still an order of magnitude larger than the median firm's Amihud measure. It is also worth noting that news days are much more infrequent for high *PSOS* firms than for low *PSOS* firms.[8] The average high *PSOS* firm has news around 6% of days, while a low *PSOS* firm has news on nearly 20% of days. Thus, the high *PSOS* firms tend to have much lower volume than the median firm on most days, punctuated occasionally by news days where the high *PSOS* firms have higher volume of both buys and sells than the median firm. Not surprisingly, these firms tend to be much less liquid, in terms of their Amihud measures, than low *PSOS* firms, since there is very little trade on most days.

The results in Table 7 also shed light on the effect of symmetric order-flow shocks on *PIN*. While high *PSOS* firms tend to have much larger *PINs* than other firms, they have only slightly larger *AdjPINs*. The reason for this lies with the symmetric order-flow shocks. For high *PSOS* firms, the periodic shocks associated with increased buy and sell volume are also associated with more volatile buy and sell volume. Thus, the increase in both buy and sell volume on days with symmetric order-flow shocks also leads to a higher probability of much larger imbalances than on a typical day. The *PIN* model is forced to identify these days as private information days while the *AdjPIN* model is not. Therefore, high *PSOS* firms have only slightly higher information asymmetry than low *PSOS* firms, as evidenced by their *AdjPIN* measures, but they have much larger estimated *PINs*. At the same time, the high *PSOS* firms also tend to be very illiquid, despite the fact that they have similar *AdjPINs* to other firms. These firms experience low levels of trade, compared to the median firm, on most days and consequently are much less liquid than other firms. Furthermore, even on news days, the high *PSOS* firms remain more illiquid than a typical firm. Therefore, *PSOS* appears to be strongly related to liquidity,

[8] Recall that *PSOS* measures the probability that a trade will come from a symmetric order-flow shock, not the absolute probability of a symmetric order-flow shock, which is often associated with public news events.

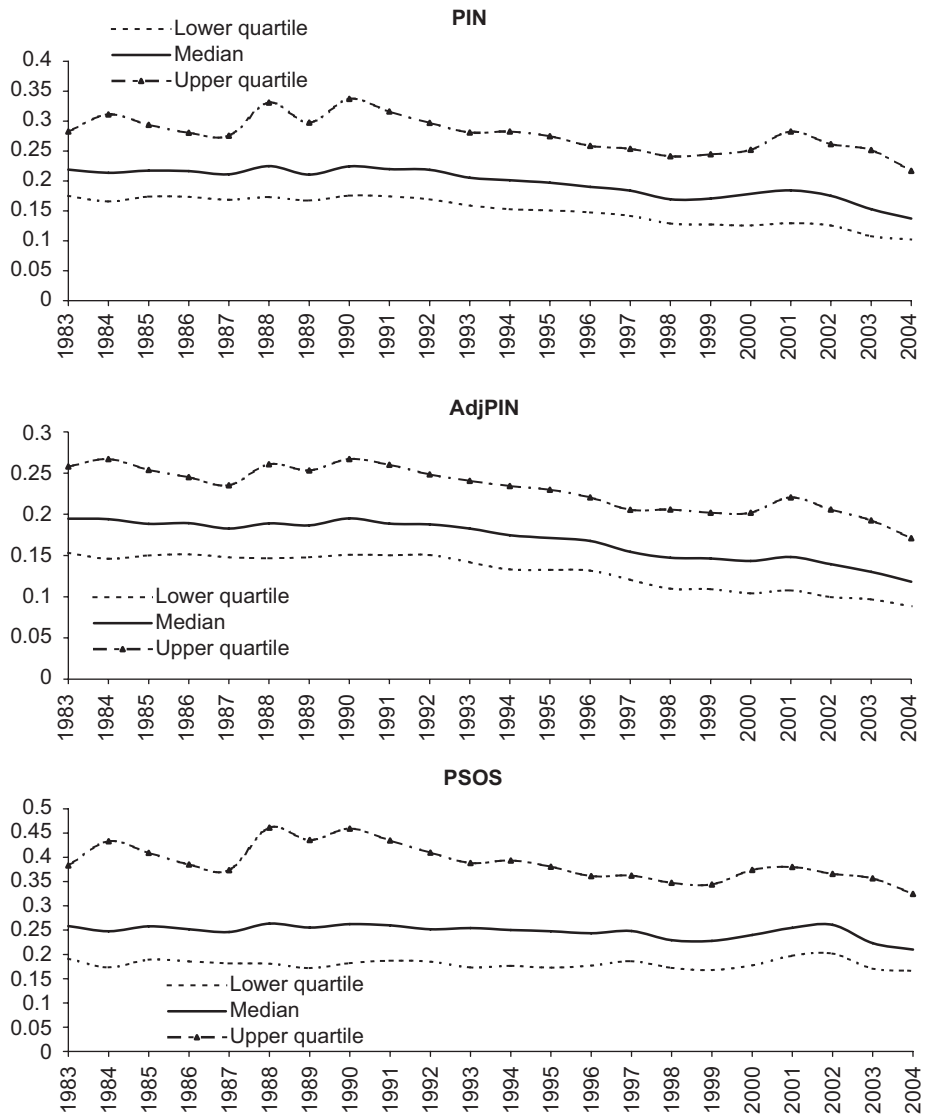


**Fig. 6.** Time series plots of *PIN*, *AdjPIN*, and *PSOS*. These figures plot the medians and quartiles of the cross-sectional distributions of *PIN* and *AdjPIN* over time. *PIN* is the probability of private information-related trade from the Easley, Hvidkjaer, and O'Hara (2002) model. *AdjPIN* is the probability of private information-related trade from the preferred extended model. *PSOS* is the probability that a given trade happens during a symmetric order-flow shock.

but for reasons apparently unrelated to information asymmetry and adverse selection.

Table 8 presents additional evidence that *PSOS* is a proxy for illiquidity. Table 8 presents coefficient estimates from annual Fama-MacBeth regressions of *PSOS* on various common proxies for illiquidity, including the natural logarithm of size, the natural logarithm of turnover, the Amihud measure, and the natural logarithm of the coefficient of variation of turnover, and opening spreads. We calculate size in year $t$ as the market value of the firm's equity at the end of December of $t-1$. We calculate turnover in year $t$ as the average daily turnover in years $t-1$ to $t-3$. The coefficient of variation of turnover in year $t$, *CV_TURNOVER*, is calculated as the coefficient of variation of daily turnover in years $t-1$ to $t-3$. We calculate opening spreads as the average, daily, opening relative spread in year $t-1$. Size and turnover are both measures of liquidity, so we expect that if *PSOS* is a proxy for illiquidity, the coefficients on size and turnover will be negative. The Amihud measure, the coefficient of variation of turnover, and bid–ask spreads are measures of illiquidity, so we expect that they will be positively related to *PSOS*. The coefficient estimates agree with these expectations in each case. *PSOS* is negatively related to size and turnover and positively related to the Amihud measure, the coefficient of variation of turnover, and spreads. Furthermore, *PSOS* is non-linearly related to each of these variables. Regression 2 in Table 8 shows the coefficients on each of the liquidity proxies squared. In each case, the squared liquidity proxies are significantly related to *PSOS*. Based on the results in Tables 7 and 8, we conclude that *PSOS* is strongly and non-linearly related to illiquidity.

**Table 6**
Correlation matrix.

The displayed correlation matrix is calculated using the cross-section of NYSE and Amex stocks for various variables of interest. *PIN* is a measure of probability of information-related trade estimated under the Easley, Hvidkjaer, and O'Hara (2002) structural microstructure model. *AdjPIN* is the probability of information-related trade estimated under the preferred extended model. *PSOS* is the probability of symmetric order-flow shocks under the preferred extended model. *PIN*, *AdjPIN*, and *PSOS* are estimated for each calendar year from 1983 to 2004. *SPREAD* in year $t$ is the average daily absolute open bid–ask spread in year $t-1$. *SIZE* is the log of the market value of firm equity from December of year $t-1$. *ILLIQ* is average daily Amihud (2002) illiquidity measure for year $t-1$. *P*-values are in parentheses.

| | *PIN* | *AdjPIN* | *PSOS* | *SPREAD* | *SIZE* | *BM* | *ILLIQ* |
|---|---|---|---|---|---|---|---|
| *PIN* | 1.0000 | 0.709 | 0.711 | 0.154 | −0.659 | 0.269 | 0.188 |
| | | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *AdjPIN* | | 1.0000 | 0.341 | 0.117 | −0.582 | 0.251 | 0.106 |
| | | | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| *PSOS* | | | 1.0000 | 0.151 | −0.578 | 0.182 | 0.178 |
| | | | | (0.000) | (0.000) | (0.000) | (0.000) |
| *SPREAD* | | | | 1.0000 | −0.214 | 0.100 | 0.190 |
| | | | | | (0.000) | (0.000) | (0.000) |
| *SIZE* | | | | | 1.0000 | −0.441 | −0.221 |
| | | | | | | (0.000) | (0.000) |
| *BM* | | | | | | 1.0000 | 0.129 |
| | | | | | | | (0.000) |
| *ILLIQ* | | | | | | | 1.0000 |

Our discussion so far suggests that symmetric order-flow shocks create differences between *AdjPIN* and *PIN*. Table 9 presents further evidence that suggests that the difference between *AdjPIN* and *PIN* is not random noise, but is related to liquidity. We explore cross-sectional differences between *PIN* and *AdjPIN* by running regressions of *AdjPIN* on *PIN* and regressions of the difference (*PIN* − *AdjPIN*) on variables related to liquidity. Table 9 presents the average difference between *PIN* and *AdjPIN* as well as the time-series average of the estimated coefficients from cross-sectional regressions of *AdjPIN* on *PIN*. This regression is useful in this context because if *PIN* is an unbiased estimate of *AdjPIN*, then the intercept in a regression of *AdjPIN* on *PIN* is zero, and the coefficient on *PIN* is equal to one. The average difference between *PIN* and *AdjPIN* is significantly positive, at around 0.04. Furthermore, the hypothesis that *PIN* is an unbiased estimator of *AdjPIN* is strongly rejected by the data. The regression's intercept is positive and highly significant. In addition, the coefficient on *PIN* is 0.52 which, while significantly different from zero, is also statistically different from one. The *t*-statistic for the test that the linear coefficient is equal to one is −49.8.

Table 9 also presents evidence that the difference between *PIN* and *AdjPIN* is related to liquidity. Table 9 presents averages of coefficient estimates from cross-sectional regressions of the difference between *PIN* and *AdjPIN* on a number of variables related to liquidity, including *PSOS*. The results suggest that the bias in *PIN* is strongly related to size, turnover, *PSOS*, and the Amihud measure. Furthermore, since both *PIN*, *AdjPIN*, and *PSOS* are non-linear functions of order flow, the bias in *PIN* could be non-linearly related to liquidity. The results in Table 9 indicate that this is indeed the case. The coefficients on squared size, turnover, and *PSOS* are large and highly significant. These results demonstrate that the difference between *PIN* and *AdjPIN* is systematically related to variables that proxy for liquidity. Thus, despite the strong positive correlation between the two measures of the probability of informed trade, positive shocks to both buy and sell order flow render *PIN* a biased measure of *AdjPIN*, and this bias is related to liquidity. In the following sections we show that this bias is important in interpreting the strong positive relation between *PIN* and average returns.

## 4. Asset pricing results

In the previous sections, we showed that the possibility of symmetric order-flow shocks creates a bias in the *PIN* measure and that despite having about average *AdjPINs*, high *PSOS* firms tend to have high estimated *PINs*. At the same time, high *PSOS* firms tend to be very illiquid. All of this suggests that cross-sectional variation in *PIN* may, in part, capture liquidity effects that are unrelated to information asymmetry. Given that a number of recent papers such as Amihud (2002), Acharya and Pedersen (2005), Sadka (2006), and Pástor and Stambaugh (2003) show that liquidity is related to average stock returns, the above evidence suggests *PIN* may be related to average returns not because it is a proxy for information asymmetry per se, but because it is a proxy for illiquidity unrelated to information asymmetry. In this section, we investigate this possibility by replicating the asset pricing tests in Easley, Hvidkjaer, and O'Hara (2002) to examine whether, after accounting for the possibility of simultaneous shocks to both buy and sell order flow, the positive relation between informed trade and average stock returns remains.

To this end, Table 10 presents firm-level Fama-MacBeth estimates of regressions of returns on $\beta$, size, book-to-market ratios (*BM*), and *PIN*, as well as regressions that include *AdjPIN* in place of *PIN*. These regressions are

**Table 7**
Trading activity on days with and without news.

This table presents the means and standard deviations of a series of variables related to trading activity on days with news and on days without news. *PSOS* is the probability of symmetric order-flow shocks estimated under the preferred extended model. *PIN* is a measure of probability of information-related trade estimated under the Easley, Hvidkjaer, and O'Hara (2002) structural microstructure model. *AdjPIN* is the probability of information-related trade estimated under the preferred extended model. *PIN*, *AdjPIN*, and *PSOS* are estimated for each calendar year from 1983 to 2004. *BUYS* is the number of buyer initiated trades on a given day. *SELLS* is the number of seller initiated trades. *IMBALANCE* is the difference between *BUYS* and *SELLS*. *TRADES* is total number of trades in a day. *ILLIQ* is the Amihud (2002) illiquidity measure. The sample includes all NYSE/Amex firms in the Chan (2003) database from 1983 to 2000, or between 379 and 670 firms per year. The Chan (2003) database is a random sample of firms representing 20–20% of all firms in the CRSP database between 1980 and 2000. Days with news are days in which a given firm is mentioned at least once in the publications covered by Dow Jones Interactive Publications Library. Stocks are sorted on their *PSOS* and the means and standard deviations of the trading related variables are calculated for each *PSOS* decile. Standard deviations are in parentheses.

| *PSOS* decile | *PSOS* | *PIN* | *Adj.PIN* | *BUYS* | *SELLS* | *BUYS* × *SELLS* | *IMBALANCE* | *TRADES* | *ILLIQ* × 100 |
|---|---|---|---|---|---|---|---|---|---|
| *Panel A: Days with news* | | | | | | | | | |
| 1 | 0.11 | 0.15 | 0.16 | 96 | 88 | 40,140 | 8 | 183 | 0.002 |
| | (0.03) | (0.05) | (0.07) | (200) | (166) | (2,541,018) | (62) | (362) | (0.132) |
| 2 | 0.16 | 0.16 | 0.16 | 97 | 87 | 25,026 | 10 | 184 | 0.001 |
| | (0.01) | (0.05) | (0.07) | (145) | (122) | (117,135) | (53) | (263) | (0.021) |
| 3 | 0.18 | 0.17 | 0.16 | 77 | 68 | 19,068 | 9 | 145 | 0.001 |
| | (0.01) | (0.05) | (0.06) | (135) | (108) | (100,596) | (47) | (240) | (0.012) |
| 4 | 0.21 | 0.18 | 0.16 | 66 | 60 | 14,559 | 7 | 126 | 0.002 |
| | (0.01) | (0.05) | (0.06) | (116) | (97) | (335,802) | (42) | (210) | (0.059) |
| 5 | 0.24 | 0.20 | 0.18 | 76 | 67 | 131,661 | 9 | 144 | 0.006 |
| | (0.01) | (0.07) | (0.09) | (410) | (317) | (5,392,776) | (125) | (722) | (0.220) |
| 6 | 0.28 | 0.22 | 0.19 | 105 | 89 | 286,929 | 16 | 195 | 0.004 |
| | (0.01) | (0.07) | (0.08) | (614) | (465) | (6,996,260) | (197) | (1,072) | (0.082) |
| 7 | 0.33 | 0.24 | 0.21 | 31 | 27 | 5,357 | 4 | 59 | 0.007 |
| | (0.02) | (0.08) | (0.10) | (76) | (63) | (92,490) | (27) | (137) | (0.113) |
| 8 | 0.40 | 0.31 | 0.25 | 57 | 49 | 22,776 | 9 | 106 | 0.021 |
| | (0.03) | (0.12) | (0.12) | (167) | (128) | (257,840) | (66) | (290) | (0.396) |
| 9 | 0.52 | 0.38 | 0.29 | 84 | 71 | 323,856 | 13 | 155 | 0.067 |
| | (0.04) | (0.14) | (0.11) | (598) | (538) | (6,579,718) | (107) | (1,133) | (1.005) |
| 10 | 0.69 | 0.41 | 0.22 | 133 | 111 | 74,741 | 22 | 244 | 0.089 |
| | (0.09) | (0.14) | (0.09) | (273) | (224) | (260,526) | (71) | (495) | (1.637) |
| *Panel B: Days without news* | | | | | | | | | |
| 1 | 0.11 | 0.15 | 0.16 | 36 | 35 | 4,665 | 1 | 71 | 0.006 |
| | (0.03) | (0.05) | (0.07) | (64) | (58) | (46,902) | (26) | (119) | (0.173) |
| 2 | 0.16 | 0.16 | 0.16 | 41 | 38 | 6,005 | 3 | 80 | 0.004 |
| | (0.01) | (0.05) | (0.07) | (75) | (63) | (54,070) | (27) | (136) | (0.064) |
| 3 | 0.18 | 0.17 | 0.16 | 32 | 30 | 3,246 | 2 | 62 | 0.003 |
| | (0.01) | (0.05) | (0.06) | (54) | (46) | (20,743) | (22) | (98) | (0.060) |
| 4 | 0.21 | 0.18 | 0.16 | 29 | 27 | 2,925 | 2 | 55 | 0.008 |
| | (0.01) | (0.05) | (0.06) | (52) | (45) | (26,495) | (22) | (96) | (0.50) |
| 5 | 0.24 | 0.20 | 0.18 | 27 | 25 | 22,766 | 2 | 52 | 0.016 |
| | (0.01) | (0.07) | (0.09) | (164) | (140) | (1,090,671) | (48) | (301) | (0.678) |
| 6 | 0.28 | 0.22 | 0.19 | 41 | 39 | 152,580 | 2 | 79 | 0.016 |
| | (0.01) | (0.07) | (0.08) | (414) | (372) | (5,240,721) | (89) | (782) | (0.40) |
| 7 | 0.33 | 0.24 | 0.21 | 12 | 12 | 2,226 | 1 | 24 | 0.023 |
| | (0.02) | (0.08) | (0.10) | (49) | (46) | (141,487) | (17) | (93) | (0.453) |
| 8 | 0.40 | 0.31 | 0.25 | 9 | 8 | 1,339 | 1 | 18 | 0.072 |
| | (0.03) | (0.12) | (0.12) | (40) | (33) | (28,052) | (15) | (73) | (1.622) |
| 9 | 0.52 | 0.38 | 0.29 | 8 | 8 | 18,208 | 1 | 16 | 0.121 |
| | (0.04) | (0.14) | (0.11) | (143) | (128) | (1,197,422) | (26) | (271) | (2.313) |
| 10 | 0.69 | 0.41 | 0.22 | 6 | 5 | 1,621 | 1 | 12 | 0.248 |
| | (0.09) | (0.14) | (0.09) | (44) | (37) | (30,182) | (11) | (80) | (5.334) |

similar to those found in Easley, Hvidkjaer, and O'Hara (2002).[9] Discussion of the computation of $\beta$, size, and *BM* can be found in Section 2. As documented in Easley, Hvidkjaer, and O'Hara (2002), the coefficient on *PIN* is large, at about 100 basis points per month, and is significant at conventional levels. However, when *AdjPIN* is substituted for *PIN*, the result changes dramatically. The coefficient on *AdjPIN* is economically small, at only 13 basis points per month, and is insignificant. Furthermore, when *PIN* and *AdjPIN* are included in the same regression, the coefficient on *PIN* remains large, positive, and significant, while the coefficient on *AdjPIN* is insignificant and negative. On the other hand, when *PSOS* is included in the regression alone it has a positive and significant coefficient. However, when both *PSOS* and *PIN* are included in the regression, the coefficients on both

[9] Note that Easley, Hvidkjaer, and O'Hara (2002) also present coefficient estimates from regressions that include additional control variables such as volume, the coefficient of variation of volume, spreads, and return volatility. Our results are robust to the inclusion of these variables.

**Table 8**
PSOS and liquidity.

This table contains time-series averages of the estimated coefficients from annual cross-sectional regressions of *PSOS* on stock characteristics related to liquidity. The regressions are run for each year from 1984 to 2004. There are between 1,542 and 2,099 firms in the regressions depending on the year. *PSOS* is the probability of symmetric order-flow shocks estimated under the preferred extended model. *PSOS* is estimated for each calendar year from 1983 to 2004. *SIZE* in year $t$ is the logarithm of the December market equity for year $t-1$. *TURNOVER* in year $t$ is the average daily turnover in years $t-1$ to $t-3$. *CVTURNOVER* in year $t$ is the coefficient of variation of daily turnover in years $t-1$ to $t-3$. *ILLIQ* is the Amihud (2002) illiquidity measure. *SPREAD* in year $t$ is the average of the daily absolute open bid–ask spread in year $t-1$. The $t$-statistics are in parentheses.

| Explanatory variables | (1) | (2) |
|---|---|---|
| Intercept | 0.5779 | 2.0343 |
| | (9.771) | (15.854) |
| *SIZE* | −0.0355 | −0.2228 |
| | (−9.839) | (−16.507) |
| *TURNOVER* | −0.0457 | −0.0585 |
| | (−19.196) | (−19.717) |
| *ILLIQ* × 100 | 0.0014 | 0.0017 |
| | (2.536) | (2.210) |
| *CV_TURNOVER* | 0.0291 | −0.0856 |
| | (6.881) | (−3.112) |
| *SPREAD* | 0.0072 | −0.0001 |
| | (4.421) | (−0.076) |
| *SIZE2* | | 0.0072 |
| | | (14.458) |
| *TURNOVER2* | | 0.0174 |
| | | (10.637) |
| *ILLIQ2* × 106 | | −0.0023 |
| | | (−2.758) |
| *CV_TURNOVER2* | | 0.0125 |
| | | (4.553) |
| *SPREAD2* | | 0.0001 |
| | | (0.941) |

**Table 9**
Factors affecting the difference between *PIN* and *AdjPIN*.

This table contains time-series averages of the estimated coefficients in annual cross-sectional regressions of *PIN* − *AdjPIN* and *AdjPIN* on a series of stock characteristics. The regressions are run for each year from 1984 to 2004. There are between 1,542 and 2,099 firms in the regressions depending on the year. *PSOS* is the probability of symmetric order-flow shocks estimated under the preferred extended model. *PIN* is a measure of probability of information-related trade estimated under the Easley, Hvidkjaer, and O'Hara (2002) structural microstructure model. *AdjPIN* is the probability of information-related trade estimated under the preferred extended model. *PIN*, *AdjPIN*, and *PSOS* are estimated for each firm-year from 1983 to 2004. *SIZE* is the logarithm of the December market equity for year $t-1$. *TURNOVER* in year $t$ is the logarithm of the average daily turnover in years $t-1$ to $t-3$. The $t$-statistics are in parentheses.

| Explanatory variables | Dependent variables | | | |
|---|---|---|---|---|
| | *PIN* − *AdjPIN* | *AdjPIN* | *PIN* − *AdjPIN* | |
| Intercept | 0.0412 | 0.0698 | −0.0313 | 0.0965 |
| | (35.202) | (19.184) | (−25.168) | (3.584) |
| *PIN* | | 0.5205 | | |
| | | (49.891) | | |
| *PSOS* | | | 0.2769 | −0.0894 |
| | | | (41.511) | (−5.402) |
| *SIZE* | | | | −0.0116 |
| | | | | (−3.110) |
| *TURNOVER* | | | | −0.0149 |
| | | | | (−11.486) |
| *ILLIQ* × 100 | | | | 0.0002 |
| | | | | (2.163) |
| $PSOS^2$ | | | | 0.4311 |
| | | | | (23.982) |
| $SIZE^2$ | | | | 0.0004 |
| | | | | (3.288) |
| $TURNOVER^2$ | | | | 0.0073 |
| | | | | (11.393) |

measures drop considerably and both are rendered insignificant. Table 10 also shows that when the difference between *AdjPIN* and *PIN* is included in the regression, it has a negative and significant coefficient. However, when the difference is in the presence of *PIN*, the coefficients on both drop dramatically and are no longer significant. This indicates that the bias in *PIN* is important for expected returns, and that this bias, rather than information asymmetry, is behind the relation between *PIN* and average returns.

In order to further examine the possibility that *PIN* is priced for liquidity reasons unrelated to the adverse selection problem created by informed trade, Table 10 also presents Fama-MacBeth regressions that include the Amihud measure. The Amihud measure is significantly positively related to average returns. However, in the presence of the Amihud measure, the coefficient on *PIN* drops substantially and is insignificantly different from zero. Similarly, when *PSOS* is included in the regression along with the Amihud measure, the coefficient on *PSOS* is reduced considerably and rendered insignificant. Overall, the evidence described above suggests that *PIN* is priced for reasons unrelated to informed trade, such as market makers' aversion to holding inventories. Naturally, the evidence in Table 10 does not address the open question in the literature of why liquidity concerns would matter for the cross-section of expected returns. (See, for instance, Constantinides, 1986.) Instead, the results in Table 10 show that if private information is identified by extreme order imbalances, then *PIN* is priced because it is a proxy for liquidity concerns unrelated to private information.

## 5. Conclusion

This study presents an extended version of the Easley, Kiefer, O'Hara, and Paperman (1996) structural microstructure model that allows for the possibility of symmetric order-flow shocks. We show that the extended model is better able to capture the variances of buys and sells as well as the large positive contemporaneous correlation between buys and sells observed in the data. We further show that the presence of symmetric order-flow shocks creates differences between the *PIN* measure and the analogous measure from the extended model, *AdjPIN*. While the two measures are correlated, *PIN* is a biased measure of *AdjPIN* and the bias is related to the tendency of a stock to have simultaneous positive shocks to both buy and sell order flow. Moreover, we show that high *PSOS* firms have high *PINs*, but only slightly above average levels of information asymmetry as measured by *AdjPIN*. At the same time, high *PSOS* firms are also very

**Table 10**
*PIN* and the cross-section of expected returns.

This table contains time-series averages of the estimated coefficients from monthly, firm-level cross-sectional regressions, 1984–2004. There are between 1,295 and 1,786 firms in the sample depending on the month. The dependent variable is the monthly return. *Beta* is post-ranking beta estimated using 40 portfolios. *SIZE* is the logarithm of the December market equity for year $t-1$, *BM* is the logarithm of book value divided by market value for year $t-1$. *PSOS* is the probability of symmetric order-flow shocks estimated under the preferred extended model. *PIN* is a measure of probability of information-related trade estimated under the Easley, Hvidkjaer, and O'Hara (2002) structural microstructure model. *AdjPIN* is the probability of information-related trade estimated under the preferred extended model. *PIN*, *AdjPIN*, and *PSOS* are estimated for each calendar year from 1983 to 2004. *ILLIQ* is the Amihud (2002) illiquidity measure for year $t-1$. The intercepts in the regressions are not reported. The *t*-statistics are in parentheses.

| Explanatory variables | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) | (12) | (13) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Beta* | 0.1745 | 0.1397 | 0.1775 | 0.1929 | 0.1926 | 0.1391 | 0.1486 | 0.1529 | 0.1532 | 0.1547 | 0.1775 | 0.1809 | 0.1427 |
| | (0.494) | (0.388) | (0.506) | (0.520) | (0.546) | (0.369) | (0.425) | (0.441) | (0.418) | (0.444) | (0.506) | (0.483) | (0.414) |
| *SIZE* | 0.0430 | 0.0043 | 0.0354 | 0.0372 | 0.0487 | 0.0647 | 0.0881 | 0.0883 | 0.0822 | 0.0902 | 0.0354 | 0.0186 | 0.0835 |
| | (0.630) | (0.064) | (0.501) | (0.639) | (0.733) | (1.216) | (1.379) | (1.337) | (1.497) | (1.436) | (0.501) | (0.324) | (1.255) |
| *BM* | 0.2675 | 0.2558 | 0.2581 | 0.2661 | 0.2645 | 0.2554 | 0.2542 | 0.2443 | 0.2457 | 0.2451 | 0.2581 | 0.2613 | 0.2461 |
| | (2.956) | (2.824) | (2.869) | (2.940) | (2.953) | (2.802) | (2.824) | (2.732) | (2.737) | (2.758) | (2.869) | (2.869) | (2.750) |
| *PIN* | 1.0043 | | 1.3650 | | 0.7455 | | 0.6481 | 0.6277 | | 0.6453 | 0.6108 | | 0.5224 |
| | (1.911) | | (2.875) | | (1.141) | | (1.168) | (1.266) | | (0.957) | (0.785) | | (0.616) |
| *AdjPIN* | | 0.1324 | −0.7541 | | | | | 0.0147 | | | | | |
| | | (0.182) | (−1.092) | | | | | (0.024) | | | | | |
| *PSOS* | | | | 0.6448 | 0.3731 | | | | 0.2455 | 0.0199 | | | |
| | | | | (2.587) | (1.215) | | | | (0.998) | (0.071) | | | |
| *ILLIQ* | | | | | | 0.0003 | 0.0003 | 0.0003 | 0.0003 | 0.0003 | | | 0.0003 |
| | | | | | | (2.989) | (2.696) | (2.543) | (2.738) | (2.581) | | | (2.583) |
| *AdjPIN-PIN* | | | | | | | | | | | −0.7541 | −1.1302 | −0.0290 |
| | | | | | | | | | | | (−1.092) | (−2.650) | (−0.047) |

illiquid, with very little trading volume, relative to the median *PSOS* firm, on most days. Most importantly, we demonstrate that while *PIN* appears to be related to average returns, the *AdjPIN* is not. We present evidence that the cross-sectional variation in *PIN* that is important for average returns relates to *PSOS* as well as to the Amihud measure.

Assuming that periods of private information can be identified by periods of abnormal order flow imbalance as motivated by sequential trade models such as Glosten and Milgrom (1985) and as assumed in empirical work such as Easley, Kiefer, O'Hara, and Paperman (1996), we conclude from our evidence that information-based trading does not affect expected stock returns. Instead we interpret our evidence as indicating that while information asymmetry does not appear to be related to average returns, microstructure and liquidity effects unrelated to information asymmetry are still important for expected returns. It is worth noting that this interpretation relies on the assumption that periods of asymmetric information can be identified as periods with abnormal order flow imbalances. It is possible the relation between private information and order flow is more complex than the one implied by sequential trade models, in which case private information could indeed be related to expected returns. However, in this case, both *PIN* and *AdjPIN* would be inappropriate proxies for information asymmetry. We leave further exploration of this matter to future research.

## Appendix A. Database construction

Quotes are limited from 8:30 am to 3:58 pm. We only consider quotes with non-zero bid price, non-zero ask price, and with ask price greater than bid price. We exclude quotes with large differences in bid and ask price. More specifically, we exclude a quote if the mid-point of the bid–ask spread falls within $5–$50 and the ratio of the bid–ask spread to the sum of the bid and ask price is greater than 0.25. We also exclude a quote if the mid-point of the bid–ask spread is greater than $50 and the ratio of the bid–ask spread to the sum of the bid and ask price is greater than 0.1.

We only consider trades with positive price. We only consider regular trades that were not corrected, changed, or signified as cancelled or in error. We only consider trades made without any stated conditions. In particular, trades under the following stated conditions are excluded: split trade, pre- and post-market close trades, average price trades, sold sale, crossing session, etc. The trades and quotes data are matched together by symbol and time. To match the quotes with trades, we use the five-second quote rule, i.e., compare the trade with quotes that are at least five seconds older. The stock symbol from ISSM/TAQ is then matched with the CRSP symbol to get CRSP PERMNO for each stock. For a CRSP observation where the trading volume is zero, we assign buys and sells a value of zero on that firm-day. We only consider common stocks. We exclude ADRs, REITs, and closed-end funds.

## Appendix B. Expected value, variance, and correlations of buys and sells

In this appendix, we sketch the proof of the formulas used to calculate the expected value of buys and sells, their variances, and covariances. The expected value of the number of buys in the original *PIN* model is equal to the sum of the probabilities of each branch of the tree in Fig. 2 multiplied by the intensity of the buy trades arrival

in the branch:

$$\mathrm{E}[B] = (1-a)\times\varepsilon_b + a\times(1-d)\times\varepsilon_b + a\times d\times(\varepsilon_b+u). \tag{10}$$

Analogously, the expected value of the seller initiated trades in the original *PIN* model is

$$\mathrm{E}[S] = (1-a)\times\varepsilon_s + a\times d\times\varepsilon_s + a\times(1-d)\times(\varepsilon_s+u). \tag{11}$$

The variances of the number of buys is $\sigma^2[B] = \mathrm{E}[B^2] - \mathrm{E}[B]^2$, calculating $\mathrm{E}[B^2]$ in the same way that we calculate $\mathrm{E}[B]$ above and using the expression above for $\mathrm{E}[B]$ we get

$$\begin{aligned}\sigma^2[B] = {} & \varepsilon_b^2 + a\times d\times u\times(1+u) + \varepsilon_b\times(1+2\times a\times d\times u)\\ & - ((1-a)\times\varepsilon_b + a\times(1-d)\times\varepsilon_b + a\times d\times(\varepsilon_b+u))^2.\end{aligned} \tag{12}$$

The variance of the seller initiated trades can be calculated in the same way and it is

$$\begin{aligned}\sigma^2[S] = {} & \varepsilon_s^2 - a\times(-1+d)\times u\times(1+u) + \varepsilon_s\times(1-2\times a\\ & \times(-1+d)\times u) - ((1-a)\times\varepsilon_s + a\times d\times\varepsilon_s\\ & + a\times(1-d)\times(\varepsilon_s+u))^2.\end{aligned} \tag{13}$$

The covariance between buys and sells is $cov[B,S] = \mathrm{E}[B\times S] - \mathrm{E}[B]\times\mathrm{E}[S]$, calculating the expected value $\mathrm{E}[B\times S]$ in the same way as above and using the above expressions for $\mathrm{E}[B]$ and $\mathrm{E}[S]$ we get

$$cov[B,S] = (au)^2\times(d-1)\times d. \tag{14}$$

The method to calculate the expected value of buys and sells, their variances, and covariances is analogous to the one used in the original *PIN* model above. These expected values, variances, and covariances in the unrestricted model are

$$\mathrm{E}[B] = \varepsilon_b + \Delta_b\times(\theta - a\times\theta + a\times\theta') + a\times d\times u_b, \tag{15}$$

$$\mathrm{E}[S] = \varepsilon_s + \Delta_s\times(\theta - a\times\theta + a\times\theta') - a\times(-1+d)\times u_s, \tag{16}$$

$$\begin{aligned}\sigma^2[B] = {} & \varepsilon_b - \Delta_b^2\times((-1+a)^2\times\theta^2 + a\times\theta'\times(-1+a\times\theta')\\ & - (-1+a)\times\theta\times(-1+2\times a\times\theta'))\\ & + a\times d\times u_b\times(1+u_b - a\times d\times u_b)\\ & + \Delta_b\times(a\times\theta'\times(1-2\times(-1+a)\times d\times u_b)\\ & + (-1+a)\times\theta\times(-1+2\times a\times d\times u_b)),\end{aligned} \tag{17}$$

$$\begin{aligned}\sigma^2[S] = {} & \varepsilon_s - \Delta_s^2\times((-1+a)^2\times\theta^2 + a\times\theta'\times(-1+a\times\theta')\\ & - (-1+a)\times\theta\times(-1+2\times a\times\theta'))\\ & - a\times(-1+d)\times u_s\times(1+(1+a\times(-1+d))\times u_s)\\ & - \Delta_s\times((-1+a)\times\theta\times(1+2\times a\times(-1+d)\times u_s)\\ & + a\times\theta'\times(-1+2\times(-1+a+d-a\times d)\times u_s)),\end{aligned} \tag{18}$$

$$\begin{aligned}cov[B,S] = {} & (-a)\times d\times u_b\times((-(-1+a))\times\Delta_s\times(\theta-\theta')\\ & - a\times(-1+d)\times u_s)\\ & - \Delta_b\times(\Delta_s\times((-1+a)^2\times\theta^2 + a\times\theta'\times(-1+a\times\theta')\\ & - (-1+a)\times\theta\times(-1+2\times a\times\theta'))\\ & + (-1+a)\times a\times(-1+d)\times(\theta-\theta')\times u_s).\end{aligned} \tag{19}$$

The expected values of the buys and sells, their variances, and covariances in the restricted extended models follow by substituting the restrictions in the expressions above. For instance, substituting the restriction $\theta = \theta'$ in the expressions above, we arrive at the formulas for the expected values, variances, and covariances of the buys and sells in the preferred extended model:

$$\mathrm{E}[B] = \varepsilon_b + \Delta_b\times\theta + a\times d\times u_b, \tag{20}$$

$$\mathrm{E}[S] = \varepsilon_s + \Delta_s\times\theta - a\times(-1+d)\times u_s, \tag{21}$$

$$\begin{aligned}\sigma^2[B] = {} & \varepsilon_b + \Delta_b\times\theta - \Delta_b^2\times(-1+\theta)\times\theta + a\times d\\ & \times u_b\times(1+u_b - a\times d\times u_b),\end{aligned} \tag{22}$$

$$\begin{aligned}\sigma^2[S] = {} & \varepsilon_s + \Delta_s\times\theta - \Delta_s^2\times(-1+\theta)\times\theta - a\times(-1+d)\times u_s\\ & \times(1+u_s - a\times u_s + a\times d\times u_s),\end{aligned} \tag{23}$$

$$\begin{aligned}cov[B,S] = {} & (-\Delta_b)\times\Delta_s\times(-1+\theta)\times\theta + a^2\\ & \times(-1+d)\times d\times u_b\times u_s.\end{aligned} \tag{24}$$

## References


Acharya, V., Pedersen, L., 2005. Asset pricing with liquidity risk. Journal of Financial Economics 77, 375–410.

Admati, A., Pfleiderer, P., 1988. A theory of intraday patterns: volume and price variability. Review of Financial Studies 1, 3–40.

Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

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

Bachelier, L., 1900. Theorie de la Speculation (Thesis). Annales Scientifiques de l'École Normale Superieure III 17, 21–86 (English translation in Cootner, P., 1964. Random Character of Stock Market Prices. MIT Press, Cambridge, MA, pp. 17–78).

Bamber, L., Barron, O., Stober, T., 1999. Differential interpretations and trading volume. Journal of Financial and Quantitative Analysis 34, 369–385.

Bernhardt, D., Hughson, E., 2002. Intraday trade in dealership markets. European Economic Review 46, 1697–1732.

Bessembinder, H., Chan, K., Seguin, P., 1996. An empirical examination of information, differences of opinion, and trading activity. Journal of Financial Economics 40, 105–134.

Boehmer, E., Grammig, J., Theissen, E., 2007. Estimating the probability of informed trading—Does trade misclassification matter? Journal of Financial Markets 10, 26–47.

Chan, W., 2003. Stock price reaction to news and no-news: drift and reversal after headlines. Journal of Financial Economics 70, 223–260.

Constantinides, G., 1986. Capital market equilibrium with transaction costs. Journal of Political Economy 94, 842–862.

Duarte, J., Han, X., Harford, J., Young, L., 2008. Information asymmetry, information dissemination and the effect of regulation FD on the cost of capital. Journal of Financial Economics 87, 24–44.

Easley, D., O'Hara, M., 1987. Price, trade size, and information in securities markets. Journal of Financial Economics 19, 69–90.

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

Easley, D., Kiefer, N., O'Hara, M., Paperman, J., 1996. Liquidity, information, and infrequently traded stocks. Journal of Finance 51, 1405–1436.

Easley, D., Kiefer, N., O'Hara, M., 1997. One day in the life of a very common stock. Review of Financial Studies 10, 805–835.

Easley, D., Hvidkjaer, S., O'Hara, M., 2002. Is Information risk a determinant of asset returns? Journal of Finance 57, 2185–2221.

Fama, E., MacBeth, J., 1973. Risk, return, and equilibrium: empirical tests. Journal of Political Economy 81, 607–636.

Gârleanu, N., Pedersen, L., 2004. Adverse selection and the required return. Review of Financial Studies 17, 643–665.

Glosten, L., Milgrom, P., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. Journal of Financial Economics 13, 71–100.
Grossman, S.J., Miller, M., 1988. Liqudity and market structure. Journal of Finance 43, 617–637.
Harris, M., Raviv, A., 1993. Differences of opinion make a horse race. Review of Financial Studies 6, 473–506.
Hughes, J., Liu, J., Liu, J., 2007. Information asymmetry, diversification, and asset pricing. The Accounting Review 82, 705–729.
Kandel, E., Pearson, N., 1995. Differential interpretation of public signals and trade in speculative markets. Journal of Political Economy 103, 831–872.
Lambert, R., Leuz, C., Verrecchia, R., 2005. Accounting information, disclosure, and the cost of capital. Unpublished working paper, University of Pennsylvania, Philadelphia, PA.
Lee, C., Ready, M., 1991. Inferring trade direction from intraday data. Journal of Finance 46, 733–746.
Pástor, L., Stambaugh, R., 2003. Liquidity risk and expected stock returns. Journal of Political Economy 111, 642–685.
Sadka, R., 2006. Momentum and post-earnings-announcement drift anomalies: the role of liquidity risk. Journal of Financial Economics 80, 309–349.
Sarkar, A., Schwartz, R., 2007. Market sidedness: insights into motives for trade initiation. Unpublished working paper, Federal Reserve Bank of New York, New York, NY.
Vega, C., 2006. Stock price reaction to public and private information. Journal of Financial Economics 82, 103–133.
Venter, J., de Jongh, D., 2004. Extending the EKOP model to estimate the probability of informed trading. Unpublished working paper, North-West University Center for Business Mathematics and Informatics, Potchefstroom, South Africa.