Identifying Information Asymmetry in Securities Markets☆

Kerry Back

Jones Graduate School of Business and Department of Economics
Rice University, Houston, TX 77005, U.S.A.

Kevin Crotty

Jones Graduate School of Business
Rice University, Houston, TX 77005, U.S.A.

Tao Li

Department of Economics and Finance
City University of Hong Kong, Kowloon, Hong Kong

Abstract

We propose and estimate a model of endogenous informed trading that is a hybrid of the PIN and Kyle models. When an informed trader trades optimally, both returns and order flows are needed to identify information asymmetry parameters. Empirically, relationships between the model's estimates and price impacts, excess kurtosis, and volatility are consistent with theory. The estimates can be used to detect information events in the time series and to characterize the information content of prices in the cross section. Relative to price impact benchmarks, a composite measure from the model compares favorably to those from other structural information asymmetry models.

☆Versions of this paper were presented under various titles at the University of Colorado, the SEC, the NYU Stern Microstructure Conference, the University of Chicago Market Microstructure and High Frequency Data Conference, the ASU Sonoran Winter Finance Conference, and the UBC Winter Finance Conference. We thank Pete Kyle, Rob Engle, Dmitry Livdan, and seminar participants for helpful comments, and we thank Slava Fos for helpful comments and for sharing his data on trading by Schedule 13D filers.

Email addresses: `Kerry.E.Back@rice.edu` (Kerry Back), `Kevin.P.Crotty@rice.edu` (Kevin Crotty), `TaoLi3@cityu.edu.hk` (Tao Li)

December 21, 2016

1. Introduction

Information asymmetry is a fundamental concept in economics, but its estimation is challenging because private information is generally unobservable. Many proxies for information asymmetry exist including bid/ask spreads, price impacts, and estimates from structural models. In this paper, we study identification of information asymmetry parameters in structural models. Structural modeling allows the econometrician to capture parameters related to the underlying economic mechanisms such as the probability and magnitude of private information events or the intensity of noise trading. Demand for plausible measures of information asymmetry is high because private information plays a key role in so many economic settings. Evidence of this demand is the large literature in finance and accounting that utilizes the probability of informed trade (PIN) measure of Easley, Kiefer, O'Hara and Paperman (1996) to proxy for information asymmetry.[1]

Our first contribution is to propose and solve a model of informed trading in securities markets that shares many features of the PIN model of Easley et al. (1996) but in which informed trading is endogenous as in Kyle (1985). We call this a hybrid PIN-Kyle model. In the paper, we study a binary signal following Easley et al. (1996), but the model can accommodate more general signal distributions.[2]

An important implication of the model is that order flows alone cannot identify information asymmetry. The intuition is quite simple. Consider, for example, a stock for which there is a large amount of private information and another for which there is only a small

[1]Some of those papers assesses whether information risk is priced. See, for example, Easley and O'Hara (2004), Duarte and Young (2009), Mohanram and Rajgopal (2009), Easley, Hvidkjaer and O'Hara (2002), Easley, Hvidkjaer and O'Hara (2010), Akins, Ng and Verdi (2012), Li, Wang, Wu and He (2009), and Hwang, Lee, Lim and Park (2013). Many other papers use PIN (and other measures) to capture a firm's information environment in a variety of applications ranging from corporate finance (e.g., Chen, Goldstein and Jiang, 2007; Ferreira and Laux, 2007) to accounting (e.g., Frankel and Li, 2004; Jayaraman, 2008).

[2]A precursor to our paper is Li (2012), which solves a continuous-time Kyle model in which the probability of an information event is less than 1 by applying filtering theory to a transformation of the aggregate order process. The filtering solution produces a stochastic differential equation for the equilibrium rather than a closed form solution. The method of proof used in this paper shares some features with the proof in Back and Crotty (2015).

amount of private information. If it is anticipated that private information is more of a concern for the first stock than for the second, then the first stock will be less liquid, other things being equal. The lower liquidity will reduce the amount of informed trading, possibly offsetting the increase in informed trading due to greater private information. In equilibrium, the amount of informed trading may be the same in both stocks, despite the difference in information asymmetry. In general, the distribution of order flows need not reflect the degree of information asymmetry when liquidity providers react to information asymmetry and informed traders react to liquidity. Thus, we provide the first theoretical explanation of why the PIN methodology, which uses order flows alone to estimate information asymmetry parameters, may not identify private information.[3]

Our second contribution is to develop novel estimates characterizing the information environment in financial markets. We structurally estimate our theoretical model for a panel of stocks and provide several validation checks that the estimated parameters are plausibly related to information asymmetry. First, reduced-form estimates of price impact are increasing in our structural estimates of the probability and magnitude of information events, as implied by theory. Second, excess kurtosis is decreasing in the ex ante probability of an information event in the model and in the data. This occurs in the model because the sensitivity of prices to orders depends on the perceived likelihood that an informed trader is present. This changing sensitivity of prices to orders means that returns are drawn from

[3]Several papers argue that PIN does not identify private information. Aktas et al. (2007) examine trading around merger announcements. They show that PIN decreases prior to announcements. In contrast, percentage spreads and the permanent price impact of trades, measured as in Hasbrouck (1991), rise before announcements, indicating the presence of information asymmetry. They describe the decline in PIN prior to announcements as a PIN anomaly. Akay et al. (2012) show that PIN is higher in the Treasury bill market than it is in markets for individual stocks. Given that it is very doubtful that informed trading in T-bills is a frequent occurrence, this is additional evidence that PIN is not measuring information asymmetry. Benos and Jochec (2007) find that PIN is higher following earnings announcements, contrary to their assumption that information asymmetry should be higher before announcements. Duarte, Hu and Young (2016) also examine earnings announcements. They estimate the parameters of the PIN model and then compute the conditional probability of an information event each day. They show that the conditional probability rises prior to announcements but stays elevated for a number of days following announcements. They show that the high post-announcement conditional probabilities are due to high turnover and argue that high turnover is misidentified as private information by the PIN model.

a mixture distribution, exhibiting excess kurtosis. Third, volatility over the latter part of a trading day is increasing in the conditional probability of an information event, where the conditioning is based on cumulative order flows over the first part of the day and our estimated parameters. This phenomenon of stochastic volatility occurs in both the model and the data.[4]

To demonstrate potential applications of the estimates, we revisit two settings in which PIN estimates have been employed. One application of PIN has been to attempt to capture time-series variation in information asymmetry. For example, Brown, Hillegeist and Lo (2004, 2009) examine changes in information asymmetry following voluntary conference calls and earnings surprises, respectively, while Duarte, Han, Harford and Young (2008) study the effect of Regulation FD on PIN and the cost of capital. We show that conditional probabilities of information events calculated using order flows and our parameter estimates rise on average around earnings announcements and also around block accumulations by Schedule 13D filers, indicating that the model does capture time series variation in information asymmetry.

The second application illustrates how estimates of the information asymmetry parameters from our model can be used to augment studies concerned with cross-sectional differences in the information content of prices. To do so, we consider the hypothesis of Chen et al. (2009) that corporate investment is more sensitive to market prices when there is more private information in prices. Our model allows us to measure the amount of private information alternatively by the frequency of private information events, by the magnitude of private information, and by the fraction of total price movement that is due to private information. We show that corporate investment is more sensitive to prices when any of these measures is higher. These measures of private information should prove useful in other settings in which researchers are interested in capturing distinct facets of the information

[4]Banerjee and Green (2015) solve a rational expectations model with myopic mean-variance investors in which investors learn whether other investors are informed. They show that variation over time in the perceived likelihood of informed trading induces volatility clustering. While their model is quite different from ours, our model also exhibits volatility clustering. Volatility follows the same pattern as Kyle's lambda, which varies over time due to variation in the market's estimate of whether an information event occurred.

environment (e.g., the amount of noise trading or the magnitude of private information).

Related structural models of informed trading include the Adjusted PIN (APIN) model of Duarte and Young (2009), the Volume-Synchronized PIN (VPIN) model of Easley, López de Prado and O'Hara (2012), and the modified Kyle model of Odders-White and Ready (2008). The APIN model allows for time variation in liquidity trading (with positively correlated buy and sell intensities), which provides a better fit to the empirical distribution of buys and sells. The VPIN model estimates buys and sells within a given time interval by assigning a fraction of total volume to buys and the remaining fraction to sells based on standardized price changes during the time interval.[5] Odders-White and Ready (OWR) analyze a Kyle model in which the probability of an information event is less than 1, as it is in our model. However, they analyze a single-period model, whereas we study a dynamic model. Unlike our dynamic model in which prices equal conditional expectations, market makers in their model only match unconditional means of prices to unconditional means of asset values.[6]

Our estimate of the probability of an information event is not positively correlated in the cross section with estimates from the other models. The divergence between the estimates is not surprising, because the models have different assumptions/implications regarding what data is required to identify the probability of an information event.[7] We also calculate a composite measure of information asymmetry in our model: the expected average lambda. This measure incorporates both the probability and magnitude of information events as

[5]Easley et al. (2011) claim that VPIN predicted the "flash crash" of May 6, 2010. This claim and some other claims regarding VPIN are challenged by Andersen and Bondarenko (2014b). See also Easley et al. (2014) and Andersen and Bondarenko (2014a).

[6]In a single-period model, because of the net order having a mixture distribution, the conditional expectation of the asset value given the net order is not a linear function of the net order. To make the model tractable, Odders-White and Ready deviate from the usual Kyle model formulation and do not require the asset price to equal its conditional expected value. Instead, they only require that unconditional expected market maker profits are zero. They find the pricing rule that is linear in the net order that has this "zero conditional expected profits on average" property. Such a pricing rule would require commitment by market makers, because it is not consistent with ex-post optimization by market makers. In contrast, pricing in our model is consistent with ex-post optimization by competitive market makers: prices equal conditional expected values.

[7]While the OWR model uses both prices and order flows for estimation, their model shares the feature of the PIN model that the unconditional order flow distribution depends on the information asymmetry parameters and hence could be used to identify information asymmetry.

well as the amount of liquidity trading. Unlike the probability of an information event, the expected average lambda from our model is positively correlated with similar measures from other models (PIN, APIN, VPIN, and the OWR lambda). Each of these measures should be increasing in the probability of an information event, so it is surprising that they are all positively correlated, given the lack of correlation of the 'probability of an information event' estimates. However, the measures are also decreasing in the amount of noise trading, and we present evidence in Section 5 that the measurement of noise trading is quite positively correlated across models, resulting in the positive correlation of the composite measures. Of course, applications of the measures generally assume that they are correlated with private information, not just inversely correlated with liquidity trading.

Theory predicts that orders have larger price impacts when information asymmetry is more severe.[8] Note that this is true in both the Kyle (1985) model upon which the hybrid and OWR models are based and the Glosten and Milgrom (1985) model upon which PIN models are based. To test this implication of theory, we compute reduced-form estimates of price impacts for our sample and regress them on estimated information asymmetry parameters from each model. Empirically, expected average lambda from the hybrid model explains a substantial amount of cross-sectional variation in price impacts. However, PIN, APIN, VPIN, and the OWR lambda are also positively related to price impacts cross-sectionally. Our hybrid model (and to a lesser extent VPIN) performs the best in a horse race.

Other related theoretical work includes Rossi and Tinn (2010), Foster and Viswanathan (1995), and Chakraborty and Yilmaz (2004). Rossi and Tinn solve a two-period Kyle model in which there are two large traders, one of whom is certainly informed and one of whom may or may not be informed. In their model, unlike ours, there are always information events. Foster and Viswanathan (1995) consider a series of single-period Kyle models in

[8]There seems to be general agreement that at least a portion of the price impact of trades is due to information asymmetry. Glosten and Harris (1988), Hasbrouck (1988), and Hasbrouck (1991) estimate models of trades and price changes in which both information asymmetry and inventory control motives are accommodated, and all three papers conclude that information asymmetry is important.

which traders choose in each period whether to pay a fee to become informed. There may be periods in which there are no informed traders. However, in their model, it is always common knowledge how many traders choose to become informed, so, in contrast to our model, there is no learning from orders about whether informed traders are present.

Chakraborty and Yilmaz (2004) study a discrete-time Kyle model in which there may or may not be an information event. Their main result is that the informed trader will manipulate (sometimes buying when she has bad information and/or selling when she has good information) if the horizon is sufficiently long. The primary difference between their model and ours is that they assume that the noise trade distribution has finite support, so market makers may incorrectly rule out a type of trader if the horizon is sufficiently long. In contrast, market makers in our model can never rule out any type of the informed trader until the end of the model, so it does not strictly pay for a low type to pretend to be a high type or vice versa.

2. The Hybrid Model

The hybrid model includes two important features of PIN models—a probability less than 1 of an information event and a binary asset value conditional on an information event—and it also includes an optimizing (possibly) informed trader, as in the Kyle (1985) model. Denote the time horizon for trading by $[0, 1]$. Assume there is a single risk-neutral strategic trader. Assume this trader receives a signal $S \in \{L, H\}$ at time 0 with probability α, where $L < 0 < H$. Let p_L and $p_H = 1 - p_L$ denote the probabilities of low and high signals, respectively, conditional on an information event. With probability $1-\alpha$, there is no information event, and the trader also knows when this happens. Let ξ denote an indicator for whether an information event has occurred ($\xi = 1$ if yes and $\xi = 0$ if no). In addition to the private information, public information can also arrive during the course of trading, represented by a martingale V. Whether there was an information event, and, if so, whether the signal was low or high becomes public information after the close of trading, producing an asset value of $V_1 + \xi S$.

In addition to the strategic trades, there are liquidity trades represented by a Brownian motion Z with zero drift and instantaneous standard deviation σ. Let X_t denote the number of shares held by the strategic trader at date t (taking $X_0 = 0$ without loss of generality), and set $Y_t = X_t + Z_t$. The processes Y and V are observed by market makers. Denote the information of market makers at date t by $\mathcal{F}_t^{V,Y}$.

One requirement for equilibrium in this model is that the price equal the expected value of the asset conditional on the market makers' information and given the trading strategy of the strategic trader:

$$P_t = \mathsf{E}\left[V_1 + \xi S \mid \mathcal{F}_t^{V,Y}\right] = V_t + \mathsf{E}\left[\xi S \mid \mathcal{F}_t^{V,Y}\right] . \tag{1}$$

We will show that there is an equilibrium in which $P_t = V_t + p(t, Y_t)$ for a function p. This means that the expected value of ξS conditional on market makers' information depends only on cumulative orders Y_t and not on the entire history of orders.

The other requirement for equilibrium is that the strategic trades are optimal. Let θ_t denote the trading rate of the strategic trader (i.e., $\mathrm{d}X_t = \theta_t \,\mathrm{d}t$). The process θ has to be adapted to the information possessed by the strategic trader, which is V, ξS, and the history of Z (in equilibrium, the price reveals Z to the informed trader). The strategic trader chooses the rate to maximize

$$\mathsf{E}\int_0^1 [V_1 + \xi S - P_t]\,\theta_t \,\mathrm{d}t = \mathsf{E}\int_0^1 [\xi S - p(t, Y_t)]\,\theta_t \,\mathrm{d}t\,, \tag{2}$$

with the function p being regarded by the informed trader as exogenous. In the optimization, we assume that the strategic trader is constrained to satisfy the "no doubling strategies" condition introduced in Back (1992), meaning that the strategy must be such that

$$\mathsf{E}\int_0^1 p(t, Y_t)^2 \,\mathrm{d}t < \infty$$

with probability 1.

Let N denote the standard normal distribution function, and let n denote the standard normal density function. Set $y_L = \sigma\,\mathrm{N}^{-1}(\alpha p_L)$ and $y_H = \sigma\,\mathrm{N}^{-1}(1-\alpha p_H)$. This means that the probability mass in the lower tail $(-\infty, y_L)$ of the distribution of cumulative liquidity trades Z_1 equals αp_L, which is the unconditional probability of bad news. Likewise, the probability mass in the upper tail (y_H, ∞) of the distribution of Z_1 equals αp_H, which is the unconditional probability of good news. Set

$$q(t,y,s) = \begin{cases} \mathsf{E}[Z_1 - Z_t \mid Z_t = y, Z_1 < y_L] & \text{if } s = L\,, \\ \mathsf{E}[Z_1 - Z_t \mid Z_t = y, y_L \le Z_1 \le y_H] & \text{if } s = 0\,, \\ \mathsf{E}[Z_1 - Z_t \mid Z_t = y, Z_1 > y_H] & \text{if } s = H\,. \end{cases} \tag{3}$$

From the standard formula for the mean of a truncated normal, we obtain the following more explicit formula for q:

$$\frac{q(t,y,s)}{\sigma\sqrt{1-t}} = \begin{cases} -\,\mathrm{n}\left(\frac{y_L-y}{\sigma\sqrt{1-t}}\right) \big/ \,\mathrm{N}\left(\frac{y_L-y}{\sigma\sqrt{1-t}}\right) & \text{if } s = L\,, \\ \left[\mathrm{n}\left(\frac{y_L-y}{\sigma\sqrt{1-t}}\right) - \mathrm{n}\left(\frac{y_H-y}{\sigma\sqrt{1-t}}\right)\right] \Big/ \left[\mathrm{N}\left(\frac{y_H-y}{\sigma\sqrt{1-t}}\right) - \mathrm{N}\left(\frac{y_L-y}{\sigma\sqrt{1-t}}\right)\right] & \text{if } s = 0\,, \\ \mathrm{n}\left(\frac{y-y_H}{\sigma\sqrt{1-t}}\right) \big/ \,\mathrm{N}\left(\frac{y-y_H}{\sigma\sqrt{1-t}}\right) & \text{if } s = H\,. \end{cases} \tag{4}$$

The equilibrium described in Theorem 1 below can be shown to be the unique equilibrium in a certain broad class, following Back (1992). The proof of Theorem 1 is given in Appendix A.[9]

Theorem 1. *There is an equilibrium in which the trading rate of the strategic trader is*

$$\theta_t = \frac{q(t, Y_t, \xi S)}{1-t}\,. \tag{5}$$

[9]The proof is based on a generalization of the Brownian bridge feature of the continuous-time Kyle model established in Back (1992). Whereas a Brownian bridge is a Brownian motion conditioned to end at a particular point, in this model (with a discrete rather than continuous distribution of the asset value) we encounter a Brownian motion conditioned only to end in a particular interval. The generalization of the Brownian bridge is established as a lemma in Appendix A.

The equilibrium asset price is $P_t = V_t + p(t, Y_t)$, where the pricing function p is given by

$$p(t, y) = L \cdot \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) + H \cdot \mathrm{N}\left(\frac{y - y_H}{\sigma\sqrt{1-t}}\right). \tag{6}$$

In this equilibrium, the process Y is a martingale given market makers' information and has the same unconditional distribution as does the liquidity trade process Z; that is, it is a Brownian motion with zero drift and standard deviation σ.

The last statement of the theorem implies that the distribution of order flows in the model does not depend on the information asymmetry parameters α, H, and L. Thus, if the model is correct, it is impossible to estimate those parameters using order flows alone. In general, the theorem suggests that it may be difficult to identify information asymmetry parameters using order flows alone, as discussed in the introduction and the next subsection. When we estimate the hybrid model, we use both order flows and returns, in contrast to the PIN model that only uses order flows.

Empirically, we test the relationship between α and price impacts of trades. Figure 1 plots the equilibrium price as a function of Y_t for two different values of α. It shows that the price is more sensitive to orders when α is larger. This is also true in the PIN model. We test the relationship for both models. To investigate further how the sensitivity of prices to orders depends on α in the hybrid model, we calculate the price sensitivity—that is, we calculate Kyle's lambda.

Theorem 2. *In the equilibrium of Theorem 1, the asset price evolves as $\mathrm{d}P_t = \mathrm{d}V_t + \lambda(t, Y_t)\,\mathrm{d}Y_t$, where Kyle's lambda is*

$$\lambda(t, y) = -\frac{L}{\sigma\sqrt{1-t}} \cdot \mathrm{n}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) + \frac{H}{\sigma\sqrt{1-t}} \cdot \mathrm{n}\left(\frac{y_H - y}{\sigma\sqrt{1-t}}\right). \tag{7}$$

Furthermore, Kyle's lambda $\lambda(t, Y_t)$ is a martingale with respect to market makers' information on the time interval $[0, 1)$.

Kyle's lambda is a stochastic process in our model, but we can easily relate the expected average lambda to α. Because lambda is a martingale, the expected average lambda is $\lambda(0,0)$. Substitute the definitions of y_L and y_H in (7) to compute[10]

$$\lambda(0,0) = -\frac{L}{\sigma}\,\mathrm{n}\left(\mathrm{N}^{-1}(\alpha p_L)\right) + \frac{H}{\sigma}\,\mathrm{n}\left(\mathrm{N}^{-1}(1-\alpha p_H)\right)\,. \tag{8}$$

Figure 2 plots the expected average lambda as a function of α for two values of H, taking $L = -H$. Doubling the signal magnitudes doubles lambda. Furthermore, the expected average lambda is increasing in α.

2.1. Nonidentifiability Using Order Flows Alone

A key result of Theorem 1 is that the aggregate order imbalance Y_1 has the same distribution as the liquidity trades Z_1 and is invariant with respect to the information asymmetry parameters.[11] Applications of the PIN model typically assume each day is a separate instance of the model and use daily buys and sells to estimate the model parameters. If our model describes reality and each day is a separate instance of the model, then the sample of daily order imbalances is a sample of i.i.d. normal random variables with mean zero and variance equal to the variance of daily liquidity trades Z_1.[12] The distribution of the sample is invariant with respect to the frequency and magnitude of information events, so the sample of daily order imbalances alone cannot identify information asymmetry.

The fact that aggregate orders Y_1 have the same distribution as liquidity trades is a consequence of the martingale property of Y (a continuous martingale with quadratic variation

[10]If information events occur for sure ($\alpha = 1$), then $\lambda(0,0) = (H - L)\,\mathrm{n}(0)/\sigma$. This is analogous to the result of Kyle (1985) that lambda is the ratio of the signal standard deviation to the standard deviation of liquidity trading. Of course, it is not quite the same as Kyle's formula, because we have a binary signal distribution, whereas the distribution is normal in Kyle (1985).

[11]This result on the nonidentifiability of information asymmetry parameters from order flows does not depend on the binary signal assumption. Internet Appendix A presents the model with a general signal distribution. The unconditional order flow distribution is the same as the distribution of noise order flows in the general model as well.

[12]Of course, we cannot know if a single day is a separate instance of either model. Many (long-lived) instances of private information may entail informed trading over multiple days (e.g., activist investors in Collin-Dufresne and Fos, 2015).

over each time interval equal to the length of the interval is automatically a Brownian motion). The martingale property of Y is equivalent to unpredictability of informed orders in our model. As mentioned before, informed orders are predictable in the PIN model, because informed traders do not react to price changes in the PIN model.[13]

Further insight into the identification issue can be gained by noting that, as in the PIN model, the unconditional distribution of the order imbalance in our model is a mixture of three conditional distributions. With probability αp_L, Y_1 is drawn from the distribution conditional on a low signal; with probability αp_H, Y_1 is drawn from the distribution conditional on a high signal; and with probability $1 - \alpha$, Y_1 is drawn from the distribution conditional on no information event. The first two distributions have nonzero means—there is an excess of sells over buys in the first and an excess of buys over sells in the second. This is also analogous to the PIN model. Thus, one might conjecture that changing α—thereby changing the likelihood of drawing from the first two distributions—will alter the unconditional distribution of Y_1. If so, then one could perhaps identify α from the distribution of Y_1. In the PIN model, it is indeed true that changing α, holding other parameters constant, alters the unconditional distribution of the order imbalance. However, it is not true in our model, because the distribution of informed trades in our model depends endogenously on α due to liquidity depending on α. With a larger alpha, the market is less liquid (see the comparative statics in Figure 2) and the informed trader trades less aggressively With endogenous informed orders, the arrival rate of informed orders depends on prior price changes as shown in Figure 3, which is not the case in the PIN model. In particular, when prices have moved

[13]The negative result on identification also holds in a more general model in which there is a predictable component of order flows. In that model,

$$Y_1 = \int_0^1 \mu_t \,\mathrm{d}t + Y_1^* \,,$$

where Y^* is the sum of informed orders and unpredictable liquidity trades, and where μ is adapted to the price process and hence adapted to Y^*. Because informed orders are unpredictable, Y^* is a martingale; therefore, it is a Brownian motion with its variance determined by the variance of liquidity trades. This implies that the distribution of Y_1 is again invariant with respect to the information asymmetry parameters.

in the direction of the news, informed orders slow down, and, when prices have moved in the opposite direction, informed orders speed up. Figure 3 shows that these changes in intensity depend on the ex ante probability of an information event α. Thus, the distributions over which we are mixing change when the mixture probabilities change, leaving the unconditional distribution of Y_1 invariant with respect to α.

The change in the conditional distributions is illustrated in Figure 4. The top and bottom panels of Figure 4 show that the strategic trader trades more aggressively when an information event occurs if an information event is less likely ($\alpha = 0.1$ versus $\alpha = 0.5$). This equilibrium reaction of informed trading to exogenous changes in the probability of information events is missing in the PIN model, in which informed trading is exogenously determined. It is a key feature of our model that results in the probability of information events being unidentified by the distribution of order imbalances. The unconditional distribution of Y_1 is standard normal for both $\alpha = 0.1$ and $\alpha = 0.5$ in Figure 4, so we cannot hope to use the unconditional distribution to recover α.

Continuing with the example in Figure 4, calculate the expected absolute order imbalance as

$$\alpha p_L \mathsf{E}\big[|Y_1| \mid \xi = 1, S = L\big] + (1-\alpha)\mathsf{E}\big[|Y_1| \mid \xi = 0\big] \; + \alpha p_H \mathsf{E}\big[|Y_1| \mid \xi = 1, S = H\big]\,,$$

with $\sigma = 1$ and $p_L = p_H = 1/2$. If $\alpha = 0.5$, then the expected absolute order imbalance is

$$\frac{1}{4} \times 1.27 \; + \; \frac{1}{2} \times 0.32 \; + \; \frac{1}{4} \times 1.27 \; = \; 0.80\,.$$

On the other hand, if $\alpha = 0.1$, then the expected absolute order imbalance is

$$\frac{1}{20} \times 2.06 \; + \; \frac{9}{10} \times 0.66 \; + \; \frac{1}{20} \times 2.06 \; = \; 0.80\,.$$

Again, we see that informed trading is more aggressive when information events occur if such events are less likely. Here, it is clear that the endogenous change in informed trading exactly offsets the exogenous change in the likelihood of information events. In other words, the endogenous changes in the distributions over which we are mixing exactly offset the changes in the mixing probabilities (this is true even with $p_L \neq p_H$). On the other hand, under the assumptions of the PIN model, the expected absolute order imbalance varies with α (see Easley et al., 2008, p. 176, for a discussion of how the absolute order imbalance is related to α and to PIN under the assumptions of the PIN model).

The previous paragraphs describe the invariance of the unconditional distribution of Y_1 with respect to α. The other important parameters governing information asymmetry are L and H. For example, if the possible signals L and H are both small in absolute value, then information asymmetry is a minor concern even if information events occur frequently. Order flows cannot identify L and H in our model. In fact, L and H do not affect even the conditional distributions shown in Figure 4; thus, they do not affect the unconditional distribution of Y_1.

Of course, identifying the information asymmetry parameters from the distribution of order imbalances is a very different issue from using order imbalances to update the probability of an information event in a particular instance of the model. Conditional on knowledge of the parameters, the order imbalance does help in estimating whether an information event occurred in a particular instance of the model; in fact, the market makers in the model update their beliefs regarding the occurrence of an information event based on the order imbalance. So, we can compute

$$\text{prob}(\text{info event} \mid Y_t, \text{parameters})\,,$$

and this probability does depend on the information asymmetry parameters. We could use this to identify the information asymmetry parameters if we had data on order imbalances *and data on whether information events occurred.* Of course, we do not have data of the latter type. Theorem 1 shows that the likelihood function of the information asymmetry parameters given only data on order imbalances is a constant function of those parameters; hence, the order imbalances alone cannot identify them.

In our empirical work, we estimate the model parameters using prices and order flows. Armed with these parameter estimates and order flow observations, we can compute conditional probabilities of an information event. We examine their predictive relation to intraday volatility and their time-series properties around earnings announcements and around Schedule 13D filer trades in Sections 3.4 and 4.1.

2.2. The Contrarian Trader Assumption

One way in which our model departs from the PIN model is that the strategic trader is present in our model even when there is no information event. When there is no information event, this trader behaves as a contrarian, selling on price increases and buying on price declines.[14] The existence of such a contrarian trader seems likely if there are always some traders who are best informed—corporate managers, for example. This would be the case if information is truly idiosyncratic to the firm. If, on the other hand, there is an industry or other aggregate component to the information, then it is possible that no one knows when no one else has information. In that case, the contrarian trader we postulate would not exist.

Our result on the nonidentifiability of information asymmetry parameters from order flows is not due to the contrarian trader assumption. In Internet Appendix B, we solve a variant of the PIN model in which contrarian traders arrive to the market when there is no information event. The contrarian traders condition their trading direction on the prevailing bid and ask quotes and the intrinsic value of the asset. The unconditional distribution of order imbalances in that model is shown in Figure 5 for three different values of α (the probability of an information event). The figure shows that the distribution depends on α;

[14] We assume the existence of such a trader because it makes the model more tractable. OWR describe the trader as also being present in their model when there is no information event, but, because the trader has no opportunity to react to price changes in their one-period model, the trader optimally chooses a zero trade in the absence of an information event.

thus, order imbalances can be used to identify information asymmetry in the PIN model even when a contrarian trader is present. Thus, the contrarian trader assumption is not the main driving force behind our nonidentifiability result. Instead, the result depends on market makers reacting to information asymmetry and to strategic traders reacting both to liquidity and to price changes. That is, order flows depend on market liquidity, which depends on information asymmetry. This creates an indirect dependence of order flows on information asymmetry that is countervailing to the direct relation.

3. Parameter Estimates

We estimate the hybrid model using trade and quote data from TAQ for NYSE firms from 1993 through 2012.[15] We sign trades as buys and sells using the Lee and Ready (1991) algorithm: trades above (below) the prevailing quote midpoint are considered buys (sells). If a trade occurs at the midpoint, then the trade is classified as a buy (sell) if the trade price is greater (less) than the previous differing transaction price.[16] We sample prices and order imbalances hourly and at the close and define order imbalances as shares bought less shares sold (denoted in thousands of shares).

We estimate the model by maximum likelihood, maintaining the standard assumptions in the literature that each day is a separate realization of the model and that parameters are constant within each year for each stock. We assume that the possible signals on each day i are proportional to the observed opening price on day i, P_{i0}. Specifically, we assume that for each firm-year, there is a parameter κ such that the possible signals on each day i are $H = -L = \kappa P_{i0}$. We also assume the public information process V is a geometric Brownian motion on each day with a constant volatility δ. Appendix B derives the likelihood function for the hybrid model under these assumptions. The likelihood function depends on the signal

[15] We require that firms have intraday trading observations for at least 200 days within the year. We also require firms have the same ticker throughout the year and experience no stock splits.

[16] Prior to 2000, quotes are lagged five seconds when matched to trades. For 2000-2002, quotes are lagged one second. From 2003 on, quotes are matched to trades in the same second. To account for quote updates within a second, we use the interpolated time technique introduced by Holden and Jacobsen (2014).

magnitude κ, the probability α of information events, the probability p_L of a negative signal conditional on an information event, the standard deviation σ of liquidity trading, and the volatility δ of public information.

3.1. Estimates of the Hybrid Model

Figure 6 displays the time-series of average estimates and the interquartile range for the cross-section of stocks for the hybrid model. The average α is almost 70% in the early part of the sample and falls to about 50% by the end of the sample, indicating that the likelihood of private information events, at least at the daily frequency we study, has fallen on average. This effect starts in 2007 and is evident in the decrease in the lower quartile of α estimates. The other components of private information events are the magnitude κ of the signal and the likelihood p_L of a bad event. Private information κ initially rises during the late 1990s, but exhibits a strong downward trend thereafter. The average p_L indicates that the distribution of information is relatively symmetric between positive and negative events. We combine these estimates into a single composite measure of information asymmetry by calculating the expected average lambda from Equation (8). The estimates indicate that the amount of private information has fallen across the twenty year sample with the exception of the late 1990s and the financial crisis.[17]

In general, the standard deviation σ of order imbalances and the volatility δ of public information appear to be roughly stationary. Despite the well-documented rise of high-frequency trading and the associated sharp increase in trading volume, the volatility of order imbalances has remained fairly stable over the twenty year sample. Like private information, public information volatility also spiked during the financial crisis. This suggests private information may be proportional to public information rather than a fixed amount.

[17] As we discuss in Section 5.3, the same pattern is seen in reduced-form price impact measures.

3.2. Testing Whether There is Always an Information Event in the Hybrid Model

Our hybrid model relaxes the assumption in Kyle (1985) that an information event occurs in each instance of the model (in each day in our implementation). A natural question is whether this relaxation is supported in the data. The Kyle framework is nested in our model by the restriction that $\alpha = 1$. Accordingly, we estimate the model with this restriction. The standard likelihood ratio test of the null that $\alpha = 1$ against the alternative that $\alpha \in [0, 1]$ is rejected for 75% of the firm-years (with a test size of 10%). However, the usual regularity conditions for the likelihood ratio test require that the restriction not be at the boundary of the parameter space. To address this issue, we bootstrap the distribution of the likelihood ratio statistic for a random sample of 100 firm-years as in Duarte and Young (2009).

Specifically, for a given firm-year, we estimate the restricted model ($\alpha = 1$) and then simulate 500 firm-years under the null using the estimated (restricted) parameters. We then estimate the restricted and unrestricted models for each simulated firm-year to obtain the distribution of the likelihood ratio under the null. The 90th percentile of this distribution is the critical value to evaluate the empirical likelihood ratio. These bootstrapped likelihood ratio tests reject the restricted Kyle model in favor of the hybrid model for 76 of the 100 randomly selected firm-years. The data thus supports the conclusion that the probability of an information event is less than 1.

3.3. Estimated Parameters and Reduced-Form Price Impacts

The model places structure on the price and order flow data, allowing the econometrician to identify components of Kyle's lambda. Of course, one can estimate a reduced-form price impact as well. As an initial test of whether our estimates relate to price impact as implied by theory, we test the comparative statics from Figure 2 that price impacts are increasing in both the probability and magnitude of information events.

We employ three estimates of the price impact of orders. The first is the 5-minute percent price impact of a given trade k as:

$$\textit{5-minute Price Impact}_k = \frac{2D_k(M_{k+5} - M_k)}{M_k}, \tag{9}$$

where M_k is the prevailing quote midpoint for trade k, M_{k+5} is the quote midpoint five minutes after trade k, and D_k equals 1 if trade k is a buy and -1 if trade k is a sell. Goyenko, Holden and Trzcinka (2009) use this measure as one of their high-frequency liquidity benchmarks in a study assessing the quality of various liquidity measures based on daily data.[18] For a given stock-day, the estimate of the percent price impact is the equal-weighted average price impact over all trades on that day. We average these daily price impact estimates for each stock-year.

We also estimate the cumulative impulse response function (Hasbrouck, 1991), which captures the permanent price impact of an order. The cumulative impulse response is calculated from a vector autoregression of log price changes and signed trades. Finally, we estimate a version of Kyle's lambda (denoted $\widehat{\lambda}_{\text{intraday}}$) using a regression of 5-minute returns on the square-root of signed volume following Hasbrouck (2009) and Goyenko, Holden and Trzcinka (2009). We estimate these for each stock day, taking the median estimate across days as the stock-year estimate.

The first panel of Table 1 reports panel regressions of the three price impact measures on α and κ. Before running the regressions, the price impacts and the structural parameters are winsorized at 1/99% and standardized to have unit standard deviation. Price impacts are positively related to each of the hybrid model parameters that measure private information (the probability α of an information event and the magnitude κ of information events). The coefficients are positive even with the inclusion of firm fixed effects, suggesting α and κ capture within-firm information asymmetry variation as well.

[18] Holden and Jacobsen (2014) show that liquidity measures such as the percent price impact can be biased when constructed from monthly TAQ data, so we follow their suggested technique in processing the data.

A summary measure of the amount of private information is the standard deviation of the signal ξS, denoted SD(ξS). In the binary signal case, SD(ξS) is:

$$2\kappa\sqrt{\alpha p_L(1-p_L)}\,. \tag{10}$$

The second panel of Table 1 shows that the estimated SD(ξS) is strongly positively correlated with the price impact estimates, as expected. Cross-sectionally, a one standard deviation increase in SD(ξS) is associated with about three-quarters of a standard deviation increase in 5-minute price impact and $\widehat{\lambda}_{\text{intraday}}$ and almost half a standard deviation increase in the cumulative impulse response measure. Variation in SD(ξS) within firm is positively correlated with within-firm variation in all three price impact measures.

3.4. Excess Kurtosis and Stochastic Volatility

In this section, we test two additional predictions of the model. The first is the relation between alpha and excess kurtosis. The second is the implication from Theorem 2 that volatility is stochastic and depends on the conditional probability of an information event.

Our model proposes a natural mechanism that causes stock returns to exhibit excess kurtosis: The sensitivity of prices to orders depends on the perceived likelihood that an informed trader is present. In turn, this depends on the cumulative order flow imbalance. If the likelihood is judged to be high, then prices are quite sensitive to orders. If the likelihood is low, then prices are relatively insensitive. Even if orders are i.i.d., this changing sensitivity of prices to orders means that returns are drawn from a mixture distribution, exhibiting excess kurtosis. This phenomenon is more extreme if the prior probability of an information event is lower; thus, the lower the prior probability of an informed trader being present, the higher is the excess kurtosis.

In Table 2, we test this implication of the model for simulated and actual data. We first simulate the model for α values ranging from 0.05 to 0.95 in 0.05 increments. For each α value, the simulated panel contains 1000 firm-years. We estimate excess kurtosis for each firm-year. The first column of Table 2 reports a regression of excess kurtosis on α. In the model, lower levels of α are associated with greater excess kurtosis.

Column two of Table 2 reports the same regression using the estimated α for the actual data. As in the model, stocks with higher alpha exhibit lower excess kurtosis. This result is robust to inclusion of controls for size, price, and volume, as well as firm and year fixed effects.

The second implication of the model that we test is that volatility depends on the conditional probability of an information event. Thus, there is stochastic volatility. In the model, market makers update their conditional probabilities of an information event, CPIE_t, as:[19]

$$\text{CPIE}_t(Y_t) = \begin{cases} \mathrm{N}\left(\frac{y_L - Y_t}{\sigma\sqrt{1-t}}\right) + \mathrm{N}\left(\frac{Y_t - y_H}{\sigma\sqrt{1-t}}\right) & \text{if } t < 1\,, \\ \mathbb{1}\left(Y_1 < y_L\right) + \mathbb{1}\left(Y_1 > y_H\right) & \text{if } t = 1\,. \end{cases} \tag{11}$$

In Table 3, we test the relation between volatility and the conditional probability of an information event. We measure volatility as absolute returns over the last three and a half hours of the trading day. CPIE is calculated for each day using the cumulative order imbalance over the first three hours of the day, along with the estimated parameters which are needed to calculate y_L and y_H. We report predictive regressions of end-of-day absolute returns on CPIE calculated from the first part of the day.

As in the previous table, the first column reports results using simulated data from the model. Higher levels of CPIE predicts higher volatility in the second part of the day. Columns two through five show that this phenomenon holds in the actual data as well. The empirical finding holds controlling for the prior day's realized absolute return as well as firm and year fixed effects. Moreover, CPIE captures more than just volatility in cumulative order flows through the first part of the day. The last column of Table 3 shows that CPIE predicts volatility even controlling for the absolute cumulative order imbalance.

[19]This formula follows from parts (B) and (C) of the lemma in Appendix A.

4. Applications

We now discuss two potential applications of the estimation procedure. A large literature uses the PIN model, as discussed previously. Broadly speaking, some of this work relates PIN estimates to times when researchers believe information events have likely occurred. Other research uses PIN to proxy for information asymmetry or price informativeness. We discuss examples of how our estimates might be useful to research of either type.

4.1. Detecting Information Events

Information asymmetry is generally unobservable, so testing performance of adverse selection measures is challenging. In this section, we study how the conditional probability of an information event as measured by our model varies in two settings considered in the literature: earnings announcements and trading by Schedule 13D filers.

4.1.1. Earnings Announcements

Many studies have examined the information environment surrounding earnings announcements. Some studies assume that information asymmetry is higher prior to information events, while others note that private ability or knowledge to interpret public information may result in adverse selection following announcements (Kim and Verrecchia, 1997). Several recent papers (Duarte et al., 2016; Brennan et al., 2016) use conditional estimates based on the EKOP and OWR models around earnings announcements.

As we discuss in Section 2.1, one can assess the probability of an information event if one observes cumulative order flows *and* knows the underlying parameters. Armed with our estimates of the parameters, we examine conditional probabilities of an information event, CPIE, on the days around earnings announcements.

Figure 7 plots the cross-sectional average of model-implied CPIE in event time around earnings announcements. The average CPIE rises significantly on day $t-1$, consistent with early leakage of some information prior to the announcement. The average CPIE is highest on days t and $t+1$, and then falls over the next week or so. The results suggest that adverse selection may actually be worse following an earnings announcement rather than before it, as discussed in Kim and Verrecchia (1997).[20]

4.1.2. Schedule 13D Filings

Collin-Dufresne and Fos (2015) examine whether various measures of adverse selection are higher during periods in which Schedule 13D filers accumulate ownership positions. These positions are generally associated with a positive stock price reaction, so these investors are privately informed. These investors must disclose days on which they traded over a sixty-day period preceding the filing date. Thus, this data provides the econometrician with a laboratory concerning informed trading. Collin-Dufresne and Fos (2015) show that measures designed to capture information asymmetry are actually lower on days when Schedule 13D filers trade. As they discuss, this could be due to endogenous trading in times of greater liquidity and due to the use of patient limit orders. The latter effect arises in part because of the filers' ability to control the timing of the private information revelation. This differs from the earnings announcement setting where an informed trader's information is valid only for an exogenous duration.

We revisit the Schedule 13D setting to assess whether the conditional probability of an information event is higher on days when these informed investors trade. Note that this setting is further from the theoretical setting we consider. The ultimate revelation of information is at least partially in the control of the informed trader. Moreover, the private information persists across trading days, so our empirical assumption that information is revealed at the end of the day is violated. Nonetheless, we consider whether the main intuition for our conditional probabilities, that order flows should be more extreme, helps reveal the presence of informed traders.

Table 4 reports average values of CPIE on days during the sixty-day disclosure window when Schedule 13D filers do or do not trade. Just over half of the firm-days with no Schedule

[20] This conclusion is also reached by Krinsky and Lee (1996) using the adverse selection component of bid-ask spreads and by Brennan et al. (2016) using conditional probabilities from the PIN model.

13D trades are identified as being event days. On the other hand, 66% of the days when Schedule 13D filers do trade are identified as event days. The increase of 7.9% is statistically significant and represents about a 14% increase in the conditional probability relative to non-13D trading days. Thus, despite the fact that trading by Schedule 13D filers is inversely correlated with the various measures of permanent price impact commonly used in the literature and employed by Collin-Dufresne and Fos (2015), we find that the trading by 13D filers is manifested in higher conditional probabilities of an information event, calculated according to our model.

We also report average CPIE for two subperiods, the first and second halves of the disclosure period (days $[t-60, t-31]$ and $[t-30, t-1]$, respectively). If block accumulation by a 13D filer is detected by other strategic traders, then both the 13D filer and the other strategic traders should trade aggressively to beat others to the market (Holden and Subrahmanyam, 1992). This is more likely to have occurred during the second subperiod, so we expect Schedule 13D filers to trade more aggressively (use more market orders rather than limit orders) in the second subperiod. Furthermore, the second subperiod includes the period after crossing the 5% threshold, after which the 13D must be filed within ten days. We certainly expect more aggressive trading during that period. As a result of these considerations, we expect signed order flow to reflect the presence of informed trade more in the second subperiod than in the first. The second and third rows of Table 4 show that this is indeed the case. There is a smaller difference of 4.9% in CPIE over the first 30 days of the block-accumulation period between Schedule 13D trading days and non-trading days. In the second half of the disclosure period however, the average CPIE is 9.8% higher on days when informed Schedule 13D filers trade than on days they do not.

4.2. Measuring the Information Content of Prices

Some studies use PIN to measure the information content of prices in order to test various economic theories. Applications in corporate finance include Chen, Goldstein and Jiang (2007), Ferreira and Laux (2007), and Bharath, Pasquariello and Wu (2009), and applications in accounting include Frankel and Li (2004), Jayaraman (2008), and Brown and Hillegeist (2007).

Here, we demonstrate how our structural estimates could be used to augment one such study. Chen et al. (2007) study how corporate managers learn from prices in making investment decisions. They find that investment sensitivity to prices (Q) is increasing with price informativeness as proxied by PIN and by $1-R^2$ from an asset pricing model. In Table 5, we replicate Chen et al. (2007) for our sample. Before running the regressions, we standardize each information environment variable to have unit standard deviation. As in Chen et al. (2007), the coefficient on Q is increasing in PIN (column 2).

To demonstrate how researchers might employ our methodology in this setting, we consider two composite measures of the information environment from the hybrid model. The first is the standard deviation of the signal (SD(ξS)) from Equation (10). We also calculate the proportion of the return variance due to private information, which we term the order-flow component of prices (OFC):

$$\frac{\mathrm{var}(\xi S)}{\mathrm{var}(\xi S)+\mathrm{var}(\mathrm{e}^{\delta B_{i1}-\delta^2/2})}=\frac{\mathrm{SD}(\xi S)^2}{\mathrm{SD}(\xi S)^2+e^{\delta^2}-1}\,. \tag{12}$$

Columns 4 and 5 of Table 5 show that investment-price sensitivity is increasing in each of these measures.

One advantage of our estimation procedure relative to PIN is that it allows us to separately estimate the probability and magnitude of information events. Investment sensitivity to prices is increasing in each of these components (column 6 of Table 5). Thus, when there are more frequent or larger episodes of private information, investment is more sensitive to prices. A one standard deviation increase in κ (the magnitude of information events) is associated with a 25% increase in investment-price sensitivity. A standard deviation change in α (the probability of an information event) has an effect about a third as large. The positive effect of α conflicts with results from decomposing PIN into the probability of an information event and the relative intensity of informed to uninformed traders (column 3). An increase in the PIN α does not lead to increased investment sensitivity to prices.

5. Comparison to Other Models

In this section, we compare the estimates of our model to those of the three structural models (PIN, APIN, OWR) and the reduced-form version of PIN (VPIN) discussed in the introduction. The estimation procedure for the other models is detailed in Internet Appendix C.

5.1. Correlations of Model Parameters

Panel A of Table 6 shows the correlations among PIN, APIN, VPIN, lambda from the OWR model (λ_{OWR}), and the expected average lambda from our model (λ_{hybrid}) – see Equation (8). All of the correlations are positive. The largest correlations with λ_{hybrid} are those of the OWR lambda and VPIN. This is perhaps not surprising since each of these estimates uses price changes in some form. The OWR lambda uses the joint distribution of returns and order flows, while VPIN signs volume using price changes.

We call PIN, APIN, VPIN, λ_{owr}, and λ_{hybrid} composite measures of information asymmetry because, with the exception of VPIN, they are functions of the underlying structural parameters.[21] We also examine the correlations of the structural parameters of the various models. Panel B of Table 6 reports correlations of the estimated probability of an information event from each model (except VPIN which does not identify α). The estimates of α for the hybrid model are negatively correlated with estimates of α from the other models. In each of the other models, the unconditional distribution of order flow imbalances changes with α, unlike in our model, so the lack of correlation of the hybrid model α with the other α's is consistent with the identification discussion in Section 2.1. The implications of the

[21] We refer to VPIN as reduced-form because it does not identify the underlying structural parameters. Rather, it proxies for PIN by separately estimating the numerator and denominator of PIN—see Internet Appendix C.4.

models for the unconditional distribution of order flow imbalances are discussed further in Internet Appendix D.[22]

The positive correlation of λ_{hybrid} with the other composite measures is somewhat surprising given that the α of the hybrid model is not positively correlated with the α's of the other models. A potential explanation lies in the estimates of noise trading. Equation (8) shows that the expected average lambda is inversely related to the volatility of noise trading. The other measures are also inversely related to noise trading (see Equations C.2, C.6, and C.4 in the Internet Appendix). Panel C of Table 6 reports correlations of the noise trading parameters of each model. We scale the PIN and APIN noise trading parameters by the estimated μ, so the fractions ε/μ and $(\varepsilon + \theta\eta)/\mu$ represent the intensity of noise trading relative to informed trading. Note that PIN and APIN are decreasing in these ratios, respectively. The noise trading parameters are positively correlated across the models. For this reason, the composite measures are positively correlated despite the lack of correlation of the estimated alphas.

5.2. Cross-Sectional Variation in Parameters

It is interesting to see how estimates of private information differ in the cross-section of firms across models. Table 7 reports average values of the estimates within market capitalization deciles. Across all the models, composite measures of information asymmetry decrease in firm size (Panel A). For the hybrid model, the average probability α of an information event decreases in firm size while the estimates for the other models are exactly the opposite, increasing in firm size (Panel B). As in the unconditional correlation analysis, the composite measures seem to behave similarly in the size cross section due to similarities in noise trading measurement. Estimates from all the models indicate more intense noise trading for larger capitalization stocks. For each of the models other than the hybrid model, the effect of the

[22] Venter and de Jongh (2006), Duarte and Young (2009), Gan, Wei and Johnstone (2014), and Duarte, Hu and Young (2016) all show that the PIN model fails to fit the empirical joint distribution of buy and sell orders.

more pronounced noise trading dominates the modest increases in α as a function of size, so these composite measures are lower for larger firms as a result of higher estimated noise trading.[23]

5.3. Relation to Price Impacts

In theory, price impacts should be larger when information asymmetry is higher. This is true for both the hybrid model and the other models. It is shown in Section 2 for the hybrid model. For the PIN model, the opening quoted spread is the product of PIN and the magnitude of the information, $H - L$.[24] In this section, we assess how cross-sectional variation in price impacts relates to the estimated composite measure from each model. For price impacts, we use the three measures described in Section 3.3.

Figure 8 plots the time-series of the cross-sectional average and interquartile range of the price impact measures and the five composite information asymmetry measures. Over the twenty year sample, price impacts initially rose over the 1990s before falling dramatically following the turn of the century, with the brief exception of the financial crisis. Note that the time-series of the hybrid model expected average lambda and the magnitude of private information, κ, exhibit similar patterns (Figure 6). The OWR lambda exhibits similar behavior. PIN, APIN, and VPIN are much less variable over time.

Table 8 reports cross-sectional regressions of the price impact measures on the composite information asymmetry measures. The information asymmetry measures are winsorized at 1/99% and standardized to have unit standard deviations. In univariate regressions (Panel A), each of the price impact measures is positively related to each of the information asymmetry measures. The hybrid model lambda and VPIN have the highest explanatory power for each of the price impact measures.

In Panel B, we include λ_{hybrid} along with the other information asymmetry measures,

[23]The OWR lambda is also a function of its estimated magnitude of private information σ_i. For both the hybrid model and the OWR model, the estimated magnitude of private information is also increasing in size.

[24]See Equation (11) of Easley et al. (1996), which assumes $p_L = p_H$.

controlling for a firm's size, price, and trading volume using the logarithm of market capitalization, the inverse stock price as of the beginning of the year, and the logarithm of trading volume over the year.[25] In each specification, λ_{hybrid} is significantly positively related to the price impact measure. APIN and the OWR lambda are generally insignificant when λ_{hybrid} is included as an explanatory variable (for the cumulative impulse response measure, the OWR lambda is marginally significant with the wrong sign). For two of the price impact measures, PIN and VPIN remain significant when λ_{hybrid} is included, though the PIN coefficient is much smaller than the λ_{hybrid} and VPIN coefficients. Furthermore, for the estimate of Kyle's lambda based on Hasbrouck (2009) and Goyenko et al. (2009), only λ_{hybrid} is a significant explanatory variable.

6. Conclusion

We propose a model of informed trading that is a hybrid of the PIN and Kyle models. Unlike the Kyle model, information events occur with probability less than one as in the PIN model, and unlike the PIN model, informed orders are endogenously determined as in the Kyle model. An important implication of the model is that both returns and order flows are needed to identify information asymmetry parameters. The reason is that order flows depend on market liquidity, which depends on information asymmetry. This is an indirect dependence of order flows on information asymmetry that is countervailing to the direct relation. This result suggests that measures of information asymmetry based solely on order flows (like PIN) may be misspecified.

We estimate the hybrid model and provide several analyses that suggest the estimates capture cross-sectional and time-series variation in information asymmetry. We illustrate two possible applications of our estimates: a new methodology to detect information events and a corporate finance application. Our model allows the econometrician to identify distinct

[25]Size, price, and volume are strongly related to price impacts (Breen, Hodrick and Korajczyk, 2002); on average, these three control variables explain 65% of the cross-sectional variation in price impact.

components of information asymmetry such as the probability and magnitude of potential information events. We hope such refinements will prove useful to future finance and accounting research.

Finally, we compare the parameter estimates to those from other structural models and to price impact measures. While composite information asymmetry measures from all of the models are positively correlated with price impacts, the measure from the hybrid model (and to a lesser extent VPIN) win in a horse race at explaining price impact measures. To a certain extent this might be expected, since the measure from the hybrid model is the expected average Kyle's lambda, and Kyle's lambda should be highly correlated with price impacts. However, the measure from the Odders-White and Ready (2008) model is also an estimate of a Kyle's lambda, and it is dominated by the hybrid model in the horse race.

Appendix A. Proofs

The process Y described in the following lemma is a variation of a Brownian bridge. It differs from a Brownian bridge in that the endpoint is not uniquely determined but instead is determined only to lie in an interval—either the lower tail $(-\infty, y_L)$, the upper tail (y_H, ∞) or the middle region $[y_L, y_H]$—depending on whether there is an information event and whether the news is good or bad. Part (C) of the lemma follows immediately from the preceding parts, because the probability (A.3) is the probability that $Y_1 \notin [y_L, y_H]$ calculated on the basis that Y is an $\mathbb{F}^Y$–Brownian motion with zero drift and standard deviation σ.

Lemma. *Let* N *denote the standard normal distribution function. Let* $\mathbb{F}^Y = \{\mathcal{F}_t^Y \mid 0 \le t \le 1\}$ *denote the filtration generated by the stochastic process* Y *defined by* $Y_0 = 0$ *and*

$$\mathrm{d}Y_t = \frac{q(t, Y_t, \xi S)}{1-t}\,\mathrm{d}t + \mathrm{d}Z_t\,. \tag{A.1}$$

Then, the following are true:

(A) Y *is an* $\mathbb{F}^Y$*–Brownian motion with zero drift and standard deviation* σ*.*

(B) *With probability one,*

$$\begin{aligned} \xi = 1 \textit{ and } S = L \quad &\Rightarrow \quad Y_1 < y_L\,, && \text{(A.2a)}\\ \xi = 0 \quad &\Rightarrow \quad y_L \le Y_1 \le y_H\,, && \text{(A.2b)}\\ \xi = 1 \textit{ and } S = H \quad &\Rightarrow \quad Y_1 > y_H\,. && \text{(A.2c)} \end{aligned}$$

(C) *For each* $t < 1$*, the probability that* $\xi = 1$ *conditional on* $\mathcal{F}_t^Y$ *is*

$$\mathrm{N}\left(\frac{y_L - Y_t}{\sigma\sqrt{1-t}}\right) + 1 - \mathrm{N}\left(\frac{y_H - Y_t}{\sigma\sqrt{1-t}}\right)\,. \tag{A.3}$$

Proof of the Lemma. Set

$$k(1,y,s) = \begin{cases} 1_{\{y<y_L\}} & \text{if } s = L\,, \\ 1_{\{y_L \le y \le y_H\}} & \text{if } s = 0\,, \\ 1_{\{y>y_H\}} & \text{if } s = H\,, \end{cases}$$

and, for $t < 1$,

$$k(t,y,s) = \begin{cases} \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) & \text{if } s = L\,, \\ \mathrm{N}\left(\frac{y_H - y}{\sigma\sqrt{1-t}}\right) - \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) & \text{if } s = 0\,, \\ \mathrm{N}\left(\frac{y - y_H}{\sigma\sqrt{1-t}}\right) & \text{if } s = H\,. \end{cases}$$

Define

$$\ell(t,y,s) = \frac{\partial \log k(t,y,s)}{\partial y}\,,$$

for $t < 1$. Then, $(1-t)\sigma^2\ell(t,y,s) = q(t,y,s)$ for $t < 1$, and the stochastic differential equation (A.1) can be written as

$$\mathrm{d}Y_t = \sigma^2\,\ell(t,Y_t,\xi S)\,\mathrm{d}t + \mathrm{d}Z_t \tag{A.4}$$

The process Y is an example of a Doob h-transform—see Rogers and Williams (2000).

To put (A.4) in a more standard form, define the two-dimensional process $\hat{Y}_t = (\xi S, Y_t)$ with random initial condition $\hat{Y}_0 = (\xi S, 0)$, and augment (A.4) with the equation $\mathrm{d}(\xi S) = 0$. The existence of a unique strong solution $\hat{Y}$ to this enlarged system follows from Lipschitz and growth conditions satisfied by ℓ. See Karatzas and Shreve (1988, Theorem 5.2.9).

The uniqueness in distribution of weak solutions of stochastic differential equations (Karatzas and Shreve, 1988, Theorem 5.3.10) implies that we can demonstrate Properties (A) and (B) by exhibiting a weak solution for which they hold. To construct such a weak solution, define a new measure $\mathbb{Q}$ on $\mathcal{F}_1$ using $k(1, Z_1, \xi S)/k(0, 0, \xi S)$ as the Radon-Nikodym derivative. The definition of k implies that $k(t, Z_t, \xi S)$ is the $\mathcal{F}_t$–conditional expectation of the indicator function $k(1, Z_1, \xi S)$, so $k(t, Z_t, \xi S)$ is a martingale on the filtration $\mathbb{F}$. By Girsanov's Theorem, the process Z^* defined by $Z_0^* = 0$ and

$$\mathrm{d}Z_t^* = -\sigma^2 \, \ell(t, Z_t, \xi S) \, \mathrm{d}t + \mathrm{d}Z_t$$

is a Brownian motion (with zero drift and standard deviation σ) on the filtration $\mathbb{F}$ relative to $\mathbb{Q}$. It follows that Z is a weak solution of (A.4) relative to the Brownian motion Z^* on the filtered probability space $(\Omega, \mathbb{F}, \mathbb{Q})$.

To establish Property (A) for the weak solution, we need to show that Z is a Brownian motion on $(\Omega, \mathbb{G}, \mathbb{Q})$. Because Z is a Brownian motion on $(\Omega, \mathbb{G}, \mathbb{P})$, it suffices to show that $\mathbb{Q} = \mathbb{P}$ when both are restricted to $\mathcal{G}_1$. This holds if for all $t_1 < \cdots < t_n \leq 1$ and all Borel B we have

$$\mathbb{P}((Z_{t_1}, \ldots, Z_{t_n}) \in B) = \mathbb{Q}((Z_{t_1}, \ldots, Z_{t_n}) \in B) \, . \tag{A.5}$$

The right-hand side of (A.5) equals

$$\mathsf{E}\left[\frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} \mathbf{1}_B(Z_{t_1}, \ldots, Z_{t_n})\right] ,$$

which can be represented as the following sum:

$$\begin{aligned} \alpha p_L \mathsf{E}\left[\frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} \mathbf{1}_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = L\right] & \\ + (1-\alpha) \mathsf{E}\left[\frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} \mathbf{1}_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi = 0\right] & \\ + \alpha p_H \mathsf{E}\left[\frac{k(1, Z_1, \xi S)}{k(0, 0, \xi S)} \mathbf{1}_B(Z_{t_1}, \ldots, Z_{t_n}) \mid \xi S = H\right] & \, . \end{aligned}$$

Using the definitions of y_L, y_H, and k, this equals

$$\begin{aligned}\mathsf{E}\left[\mathbf{1}_{\{Z_1<y_L\}}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n}) \mid \xi S = L\right] \\ + \mathsf{E}\left[\mathbf{1}_{\{y_L\le Z_1\le y_L\}}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n}) \mid \xi = 0\right] \\ + \mathsf{E}\left[\mathbf{1}_{\{Z_1>y_H\}}\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n}) \mid \xi S = H\right] .\end{aligned}$$

The $\mathbb{P}$–independence of Z and ξS imply that the conditional expectations equal the unconditional expectations, so adding the three terms gives

$$\mathsf{E}\left[\mathbf{1}_B(Z_{t_1},\ldots,Z_{t_n})\right] = \mathbb{P}((Z_{t_1},\ldots,Z_{t_n}) \in B)\,.$$

This completes the proof that Z is a Brownian motion on $(\Omega, \mathbb{G}, \mathbb{Q})$.

To establish Property (B) for the weak solution of (A.4), we need to show that

$$\mathbb{Q}(Z_1 < y_L \mid \xi S = L) = 1\,, \tag{A.6a}$$

$$\mathbb{Q}(y_L \le Z_1 \le y_H \mid \xi = 0) = 1\,, \tag{A.6b}$$

$$\mathbb{Q}(Z_1 > y_H) \mid \xi S = H) = 1\,. \tag{A.6c}$$

Consider (A.6a). We have

$$\begin{aligned}\mathbb{Q}(\xi S = L) &= \mathsf{E}\left[\frac{k(1,Z_1,\xi S)}{k(0,0,\xi S)}\mathbf{1}_{\{\xi S=L\}}\right] \\ &= \mathsf{E}\left[\frac{k(1,Z_1,L)}{k(0,0,L)}\mathbf{1}_{\{\xi S=L\}}\right] \\ &= \mathsf{E}\left[\mathbf{1}_{\{Z_1<y_L\}}\mathbf{1}_{\{\xi S=L\}}\right]\big/\,\alpha p_L \\ &= \alpha p_L\,,\end{aligned}$$

using the definition of k for the third equality and the $\mathbb{P}$–independence of Z and ξS for the last equality. By similar reasoning,

$$\begin{aligned}
\mathbb{Q}(Z_1 < y_L, \xi S = L) &= \mathsf{E}\left[\frac{k(1, Z_1, \xi S)}{k(0,0,\xi S)} \mathbf{1}_{\{Z_1<y_L\}} \mathbf{1}_{\{\xi S = L\}}\right] \\
&= \mathsf{E}\left[\frac{k(1, Z_1, L)}{k(0,0,L)} \mathbf{1}_{\{Z_1<y_L\}} \mathbf{1}_{\{\xi S = L\}}\right] \\
&= \mathsf{E}\left[\mathbf{1}_{\{Z_1<y_L\}} \mathbf{1}_{\{\xi S = L\}}\right] \big/ \alpha p_L \\
&= \alpha p_L\,.
\end{aligned}$$

Thus,

$$\mathbb{Q}(Z_1 < y_L \mid \xi S = L) = \frac{\mathbb{Q}(Z_1 < y_L, \xi S = L)}{\mathbb{Q}(\xi S = L)} = \frac{\alpha p_L}{\alpha p_L} = 1\,.$$

Conditions (A.6b) and (A.6c) can be verified by the same logic. □

Proof of Theorem 1. It is explained in the text why the equilibrium condition (1) holds. It remains to show that the strategy (5) is optimal for the informed trader. Let $\mathbb{G} \stackrel{\text{def}}{=} \{\mathcal{G}_t \mid 0 \leq t \leq T\}$ denote the completion of the filtration generated by Z, form the enlarged filtration with σ–fields $\mathcal{G}_t \vee \sigma(\xi S)$, and let $\mathbb{F} \stackrel{\text{def}}{=} \{\mathcal{F}_t \mid 0 \leq t \leq T\}$ denote the completion of the enlarged filtration. The filtration $\mathbb{F}$ represents the informed trader's information.

Define

$$\begin{aligned}
J(1, y, L) &= -L(y - y_L)1_{\{y>y_L\}} + H(y - y_H)1_{\{y>y_H\}}\,, \\
J(1, y, 0) &= -L(y_L - y)1_{\{y<y_L\}} + H(y - y_H)1_{\{y>y_H\}}\,, \\
J(1, y, H) &= -L(y_L - y)1_{\{y<y_L\}} + H(y_H - y)1_{\{y<y_H\}}\,.
\end{aligned}$$

For $t < 1$ and $s \in \{L, 0, H\}$, set $J(t, y, s) = \mathsf{E}[J(t, Z_1, s) \mid Z_t = y]$. Then, $J(t, Z_t, \xi S)$ is an $\mathbb{F}$–martingale, so it has zero drift. From Itô's formula, its drift is

$$\frac{\partial}{\partial t} J(t, Z_t, \xi S) + \frac{1}{2}\sigma^2 \frac{\partial^2}{\partial z^2} J(t, Z_t, \xi S)\,.$$

Equating this to zero, Itô's formula implies

$$J(1, Y_1, \xi S) = J(0, 0, \xi S) + \int_0^1 \mathrm{d}J(t, Y_t, \xi S) = J(0, 0, \xi S) + \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y} \, \mathrm{d}Y_t \, .$$

Therefore,

$$\mathsf{E}[J(1, Y_1, \xi S) - J(0, 0, \xi S)] = \mathsf{E} \int_0^1 \frac{\partial J(t, Y_t, \xi S)}{\partial y} \, \mathrm{d}Y_t \, . \tag{A.7}$$

To calculate $\partial J(t, y, s)/\partial y$, use the fact that, by independent increments,

$$J(t, y, s) = \mathsf{E}[J(t, Z_1, s) \mid Z_t = y] = \mathsf{E}[J(t, Z_1 - Z_t + y, s)]$$

to obtain

$$\frac{\partial J(t, y, s)}{\partial y} = \mathsf{E}\left[\frac{\partial}{\partial y} J(t, Z_1 - Z_t + y, s)\right] \, .$$

Now, note that, for any real number a excluding the kinks at $y_L - y$ and $y_H - y$,

$$\begin{aligned}
\frac{\partial}{\partial y} J(1, a + y, L) &= -L 1_{\{a > y_L - y\}} + H 1_{\{a > y_H - y\}} \, , \\
\frac{\partial}{\partial y} J(1, a + y, 0) &= L 1_{\{a < y_L - y\}} + H 1_{\{a > y_H - y\}} \, , \\
\frac{\partial}{\partial y} J(1, a + y, H) &= L 1_{\{a < y_L - y\}} - H 1_{\{a < y_H - y\}} \, .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\partial J(t, y, L)}{\partial y} &= -L \, \mathrm{N}\left(\frac{y - y_L}{\sigma\sqrt{1-t}}\right) + H \, \mathrm{N}\left(\frac{y - y_H}{\sigma\sqrt{1-t}}\right) \, , \\
\frac{\partial J(t, y, 0)}{\partial y} &= L \, \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) + H \, \mathrm{N}\left(\frac{y - y_H}{\sigma\sqrt{1-t}}\right) \, , \\
\frac{\partial J(t, y, H)}{\partial y} &= L \, \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) - H \, \mathrm{N}\left(\frac{y_H - y}{\sigma\sqrt{1-t}}\right) \, .
\end{aligned}$$

Now, the definition (6) gives us

$$\frac{\partial J(t,y,s)}{\partial y} = p(t,y) - s$$

for all $s \in \{L, 0, H\}$. Substituting this into (A.7) gives us

$$\mathsf{E}[J(1,Y_1,\xi S) - J(0,0,\xi S)] = \mathsf{E}\int_0^1 [p(t,Y_t) - \xi S]\,\mathrm{d}Y_t\,. \tag{A.8}$$

The "no doubling strategies" condition implies that $\int p\,\mathrm{d}Z$ is a martingale, so the right-hand side of this equals

$$\mathsf{E}\int_0^1 [p(t,Y_t) - \xi S]\theta_t\,\mathrm{d}t\,.$$

Rearranging produces

$$\mathsf{E}\int_0^1 [\xi S - p(t,Y_t)]\theta_t\,\mathrm{d}t = \mathsf{E}[J(0,0,\xi S) - J(1,Y_1,\xi S)] \leq \mathsf{E}[J(0,0,\xi S)]\,,$$

using the fact that $J(1,y,s) \geq 0$ for all (y,s) for the inequality. Thus, $\mathsf{E}[J(0,0,\xi S)]$ is an upper bound on the expected profit, and the bound is achieved if and only if $J(1,Y_1,\xi S) = 0$ with probability one. By the definition of $J(1,y,s)$, this is equivalent to $Y_1 < y_L$ with probability one when $\xi S = L$, $y_L \leq Y_1 \leq y_H$ with probability one when $\xi = 0$, and $Y_1 > y_H$ with probability one when $\xi S = H$. By part (B) of the proposition, the strategy (5) is therefore optimal. □

Proof of Theorem 2. By Itô's formula and the fact that $(\mathrm{d}Y)^2 = (\mathrm{d}Z)^2 = \sigma^2\,\mathrm{d}t$, we have

$$\mathrm{d}p(t,Y_t) = \left(p_t(t,Y_t) + \frac{1}{2}\sigma^2 p_{yy}(t,Y_t)\right)\mathrm{d}t + p_y(t,Y_t)\,\mathrm{d}Y_t\,,$$

where we use subscripts to denote partial derivatives. Both Y and $p(t,Y_t)$ are martingales with respect to the market makers' information, so the drift term must be zero. That can also be verified by direct calculation of the partial derivatives, using the formula (6) for

$p(t, y)$. Thus,

$$\mathrm{d}p(t, Y_t) = p_y(t, Y_t)\,\mathrm{d}Y_t\,.$$

A direct calculation based on the formula (6) for $p(t, y)$ shows that $p_y(t, y) = \lambda(t, y)$ defined in (7).

To see that $\lambda(t, Y_t)$ is a martingale for $t \in [0, 1)$, with respect to market makers' information, we can calculate, for $t < u < 1$,

$$\begin{aligned}\mathsf{E}[\lambda(u, Y_u) \mid Y_t = y] = -\frac{L}{\sigma\sqrt{1-u}} \cdot \int_{-\infty}^{\infty} \mathrm{n}\left(\frac{y_L - y'}{\sigma\sqrt{1-u}}\right) f(y' \mid u-t, y)\mathrm{d}y' \\ + \frac{H}{\sigma\sqrt{1-u}} \cdot \int_{-\infty}^{\infty} \mathrm{n}\left(\frac{y_H - y'}{\sigma\sqrt{1-u}}\right) f(y' \mid u-t, y)\mathrm{d}y'\,,\end{aligned}$$

where $f(\cdot \mid \tau, y)$ denotes the normal density function with mean y and variance $\sigma^2\tau$. A straightforward calculation shows that this equals $\lambda(t, y)$. For example, to evaluate the first term, use the fact that

$$\begin{aligned}&\frac{1}{\sigma\sqrt{1-u}}\,\mathrm{n}\left(\frac{y_L - y'}{\sigma\sqrt{1-u}}\right) f(y' \mid u-t, y) \\ &\qquad = \frac{1}{\sigma\sqrt{1-t}}\,\mathrm{n}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) \times \frac{1}{\sqrt{2\pi\sigma^2(1-u)(u-t)/(1-t)}} \\ &\qquad\qquad \times \exp\left(-\left(\frac{1-t}{2(1-u)(u-t)\sigma^2}\right)\left(y' - \frac{(1-u)y + (u-t)y_L}{1-t}\right)^2\right)\,,\end{aligned}$$

which integrates to

$$\frac{1}{\sigma\sqrt{1-t}}\,\mathrm{n}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right)\,,$$

because the other factors constitute a normal density function. $\square$

Appendix B. Hybrid Model Likelihood Function

Assume the trading period $[0, 1]$ corresponds to a day. This implies that any private information becomes public before trading opens on the following day.[26] We can estimate the model parameters using intraday price and order flow information. If we assume further that the model parameters are stable over time, then the price and order flow information from multiple days can be merged to estimate the parameters with greater precision.

The opening price on each day i is $P_{i0} \stackrel{\text{def}}{=} \mathsf{E}[V_{i1} + \xi_i S_i] = V_{i0}$. To obtain stationarity, we assume that the signal S_i on day i is proportional to the observed opening price P_{i0}. This construction causes the pricing function to be day-specific, and we denote it by $p_i(t, y)$. In fact,

$$p_i(t, y) = P_{i0} \times p(t, y)$$

where $p(t, y)$ is defined in Theorem 1. We specify that $H = -L = \kappa$ in the empirical implementation. Under this specification, the pricing function expressed in returns is given by:

$$p(t, Y_t) = \begin{cases} -\kappa 1_{\{Y_t < y_L\}} + \kappa 1_{\{Y_t > y_H\}} & \text{if } t = 1\,, \\ -\kappa F(y_L | t, Y_t) + \kappa(1 - F(y_H | t, Y_t)) & \text{if } t < 1 \end{cases}$$

where $F(y|t, Y_t)$ is the normal distribution function with mean Y_t and variance $(1 - t)\sigma^2$.

The price at time t on day i is $V_{it} + p_i(t, Y_{it})$, so the gross return through time t is

$$\frac{P_{it}}{P_{i0}} = \frac{V_{it}}{V_{i0}} + \frac{p_i(t, Y_{it})}{P_{i0}} = \frac{V_{it}}{V_{i0}} + p(t, Y_{it})\,. \tag{B.1}$$

Assume

$$\frac{\mathrm{d}V_{it}}{V_{it}} = \delta\, \mathrm{d}B_{it}$$

[26]In contrast to Odders-White and Ready (2008), our estimation does not use overnight returns. In our theoretical model, private information that is made public after the close of trading is incorporated into prices before trading ends (convergence to strong-form efficiency). Thus, overnight returns in our model are due to arrival of new public information, which does not aid in estimating the model.

for a constant δ and a Brownian motion B_i, so we have

$$\frac{P_{it}}{P_{i0}} = p(t, Y_{it}) + \mathrm{e}^{\delta B_{it} - \delta^2 t/2}\,.$$

Assume the price and order imbalance are observed at times $t_1, \ldots, t_{k+1}$ each day with $t_{k+1} = 1$ being the close and the other times being equally spaced: $t_j = j\Delta$ for $\Delta > 0$ and $j \leq k$. Let P_{ij} denote the observed price and Y_{ij} the observed order imbalance at time t_j on date i. Define

$$\Gamma = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ k \\ 1/\Delta \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 2 & \cdots & 2 & 2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 2 & \cdots & k & k \\ 1 & 2 & \cdots & k & 1/\Delta \end{pmatrix}.$$

On each day i, the vector $Y_i = (Y_{i,t_1}, \ldots, Y_{i,t_{k+1}})'$ is normally distributed with mean 0 and covariance matrix $\sigma^2 \Delta \Sigma$. Set

$$U_{ij} = \log\left(\frac{P_{ij}}{P_{i0}} - p(t_j, Y_{ij})\right) \tag{B.2}$$

and $U_i = (U_{i1}, \ldots, U_{i,k+1})'$. The density function of $(P_{i1}/P_{i0}, \ldots, P_{i,k+1}/P_{i0})$ conditional on Y_i is

$$f(U_{i1}, \ldots U_{i,k+1})\mathrm{e}^{-\sum_{j=1}^{k+1} U_{ij}}\,,$$

where f denotes the multivariate normal density function with mean vector $-(\delta^2\Delta/2)\Gamma$ and covariance matrix $\delta^2 \Delta \Sigma$.

Let $\mathcal{L}_i$ denote the log-likelihood function for day i. Dropping terms that do not depend on the parameters, we have

$$- \mathcal{L}_i = (k+1)\log\sigma + \frac{1}{2\sigma^2\Delta} Y_i' \Sigma^{-1} Y_i + (k+1)\log\delta + \frac{1}{2\delta^2\Delta}\left(U_i + \frac{\delta^2\Delta}{2}\Gamma\right)' \Sigma^{-1}\left(U_i + \frac{\delta^2\Delta}{2}\Gamma\right) + \sum_{j=1}^{k+1} U_{ij}\,.$$

Using the facts that $\Gamma'\Sigma^{-1} = (0, \ldots, 0, 1)$ and $\Gamma'\Sigma^{-1}\Gamma = 1/\Delta$, this simplifies to

$$- \mathcal{L}_i = (k+1)\log\sigma + \frac{1}{2\sigma^2\Delta} Y_i' \Sigma^{-1} Y_i + (k+1)\log\delta + \frac{1}{2\delta^2\Delta} U_i' \Sigma^{-1} U_i + \frac{1}{2} U_{i,k+1} + \frac{\delta^2}{8} + \sum_{j=1}^{k+1} U_{ij}\,.$$

Hence, the log-likelihood function $\mathcal{L}$ for an observation period of n days satisfies

$$- \mathcal{L} = n(k+1)\log\sigma + \frac{1}{2\sigma^2\Delta}\sum_{i=1}^{n} Y_i' \Sigma^{-1} Y_i + n(k+1)\log\delta + \frac{1}{2\delta^2\Delta}\sum_{i=1}^{n} U_i' \Sigma^{-1} U_i + \frac{n\delta^2}{8} + \sum_{i=1}^{n}\left(\sum_{i=1}^{k} U_{ik} + \frac{3}{2} U_{i,k+1}\right). \quad \text{(B.3)}$$

We minimize (B.3) in α, κ, p_L, σ, and δ (note that κ and p_L enter $\mathcal{L}$ because they affect the function p that enters $\mathcal{L}$ via (B.2)). We sample every hour and at the close, so $\Delta = 1/6.5$.

References

Akay, O., Cyree, K.B., Griffiths, M.D., Winters, D.B., 2012. What does PIN identify? Evidence from the T-bill market. Journal of Financial Markets 15, 29–46.

Akins, B., Ng, J., Verdi, R.S., 2012. Investor competition over information and the pricing of information asymmetry. The Accounting Review 87, 35–58.

Aktas, N., de Bodt, E., Declerck, F., Van Oppens, H., 2007. The PIN anomaly around M&A announcements. Journal of Financial Markets 10, 160–191.

Andersen, T., Bondarenko, O., 2014a. Reflecting on the VPIN dispute. Journal of Financial Markets 17, 53–64.

Andersen, T., Bondarenko, O., 2014b. VPIN and the flash crash. Journal of Financial Markets 17, 1–46.

Back, K., 1992. Insider trading in continuous time. Review of Financial Studies 5, 387–409.

Back, K., Crotty, K., 2015. The informational role of stock and bond volume. Review of Financial Studies 28, 1381–1427.

Banerjee, S., Green, B., 2015. Signal or noise? Uncertainty and learning about whether other traders are informed. Journal of Financial Economics 117, 398–423.

Benos, E., Jochec, M., 2007. Testing the PIN variable. Working Paper, University of Illinois.

Bharath, S.T., Pasquariello, P., Wu, G., 2009. Does asymmetric information drive capital structure decisions? Review of Financial Studies 22, 3211–3243.

Breen, W.J., Hodrick, L.S., Korajczyk, R.A., 2002. Predicting equity liquidity. Management Science 48, 470–483.

Brennan, M.J., Huh, S.W., Subrahmanyam, A., 2016. High-Frequency Measures of Informed Trading and Corporate Announcements. Working Paper,. UCLA.

Brown, S., Hillegeist, S.A., 2007. How disclosure quality affects the level of information asymmetry. Review of Accounting Studies 12, 443–477.

Brown, S., Hillegeist, S.A., Lo, K., 2004. Conference calls and information asymmetry. Journal of Accounting and Economics 37, 343–366.

Brown, S., Hillegeist, S.A., Lo, K., 2009. The effect of earnings surprises on information asymmetry. Journal of Accounting and Economics 47, 208–225.

Chakraborty, A., Yilmaz, B., 2004. Manipulation in market order models. Journal of Financial Markets 7, 187–206.

Chen, L., Collin-Dufresne, P., Goldstein, R.S., 2009. On the relation between the credit spread puzzle and the equity premium puzzle. Review of Financial Studies 22, 3367–3409.

Chen, Q., Goldstein, I., Jiang, W., 2007. Price informativeness and investment sensitivity to stock price. Review of Financial Studies 20, 619–650.

Collin-Dufresne, P., Fos, V., 2015. Do prices reveal the presence of informed trading? Journal of Finance 70, 1555–1582.

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.

Duarte, J., Hu, E., Young, L., 2016. What Does the PIN Model Identify as Private Information? Working Paper. Rice University and University of Washington.

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

Easley, D., Engle, R.F., O'Hara, M., Wu, L., 2008. Time-varying arrival rates of informed and uninformed trades. Journal of Financial Econometrics 6, 171–207.

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

Easley, D., Hvidkjaer, S., O'Hara, M., 2010. Factoring information into returns. Journal of Financial and Quantitative Analysis 45, 293–309.

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

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

Easley, D., López de Prado, M., O'Hara, M., 2011. The microstructure of the "flash crash": Flow toxicity, liquidity crashes, and the probability of informed trading. Journal of Portfolio Management 37, 118–128.

Easley, D., López de Prado, M., O'Hara, M., 2012. Flow toxicity and liquidity in a high-frequency world. Review of Financial Studies 25, 1457–1493.

Easley, D., López de Prado, M., O'Hara, M., 2014. VPIN and the flash crash: A rejoinder. Journal of Financial Markets 17, 47–52.

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

Ferreira, M.A., Laux, P.A., 2007. Corporate governance, idiosyncratic risk, and information flow. Journal of Finance 62, 951–989.

Foster, F.D., Viswanathan, S., 1995. Can speculative trade explain the volume-volatility relation? Journal of Business & Economic Statistics 13, 379–396.

Frankel, R., Li, X., 2004. Characteristics of a firm's information environment and the information asymmetry between insiders and outsiders. Journal of Accounting and Economics 37, 229–259.

Gan, Q., Wei, W.C., Johnstone, D., 2014. Does the probability of informed trading model fit empirical data? Working Paper.

Glosten, L.R., Harris, L.E., 1988. Estimating the components of the bid/ask spread. Journal of Financial Economics 21, 123–142.

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

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

Hasbrouck, J., 1988. Trades, quotes, inventories, and information. Journal of Financial Economics 22, 229–252.

Hasbrouck, J., 1991. Measuring the information content of stock trades. Journal of Finance 46, 179–207.

Hasbrouck, J., 2009. Trading costs and returns for u.s. equities: Estimating effective costs from daily data. Journal of Finance 64, 14451477.

Holden, C.W., Jacobsen, S., 2014. Liquidity measurement problems in fast, competitive markets: Expensive and cheap solutions. Journal of Finance 69, 1747–1785.

Holden, C.W., Subrahmanyam, A., 1992. Long-lived private information and imperfect competition. Journal of Finance , 247–270.

Hwang, L.S., Lee, W.J., Lim, S.Y., Park, K.H., 2013. Does information risk affect the implied cost of equity capital? an analysis of PIN and adjusted PIN. Journal of Accounting and Economics 55, 148–167.

Jayaraman, S., 2008. Earnings volatility, cash flow volatility, and informed trading. Journal of Accounting Research 46, 809–851.

Karatzas, I., Shreve, S.E., 1988. Brownian Motion and Stochastic Calculus. Springer-Verlag, New York.

Kim, O., Verrecchia, R.E., 1997. Pre-announcement and event-period private information. Journal of Accounting and Economics 24, 394–419.

Krinsky, I., Lee, J., 1996. Earnings announcements and the components of the bid-ask spread. Journal of Finance 51, 1523–1535.

Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1336.

Lee, C.M., Ready, M.J., 1991. Inferring trade direction from intraday data. Journal of Finance 46, 733–746.

Li, H., Wang, J., Wu, C., He, Y., 2009. Are liquidity and information risks priced in the Treasury bond market? Journal of Finance 64, 467–503.

Li, T., 2012. Insider trading with uncertain informed trading. Working Paper, City University of Hong Kong.

Mohanram, P., Rajgopal, S., 2009. Is PIN priced risk? Journal of Accounting and Economics 47, 226–243.

Newey, W.K., West, K.D., 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55, 703–708.

Odders-White, E.R., Ready, M.J., 2008. The probability and magnitude of information events. Journal of Financial Economics 87, 227–248.

Rogers, L.C.G., Williams, D., 2000. Diffusions, Markov Processes and Martingales: Vol. 2: Itô Calculus. 2nd ed., Cambridge University Press, Cambridge.

Rossi, S., Tinn, K., 2010. Man or machine? Rational trading without information about fundamentals. Working Paper.

Venter, J.H., de Jongh, D., 2006. Extending the EKOP model to estimate the probability of informed trading. Studies in Economics and Econometrics 30, 25–39.

Table 1: Price Impact and Probability and Magnitude of Information Events
Panel regressions of price impacts on the estimated probability of an information event α and the magnitude of an information event κ (Panel A) and the standard deviation of the signal (SD(ξS)) (Panel B). The dependent variables are the 5-minute price impact, the cumulative impulse response estimated following Hasbrouck (1991), and an estimate of Kyle's lambda ($\widehat{\lambda}_{\text{intraday}}$) using a regression of 5-minute returns on the square-root of signed volume following Hasbrouck (2009) and Goyenko et al. (2009). All variables are standardized to have a unit standard deviation. Standard errors are clustered by firm and year. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Panel A. Probability and Magnitude of Information Events

	5-Minute Price Impact		Cumulative Impulse Response		$\widehat{\lambda}_{\text{intraday}}$	
	(1)	(2)	(3)	(4)	(5)	(6)
α	0.21***	0.09***	0.16***	0.06***	0.22***	0.12***
	(4.73)	(3.93)	(3.66)	(2.93)	(4.58)	(3.98)
κ	0.62***	0.41***	0.45***	0.26***	0.72***	0.55***
	(20.00)	(10.77)	(11.00)	(6.58)	(10.58)	(8.61)
Constant	0.14***	-0.25***	0.48***	-0.54***	-0.57***	0.25***
	(7.12)	(-6.10)	(19.64)	(-12.78)	(-16.80)	(3.67)
Observations	19940	19940	19940	19940	19940	19940
R^2	0.605	0.810	0.630	0.833	0.396	0.658
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes

Panel B. Unconditional Signal Standard Deviation

	5-Minute Price Impact		Cumulative Impulse Response		$\widehat{\lambda}_{\text{intraday}}$	
	(1)	(2)	(3)	(4)	(5)	(6)
SD(ξS)	0.73***	0.52***	0.55***	0.36***	0.84***	0.69***
	(27.42)	(18.45)	(13.97)	(8.83)	(11.07)	(11.55)
Constant	0.07***	-0.19***	0.42***	-0.47***	-0.65***	0.33***
	(5.16)	(-6.64)	(21.29)	(-12.21)	(-16.48)	(5.65)
Observations	19940	19940	19940	19940	19940	19940
R^2	0.635	0.825	0.659	0.845	0.440	0.684
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes

Table 2: Excess Kurtosis and Probability of an Information Event
Panel regressions of excess kurtosis of returns relative to the estimated probability of an information event α. The first column uses simulated data from the model using α values ranging from 0.05 to 0.95 in 0.05 increments. For each α value, the panel contains 1000 simulated firm-years. Excess kurtosis is estimated for each firm-year of 252 days. The other parameters are $\kappa = 0.02$, $\sigma = 0.1$, $p_L = 0.5$, and $\delta = 0.02$. The remaining columns use excess kurtosis and estimated α for NYSE firms. Standard errors are clustered by firm and year. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

	(1)	(2)	(3)	(4)
α	-0.62***	-0.31**	-0.26*	-0.26*
	(-14.21)	(-2.10)	(-1.93)	(-1.81)
Log(Size)			0.18***	0.11
			(2.85)	(1.34)
Inverse Price			0.09	0.30
			(0.20)	(1.10)
Log(Volume)			-0.20***	-0.05
			(-3.33)	(-0.48)
Observations	19000	19940	19940	19940
R^2	0.275	0.139	0.144	0.269
Year FE	N/A	Yes	Yes	Yes
Firm FE	No	No	No	Yes
Data	Simulated	Actual	Actual	Actual

Table 3: End-of-day Volatility and Conditional Probability of an Information Event
Panel regressions of end-of-day absolute returns as a function of the conditional probability of an information event (CPIE). The dependent variable is the absolute return over the last three and a half hours of the day. The conditional probabilities are based on the cumulative order flow over the first three hours of the day as defined in Equation (11). The first column uses simulated data from the model using α values ranging from 0.05 to 0.95 in 0.05 increments. For each α value, the panel contains 1000 simulated firm-years. The other parameters are $\kappa = 0.02$, $\sigma = 0.1$, $p_L = 0.5$, and $\delta = 0.02$. The remaining columns use daily data for NYSE firms. Standard errors are clustered by firm and year. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

	(1)	(2)	(3)	(4)
CPIE	28.05***	14.77***	7.70***	4.44***
	(4.81)	(11.43)	(7.07)	(5.34)
Lag Abs Ret		0.21***	0.15***	0.15***
		(16.84)	(10.45)	(10.39)
Abs OIB				4.49***
				(17.53)
Constant	112.60***	50.57***	55.00***	54.62***
	(31.61)	(41.35)	(57.34)	(52.87)
Observations	4788000	4908887	4908887	4908887
R^2	0.009	0.091	0.134	0.135
Year FE	N/A	Yes	Yes	Yes
Firm FE	N/A	No	Yes	Yes
Data	Simulated	Actual	Actual	Actual

Table 4: Conditional Probability of an Information Event on Days when Schedule 13D Filers Trade
Average levels of the conditional probability of an information event (CPIE) on days when Schedule 13D filers do or do not trade. CPIE is defined in Equation (11). The sample contains trading days in the sixty-day disclosure period prior to a Schedule 13D filing date for NYSE firms in the sample of Collin-Dufresne and Fos (2015). The first column reports the average CPIE on days when Schedule 13D filers trade. The second column reports the average CPIE on days when Schedule 13D filers do not trade. The third column reports the differences between the two types of days. We report the analysis for two subperiods, the first and second halves of the disclosure period (days $[t-60, t-31]$ and $[t-30, t-1]$, respectively). Standard errors are clustered by event. t statistics of the differences are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

	Days with Informed Trading	Days with No Informed Trading	Difference
	(1)	(2)	(3)
Full Disclosure Window: Days $[t-60, t-1]$			
CPIE	65.6	57.6	7.9***
			(4.73)
1st Half of Disclosure Window: Days $[t-60, t-31]$			
CPIE	62.5	57.5	4.9***
			(2.03)
2nd Half of Disclosure Window: Days $[t-30, t-1]$			
CPIE	67.6	57.8	9.8***
			(5.03)

Table 5: Investment-Price Sensitivity and the Information Content of Prices

Panel regressions of corporate investment (capital expenditures) on market-to-book of assets (Q) and the information content of prices following Chen et al. (2007). PIN is the probability of informed trading from Easley et al. (1996). SD(ξS) is the standard deviation of the signal ξS as in Equation (10). OFC is the proportion of return variance due to private information (the order-flow component of prices) as in Equation (12). α is the probability of an information event in either the PIN or hybrid model. κ_{hybrid} is the magnitude of an information event and σ_{hybrid} is the standard deviation of noise trading from the hybrid model. ε/μ is the ratio of the uninformed to informed trading intensities from PIN. CF is firm cash flows. RET is the cumulative return over the next three years. INV ASSET is the inverse of the book value of assets. Standard errors are clustered by firm and year. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

	(1)	(2)	(3)	(4)	(5)	(6)
Q	1.59***	1.17***	2.08***	1.17***	1.29***	1.12***
	(8.36)	(4.66)	(7.35)	(4.84)	(5.67)	(3.69)
$Q \times$ PIN		0.19***				
		(2.61)				
$Q \times \alpha_{\text{PIN}}$			-0.01			
			(-0.11)			
$Q \times \frac{\varepsilon}{\mu}$			-0.30***			
			(-2.71)			
$Q \times$ SD(ξS)				0.26***		
				(3.27)		
$Q \times$ OFC					0.20**	
					(2.28)	
$Q \times \alpha_{\text{hybrid}}$						0.10***
						(2.87)
$Q \times \kappa_{\text{hybrid}}$						0.29***
						(3.64)
$Q \times \sigma_{\text{hybrid}}$						-0.22**
						(-1.99)
CF	7.71***	7.73***	7.89***	7.85***	7.98***	7.66***
	(5.44)	(5.46)	(5.55)	(5.55)	(5.54)	(5.56)
RET	-0.18	-0.18	-0.19*	-0.16	-0.19*	-0.19*
	(-1.59)	(-1.56)	(-1.68)	(-1.56)	(-1.70)	(-1.76)
INV ASSET	0.56***	0.52**	0.51**	0.56***	0.53**	0.49**
	(2.70)	(2.56)	(2.50)	(2.73)	(2.57)	(2.39)
PIN		-0.23***				
		(-2.72)				
α_{PIN}			0.01			
			(0.16)			
$\frac{\varepsilon}{\mu}$			0.32**			
			(2.35)			
SD(ξS)				-0.55***		
				(-4.14)		
OFC					-0.21	
					(-1.61)	
α_{hybrid}						-0.11**
						(-2.02)
κ_{hybrid}						-0.52***
						(-4.05)
σ_{hybrid}						-0.28
						(-1.24)
Adjusted R^2	0.745	0.745	0.746	0.746	0.745	0.748
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes

Table 6: Structural Parameter Correlations

Correlations of structural parameters from the hybrid and other structural PIN models. For all models, $\alpha =$ probability of an information event. For the hybrid model, λ_{hybrid} is the expected average lambda $\lambda(0,0)$ based on Equation (8). PIN, APIN, and VPIN are the probabilities of informed trading estimated using the methodologies in Easley et al. (1996), Duarte and Young (2009), and Easley et al. (2012), respectively. λ_{OWR} is the estimate of Kyle's lambda from Odders-White and Ready (2008). σ_{hybrid} and σ_u are the standard deviations of noise trading from the hybrid and OWR models, respectively. ε/μ and $(\varepsilon+\theta\eta)/\mu$ are the ratios of the informed to uninformed trading intensities from the PIN and APIN models, respectively.

Panel A. Composite Measures

	λ_{hybrid}	PIN	λ_{OWR}	APIN	VPIN
λ_{hybrid}	1.00				
PIN	0.35	1.00			
λ_{OWR}	0.55	0.17	1.00		
APIN	0.42	0.58	0.19	1.00	
VPIN	0.56	0.42	0.26	0.48	1.00

Panel B. Probability of an Information Event

	α_{hybrid}	α_{PIN}	α_{OWR}	α_{APIN}	VPIN
α_{hybrid}	1.00				
α_{PIN}	-0.08	1.00			N/A
α_{OWR}	-0.10	0.05	1.00		
α_{APIN}	-0.01	0.25	0.04	1.00	

Panel C. Noise Trading

	σ_{hybrid}	$\frac{\varepsilon}{\mu}$	σ_u	$\frac{\varepsilon+\theta\eta}{\mu}$	VPIN
σ_{hybrid}	1.00				
$\frac{\varepsilon}{\mu}$	0.57	1.00			N/A
σ_u	0.92	0.51	1.00		
$\frac{\varepsilon+\theta\eta}{\mu}$	0.53	0.83	0.48	1.00	

Table 7: Structural Parameter Estimates and Market Capitalization

Average values of parameter estimates within each NYSE market capitalization decile (formed annually). For all models, α = probability of an information event. For the hybrid model, λ_{hybrid} is the expected average lambda $\lambda(0,0)$ based on Equation (8). PIN, APIN, and VPIN are the probabilities of informed trading estimated using the methodologies in Easley et al. (1996), Duarte and Young (2009), and Easley et al. (2012), respectively. λ_{OWR} is the estimate of Kyle's lambda from Odders-White and Ready (2008). σ_{hybrid} and σ_u are the standard deviations of noise trading from the hybrid and OWR models, respectively. ε/μ and $(\varepsilon+\theta\eta)/\mu$ are the ratios of the informed to uninformed trading intensities from the PIN and APIN models, respectively.

Panel A. Composite Measures

	λ_{hybrid}	PIN	λ_{OWR}	APIN	VPIN
1 (Small)	0.199	0.18	0.139	0.15	0.28
2	0.144	0.15	0.089	0.13	0.27
3	0.109	0.14	0.068	0.12	0.25
4	0.085	0.13	0.058	0.12	0.24
5	0.066	0.13	0.049	0.11	0.23
6	0.052	0.12	0.040	0.10	0.23
7	0.042	0.12	0.034	0.10	0.22
8	0.035	0.11	0.032	0.09	0.21
9	0.025	0.09	0.024	0.08	0.20
10 (Large)	0.020	0.08	0.020	0.07	0.18

Panel B. Probability of an Information Event

	α_{hybrid}	α_{PIN}	α_{OWR}	α_{APIN}	VPIN
1 (Small)	0.72	0.31	0.11	0.41	
2	0.70	0.33	0.12	0.44	
3	0.67	0.34	0.12	0.44	
4	0.67	0.35	0.12	0.45	
5	0.65	0.36	0.14	0.45	N/A
6	0.63	0.36	0.14	0.45	
7	0.62	0.38	0.16	0.46	
8	0.59	0.38	0.17	0.46	
9	0.57	0.39	0.18	0.46	
10 (Large)	0.52	0.39	0.23	0.47	

Panel C. Noise Trading

	σ_{hybrid}	$\frac{\varepsilon}{\mu}$	σ_u	$\frac{\varepsilon+\theta\eta}{\mu}$	VPIN
1 (Small)	0.06	0.73	0.04	1.24	
2	0.06	0.94	0.04	1.51	
3	0.07	1.07	0.05	1.70	
4	0.08	1.18	0.06	1.83	
5	0.09	1.28	0.08	1.97	N/A
6	0.11	1.38	0.09	2.08	
7	0.12	1.55	0.11	2.25	
8	0.15	1.74	0.14	2.50	
9	0.19	2.12	0.19	2.82	
10 (Large)	0.29	2.64	0.33	3.43	

Table 8: Price Impact and Composite Information Asymmetry Measures

This table reports Fama and MacBeth (1973) cross-sectional regressions of price impact estimates on structural parameters from the hybrid and other PIN models. The dependent variables are the 5-minute price impact, the cumulative impulse response estimated following Hasbrouck (1991), and an estimate of Kyle's lambda ($\widehat{\lambda}_{\text{intraday}}$) using a regression of 5-minute returns on the square-root of signed volume following Hasbrouck (2009) and Goyenko et al. (2009). Panel A reports univariate regressions. Panel B reports multivariate regressions controlling for market capitalization, the inverse of the stock price, and log volume. All variables are winsorized and standardized to have a unit standard deviation. Standard errors are adjusted for serial correlation following Newey and West (1987) with 5 lags. t statistics are in parentheses, and statistical significance is represented by * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

Panel A. Univariate

	5-Minute Price Impact					Cumulative Impulse Response					$\widehat{\lambda}_{\text{intraday}}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
λ_{hybrid}	0.47***					0.48***					0.35***				
	(9.23)					(4.36)					(13.27)				
PIN		0.37***					0.32***					0.31***			
		(9.38)					(5.28)					(5.50)			
λ_{OWR}			0.26***					0.26***					0.20***		
			(5.26)					(3.33)					(7.17)		
APIN				0.43***					0.36***					0.41***	
				(8.18)					(7.21)					(4.08)	
VPIN					0.41***					0.38***					0.35**
					(3.95)					(5.72)					(2.23)
Constant	0.05	0.06	0.08	0.09	0.08	0.07	0.08	0.12	0.07	0.07	-0.03	-0.00	-0.02	0.05	0.06
	(0.24)	(0.28)	(0.31)	(0.46)	(0.40)	(0.23)	(0.27)	(0.35)	(0.27)	(0.26)	(-0.48)	(-0.03)	(-0.23)	(0.50)	(0.75)
R^2	0.315	0.200	0.098	0.255	0.356	0.418	0.204	0.120	0.263	0.396	0.192	0.115	0.066	0.153	0.185

Table 8: (continued) Price Impact and Composite Information Asymmetry Measures

Panel B. Multivariate

	5-Minute Price Impact					Cumulative Impulse Response					$\widehat{\lambda}_{\text{intraday}}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)
λ_{hybrid}	0.20***	0.20***	0.20***	0.17***	0.17***	0.22***	0.23***	0.22***	0.19***	0.20***	0.15***	0.15***	0.15***	0.14***	0.14***
	(9.03)	(12.41)	(9.06)	(13.78)	(15.53)	(6.43)	(5.74)	(6.60)	(7.01)	(6.07)	(6.48)	(5.89)	(6.49)	(4.48)	(4.25)
PIN	0.04***				0.03***	0.03***				0.03**	0.01				0.01
	(3.50)				(3.49)	(3.04)				(2.37)	(0.91)				(0.68)
λ_{OWR}		-0.00			-0.00		-0.02			-0.01		0.00			0.00
		(-0.30)			(-0.23)		(-1.60)			(-1.65)		(0.90)			(0.44)
APIN			0.04		0.01			0.01		0.00			0.03		0.01
			(1.46)		(0.63)			(1.06)		(0.60)			(1.60)		(1.46)
VPIN				0.22***	0.21***				0.19***	0.18***				0.04	0.04
				(3.95)	(3.75)				(6.45)	(6.51)				(0.33)	(0.32)
Log(Size)	-0.22**	-0.23**	-0.23**	-0.22**	-0.22**	-0.20**	-0.20**	-0.20**	-0.20**	-0.20**	-0.13**	-0.13**	-0.13**	-0.12*	-0.12*
	(-2.77)	(-2.81)	(-2.77)	(-2.44)	(-2.43)	(-2.79)	(-2.79)	(-2.75)	(-2.54)	(-2.54)	(-2.57)	(-2.58)	(-2.60)	(-2.07)	(-2.08)
Inverse Price	4.66***	4.73***	4.66***	4.67***	4.60***	2.94***	2.99***	2.96***	3.51***	3.46***	6.80***	6.81***	6.76***	6.32***	6.29***
	(11.47)	(11.60)	(10.59)	(6.61)	(6.52)	(5.02)	(5.02)	(4.88)	(3.43)	(3.43)	(3.35)	(3.36)	(3.34)	(3.14)	(3.14)
Log(Volume)	0.01	-0.00	0.02	0.11***	0.12***	-0.07***	-0.07***	-0.06***	0.04**	0.04**	0.02	0.02	0.03	0.01	0.02
	(0.16)	(-0.00)	(0.44)	(4.24)	(5.30)	(-6.12)	(-8.14)	(-4.91)	(2.13)	(2.21)	(0.30)	(0.21)	(0.39)	(0.30)	(0.50)
Constant	3.14**	3.30**	3.20**	2.38	2.28	3.22**	3.35**	3.32**	2.63*	2.54*	1.51***	1.54***	1.46***	1.41*	1.37*
	(2.60)	(2.75)	(2.51)	(1.65)	(1.59)	(2.42)	(2.47)	(2.37)	(1.95)	(1.92)	(4.01)	(4.44)	(3.96)	(1.85)	(1.75)
R^2	0.727	0.727	0.728	0.744	0.749	0.734	0.734	0.734	0.751	0.754	0.630	0.631	0.631	0.641	0.644

Figure 1: Price and Order Imbalance
The equilibrium price $V_t + p(t, Y_t)$ as a function of the order imbalance Y_t at $t = 0.5$ for two values of the probability α of an information event, when $\xi S = H$, $V_t = 50$, $H = 10$, $L = -10$, $\sigma = 1$, and $p_H = p_L = 1/2$.

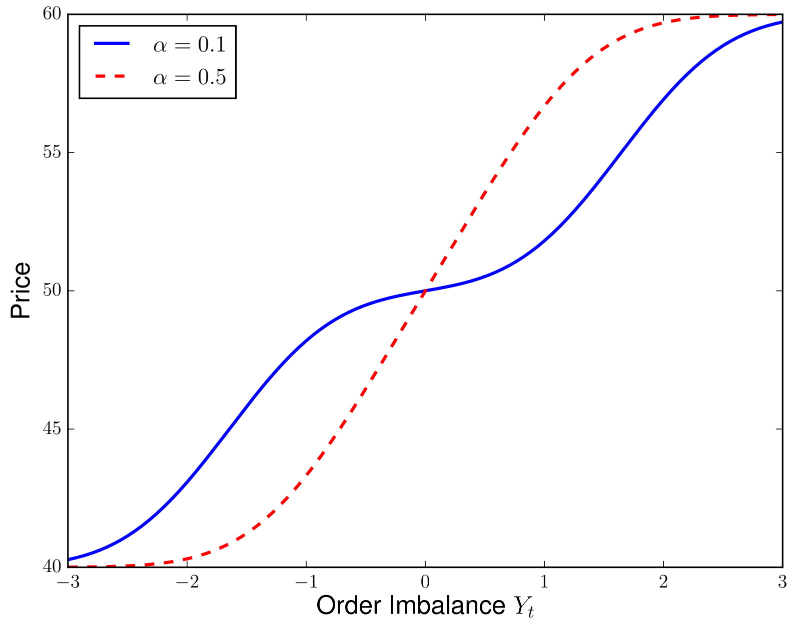


Figure 2: Expected Average Lambda
Expected average lambda (8) as a function of α for two different values of $H - L$, taking $|L| = H$, when $\sigma = 1$ and $p_L = p_H = 1/2$.

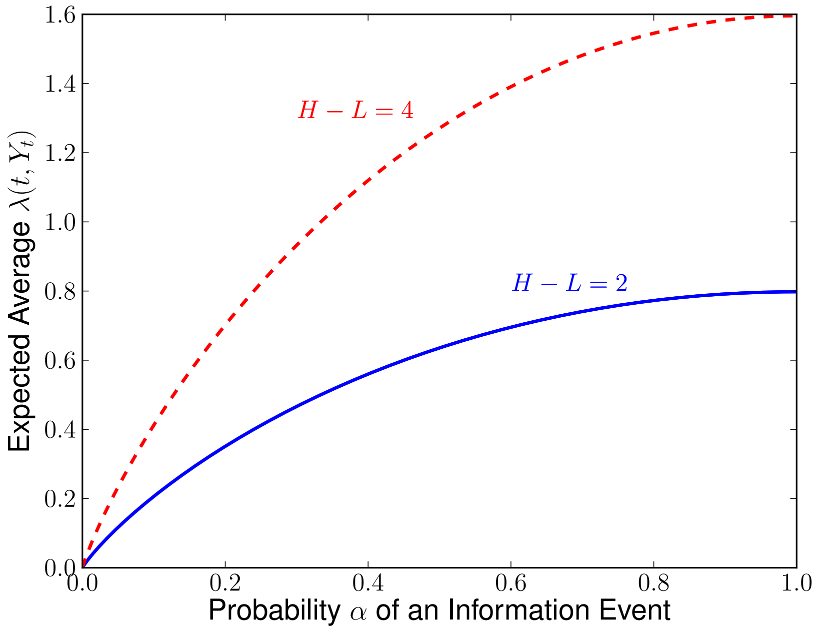


Figure 3: Equilibrium Informed Trading and Price

The equilibrium informed trading rate θ_t as a function of the price $V_t + p(t, Y_t)$ at $t = 0.5$ for two values of the probability α of an information event, when $\xi S = H$, $V_t = 50$, $H = 10$, $L = -10$, $\sigma = 1$, and $p_H = p_L = 1/2$.

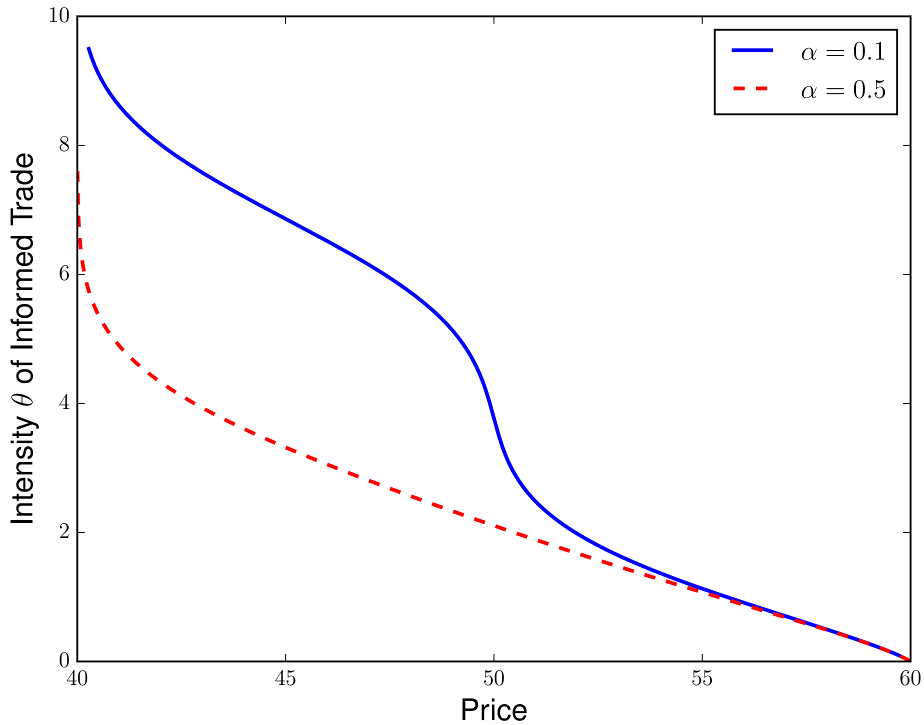


Figure 4: Conditional Order Imbalance Distributions
The density function of the net order flow Y_1 conditional on a low signal, no information event, and a high signal for two values of the probability α of an information event, when $\sigma = 1$ and $p_L = p_H = 1/2$.

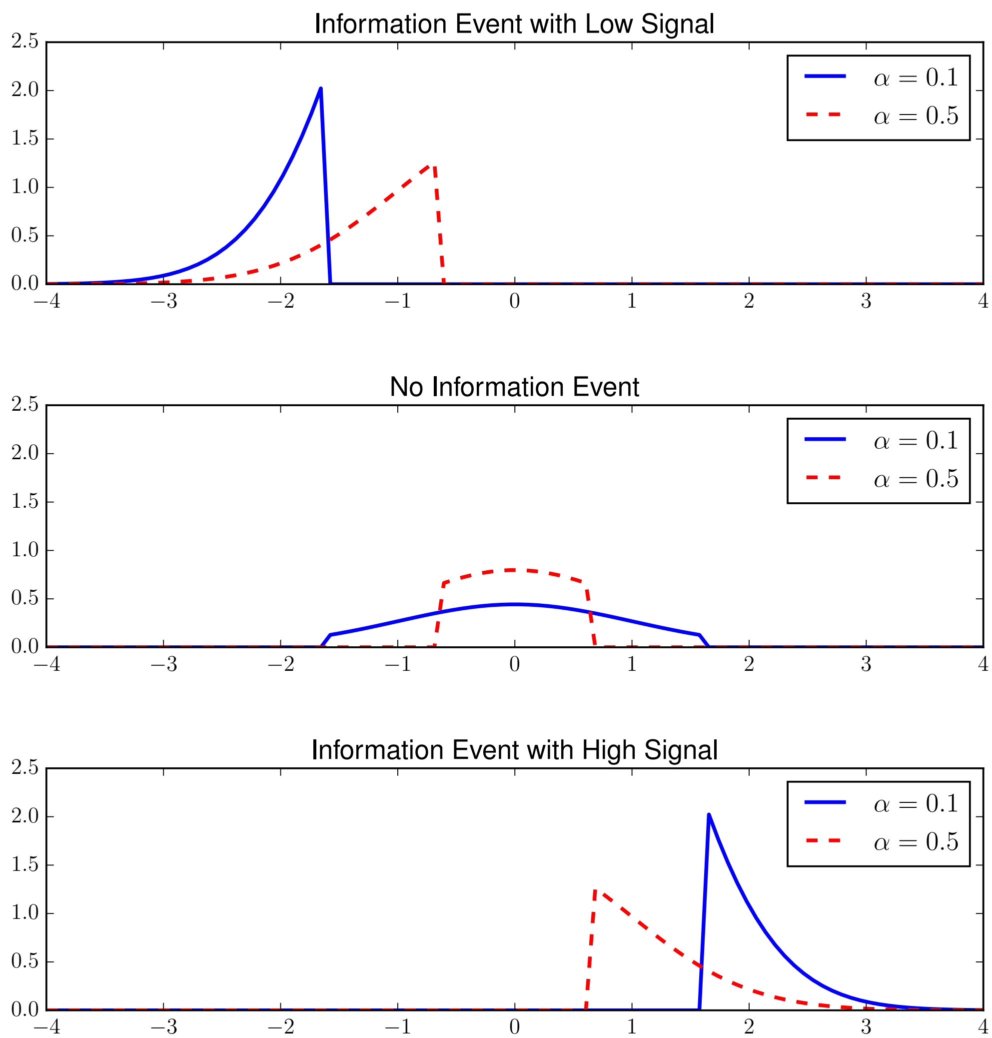
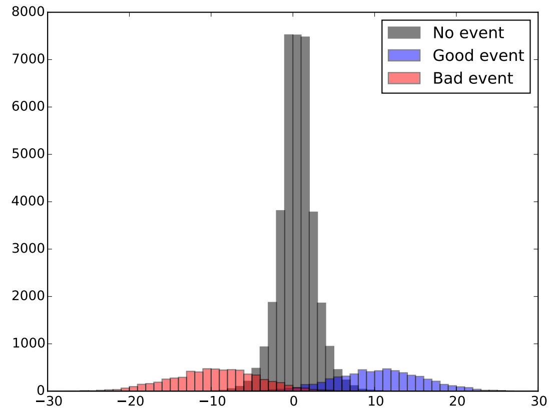
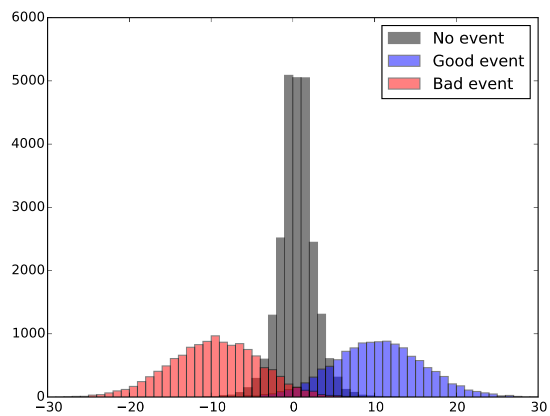


Figure 5: Order Imbalance Distribution for a PIN Model with Contrarian Traders
This figure plots the simulated distribution of order imbalances for a variant of the Easley et al. (1996) model in which contrarian traders arrive in the event of no information as described in Internet Appendix B. Order imbalance is the number of buys minus number of sells. The histograms plot 50,000 instances of the model. The parameter values are $\alpha \in \{0.25, 0.5, 0.75\}$, $p_L = 0.5$, $\varepsilon = 10$, $\mu = 10$, $L = -1$, $H = 1$, $V^* = 0$.

(a) $\alpha = 0.25$



(b) $\alpha = 0.50$



(c) $\alpha = 0.75$

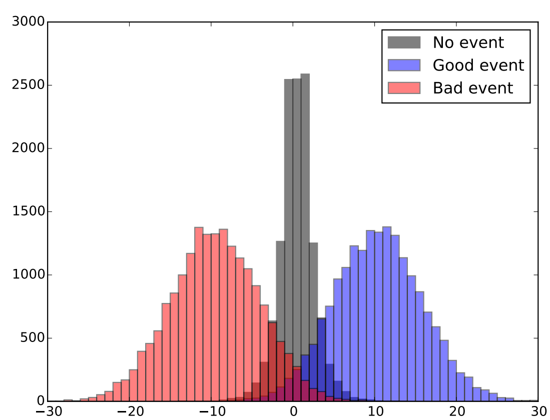
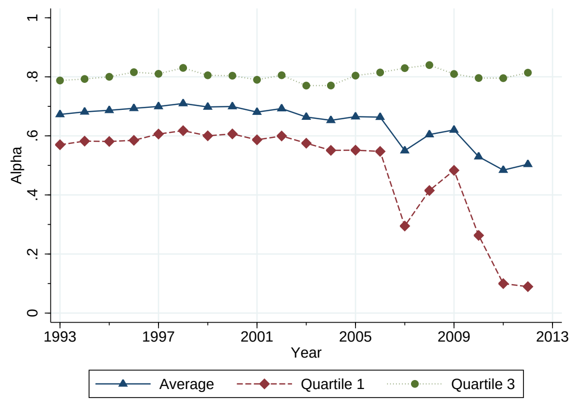


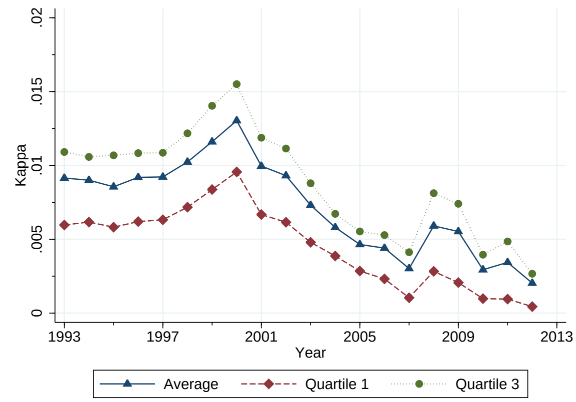
Figure 6: Time Series of Hybrid Model Estimates

The annual cross-sectional mean, 25th and 75th percentiles of parameter estimates for the hybrid model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using prices and order imbalances in six hourly intraday bins and at the close. The model parameters are α = probability of an information event, κ = signal scale parameter, σ = standard deviation of liquidity trading, δ = volatility of public information, and p_L = probability of a negative event. λ_{hybrid} is the expected average lambda $\lambda(0,0)$ based on Equation (8).

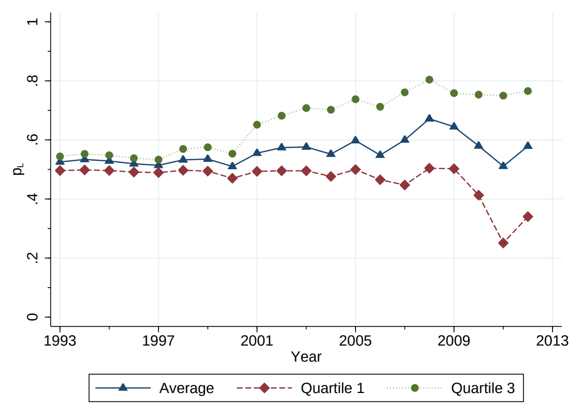
(a) α



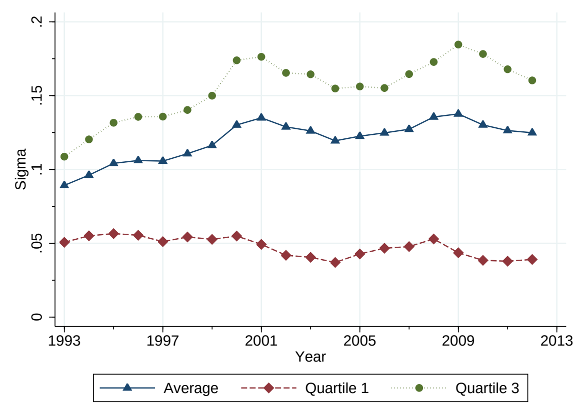
(b) κ



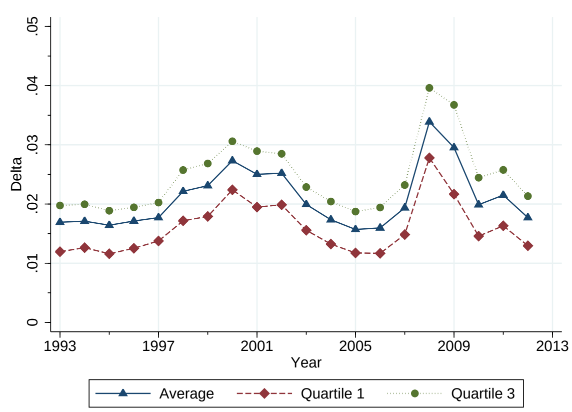
(c) p_L



(d) σ



(e) δ



(f) λ_{hybrid}

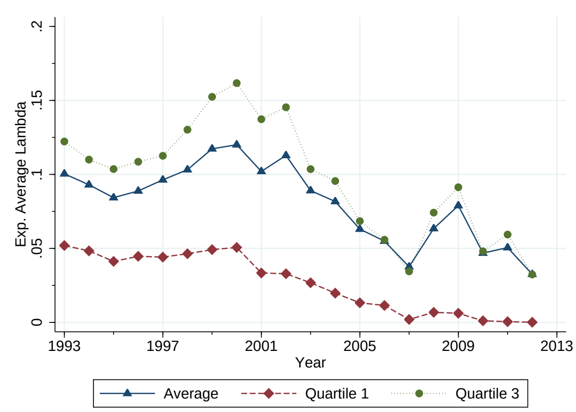


Figure 7: Conditional Probability of an Information Event around Earnings Announcements
The figure reports averages of the end-of-day conditional probability of an information event, CPIE, defined in Equation (11), in event time around earnings announcements. CPIE is calculated using the estimated parameters and order flows. Dashed lines indicate the 95% confidence interval.

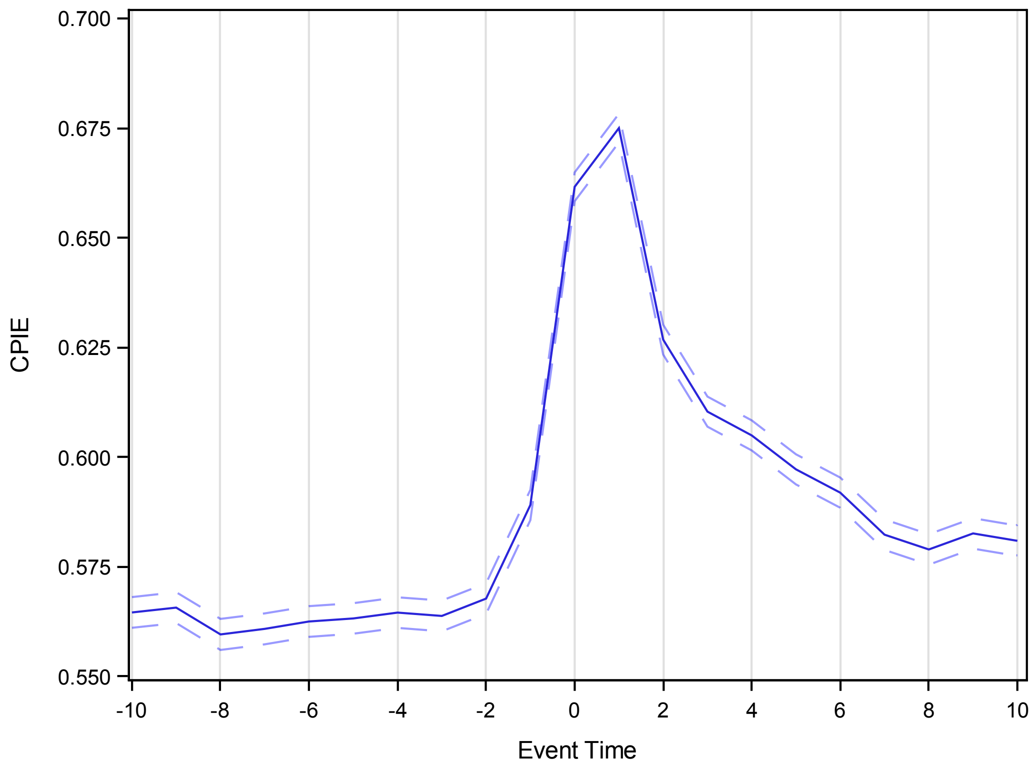
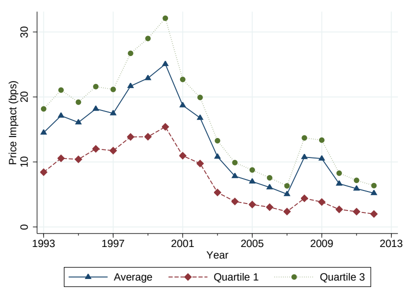
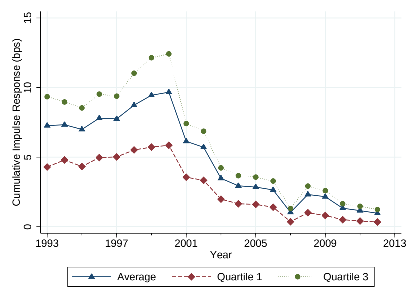


Figure 8: Time Series of Price Impact Estimates and Composite Measures
The annual cross-sectional mean, 25th and 75th percentiles of 5-minute price impacts and composite information asymmetry measures. Five-minute price impacts are estimated daily and averaged annually for each stock-year for NYSE stocks from 1993 through 2012.The stock-year estimates of the cumulative impulse response and $\lambda_{\text{intraday}}$ are the medians of daily estimates. λ_{hybrid} is the expected average lambda $\lambda(0,0)$ based on Equation (8). PIN, APIN, and VPIN are the probabilities of informed trading estimated using the methodologies in Easley et al. (1996), Duarte and Young (2009), and Easley et al. (2012), respectively. λ_{OWR} is the estimate of Kyle's lambda from Odders-White and Ready (2008).

(a) 5-Minute Price Impact

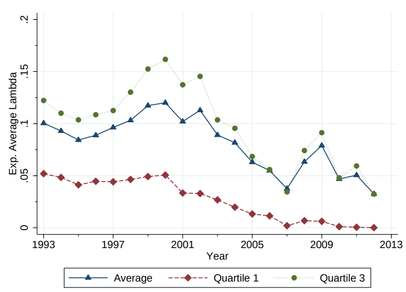


(b) Cumulative Impulse Response

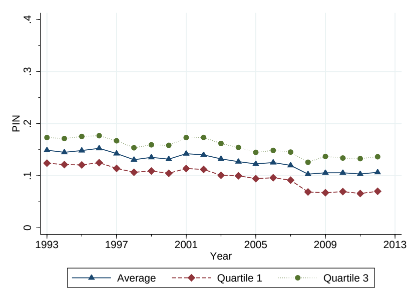


(c) $\lambda_{\text{intraday}}$

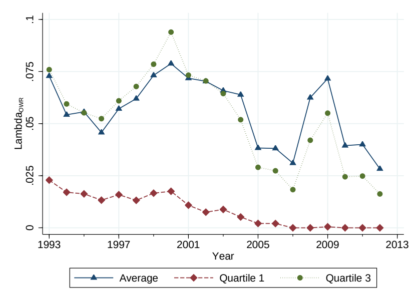
(d) λ_{hybrid}



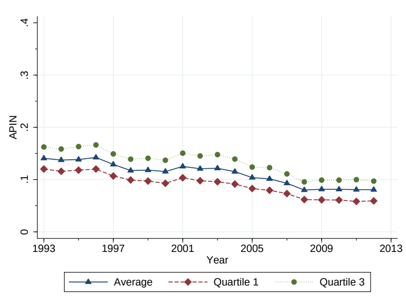
(e) PIN



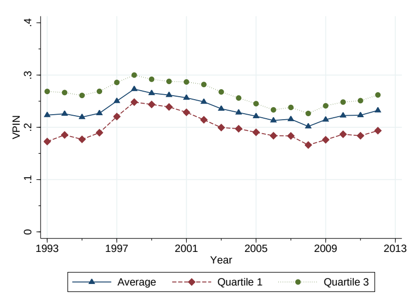
(f) λ_{OWR}



(g) APIN



(h) VPIN



Internet Appendix to

Identifying Information Asymmetry in Securities Markets

Internet Appendix A. Hybrid Model with General Signal Distribution

We present the hybrid model with a general signal distribution. For simplicity, we omit public news arrival, which is straightforward to add as in the paper.

Internet Appendix A.1. Model

Assume the single strategic trader receives a signal S at time 0 with probability α. The value of the asset at the end of the day conditional on all available information is $V_1 + \xi S$. The standard continuous-time Kyle (1985) model is a special case of this model in which $\alpha = 1$, V is constant, and S is normally distributed.

Assume the signal S has a continuous distribution function G. Set $\underline{s} = \inf\{s \mid G(s) > 0\}$ and $\bar{s} = \sup\{s \mid G(s) < 1\}$. Assume $-\infty \leq \underline{s} < 0 < \bar{s} \leq \infty$. Assume G is strictly increasing on $(\underline{s}, \bar{s})$ except possibly on some interval containing zero. If there is such an interval with zero in its interior, then there is zero probability of very small good or bad news. Including this feature in the model would make it possible to ensure that information events are nontrivial. Under these assumptions, G^{-1} is uniquely defined on $(0, 1)$, except possibly at $G(0)$.

Internet Appendix A.2. Brownian Bridge

Let F denote the distribution function of the normally distributed variable Y_1. Set $y_L = F^{-1}(\alpha G(0))$ and $y_H = F^{-1}(1 - \alpha + \alpha G(0))$. This means that

$$\alpha \, \text{prob}(S \leq 0) = \text{prob}(Y_1 \leq y_L)\,,$$

and

$$\alpha \, \text{prob}(S > 0) = \text{prob}(Y_1 > y_H)\,.$$

Thus, the unconditional probability of bad news is equal to the probability that $Y_1 \le y_L$, and the unconditional probability of good news is equal to the probability that $Y_1 > y_H$.

Set

$$q(t,y,s) = \begin{cases} F^{-1}(\alpha G(s)) - y & \text{if } G(s) < G(0)\,, \\ \mathsf{E}[Y_1 \mid Y_t = y, y_L \le Y_1 \le y_H] - y & \text{if } G(s) = G(0)\,, \\ F^{-1}(1-\alpha+\alpha G(s)) - y & \text{if } G(s) > G(0)\,. \end{cases} \tag{A.1}$$

Note that if $G(s) < G(0)$, then $y \stackrel{\text{def}}{=} F^{-1}(\alpha G(s))$ satisfies

$$F(y) = \alpha G(s) < \alpha G(0) = F(y_L)\,.$$

Thus, the function $s \mapsto F^{-1}(\alpha G(s))$ maps $\{s \mid G(s) < G(0)\}$ to $\{y \mid y < y_L\}$. Symmetrically, the function $s \mapsto F^{-1}(1-\alpha+\alpha G(s))$ maps $\{s \mid G(s) > G(0)\}$ to $\{y \mid y > y_H\}$.

Lemma. *Let* N *denote the standard normal distribution function. Let* $\mathbb{F}^Y = \{\mathcal{F}_t^Y \mid 0 \le t \le 1\}$ *denote the filtration generated by the stochastic process* Y *defined by* $Y_0 = 0$ *and*

$$\mathrm{d}Y_t = \frac{q(t, Y_t, \xi S)}{1-t}\,\mathrm{d}t + \mathrm{d}Z_t\,. \tag{A.2}$$

Then, the following are true:

(A) Y *is an* $\mathbb{F}^Y$*–Brownian motion with zero drift and standard deviation* σ*.*

(B) *With probability one,*

$$\xi = 1 \textit{ and } S < 0 \quad \Rightarrow \quad Y_1 = F^{-1}(\alpha G(S)) < y_L\,, \tag{A.3a}$$

$$\xi = 0 \quad \Rightarrow \quad y_L \le Y_1 \le y_H\,, \tag{A.3b}$$

$$\xi = 1 \textit{ and } S > 0 \quad \Rightarrow \quad Y_1 = F^{-1}(1-\alpha+\alpha G(S)) > y_H\,. \tag{A.3c}$$

(C) *For each $t < 1$, the probability that $\xi = 1$ conditional on $\mathcal{F}_t^Y$ is*

$$\mathrm{N}\left(\frac{y_L - Y_t}{\sigma\sqrt{1-t}}\right) + 1 - \mathrm{N}\left(\frac{y_H - Y_t}{\sigma\sqrt{1-t}}\right). \tag{A.4}$$

The process Y described in the lemma is a variation of a Brownian bridge. It differs from a Brownian bridge in that the endpoint is not uniquely determined when there is no information event ($\xi = 0$). Part (C) of the proposition follows immediately from the preceding parts, because the probability (A.4) is the probability that $Y_1 \notin [y_L, y_H]$ calculated on the basis that Y is an $\mathbb{F}^Y$–Brownian motion with zero drift and standard deviation σ.

Internet Appendix A.3. Equilibrium

Let $f(\cdot \mid t, y)$ denote the density function of Y_1 conditional on $Y_t = y$, that is, the normal density function with mean y and variance $(1-t)\sigma^2$.

Theorem. *There is an equilibrium in which the trading rate of the strategic trader is*

$$\theta_t = \frac{q(t, Y_t, \xi S)}{1-t}. \tag{A.5}$$

The equilibrium asset price is $P_t = V_t + p(t, Y_t)$, where the pricing function p is given by

$$p(t,y) = \int_{-\infty}^{y_L} G^{-1}\left(\frac{F(z)}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z + \int_{y_H}^{\infty} G^{-1}\left(\frac{F(z) - 1 + \alpha}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z. \tag{A.6}$$

The asset price evolves as $\mathrm{d}P_t = \mathrm{d}V_t + \lambda(t, Y_t)\,\mathrm{d}Y_t$, where Kyle's lambda is

$$\begin{aligned}\lambda(t,y) = \frac{1}{\sigma^2(1-t)} &\int_{-\infty}^{y_L} (z-y) G^{-1}\left(\frac{F(z)}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z \\ &+ \frac{1}{\sigma^2(1-t)} \int_{y_H}^{\infty} (z-y) G^{-1}\left(\frac{F(z) - 1 + \alpha}{\alpha}\right) f(z \mid t, y)\,\mathrm{d}z. \end{aligned}\tag{A.7}$$

There is convergence to strong-form efficiency in the sense that $\lim_{t\to 1} P_t = V_1 + \xi S$ with probability one.

The probability that an information event occurred, conditional on the market's information at any date $t < 1$, is given by (A.4). The probability is generally an increasing function of the absolute net order imbalance at t; more precisely, it is an increasing function of the distance of the net order imbalance from the midpoint of y_L and y_H. The strong-form efficiency condition means that the market learns by the close of trading whether the strategic trader is informed and, if so, what her information is. From the proposition, we know that if $\xi = 1$ and $S < 0$, then

$$Y_t \to F^{-1}(\alpha G(S)) < y_L \tag{A.8a}$$

with probability one as $t \to 1$. On the other hand, if $\xi = 1$ and $S > 0$, then

$$Y_t \to F^{-1}(1 - \alpha + \alpha G(S)) > y_H \tag{A.8b}$$

with probability one. In each case, the market learns S from Y as $t \to 1$. If the strategic trader is uninformed ($\xi = 0$), then

$$y_L \le \liminf_{t \to 1} Y_t \le \limsup_{t \to 1} Y_t \le y_H\,, \tag{A.8c}$$

and the difference between P_t and V_t converges to zero as $t \to 1$.

The proofs of the lemma and theorem are similar to those in the paper and are available upon request.

Internet Appendix B. EKOP with Contrarian

The primary difference between the hybrid model and the PIN model is that, in the former model, the strategic trader endogenously trades based on liquidity in the market. A second difference is that the strategic trader acts as a contrarian in the absence of an event. We now present evidence that the result on identification of information asymmetry parameters does not result from this assumption.

We analyze an alteration of the original EKOP Glosten-Milgrom model to include the presence of contrarian informed traders on non-event days, as in the hybrid Kyle model. However, we maintain the assumption of exogenous trading by these contrarians. Contrarians have Poisson arrival rate μ and buy the asset if $a(t) < V^*$, sell the asset if $b(t) > V^*$, and refrain from trade if the known value is within the spread.

Let $\mathbb{1}^{over}$ be an indicator variable for $b(t) > V^*$. This is an indicator for whether a contrarian finds the asset over-priced on a non-event day n. Let $\mathbb{1}^{under}$ be an indicator variable for $a(t) < V^*$. This is an indicator for whether a contrarian finds the asset under-priced on a non-event day n. Let $\mathbb{1}^{inside}$ be an indicator variable for $V^* \in [b(t), a(t)]$. This is an indicator for when a contrarian on non-event days finds it optimal not to trade on a non-event day n due to the spread.

Internet Appendix B.1. Bid prices

Following Section I.B of EKOP, the market maker's posterior probability of no news at time t conditional on a sell order arriving S_t is:

$$\Pr(n|S_t) = P_n(t|S_t) = \frac{\Pr(S_t|n)\Pr(n)}{\Pr(S_t|n)\Pr(n) + \Pr(S_t|g)\Pr(g) + \Pr(S_t|b)\Pr(b)} \tag{B.1}$$

$$= \frac{(\varepsilon + \mathbb{1}^{over}\mu)P_n(t)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over}P_n(t)\right)} \tag{B.2}$$

The posterior probability for bad news conditional on a sell order arriving S_t is:

$$\Pr(b|S_t) = P_b(t|S_t) = \frac{\Pr(S_t|b)\Pr(b)}{\Pr(S_t|n)\Pr(n) + \Pr(S_t|g)\Pr(g) + \Pr(S_t|b)\Pr(b)} \tag{B.3}$$

$$= \frac{(\varepsilon + \mu)P_b(t)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over}P_n(t)\right)} \tag{B.4}$$

The posterior probability for good news conditional on a sell order arriving S_t is:

$$\Pr(g|S_t) = P_g(t|S_t) = \frac{\Pr(S_t|g)\Pr(g)}{\Pr(S_t|n)\Pr(n) + \Pr(S_t|g)\Pr(g) + \Pr(S_t|b)\Pr(b)} \tag{B.5}$$

$$= \frac{\varepsilon P_g(t)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over}P_n(t)\right)} \tag{B.6}$$

Then the bid price will be

$$b(t) = V^* \cdot P_n(t|S_t) + L \cdot P_b(t|S_t) + H \cdot P_g(t|S_t) \tag{B.7}$$

$$= \frac{V^* \cdot (\varepsilon + \mathbb{1}^{over}\mu)P_n(t) + L \cdot (\varepsilon + \mu)P_b(t) + H \cdot \varepsilon P_g(t)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over}P_n(t)\right)} \tag{B.8}$$

Let b_0 denote the value of $b(t)$ when we substitute $1^{over} = 0$ into the formula and let b_1 denote the value of $b(t)$ when we substitute $1^{over} = 1$. Define $\underline{\mathrm{p}}$ as

$$\underline{\mathrm{p}} = \frac{\varepsilon P_g(t)}{\varepsilon P_g(t) + [\varepsilon + \mu]P_b(t)}.$$

Then

$$b_0 = V^* + [\underline{\mathrm{p}}H + (1 - \underline{\mathrm{p}})L - V^*] \times \frac{(\varepsilon + \mu)P_b + \varepsilon P_g}{\varepsilon + \mu P_b}$$

$$b_1 = V^* + [\underline{\mathrm{p}}H + (1 - \underline{\mathrm{p}})L - V^*] \times \frac{(\varepsilon + \mu)P_b + \varepsilon P_g}{\varepsilon + \mu P_b + \mu P_n}$$

Note that the formulas for b_0 and b_1 are the same except that the denominator in the fraction is larger for b_1, so the fraction is larger for b_0. This shows that

$$\underline{\mathrm{p}}H + (1 - \underline{\mathrm{p}})L - V^* > 0 \Rightarrow b_0 > b_1 > V^*$$

$$\underline{\mathrm{p}}H + (1 - \underline{\mathrm{p}})L - V^* < 0 \Rightarrow b_0 < b_1 < V^*$$

So, $b(t) = b_1$ in the former case ($\mathbb{1}^{over} = 1$), and $b(t) = b_0$ in the latter case ($\mathbb{1}^{over} = 0$).

Internet Appendix B.2. Ask prices

The market maker's posterior probability of no news at time t conditional on a buy order arriving B_t is:

$$\Pr(n|B_t) = P_n(t|B_t) = \frac{\Pr(B_t|n)\Pr(n)}{\Pr(B_t|n)\Pr(n) + \Pr(B_t|g)\Pr(g) + \Pr(B_t|b)\Pr(b)} \tag{B.9}$$

$$= \frac{(\varepsilon + \mathbb{1}^{under}\mu)P_n(t)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under}P_n(t)\right)} \tag{B.10}$$

The posterior probability for bad news conditional on a buy order arriving B_t is:

$$\Pr(b|B_t) = P_b(t|B_t) = \frac{\Pr(B_t|b)\Pr(b)}{\Pr(B_t|n)\Pr(n) + \Pr(B_t|g)\Pr(g) + \Pr(B_t|b)\Pr(b)} \tag{B.11}$$

$$= \frac{\varepsilon P_b(t)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under}P_n(t)\right)} \tag{B.12}$$

The posterior probability for good news conditional on a buy order arriving B_t is:

$$\Pr(g|B_t) = P_g(t|B_t) = \frac{\Pr(B_t|g)\Pr(g)}{\Pr(B_t|n)\Pr(n) + \Pr(B_t|g)\Pr(g) + \Pr(B_t|b)\Pr(b)} \tag{B.13}$$

$$= \frac{(\varepsilon + \mu)P_g(t)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under}P_n(t)\right)} \tag{B.14}$$

Then the ask price will be

$$a(t) = V^* \cdot P_n(t|B_t) + L \cdot P_b(t|B_t) + H \cdot P_g(t|B_t) \tag{B.15}$$
$$= \frac{V^* \cdot (\varepsilon + \mathbb{1}^{under}\mu)P_n(t) + L \cdot \varepsilon P_b(t) + H \cdot (\varepsilon + \mu)P_g(t)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under}P_n(t)\right)} \tag{B.16}$$

Let a_0 denote the value of $a(t)$ when we substitute $\mathbb{1}^{under} = 0$ into the formula and let a_1 denote the value of $a(t)$ when we substitute $\mathbb{1}^{under} = 1$. Define $\bar{p}$ as

$$\bar{p} = \frac{\varepsilon P_b(t)}{\varepsilon P_b(t) + [\varepsilon + \mu]P_g(t)}.$$

Then

$$a_0 = V^* + [\bar{p}L + (1 - \bar{p})H - V^*] \times \frac{(\varepsilon + \mu)P_b + \varepsilon P_g}{\varepsilon + \mu P_b}$$

$$a_1 = V^* + [\bar{p}L + (1 - \bar{p})H - V^*] \times \frac{(\varepsilon + \mu)P_b + \varepsilon P_g}{\varepsilon + \mu P_b + \mu P_n}$$

Note that the formulas for a_0 and a_1 are the same except that the denominator in the fraction is larger for a_1, so the fraction is larger for a_0. This shows that

$$\bar{p}L + (1 - \bar{p})H - V^* > 0 \Rightarrow a_0 > a_1 > V^*$$

$$\bar{p}L + (1 - \bar{p})H - V^* < 0 \Rightarrow a_0 < a_1 < V^*$$

So, $a(t) = a_0$ in the former case ($\mathbb{1}^{under} = 0$), and $a(t) = a_1$ in the latter case ($\mathbb{1}^{under} = 1$).

Internet Appendix B.3. Updating probabilities and prices between arrival of traders

$P_i(t)$ denotes the probability of an event i day ($i \in \{n, g, b\}$) conditional on information up to time t. This includes both past trades and the absence of trades. We need to calculate the updating about day type over intervals without trades. Let N_t denote the absence of buys or sells at time t. The market maker's posterior probability of no news at time t conditional on no order arriving N_t is:

$$
\begin{aligned}
P_n(t|N_t) &= \frac{\Pr(N_t|n)\Pr(n)}{\Pr(N_t|n)\Pr(n) + \Pr(N_t|g)\Pr(g) + \Pr(N_t|b)\Pr(b)} && \text{(B.17)}\\
&= \frac{\left(1 - 2\varepsilon\,\mathrm{d}t - (1 - \mathbb{1}^{inside})\mu\,\mathrm{d}t\right) P_n(t)}{1 - (\mu + 2\varepsilon)\,\mathrm{d}t + P_n(t)\mathbb{1}^{inside}\mu\,\mathrm{d}t} && \text{(B.18)}\\
&= \frac{\left(1 - (\mu + 2\varepsilon)\,\mathrm{d}t + \mathbb{1}^{inside}\mu\,\mathrm{d}t\right) P_n(t)}{1 - (\mu + 2\varepsilon)\,\mathrm{d}t + P_n(t)\mathbb{1}^{inside}\mu\,\mathrm{d}t} && \text{(B.19)}
\end{aligned}
$$

The posterior probability for bad news conditional on no order arriving N_t is:

$$
\begin{aligned}
P_b(t|N_t) &= \frac{\Pr(N_t|b)\Pr(b)}{\Pr(N_t|n)\Pr(n) + \Pr(N_t|g)\Pr(g) + \Pr(N_t|b)\Pr(b)} && \text{(B.20)}\\
&= \frac{\left(1 - (\mu + 2\varepsilon)\,\mathrm{d}t\right) P_b(t)}{1 - (\mu + 2\varepsilon)\,\mathrm{d}t + P_n(t)\mathbb{1}^{inside}\mu\,\mathrm{d}t} && \text{(B.21)}
\end{aligned}
$$

The posterior probability for good news conditional on no order arriving N_t is:

$$
\begin{aligned}
P_g(t|N_t) &= \frac{\Pr(N_t|g)\Pr(g)}{\Pr(N_t|n)\Pr(n) + \Pr(N_t|g)\Pr(g) + \Pr(N_t|b)\Pr(b)} && \text{(B.22)}\\
&= \frac{\left(1 - (\mu + 2\varepsilon)\,\mathrm{d}t\right) P_g(t)}{1 - (\mu + 2\varepsilon)\,\mathrm{d}t + P_n(t)\mathbb{1}^{inside}\mu\,\mathrm{d}t} && \text{(B.23)}
\end{aligned}
$$

Because the informed traders do not trade when the value is within the spread on non-event days, market makers update slightly more towards the occurrence of a non-event day relative to good or bad events in the absence of trade when V^* falls within the spread.

Internet Appendix B.4. Expected values and spreads

The expected value of the asset conditional on the history of trades and prices:

$$
\mathsf{E}_t[V] = V^* \cdot P_n(t) + L \cdot P_b(t) + H \cdot P_g(t) \tag{B.24}
$$

Substituting into the bid and ask equations:

$$b(t) = \mathsf{E}_t[V] - \frac{\mu\left(P_b(t) + \mathbb{1}^{over} P_n(t)\right)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over} P_n(t)\right)}\left(\mathsf{E}_t[V] - L\right) \tag{B.25}$$

$$a(t) = \mathsf{E}_t[V] + \frac{\mu\left(P_g(t) + \mathbb{1}^{under} P_n(t)\right)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under} P_n(t)\right)}\left(H - \mathsf{E}_t[V]\right) \tag{B.26}$$

When the bid (and expected asset value) is above V^* (i.e., $\mathbb{1}^{over} = 1$), market-makers lower the bid beyond the level in EKOP to protect against selling by a contrarian informed trader. Similarly, when the ask (and expected asset value) is below V^* (i.e., $\mathbb{1}^{under} = 1$), then the ask is above the EKOP ask as market-makers protect against buying by a contrarian informed trader. The resulting bid-ask spread is:

$$a(t) - b(t) = \frac{\mu\left(P_g(t) + \mathbb{1}^{under} P_n(t)\right)}{\varepsilon + \mu\left(P_g(t) + \mathbb{1}^{under} P_n(t)\right)}\left(H - \mathsf{E}_t[V]\right) + \frac{\mu\left(P_b(t) + \mathbb{1}^{over} P_n(t)\right)}{\varepsilon + \mu\left(P_b(t) + \mathbb{1}^{over} P_n(t)\right)}\left(\mathsf{E}_t[V] - L\right) \tag{B.27}$$

When the expected asset value (and bid) is above V^* (i.e., $\mathbb{1}^{over} = 1$), then the spread is:

$$\frac{\mu P_g(t)}{\varepsilon + \mu P_g(t)}\left(H - \mathsf{E}_t[V]\right) + \frac{\mu\left(P_b(t) + P_n(t)\right)}{\varepsilon + \mu\left(P_b(t) + P_n(t)\right)}\left(\mathsf{E}_t[V] - L\right) \tag{B.28}$$

When the expected asset value (and ask) is below V^* (i.e., $\mathbb{1}^{under} = 1$), the spread is

$$\frac{\mu\left(P_g(t) + P_n(t)\right)}{\varepsilon + \mu\left(P_g(t) + P_n(t)\right)}\left(H - \mathsf{E}_t[V]\right) + \frac{\mu P_b(t)}{\varepsilon + \mu P_b(t)}\left(\mathsf{E}_t[V] - L\right) \tag{B.29}$$

Internet Appendix B.5. Distribution of Order Imbalances and Identification

We simulate the model to characterize the end-of-day distribution of order imbalances. We discretize the day ($T = 1$) into 1000 equal-spaced bins and determine at each bin whether a buy order, a sell order, or no order arrives. The probabilities of each of these events differ based on the type of day realized ($i \in \{n, g, b\}$) and on the price path for non-event days.

The assumption of contrarian informed traders for non-event days does not change the ability of the econometrician to identify information asymmetry parameters from the distribution of order imbalances. The unconditional distribution of order imbalances in the EKOP model with contrarians consists of three conditional distributions. On good or bad event days, the conditional distributions have positive or negative order imbalances on average as in the standard EKOP model. These are distributed Skellam as in the original PIN model. The distribution of order imbalances conditional on a non-event day are more balanced. However, this is no longer Skellam since the arriving informed traders may either buy, sell, or abstain from trade based on prices. However, the general intuition of the EKOP identification holds. $1 - \alpha$ is estimated as the mass of balanced trade corresponding to the non-event days. p_L is estimated using the mass of days with sell order imbalances relative to the mass of days with buy order imbalances. The location of each of these Skellam distributions is used to determine μ, while ε is identified based on the variance of each of the conditional distributions.

Internet Appendix C. Likelihoods and Estimates of Other Models

Internet Appendix C.1. PIN Model

The likelihood of the PIN model is:

$$L(B,S|\alpha,p_L,\mu,\varepsilon)=\prod_{t=1}^{T}\left\{\begin{array}{l}(1-\alpha)\left[\exp\left(-2\varepsilon\right)\frac{\varepsilon^{B_t+S_t}}{B_t!S_t!}\right]\\ +\alpha p_L\left[\exp\left(-(\mu+2\varepsilon)\right)\frac{(\mu+\varepsilon)^{S_t}\varepsilon^{B_t}}{B_t!S_t!}\right]\\ +\alpha(1-p_L)\left[\exp\left(-(\mu+2\varepsilon)\right)\frac{(\mu+\varepsilon)^{B_t}\varepsilon^{S_t}}{B_t!S_t!}\right]\end{array}\right\} \tag{C.1}$$

where B_t (S_t) is the number of buys (sells) on day t, α is the probability of an information event, p_L is the probability that an information event is bad news, and μ and ε are the arrival rates of informed and uninformed traders. PIN, the probability of informed trade, is given by the formula:

$$\text{PIN}=\frac{\alpha\mu}{\alpha\mu+2\varepsilon}. \tag{C.2}$$

Figure C.1 displays the time series of average parameter estimates for the PIN model. The average estimated α is much lower than in the hybrid model at 30 to 40%. The uninformed trading intensity ε and informed trading intensity μ each rise markedly in the mid-2000's reflecting the dramatic rise in trading volume. The average estimated PIN falls from about 15% in 1993 to 10% in 2012.

Internet Appendix C.2. Odders-White and Ready Model (OWR)

The parameter vector for the Odders-White/Ready model is $\Theta=(\alpha,\sigma_u,\sigma_z,\sigma_i,\sigma_{p,d},\sigma_{p,o})$ where α is the probability of an information event, σ_u is the standard deviation of noise trading, σ_z is the volatility of the noise in the order flow observed by the econometrician, and σ_i is the standard deviation of the normally distributed private information. $\sigma_{p,d}$ and $\sigma_{p,o}$ are the standard deviations of the intraday and overnight returns. The likelihood of the

Odders-White/Ready model is:

$$L(y_{e,t}, r_{d,t}, r_{o,t}|\Theta) = \prod_{t=1}^{T} \left\{ \begin{array}{l} (1-\alpha) f_N(y_{e,t}, r_{d,t}, r_{o,t}; \Theta) \\ +\alpha f_E(y_{e,t}, r_{d,t}, r_{o,t}; \Theta) \end{array} \right\} \tag{C.3}$$

where $y_{e,t}$ is the order flow observed on day t, $r_{d,t}$ is the intraday return, and $r_{o,t}$ is the overnight return. f_N and f_E are multivariate normal densities conditional on no event or an event occurring, respectively. Both f_N and f_E are mean zero with the following variances and covariances. Conditional on no event, they are:

$$\begin{aligned} \text{var}(y_{e,t}) &= \sigma_u^2 + \sigma_z^2\,, \\ \text{var}(r_{d,t}) &= \sigma_{p,d}^2 + \alpha\sigma_i^2/4\,, \\ \text{var}(r_{o,t}) &= \sigma_{p,o}^2 + \alpha\sigma_i^2/4\,, \\ \text{cov}(r_{d,t}, r_{o,t}) &= -\alpha\sigma_i^2/4\,, \\ \text{cov}(r_{d,t}, y_{e,t}) &= \alpha^{1/2}\sigma_i\sigma_u/2\,, \\ \text{cov}(r_{o,t}, y_{e,t}) &= -\alpha^{1/2}\sigma_i\sigma_u/2\,. \end{aligned}$$

Conditional on an event, they are:

$$\begin{aligned} \text{var}(y_{e,t}) &= (1+1/\alpha)\sigma_u^2 + \sigma_z^2\,, \\ \text{var}(r_{d,t}) &= \sigma_{p,d}^2 + (1+\alpha)\sigma_i^2/4\,, \\ \text{var}(r_{o,t}) &= \sigma_{p,o}^2 + (1+\alpha)\sigma_i^2/4\,, \\ \text{cov}(r_{d,t}, r_{o,t}) &= (1-\alpha)\sigma_i^2/4\,, \\ \text{cov}(r_{d,t}, y_{e,t}) &= \alpha^{-1/2}\sigma_i\sigma_u/2 + \alpha^{1/2}\sigma_i\sigma_u/2\,, \\ \text{cov}(r_{o,t}, y_{e,t}) &= \alpha^{-1/2}\sigma_i\sigma_u/2 - \alpha^{1/2}\sigma_i\sigma_u/2\,. \end{aligned}$$

The OWR λ is:

$$\lambda_{\text{OWR}} = \frac{\alpha^{1/2}\sigma_i}{2\sigma_u}\,. \tag{C.4}$$

We measure $r_{d,t}$ and $r_{o,t}$ as open-to-VWAP (all on day t) and VWAP-to-open (from day t to day $t+1$) returns. As in the hybrid model, $y_{e,t}$ is total share imbalance in thousands of shares. Figure C.2 displays the time series of average parameter estimates for the OWR model. All three of the return variables, σ_i, $\sigma_{p,d}$, and $\sigma_{p,o}$, rise during the late 1990's and the financial crisis.

Internet Appendix C.3. Adjusted PIN Model (APIN)

The likelihood of the Duarte-Young model is:

$$L(B,S|\alpha,p_L,\mu,\varepsilon,\theta,\eta) = \prod_{t=1}^{T} \left\{ \begin{array}{l} (1-\alpha)(1-\theta)\left[\exp\left(-2\varepsilon\right)\frac{\varepsilon^{B_t+S_t}}{B_t!S_t!}\right] \\ (1-\alpha)\theta\left[\exp\left(-2(\varepsilon+\eta)\right)\frac{(\varepsilon+\eta)^{B_t+S_t}}{B_t!S_t!}\right] \\ +\alpha(1-\theta)p_L\left[\exp\left(-(\mu+2\varepsilon)\right)\frac{(\mu+\varepsilon)^{S_t}\varepsilon^{B_t}}{B_t!S_t!}\right] \\ +\alpha\theta p_L\left[\exp\left(-(\mu+2\varepsilon+2\eta)\right)\frac{(\mu+\varepsilon+\eta)^{S_t}(\varepsilon+\eta)^{B_t}}{B_t!S_t!}\right] \\ +\alpha(1-\theta)(1-p_L)\left[\exp\left(-(\mu+2\varepsilon)\right)\frac{(\mu+\varepsilon)^{B_t}\varepsilon^{S_t}}{B_t!S_t!}\right] \\ +\alpha\theta(1-p_L)\left[\exp\left(-(\mu+2\varepsilon+2\eta)\right)\frac{(\mu+\varepsilon+\eta)^{B_t}(\varepsilon+\eta)^{S_t}}{B_t!S_t!}\right] \end{array} \right\} \tag{C.5}$$

where B_t (S_t) is the number of buys (sells) on day t, α is the probability of an information event, p_L is the probability that an information event is bad news, μ and ε are the arrival rates of informed and uninformed traders, θ is the probability of a shock to buy and sell intensities, and η is the increment to buy and sell intensities when such a symmetric order flow shock occurs. We calculate Adjusted PIN using the formula:

$$\text{APIN} = \frac{\alpha\mu}{\alpha\mu + 2\varepsilon + 2\theta\eta}. \tag{C.6}$$

Figure C.3 displays the time series of average parameter estimates for the DY model. The parameters exhibit similar time-series dynamics to their counterparts in the PIN model.

Internet Appendix C.4. Volume-Synchronized PIN Measure (VPIN)

VPIN of Easley et al. (2012) builds on the intuition of the EKOP model that the numerator in PIN is the expected order imbalance while the denominator is expected volume. In order to estimate each of these components, the trading day is divided into equal size volume bins occurring in volume time τ. Let n denote the number of volume bins and V denote the volume in a single bin. For every volume bin τ, volume is signed to buying or selling volume based on the price change occurring over that bin. Let $t(\tau)$ denote the clock time corresponding to volume time τ and $\mathrm{N}(\cdot)$ denote the standard normal cumulative distribution function. Then volume in bin τ is assigned to buying and selling activity, respectively, as:

$$\begin{aligned} V_\tau^B &= \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot \mathrm{N}\left(\frac{P_i - P_{i-1}}{\sigma_{\Delta P}}\right) \\ V_\tau^S &= \sum_{i=t(\tau-1)+1}^{t(\tau)} V_i \cdot \left[1 - \mathrm{N}\left(\frac{P_i - P_{i-1}}{\sigma_{\Delta P}}\right)\right], \end{aligned}$$

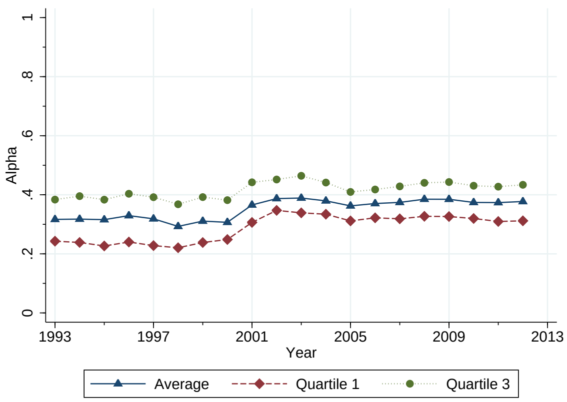
where the summation is over the number of 1-minute time intervals contained within volume bin τ, V_i is the volume in time bin i, $P_i - P_{i-1}$ is the price change over time bin i, and $\sigma_{\Delta P}$ is an estimate of the standard deviation of price changes within the day. We estimate VPIN using $n = 20$ volume bins per day. Volume-synchronized PIN is then defined as:

$$\mathrm{VPIN} = \frac{\sum_{\tau=1}^{n} \left|V_\tau^B - V_\tau^S\right|}{nV}. \tag{C.7}$$

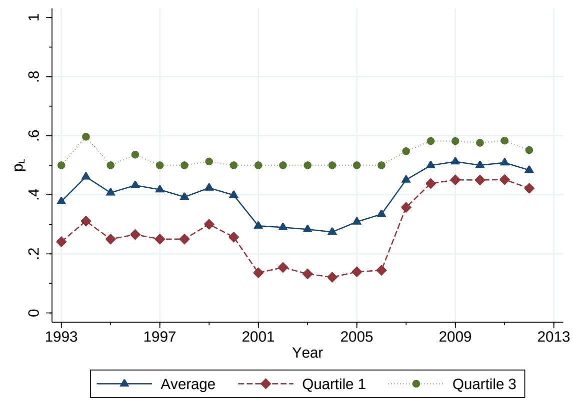
We calculate VPIN each day and average across days to create an average VPIN for each firm-year.

Figure C.1: Time Series of PIN Model Estimates
The annual cross-sectional mean, 25th and 75th percentiles of parameter estimates for the Easley et al. (1996) model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using daily buys and sells. The model parameters are α = probability of an information event, p_L = probability of a negative event, ε = Poisson intensity of uninformed trades, μ = Poisson intensity of informed trades, and PIN = Probability of informed trade.

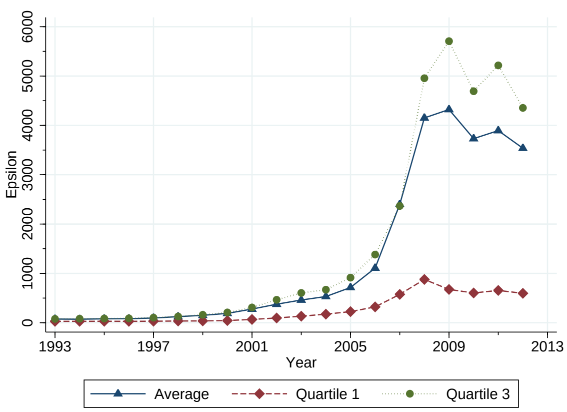
(a) α



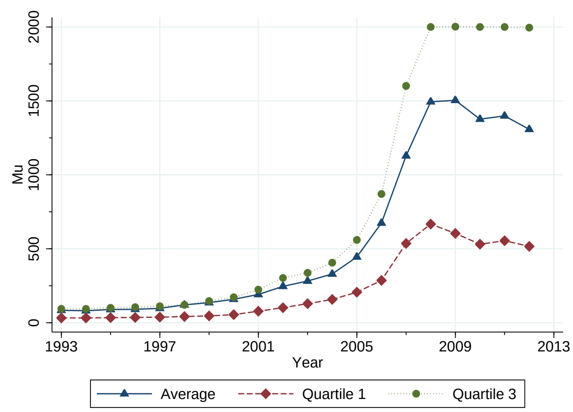
(b) p_L



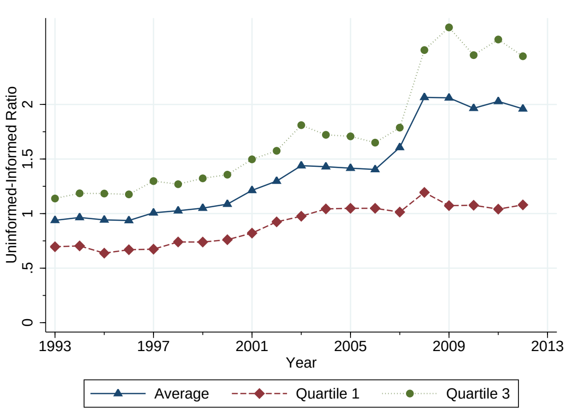
(c) ε



(d) μ



(e) $\frac{\varepsilon}{\mu}$



(f) PIN

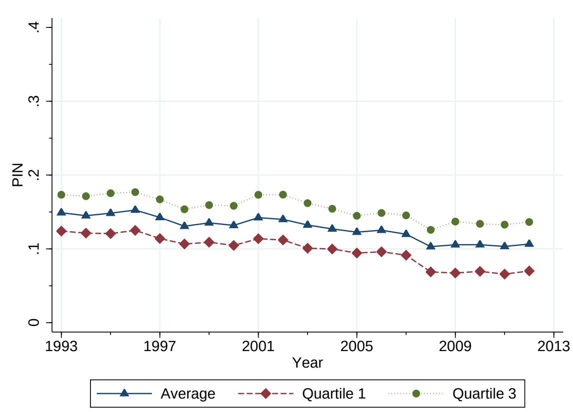
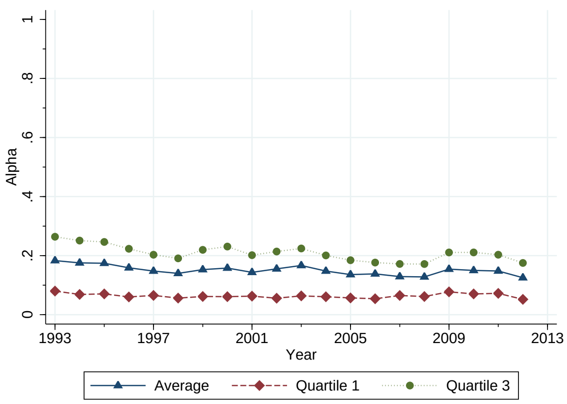
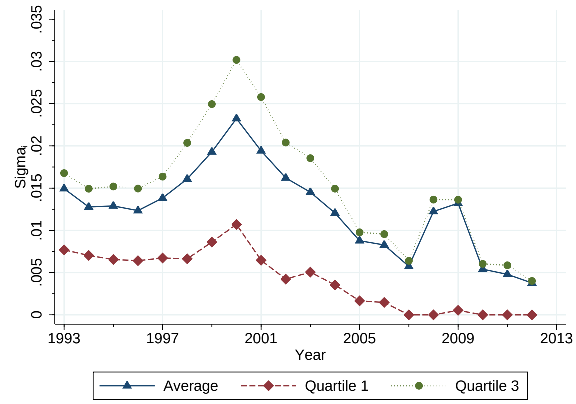


Figure C.2: Time Series of Odders-White and Ready Model Estimates
This figure plots the annual cross-sectional mean, 25th and 75th percentiles of parameter estimates for the Odders-White and Ready (2008) model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using daily order imbalances, intraday open-to-VWAP returns, and overnight VWAP-to-open returns. The model parameters are $\alpha =$ probability of an information event, $\sigma_i =$ the standard deviation of the mean zero, normally distributed private information conditional on an information event, $\sigma_u =$ the standard deviation of the mean zero, normally distributed net order flow from uninformed traders, $\sigma_{p_D} =$ the standard deviation of mean zero, normally distributed intraday public news, $\sigma_{p_O} =$ the standard deviation of mean zero, normally distributed overnight public news, and $\lambda =$ the price impact. Estimates of σ_z, the error with which the econometrician observes order flow is suppressed for space.

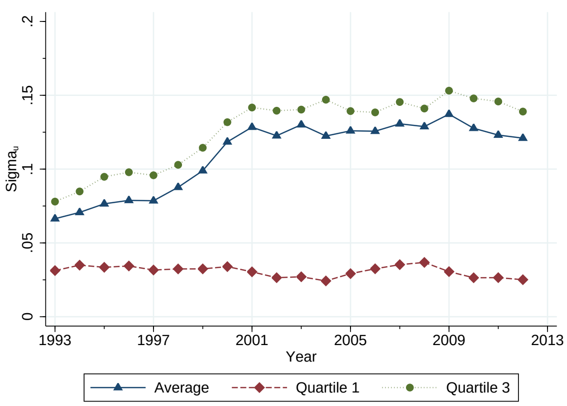
(a) α



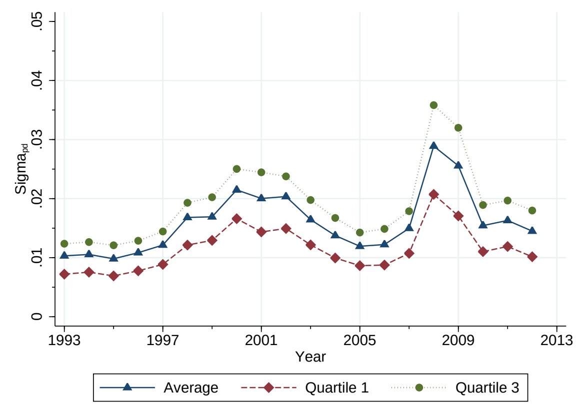
(b) σ_i



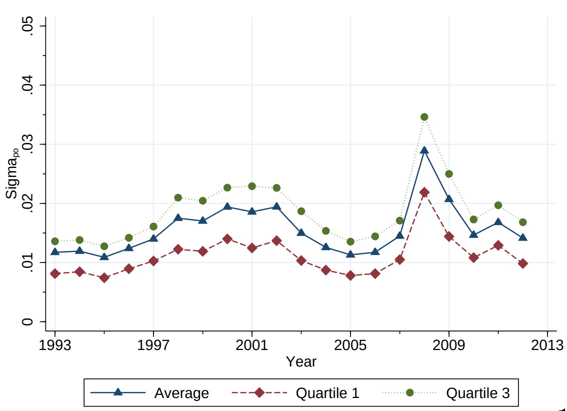
(c) σ_u



(d) σ_{p_D}



(e) σ_{p_O}



(f) λ_{OWR}

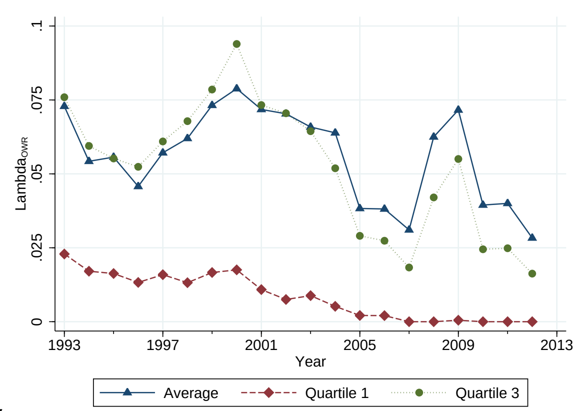
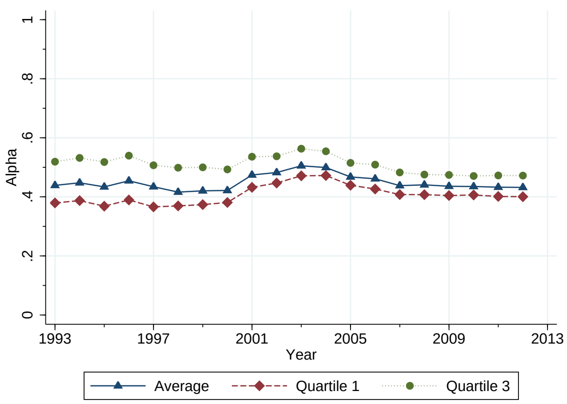


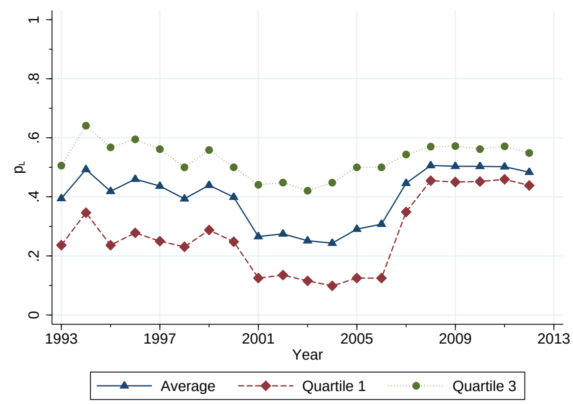
Figure C.3: Time Series of Adjusted PIN Model Estimates

This figure plots the annual cross-sectional mean, 25th and 75th percentiles of parameter estimates for the Duarte and Young (2009) model. The model is estimated on a stock-year basis for NYSE stocks from 1993 through 2012 using daily buys and sells. The model parameters are $\alpha =$ probability of an information event, $p_L =$ probability of a negative event, $\varepsilon =$ Poisson intensity of uninformed trades, $\mu =$ Poisson intensity of informed trades, $\theta =$ probability of a shock to buy and sell intensities, $\eta =$ increment to buy and sell intensities when a symmetric order flow shock occurs, and APIN $=$ Probability of informed trade.

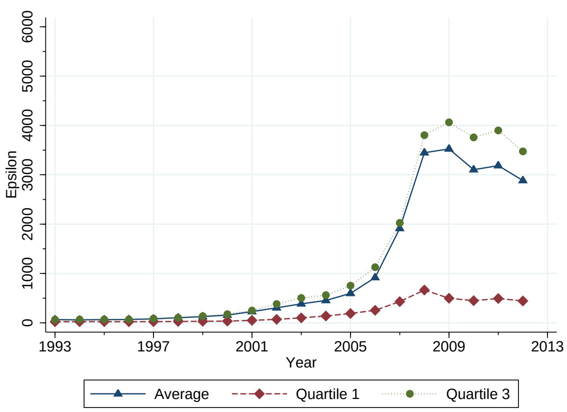
(a) α



(b) p_L



(c) ε



(d) μ

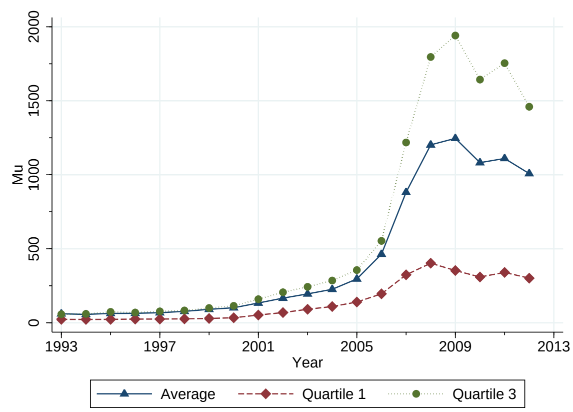
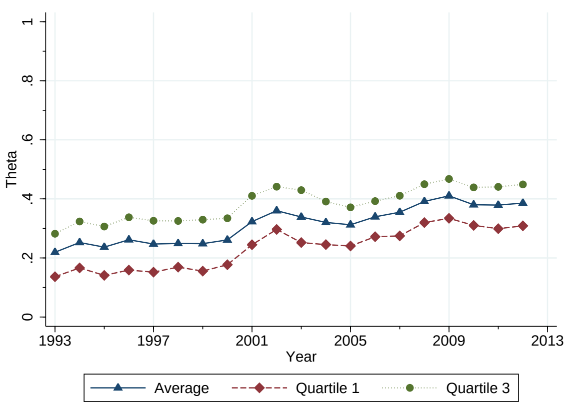
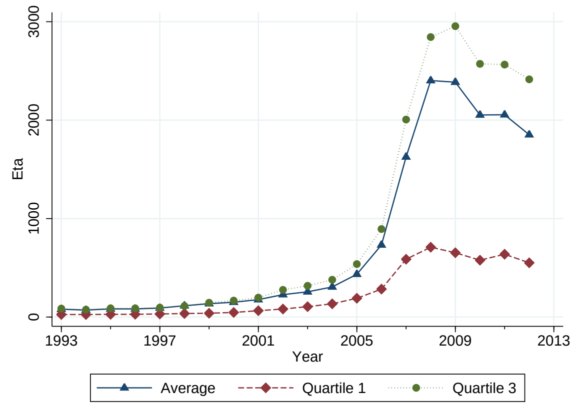


Figure C.3: (continued) Time Series of Adjusted PIN Model Estimates

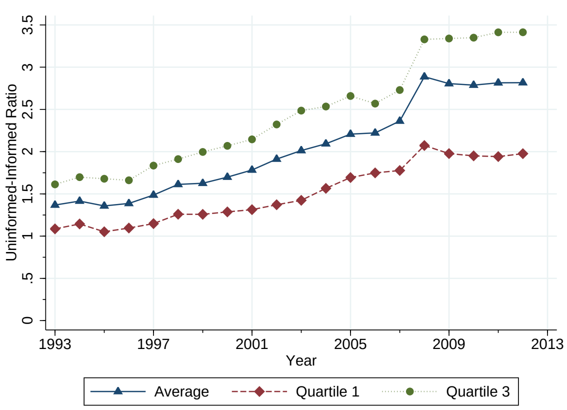
(e) θ



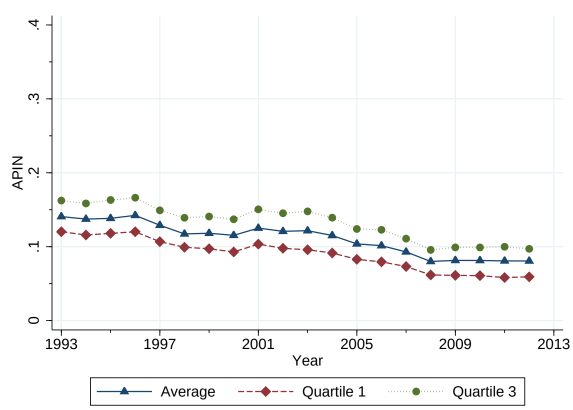
(f) η



(g) $\frac{\varepsilon+\theta\eta}{\mu}$



(h) APIN



Internet Appendix D. Empirical and Theoretical Order Flow Distributions

Each of the models have different implications for the unconditional distribution of order imbalances. For all four structural models, the order flow distribution is a mixture distribution. Figure D.1 shows how the distributions can differ based on the underlying parameter values, plotting the model-implied order imbalance distributions based on the estimates for the smallest and largest NYSE firm deciles. Under the hybrid model, end-of-day order flows are normally distributed with standard deviation σ. Under the OWR model, order flows are a mixture of two normal distributions, one for non-event days and a higher variance one for event days. Both of the Kyle-based models result in unimodal order flow distributions. On the other hand, the PIN and Adjusted PIN models imply order imbalance distributions that can be trimodal. Indeed, this is generally the case for order imbalances implied by structural estimates of the PIN and adjusted PIN models. The PIN and adjusted PIN models must fit volume as well as order imbalances since the input data are buy and sell volumes. On the other hand, the hybrid and OWR models need only fit the order flow distribution.

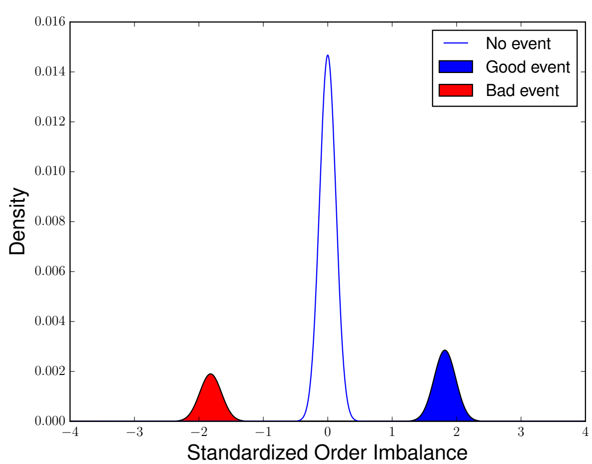
How do the model-implied order imbalance distributions compare to those found empirically? Figure D.2 shows the empirical standardized order imbalance distributions for the smallest and largest NYSE size deciles in our sample. The figure displays both share and trade imbalances since these are the underlying data for the Kyle-based and PIN models, respectively. The empirical distributions do not exhibit strong multimodal behavior. This is more consistent with the modeling assumption of the Kyle-based models than that of the PIN models.

Figure D.1: Model-implied Order Imbalance Distributions and Market Capitalization
The mixture distributions of standardized order imbalances implied by structural estimates from the hybrid and PIN models for the smallest and largest size deciles. Order imbalances are standardized by the standard deviation of order imbalances. For the hybrid model, the order imbalance variance is σ^2. For the PIN model, the order imbalance variance is $2\varepsilon + \alpha\mu(1+\mu) - (\alpha\mu(1-2p_L))^2$. For the APIN model, the order imbalance variance is $2(\varepsilon + \theta\eta) + \alpha\mu(1+\mu) - (\alpha\mu(1-2p_L))^2$. For the OWR model, the order imbalance variance is σ_u^2. For the hybrid and OWR model, the order imbalances are measures in shares. For the PIN and APIN model, the order imbalances are measures in number of trades. The parameters for each size decile are based on the structural estimates, some of which are reported in Table 7.

(a) Smallest Size Decile (Hybrid)

(b) Largest Size Decile (Hybrid)

(c) Smallest Size Decile (PIN)



(d) Largest Size Decile (PIN)

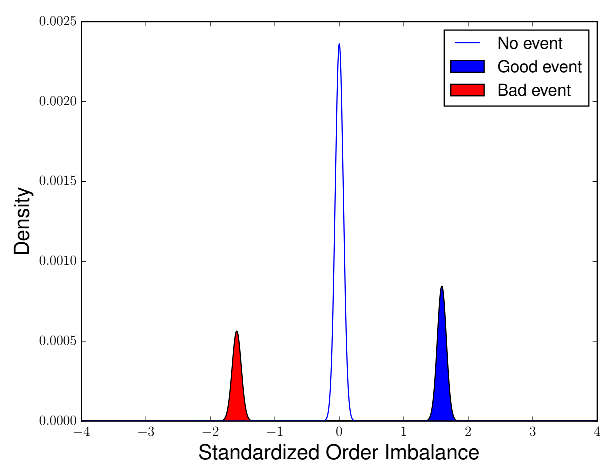
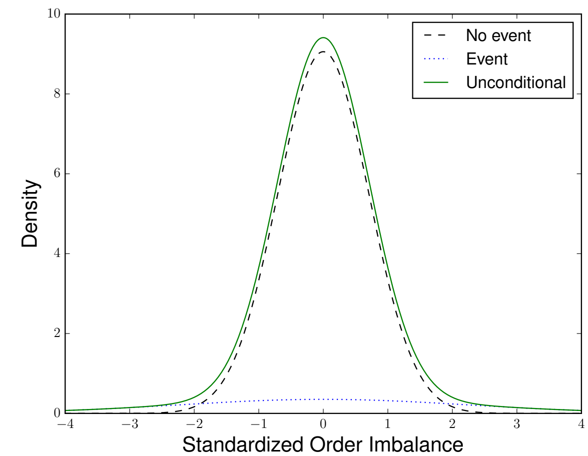
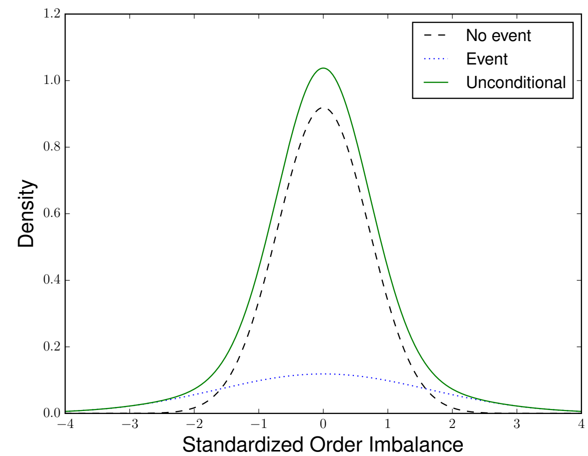


Figure D.1: (continued) Model-implied Order Imbalance Distributions and Market Capitalization

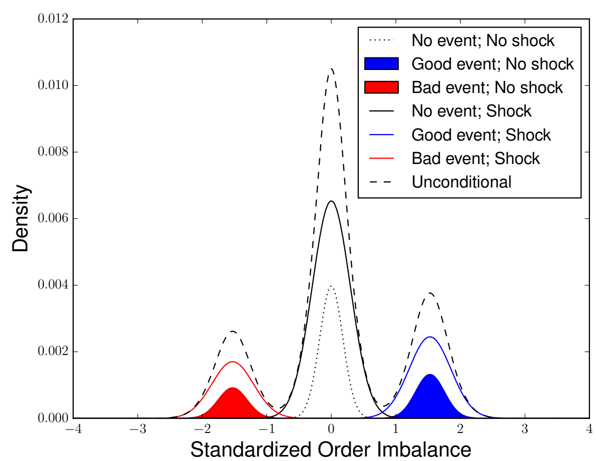
(e) Smallest Size Decile (OWR)



(f) Largest Size Decile (OWR)



(g) Smallest Size Decile (APIN)



(h) Largest Size Decile (APIN)

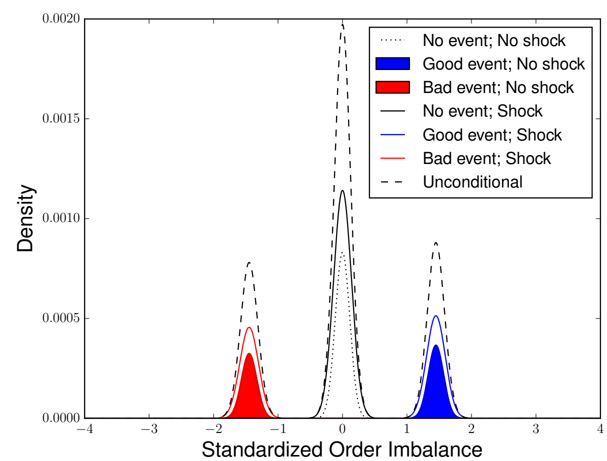
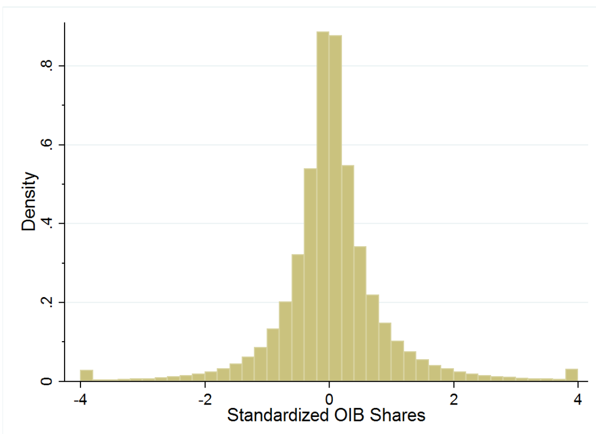
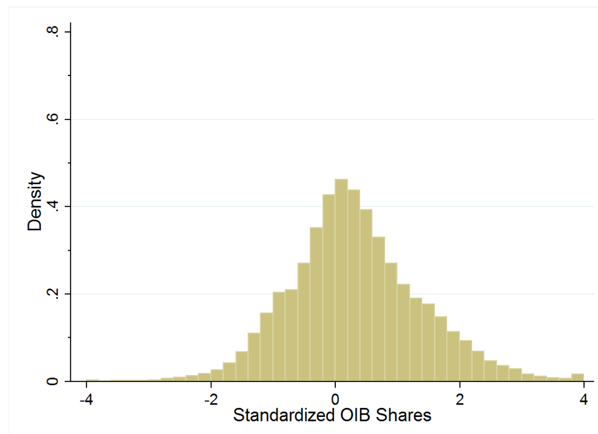


Figure D.2: Empirical Order Imbalance Distributions and Market Capitalization
The distributions of daily standardized order imbalances for the smallest and largest size deciles. For each firm-year, daily order imbalances are standardized by the firm-year standard deviation. The hybrid model is estimated using order imbalances measured in shares (top row) and the PIN models are estimated using order imbalances measured in number of trades (bottom row).

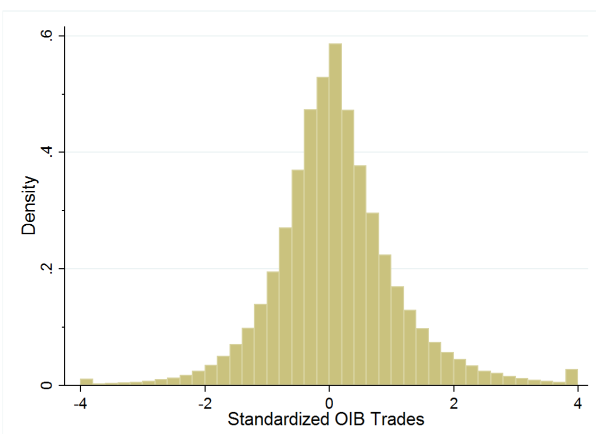
(a) Smallest Size Decile (Empirical Shares)



(b) Largest Size Decile (Empirical Shares)



(c) Smallest Size Decile (Empirical Trades)



(d) Largest Size Decile (Empirical Trades)