What does the PIN model identify as private information?

Jefferson Duarte, Edwin Hu, and Lance Young[*]

May 1st, 2015

Abstract

Some recent papers suggest that the Easley and O'Hara (1987) probability of informed trade (PIN) model fails to capture private information. We investigate this issue by comparing the PIN model with the Duarte and Young (2009) (DY) and Odders-White and Ready (2008) (OWR) models of private information arrival. We find that the PIN and DY models fail to capture private information because they mistakenly associate variations in turnover with the arrival of private information. On the other hand, the OWR model, which uses returns along with order flow imbalance, seems to plausibly identify the arrival of private information.

Keywords: Liquidity; Information Asymmetry

[*] We thank Kerry Back, Kevin Crotty, Gustavo Grullon, Sahn-Wook Huh, Avi Wohl and seminar participants at Rice University and Texas A&M University for helpful comments. We thank Elaine Brewer, Frank Gonzalez, Judy Hua, and Edward Martinez for computational support. Duarte and Hu are with the Jesse H. Jones School of Business at Rice University. Young is with the Michael G. Foster School of Business at the University of Washington. E-mails: jefferson.duarte@rice.edu (Duarte), eh7@rice.edu (Hu), and youngla@u.washington.edu (Young).

The Probability of Informed Trade (PIN) model, developed in a series of seminal papers including, Easley and O'Hara (1987), Easley, Kiefer, O'Hara, and Paperman (1996), and Easley, Kiefer, and O'Hara (1997), has been used extensively in accounting, corporate finance and asset pricing literature as a measure of information asymmetry.[1] The PIN model is based on the notion, originally developed by Glosten and Milgrom (1985), that periods of informed trade can be identified by abnormally large order flow imbalances. Recently, however, several papers have called into question the model's ability to identify informed trade because PINs tend to be at their lowest when information asymmetry should be at its highest (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), and Collin-Dufresne and Fos (2014a)).

We conduct an empirical examination of the PIN model to identify what might cause difficulties in its ability to identify informed trade. This exercise is important because the various possibilities imply very different agendas for this growing area of research. If PIN fails because its model does not fit the data well, then the PIN model could, in principle, be corrected by extending it to better fit the order flow data. Alternatively, it could be that net order flow itself is such a poor indicator of private information that no model based on order flow alone is capable of identifying informed trade, no matter how well it fits the data. Indeed, the theoretical work of Back, Crotty, and Li (2014) and the results in Kim and Stoll (2014) suggest that by itself, order flow imbalance is insufficient to isolate information events. If this is indeed the case, there is little to be gained by extending the PIN model to better fit the order flow data. Instead, a different approach involving variables other than order flow is necessary to generate useful inferences about the arrival of informed trade.

To address these possibilities, we compare the PIN model with two other models: one developed by Duarte and Young (2009) (the DY model), and the other by Odders-White and

[1] A Google scholar search reveals that this series of PIN papers has been cited more than 3,500 times as of this writing. Examples of papers that use PIN in the finance and accounting literature include Duarte, Han, Harford, and Young (2008), Bakke and Whited (2010), Da, Gao, and Jagannathan (2011), and Akins, Ng, and Verdi (2012).

Ready (2008) (the OWR model). Duarte and Young show that while the PIN model implies that the covariance between the daily number of buyer initiated trades (buys) and seller initiated trades (sells) is negative, it is, in fact, strongly positive for almost every publicly traded U.S. firm. The authors propose an alternative PIN measure ($Adj.\ PIN$) that is also based on Glosten and Milgrom (1985) but accounts for the positive correlation between buys and sells and thus improves the fit of the model. By comparing the ability of the DY and PIN models to identify private information, we can examine how a better fitting model of order flow helps in identifying private information. Odders-White and Ready develop a measure of private information based on Kyle (1985) rather than Glosten and Milgrom (1985). The OWR model uses intraday and overnight returns, along with order imbalance, to identify private information events. We compare the OWR and the PIN/DY models to highlight the importance of looking beyond order flow when identifying private information.

To compare these models, we examine the conditional probability of an information event ($CPIE$) implied by each model. To compute $CPIE$s, we first estimate each model's parameters using an entire year of data, and then use the observed market data (buys, sells, returns) to estimate the posterior or model-implied probability of an information event for each day in our sample. While the PIN and DY models allow for a calculation of the probability of informed trade, the OWR model does not. However, all three models have a parameter that controls the unconditional probability of an information event on a given day (α) and allow for the calculation of $CPIE$. As a result, we can compare how each model identifies private information through its $CPIE$.

We then examine how each of the three models identifies private information. We do this in two ways. First, we regress the PIN, DY, and OWR models' $CPIE$s ($CPIE_{PIN}$, $CPIE_{DY}$, $CPIE_{OWR}$ respectively) for each firm-year on the variables that, according to each model, are associated with the arrival of private information. In theory, variations in $CPIE_{PIN}$ and $CPIE_{DY}$ result from variation in absolute order imbalance, while variation in $CPIE_{OWR}$ is the result of variation in the squared and interaction terms of intraday returns, overnight returns, and order flow imbalances. In practice, however, the models may produce poor descriptions of the data, and model misspecification can affect the way they actually identify private information. Therefore, the hypothesis that we test in our regressions is whether observed variation in each model's $CPIE$ is consistent with each model's theory.

Second, we use estimates of each model's $CPIE$ to conduct event studies using a sample of target firms in mergers and acquisitions. We have two sets of hypotheses for our event studies. Our first event study hypothesis is similar to the one in our regression tests. The hypothesis is that variation in each model $CPIE$ in the event window is consistent with the theory behind each model. It is important to note that we test this hypothesis without any assumptions about whether or not private information arrives around the announcement. For our second event study test, we adopt the working hypothesis that private information should arrive prior to the event, rather than after the event.[2] Under this hypothesis, we expect that if a model correctly identifies informed trade, its $CPIE$ will increase prior to the announcement. We also anticipate that informed trading, and hence $CPIE$s, will decline rapidly after the announcement, when investors have the same (now public) information.

Our results are as follows. We find that the PIN model primarily identifies information events based on volume rather than absolute order flow imbalance. In regressions of $CPIE_{PIN}$ on absolute order imbalance, turnover, and their squared terms, turnover and turnover squared account for, on average, around 65% of the overall R^2. This is a problem because the PIN model mechanically associates increases in turnover with the arrival of private information although turnover can vary for reasons unrelated to private information.[3] This problem becomes more pronounced late in the sample because the model breaks down with the increase in both the level and volatility of turnover. For example, after 2006, the

[2] There is considerable evidence suggesting the possibility of high asymmetric information prior to important announcements. See for example Meulbroek (1992) and Hendershott, Livdan, and Schurhoff (2014).

[3] For instance, turnover can increase with disagreement (e.g. Kandel and Pearson (1995), and Banerjee and Kremer (2010)). Furthermore, recent work by Chae (2005) indicates that turnover may either decrease or increase due to asymmetric information.

PIN model suggests that 90% of the observed daily order flows for the median stock have a near-zero probability (i.e. smaller than 10^{-10}) of occurring. In addition, in our event studies, we find that $CPIE_{PIN}$ is higher after the announcement, due to the higher levels of volume in the post-announcement period. The intuition behind these results is that under the PIN model increases in volume can only come about through the arrival of private information. When confronted with actual data where turnover can change without the arrival of private information, the model mechanically interprets turnover shocks as periods of private information arrival, even when the order imbalance is zero.

We find that the DY model also identifies a relatively large proportion of variation in $CPIE$ with variation in turnover instead of order imbalance. In regressions of $CPIE_{DY}$ on absolute order imbalance, turnover, and their squares, turnover accounts for, on average, 40% of the overall R^2. Furthermore, the DY model also breaks down late in the sample. As with the PIN model, for the median stock in our sample, the DY model suggests that about 90% of daily order flows have a near-zero probability of occurring after 2006. Moreover, in our event study of M&A targets, we find that like $CPIE_{PIN}$, the $CPIE_{DY}$ increases before the event, but remains elevated after the event, largely due to the increased volume in the post-event period.

In contrast, we find that the OWR model generally behaves as our hypotheses suggest it should. We find that the R^2 of a regression of $CPIE_{OWR}$ on the squared and interaction terms of intraday returns, overnight returns, and order flow imbalances is around 80%. In contrast to the PIN and DY models, the behavior of the OWR model does not change in the latter part of the sample because it relies on order imbalance, which remains stable over time, as opposed to the levels of buyer and seller initiated trades, which are subject to (common) stochastic trends. Furthermore, we find that variation in the $CPIE_{OWR}$ around merger announcements is almost completely due to variation in both the squares and the interactions of intraday and overnight returns instead of variation in the square and interactions of order imbalance. Hence, under the working hypothesis that private information about merger targets arrives in the market before the announcement of the merger, our findings suggests that variables other than order flow may be significant in identifying private information. In addition, we find the $CPIE_{OWR}$ increases before the announcement and decreases rapidly after the announcement, which suggests, under our working hypothesis, that the OWR model identifies the arrival of informed trade around events in a sensible way.

Overall, our findings suggest that PIN and *Adj.* PIN are poor proxies for private information, while the OWR model's $CPIE$ or its α may be reliable proxies for private information. It seems implausible that PIN and *Adj.* PIN correctly capture private information late in the sample since they are based on models that cannot account for 90% of the median stock's observed order flow. Moreover, the mechanical relation between turnover and $CPIE_{PIN}$ suggests that PIN does not properly identify private information. The performance of the OWR model, on the other hand, suggests that its α or $CPIE$ may be promising measures of information asymmetry.

Our paper is related to a growing literature that analyzes the extent to which PIN actually captures information asymmetry.[4] Most of this research attempts to do so by estimating $PINs$ around events and testing whether PIN is higher before than after an announcement. Using this approach, Collin-Dufresne and Fos (2014a) find that PIN and other adverse selection measures are lower when Schedule 13D filers trade. Collin-Dufresne and Fos partially attribute this finding to informed traders choosing to trade in periods of high liquidity and turnover (see also Collin-Dufresne and Fos (2014b)). Similarly, Aktas, de Bodt, Declerck, and Van Oppens (2007) find that PIN is higher after merger announcements than before, partially as a result of increases in PIN model's α.[5] Easley, Engle, O'Hara, and Wu (2008)

[4] A series of papers addresses the pricing of information asymmetry in the cross section of stock returns (e.g. Easley, Hvidkjaer, and O'Hara (2002), Duarte and Young (2009), Mohanram and Rajgopal (2009), Lambert and Leuz (2012), and Lai, Ng, and Zhang (2014)). In contrast to this paper, Duarte and Young (2009) question whether PIN is priced because it identifies private information or because it is related to illiquidity. We do not address the pricing of PIN, instead we aim to understand how the PIN, DY, and OWR models identify private information.

[5] In addition, Benos and Jochec (2007) find that PIN is higher after earnings announcements than before the announcements.

critique this line of research, noting that PIN is a stock characteristic rather than a measure of the extent to which private information is present in a given calendar time period.[6] To address this critique, Easley, Engle, O'Hara, and Wu (2008) develop an extension of the original model in which PIN is time-varying, and in a paper contemporaneous to ours, Brennan, Huh, and Subrahmanyam (2015) use conditional probabilities similar to $CPIE_{PIN}$.

We contribute to this literature in three ways. First, our results indicate that these previously identified PIN anomalies are partially related to the strong connection between $CPIE_{PIN}$ and turnover that we document. Our findings cannot speak to the possibility that informed traders sometimes choose to trade in periods of high volume as suggested by Collin-Dufresne and Fos (2014b), instead we show that the PIN model often mechanically attributes increases in turnover with the arrival of private information. Second, we show that event studies that use daily measures of private information (e.g. Easley, Engle, O'Hara, and Wu (2008)) can be misleading if variation in these measures around event announcements is due to variables not necessarily related to information asymmetry. For instance, Brennan, Huh, and Subrahmanyam (2015) interpret the fact that their $CPIE_{PIN}$ measures are higher after earnings announcements than before as evidence of informed trading. We show that $CPIE_{PIN}$ is closely related to volume (controlling for order flow imbalance), therefore their findings may be due to the fact that turnover is higher after announcements. Third, we show that the OWR model represents a promising model for potential future use in the wide variety of contexts where proxies for information asymmetry are needed.

The remainder of the paper is outlined as follows. Section 1 briefly describes the intuition behind the PIN, DY, and OWR models. Section 2 outlines the data we use for our empirical results. Section 3 describes the maximum likelihood procedure. Section 4 analyzes how each

[6] Easley, Lopez de Prado, and O'Hara (2012) develop the volume-synchronized probability of informed trading or $VPIN$. We do not consider $VPIN$ in this paper because, as Easley, Lopez de Prado, and O'Hara (2012) point out, $VPIN$ is a measure of order flow toxicity at high frequencies rather than a stock characteristic that measures adverse selection at lower frequencies as PIN is widely used in the finance and accounting literature. Moreover, Andersen and Bondarenko (2014) provide detailed critique of the $VPIN$ measure.

model identifies information events. Section 5 examines the three models' ability to identify private information arrival around merger announcements. Section 6 concludes.

1 Description of the estimated models

In this section, we outline the intuition behind the PIN, DY, and OWR models.[7] We then discuss how to compute the $CPIE$ for each using the data and the model parameters. Our analyses focus on the probability of an information event that an econometrician using each model would infer on a particular day after seeing the 'market' data for that day. For the PIN and DY models, the 'market' data consist of the numbers of buy and sell orders. For the OWR model, the data consist of the intraday and overnight returns and the order flow imbalance.

1.1 The PIN model

The Easley, Kiefer, O'Hara, and Paperman (1996) PIN model posits the existence of a liquidity provider who receives buy and sell orders from both informed traders and uninformed traders. At the beginning of each day, the informed traders receive a private signal with probability α. If the private signal is positive, buy orders from informed and uninformed traders arrive following a Poisson distribution with intensity $\mu + \epsilon_B$, while sell orders come only from the uninformed traders and arrive with intensity ϵ_S. If the private signal is negative, sell orders from informed and uninformed traders arrive following a Poisson distribution with intensity $\mu + \epsilon_S$, while buy orders come only from the uninformed traders and arrive with intensity ϵ_B. If the informed traders receive no private signal, they do not trade; thus, all buy and sell orders come from the uninformed traders and arrive with intensity ϵ_B and ϵ_S, respectively. Fig.1 Panel A shows a tree diagram of this model.

[7] For a more detailed description of the PIN model, see Easley, Kiefer, O'Hara, and Paperman (1996). A more detailed description of the DY model is in Duarte and Young (2009). For a more detailed discussion of the OWR model, see Odders-White and Ready (2008).

The difference in arrival rates captures the intuition that on days with positive private information, the arrival rate of buy orders increases over and above the normal rate of noise trading because informed traders enter the market to place buy orders. Similarly, the arrival rate of sell orders rises when the informed traders seek to sell based on their negative private signals. Therefore, the PIN model identifies the arrival of private information through increases in the absolute value of the order imbalance. The model also ties large variations in turnover to the arrival of private information. To see this, note that the expected number of buys plus sells on days with private information is $\epsilon_B + \epsilon_S + \mu$, while the expected number of trades on days without private information is $\epsilon_B + \epsilon_S$. Thus, under the PIN model, private information is necessarily the cause of any variation in average turnover.

1.1.1 $CPIE_{PIN}$

We estimate the PIN model numerically via maximum likelihood. Let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock i on day t and $\Theta_{PIN,i} = (\alpha_i, \mu_i, \epsilon_{B_i}, \epsilon_{S_i}, \delta_i)$ represent the vector of the PIN model parameters for stock i. Let $D_{PIN,i,t} = [\Theta_{PIN,i}, B_{i,t}, S_{i,t}]$. The likelihood function of the Easley, Kiefer, O'Hara, and Paperman (1996) model is $\prod_{t=1}^{T} L(D_{PIN,i,t})$, where

$$L(D_{PIN,i,t}) = L_{NI}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t}) + L_{I^-}(D_{PIN,i,t}). \tag{1}$$

$L_{NI}(D_{PIN,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information; $L_{I^+}(D_{PIN,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with positive information; and $L_{I^-}(D_{PIN,i,t})$ is the likelihood on a day with negative information. The likelihood equation shows that at each node of the tree in Fig. 1 Panel A, buys and sells arrive according to independent Poisson distributions, with the intensity parameters differing according to the node of the tree. See Internet Appendix A for the formulas for the likelihood functions.

Using the PIN model, for each stock-day, we compute the probability of an information event conditional both on the model parameters and on the observed total number of buys and sells. Specifically, let the indicator $I_{i,t}$ take the value of one if an information event occurs for stock i on day t, and zero otherwise. For the PIN model, we compute $CPIE_{PIN,i,t} = P\left[I_{i,t} = 1 | D_{PIN,i,t}\right]$. This probability is given by

$$CPIE_{PIN,i,t} = \frac{L_{I^-}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t})}{L(D_{PIN,i,t})}, \tag{2}$$

$CPIE_{PIN,i,t}$ represents the econometrician's posterior probability of an information event given the data observed on that day, and assuming that he or she knows the underlying model's parameters.

Note that if we condition down with respect to the data, $CPIE_{PIN,i,t}$ reduces to the model's unconditional probability of information events (α_i). The unconditional probability represents the econometrician's beliefs about the likelihood of an information event before seeing any actual orders or trades. In the absence of buy and sell data, an econometrician would assign a probability α_i to an information event for stock i on day t, where $\alpha_i = E[CPIE_{PIN,i}]$ and the expectation is taken with respect to the joint distribution of $B_{i,t}$ and $S_{i,t}$. The PIN of a stock, defined as $\frac{\alpha\mu}{\alpha\mu+\varepsilon_B+\varepsilon_S}$, is the unconditional probability that any given trade is initiated by an informed trader. $CPIE$ and PIN are linked via the unconditional probability of an information event, α, which is also the unconditional expectation of $CPIE$.

1.2 The DY model

Duarte and Young (2009) extend the PIN model to address some of its shortcomings in matching the order flow data. Specifically, the authors note that the PIN model implies that the number of buys and sells are negatively correlated; however, in the data the correlation between the number of buys and sells is overwhelmingly positive. To correct this problem, the DY model partially disentangles turnover variation from private information arrival. As in the PIN model, the DY model posits that at the beginning of each day, informed investors receive a private signal with probability α. If the private signal is positive, buy orders from the informed traders arrive according to a Poisson distribution with intensity μ_B. If the private signal is negative, informed sell orders arrive according to a Poisson distribution with intensity μ_S. If the informed traders receive no private signal, they do not trade.

In contrast to the PIN model, the DY model allows for symmetric order flow shocks. These shocks increase both the number of buyer- and seller-initiated trades but are unrelated to private information events. Symmetric order flow shocks can happen for a variety of reasons, such as disagreement among traders about the interpretation of public news. Alternatively, liquidity shocks may occur that cause investors holding different collections of assets to simultaneously rebalance their portfolios, resulting in increases to both buys and sells. Regardless of the mechanism, symmetric order flow shocks arrive on any given day with probability θ. On days with symmetric order flow shocks, both the number of buyer- and seller-initiated trades increase by amounts drawn from independent Poisson distributions with intensity Δ_B or Δ_S, respectively. Buy and sell orders from uninformed traders arrive according to a Poisson distribution with intensities ϵ_B $(\epsilon_B + \Delta_B)$ and ϵ_S $(\epsilon_S + \Delta_S)$ on days without (with) symmetric order flow shocks. Fig. 1 Panel B shows the structure of the DY model.

Under the DY model, turnover can increase due to either symmetric order flow shocks or the arrival of private information. To see this, note that the expected number of buys plus sells on days with positive (negative) information and without symmetric order flow shocks is $\epsilon_B + \epsilon_S + \mu_B$ $(\epsilon_B + \epsilon_S + \mu_S)$; the expected number of trades on days with symmetric order flow shocks and without private information shocks is $\epsilon_B + \epsilon_S + \Delta_B + \Delta_S$, and the expected number of trades is $\epsilon_B + \epsilon_S$ on days without either.

1.2.1 $CPIE_{DY}$

As with the PIN model, we estimate the extended model numerically via maximum likelihood. Let $\Theta_{DY,i} = (\alpha_i, \mu_{B_i}, \mu_{S_i}, \epsilon_{B_i}, \epsilon_{S_i}, \delta_i, \theta_i, \Delta_{B_i}, \Delta_{S_i})$ be the vector of parameters of the DY model for stock i. Let $B_{i,t}$ and $S_{i,t}$ be the number of buys and sells, respectively, for stock i on day t. Let $D_{DY,i,t} = [B_{i,t}, S_{i,t}, \Theta_{DY,i}]$. The likelihood function of the extended model is $\prod_{t=1}^{T} L(D_{DY,i,t})$:

$$\begin{aligned} L(D_{DY,i,t}) &= L_{NI,NS}(D_{DY,i,t}) + L_{NI,S}(D_{DY,i,t}) + L_{I^-,NS}(D_{DY,i,t}) \\ &\quad + L_{I^-,S}(D_{DY,i,t}) + L_{I^+,NS}(D_{DY,i,t}) + L_{I^+,S}(D_{DY,i,t}) \end{aligned} \tag{3}$$

where $L_{NI,NS}(D_{DY,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information or a symmetric order flow shock; $L_{NI,S}(D_{DY,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day without private information but with a symmetric order flow shock; $L_{I^-,NS}$ ($L_{I^-,S}$) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information and without (with) a symmetric order flow shock; and $L_{I^+,NS}$ ($L_{I^+,S}$) is the probability on a day with positive information and without (with) a symmetric order flow shock. Analogous to the original PIN model, each term in the likelihood function corresponds to a branch in the tree in Fig. 1, Panel B. See Internet Appendix B for the formulas for the likelihood functions.

As with the PIN model, for each stock-day, we compute the probability of an information event conditional on both the model parameters and on the number of buys and sells observed that day. Specifically, let the indicator $I_{i,t}$ take the value of one if an information event occurs for stock i on day t and zero otherwise. We compute $CPIE_{DY,i,t} = P\left[I_{i,t} = 1 | D_{DY,i,t}\right]$ as:

$$CPIE_{DY,i,t} = \frac{L_{I^+,NS}(D_{DY,i,t}) + L_{I^+,S}(D_{DY,i,t}) + L_{I^-,S}(D_{DY,i,t}) + L_{I^-,NS}(D_{DY,i,t})}{L(D_{DY,i,t})} \tag{4}$$

Analogous to the PIN model, the *Adj. PIN* of a stock is $\frac{\alpha(\delta\mu_B+(1-\delta)\mu_S)}{\alpha(\delta\mu_B+(1-\delta)\mu_S)+\varepsilon_B+\varepsilon_S+\theta(\Delta_B+\Delta_S)}$. This is the unconditional probability that any given trade is initiated by an informed trader. $CPIE_{DY}$ and *Adj. PIN* are linked via the unconditional probability of an information event, α, which is also the unconditional expectation of $CPIE_{DY}$.

To illustrate how the $CPIE_{PIN}$ and $CPIE_{DY}$ work, we present a stylized example of the PIN and DY models in Fig. 2. We plot simulated and real order flow data for Exxon-Mobil during 1993, with buys on the horizontal axis and sells on the vertical axis. Real data are marked as +, and simulated data as transparent dots. The real data are shaded according to the model-specific $CPIE$, with lighter points (+) representing low and darker points (+) high $CPIE$s.

Panel A of Fig. 2 illustrates the central intuition behind the PIN model. The simulated data comprise three types of days, which create three distinct clusters. Two of the clusters are made up of days characterized by relatively large order flow imbalance, with a large number of sells (buys) and relatively few buys (sells). The third group of days has relatively low numbers of buys and sells; these days have no private information arrival. Generalizing from this figure, days with large order flow imbalances (i.e. the number of buys minus the number of sells) are likely to correspond to informed traders entering the market. The real data, on the other hand, show no distinct clusters. Furthermore, an econometrician naively using the model to identify days with private information would mistakenly classify those with large turnover (i.e. the days in the northeast corner of the panel) as days with private information events.

Panel B shows the same graph for the DY model. The DY model generates six data clusters, greatly improving upon the PIN model's coverage of the data. The two clusters on the dotted line in Panel B are not related to private information, but the other four clusters are. An econometrician using the DY model, moving along the dotted line, would observe that high turnover–considered information days under the PIN model–are no longer classified as such, because higher turnover may be driven by symmetric order flow shocks under the DY model. Instead, the DY model identifies private information when moving away from the dotted line; when buys are greater than sells and vice versa.

1.3 The OWR model

Odders-White and Ready (2008) extends Kyle (1985) by allowing for days with no information events. The OWR model identifies the arrival of private information through both the price response to order imbalances and the subsequent responses to the public revelation of private information (which they assume arrives overnight). In the OWR model, stock prices respond to order imbalances because there is always the risk of an information event; however, when this order flow is entirely based on of noise traders, the price response is subsequently reversed.

Specifically, the OWR model assumes that each day the econometrician observes three variables: a noisy measure of the day's total order imbalance due to informed and noise traders y_e, the intraday return (measured from the open to the volume-weighted average price (VWAP)) r_d, and the overnight return (measured from the VWAP to the open the following day), r_o.[8] The econometrician can make inferences about the probability of an information event by observing y_e, r_d and r_o because the covariance matrix of the three variables differs between days when private information arrives and those when only public information is available.

To see this, consider the variance of the observed order flow, y_e. If no information event occurs, then $Var(y_e)$ is composed of only the variances of the uninformed order flow and the noise in the data. However, if an event occurs, $Var(y_e)$ increases because the order flow reflects at least some informed trading. Similarly, $Var(r_d)$ is higher for an information event, because it reflects the market maker's partial reaction to the day's increased order flow. Since the private signal is revealed after trading closes, $Var(r_o)$ increases in the wake of an information event, as it reflects the remainder of the market maker's partial reaction to the informed trade component in the order flow. Likewise, information events make $cov(y_e, r_d)$ and $cov(y_e, r_o)$ rise. The higher covariance between order flow and intraday returns occurs because, for an information event, both order flow and the intraday return (partially) reflect the impact of informed trading. Along these same lines, because the market maker cannot perfectly separate the informed from the uninformed order flows, she is unable to fully adjust the price during the day to reflect the informed trader's private signal. However, since the private signal is assumed to be fully reflected in prices after the close, for an information event, $cov(y_e, r_o)$ is higher because the overnight returns incorporate any additional reaction to the private signal that was not captured in prices during the day. Finally, $cov(r_o, r_d)$ is positive for information events, reflecting the fact that the information

[8] We suppress the t subscript for ease of exposition.

event is not completely captured in prices during the day. In contrast, $cov(r_o, r_d)$ is negative in the absence of an information event, as the market marker's reaction to the noise trade during the day is reversed after she learns that there was no private signal.

1.3.1 $CPIE_{OWR}$

As with the PIN and DY models, we estimate the OWR model numerically via maximum likelihood. Let $\Theta_{OWR,i}$ represent the vector of OWR model parameters for stock i, $r_{d,i,t}$ and $r_{o,i,t}$ represent the intraday and overnight returns for stock i on day t, and $y_{e,i,t}$ represent the order flow imbalance for stock i on day t. Let $D_{OWR,i,t} = [\Theta_{OWR,i}, y_{e,i,t}, r_{d,i,t}, r_{o,i,t}]$. The likelihood function of the OWR model is $\prod_{t=1}^{T} L(D_{OWR,i,t})$ where:

$$L(D_{OWR,i,t}) = L_{NI}(D_{OWR,i,t}) + L_I(D_{OWR,i,t}), \tag{5}$$

where L_{NI} (L_I) is the likelihood of observing $y_{e,i,t}$, $r_{d,i,t}$, and $r_{o,i,t}$ on a day without (with) an information event. See Internet Appendix C for the formulas for the likelihood functions.

The probability of an information event, conditional on $D_{OWR,i,t}$ is therefore $CPIE_{OWR,i,t} = P\left[I_{i,t} = 1 | D_{OWR,i,t}\right]$. This probability is given by:

$$CPIE_{OWR,i,t} = \frac{L_I(D_{OWR,i,t})}{L(D_{OWR,i,t})} \tag{6}$$

As in the PIN and DY models, if we condition down with respect to the data, $CPIE_{OWR,i,t}$ reduces to the model's unconditional probability of information events (α_i).

In contrast to the PIN and DY models, the OWR model does not contain a direct analog to the probability of informed trading. To understand this result, note that the probability of informed trade in the PIN and DY models is given by the ratio of the expected number of informed trades to the expected total number of trades on a given day. Since the OWR model does not make assumptions about the number of trades, it is mute regarding this ratio. This may appear to be a limitation of the OWR model, but we argue that it is actually an advantage. This is because, as we show in Section 4, the fact the OWR does not make assumptions about the level of buys and sells allows it to disentangle variations in turnover from the arrival of informed trading. Moreover, the parameter α in the OWR model is a measure of information asymmetry in a given stock.

2 Data

To estimate the PIN, DY, and OWR models, we need daily data on the number of buyer- and seller-initiated trades for each firm in the sample. We collect trades and quotes data for stocks between 1993 and 2012 from the NYSE TAQ database. We require that the stocks in our sample have only one issue (i.e. one `PERMNO`), are common stock (share code 10 or 11), are listed on the NYSE (exchange code 1), and have at least 200 days worth of non-missing observations for the year. Our sample contains 1,060 stocks per year on average, with a minimum number stocks in 2012 of 934, and a maximum number of stocks in 1997 of 1,155. As a result of our sample selection, about 36% (25%) of the stocks in our sample are in the top (bottom) three Fama-French size deciles. For each stock in the sample, we classify each day's trades as either buys or sells, following the Lee and Ready (1991) algorithm. In our analysis, we include turnover, defined as the sum of daily buys and sells divided by the number of shares outstanding. Internet Appendix D describes the computation of the number of buys and sells.

We estimate both the PIN and DY models using the daily total number of buys and sells ($B_{i,t}$ and $S_{i,t}$). The OWR model also requires intraday and overnight returns as well as order imbalances. Following Odders-White and Ready (2008) we compute the intraday return at day t as the volume-weighted average price (VWAP) at t minus the opening quote midpoint at t plus dividends at time t, all divided by the opening quote midpoint at time t.[9] We compute the overnight return at t as the opening quote midpoint at $t+1$ minus the VWAP at t, all divided by the opening quote midpoint at t. The total return, or sum of the

[9] The opening quote midpoint is not available in TAQ many instances. When the opening quote midpoint is not available, we use the matched quote of the first trade in the day as a proxy for the opening quote.

intraday and overnight returns is the open-to-open return from t to $t+1$. We compute order imbalance (y_e) as the daily total number of buys minus the total number of sells, divided by the total number of buys plus sells. We follow Odders-White and Ready and remove systematic effects from returns to obtain measures of unexpected overnight and intraday returns ($r_{o,i,t}$ and $r_{d,i,t}$). Internet Appendix D describes how we compute $r_{o,i,t}$ and $r_{d,i,t}$.

Like Odders-White and Ready (2008), we remove days around unusual distributions or large dividends, as well as CUSIP or ticker changes. We also drop days for which we are missing overnight returns ($r_{o,i,t}$), intraday returns ($r_{d,i,t}$), order imbalance (y_e), buys (B), or sells (S). Our empirical procedures follow those of Odders-White and Ready with two exceptions. First, OWR estimate y_e as the idiosyncratic component of net order flow divided by shares outstanding. We do not follow the same procedure as OWR in defining y_e because we find that estimating y_e as we do results in less noisy estimates. Specifically, we find that y_e defined as buys minus sells divided by shares outstanding, as in Odders-White and Ready (2008), suffers from scale effects late in the sample, when order flow is several orders of magnitude larger than shares outstanding. Second, Odders-White and Ready remove a whole trading year of data surrounding distribution events, but we only remove one trading week [-2,+2] around these events.

For the event study portion of our analysis, we examine merger and acquisition (M&A) targets. We collect M&A dates from Thompson Reuter's SDC database. If the event occurs on a non-trading day, then we use the next available trading day as the event date. We require that firms have a `PERMNO`, which we get by matching six-digit CUSIPs from SDC to CRSP. Our event sample includes on average 673 merger targets per year, with a minimum of 231 in 2012 and a maximum of 1,030 in 1998. Table 1 contains summary statistics of all the variables used to estimate the models. Panel A gives summary statistics of our entire sample, and Panel B displays the summary statistics for the days of merger announcements.

3 Estimation of the models

For every firm-year in our sample, we estimate the PIN, DY, and OWR models by maximizing the likelihood functions introduced in Section 1. The estimation procedure is similar to that used in Duarte and Young (2009). The parameter estimates are used for computing the $CPIE$s used in Section 4. Internet Appendices E and F provide details about the maximum likelihood procedure and the calculation of $CPIE$s.

Table 2 contains summary statistics for the parameter estimates from each model. Panel A displays the summary statistics of the PIN model parameters, Panel B of the DY model, and Panel C of the OWR model. Table 2 also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE$. We see that the mean $CPIE$ behaves exactly like α for all of the three models. Hence, changes in $CPIE$ and changes in the estimated alphas are analogous. Fig. 3 shows how the distribution of α changes over time. The distribution of the PIN and DY α parameters in the early part of the sample is similar to that in Duarte and Young (2009). On the other hand, the estimated OWR α parameters are in general higher than those in Odders-White and Ready (2008). This is due to the fact that our definition of y_e is different from that in Odders-White and Ready (2008) (see the discussion in Section 2 above).[10] Interestingly, the PIN model α increases over time, with the median PIN α rising from about 30% in 1993 to 50% in 2012.[11]α parameters from 1993 to 2012 is comparable to that in Brennan, Huh, and Subrahmanyam (2015). Neither the DY nor the OWR α changes as much as the PIN α, remaining within the 40% to 50% range. Panels A and B of Fig. 4 plot the time series of PIN and $Adj.\ PIN$ respectively. Note that both PIN and $Adj.\ PIN$ decrease over time in spite of the fact that α increases. This happens because, according to these models, the intensity of noise trading is increasing over

[10] In fact, we get α estimates close to those reported in Odders-White and Ready (2008) if we define y_e in the same way that they do.

[11] The increase in our estimated PIN model α parameters is somewhat larger than that in Brennan, Huh, and Subrahmanyam (2015). This small difference arises because Brennan, Huh, and Subrahmanyam (2015) have a larger number of stocks per year due to the fact that we apply sample filters similar to those in Odders-White and Ready (2008). In fact, without these filters, the increase in our estimated PIN model

time while the intensity of informed trading remains flat as shown in Panels C and D of Fig. 4. It is important to note, however, that the time series pattern of the model parameters in Figures 3 and 4 has no implication for how each of the models identifies private information.

We also estimate the parameter vectors $\Theta_{PIN,i}$, $\Theta_{DY,i}$, and $\Theta_{OWR,i}$ in the period $t \in [-312, -60]$ before a merger announcement. These parameter estimates are used to compute the $CPIE$s in Section 5. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 2 and in Figures 3 and 4.

4 How does each model identify private information?

In theory, the PIN and DY models identify information events from changes in the absolute net order flow, while the OWR model identifies such events from covariation in net order flow (y_e) and overnight and intraday returns ($r_{o,i,t}$ and $r_{d,i,t}$). In practice, however, the models may produce poor descriptions of the data, and model misspecification can affect the way they actually identify private information.

To analyze how each model identifies private information in practice, we compare results from data created by simulating the models to results from real data. To create the simulated data, we first estimate the parameters of each model for each firm-year in our sample. Then, for each firm-year, we generate 1,000 artificial firm-years' worth of data (i.e. $B_{i,t}$ and $S_{i,t}$ for the PIN and DY models; $r_{o,i,t}$, $r_{d,i,t}$ and $y_{e,i,t}$ for the OWR model) using the estimated parameters. We then compute the $CPIE_{PIN,i,t}$, $CPIE_{DY,i,t}$, and $CPIE_{OWR,i,t}$ for each trading day in a simulated trading year and regress these $CPIE$s on the variables that should, in theory, identify information events in each model. The results of the regressions using simulated data reveal how each model should perform if the data conform to the model. The simulated data regressions also allow us to build empirical distributions of the R^2s of the regressions of $CPIE$s on the variables that identify information events in each model. For each model, we use the empirical distribution of the R^2s to test the null hypothesis that the real data conform to the model.

4.1 How does the PIN model identify private information?

Panel A of Table 3 presents the results of yearly multivariate regressions of $CPIE_{PIN}$ on absolute order flow imbalance ($|B - S|$) and $|B - S|^2$. We add squared terms to these regressions to account for nonlinearities in the relation between $CPIE_{PIN}$ and $|B - S|$. We average the simulated results for each `PERMNO`-Year and report in Panel A of Table 3 the median coefficient estimates and $t-$stats. The coefficients are standardized so they represent the increase in $CPIE_{PIN}$ due to a one standard deviation increase in the corresponding independent variable. We also report the average of the median, the 5^{th}, and the 95^{th} percentiles of the empirical distribution of R^2s of these regressions generated by the 1,000 simulations. In general, the coefficients are highly statistically significant, and the R^2s are high, confirming the theoretical implication that absolute order imbalance can be used to infer the arrival of private information under the PIN model.

The columns of Table 3 labeled as '$R^2_{inc.}$' include statistics on the increase in the R^2 that is due to the inclusion of turnover ($turn$) and turnover squared ($turn^2$) in these regressions. $R^2_{inc.}$ is equal to the difference between the R^2 of the extended regression model with turnover terms and the R^2 of the regression with only order imbalance terms. We report the average of the median, the 5^{th}, and the 95^{th} percentiles of the R^2_{inc}s of these regressions across the 1,000 simulations. These incremental R^2s are relatively low, with an average value of around 10%, which implies that, under the data generating process of the model, turnover has only modest incremental power in explaining $CPIE_{PIN}$ once controlling for absolute order imbalance and its square.

The picture that emerges from these regressions is that if the PIN model were a perfectly accurate representation of trading activity, $CPIE_{PIN}$ is determined by the order flow imbalance on each day. Since the order flow imbalance and turnover on any given day are highly positively correlated, turnover adds little to explain $CPIE_{PIN}$ in theory.

Panel B of Table 3 reports regression results for the real rather than simulated data. With the real data, the picture is very different. The R^2s of the regressions of $CPIE_{PIN}$ on $|B-S|$ and $|B-S|^2$ are much smaller than those in the simulations. We test the hypothesis that the real data R^2s and $R^2_{inc.}$s are consistent with those generated by the PIN model. Panel B reports the average across all stocks' p-value (the probability of observing an R^2 in the simulations at least as small as what we observe in the data), and the frequency that we reject the null at the 5% level implied by the distribution of simulated R^2s. The PIN model is rejected in about 98% of the stock-years in our sample, and there is on average less than a 1% chance of the PIN model generating R^2s as low as what we see in the data. On the other hand, the incremental R^2s of turnover are much higher than those in the Panel A. The incremental R^2 increases over time with a value of about 36% in 1993, but nearly 46% in 2012. This implies that turnover and turnover squared explain a much larger degree of variation in $CPIE_{PIN}$ than order imbalance. In fact, the average ratio of the median R^2s, $R^2/(R^2+R^2_{inc})$, is about 65%. The difference arises because, in the real data, absolute order flow and turnover are only weakly correlated. For instance, large absolute order flow imbalances are possible when turnover is below average, and vice versa. Under the PIN model, however, the two are highly correlated.

4.2 How does the DY model identify private information?

Table 4 presents a set of regressions for the DY model similar to those in Table 3. The dependent variable in Table 4 is $CPIE_{DY}$. The right-hand side variables are the absolute order imbalance adjusted for buy/sell correlations (|adj. OIB|), turnover and their squared terms. We define the adjusted absolute order imbalance as the absolute value of the residual from a regression of buys on sells. We use this measure to analyze the DY model because, as Fig. 2 suggests, the DY model implies that days with information events are far from the dashed line in this figure.[12] Turnover, as before, is defined as the sum of buys and sells. We report median coefficient estimates and $t-$stats across all firms within a particular year. The coefficients are standardized as above. We report the average of the median, the 5^{th}, and the 95^{th} percentiles of the R^2s and R^2_{inc}s.

As with the $CPIE_{PIN}$, in theory, turnover has little additional power in explaining $CPIE_{DY}$. The incremental R^2s in Table 4 Panel A are low with an average value close to 4%. This is smaller than the average incremental R^2s of the PIN model in Panel A of Table 3. The intuition for this result is that the DY model disentangles turnover and order flow shocks by including the possibility of symmetric order flow shocks. Buying and selling activity can simultaneously be higher than average, but this is not indicative of private information unless there is a large order flow imbalance.

Panel B of Table 4 reports regression results for the real, rather than simulated, data. The DY model behaves very differently when using real data as opposed to data generated from the model. The R^2s for the real data are much lower than those in the simulated data, declining from 35% in 1993 to 12% in 2012. The p-values (frequency of rejection) also decreases (increases) over time. For example, in 1993, our hypothesis test based on R^2 rejects the model at 5% significance for 81% of the stocks, while in 2012 this percentage increases to around 95%. The incremental R^2 indicate that turnover and turnover squared explain a large degree of variation in $CPIE_{DY}$. Indeed, the average ratio of the median R^2s, $R^2/(R^2 + R^2_{inc})$, is about 40%.

4.3 How does the OWR model identify private information?

As we saw in Section 1, the OWR model identifies private information from the covariance matrix of the three variables in the model $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$. Therefore, to analyze how the OWR model identifies private information, we run the regression of $CPIE_{OWR}$ on the

[12] Our results are qualitatively similar if we use absolute order imbalance instead of adjusted absolute order imbalance.

squared and interaction terms of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$:

$$CPIE_{OWR,i,t} = \beta_{0,1}+\beta_{1,1}y^2_{e,i,t}+\beta_{1,2}r^2_{d,i,t}+\beta_{1,3}r^2_{o,i,t}+\beta_{1,4}y_{e,i,t}r_{d,i,t}+\beta_{1,5}y_{e.i,t}r_{o,i,t}+\beta_{1,6}r_{d,i,t}r_{o,i,t}+u_{i,t}. \tag{7}$$

Panel A of Table 5 presents median coefficient estimates, t-stats, and three percentiles of R^2s across all firms within a particular year using simulated data. The results highlight the intuition behind the model. The probability of an information event on any given day is increasing in the square of intraday returns, the interaction between imbalance and intraday (or overnight) returns, and the interaction between intraday and overnight returns. The coefficient estimates on the square of the order imbalance and on the square of overnight returns are too small to be precisely measured. The high R^2s indicate that, practically speaking, the square of intraday returns, the interaction between intraday and overnight returns and the interaction between intraday returns and order flow imbalance are sufficient to explain a large part of the variation in $CPIE_{OWR}$.

Panel B of Table 5 shows the median coefficient estimates, t-stats, and the results of the hypothesis tests based on R^2s across all firms within a particular year using real data. Unlike the PIN and DY models, the coefficient estimates are consistent across the simulated and real data. For instance in simulated data regressions in Panel A, 2008 is the only year in which y_e^2 is the most important term. In the real data regressions in Panel B, 2008 is also the only year in which y_e^2 is the most important term, indicating that the model matches the features of the data quite well, even for clear outliers like 2008. Furthermore, as with the simulated data regressions, the high median R^2s indicate that a large part of the variation in $CPIE_{OWR}$ is explained by the squared and interaction terms of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$ as implied by the model. The average across years of the R^2s in Panel B is about 83% and these R^2s increase over time, reaching 90% in 2012. Moreover, we reject the null hypothesis that the R^2s observed in the real data are consistent with the OWR model at 5% level for about 40% of the sample in 1993 and for about 8% of the sample in 2012.

The high R^2s in Panel B imply that, in principle, any variable unrelated to private information under the OWR model has only a small incremental value in explaining the $CPIE_{OWR}$. To see this note that the typical R^2 in Panel B is around 85%. This suggests that any additional regressor, even if it explained 100% of the residual variation in the regressions in Panel B, could only marginally improve the R^2 from 85% to 100%. Note that in the case of the PIN and DY models, our results show that turnover, which in principle is a poor measure of private information, largely drives the PIN and DY models' identification of private information. In contrast, under the OWR model the variables related to private information in the model (squares and interactions of y_e, r_o, and r_d) can explain a fairly large amount of the variation in $CPIE_{OWR}$. As a result, any variable that is not related to private information in the OWR model can only explain a relatively small fraction of the variation in $CPIE_{OWR}$.

4.4 Discussion of results

The results in Table 3 indicate that the PIN model frequently identifies private information in periods of increased volume, as opposed to periods of large order imbalances. This is especially true in the later portion of the sample, as order imbalance becomes less important in explaining $CPIE_{PIN}$ over time. The results in Table 4 indicates that the DY model also identifies private information as periods of high volume, particularly later in the sample.

To illustrate why both the DY and the PIN models fail to describe the order flow data, Fig. 5, Panels A and B plot the simulated and real PIN and DY data for Exxon in 2012. Note that the PIN model's three clusters do not fit the majority of the data. Unlike Fig. 2, the PIN model in Fig. 5 essentially classifies days with above average turnover as being private information days (i.e. $CPIE_{PIN}$ equal to one) and days with below average turnover as days without private information (i.e. $CPIE_{PIN}$ equal to zero) with no intermediate values. This occurs because these data points are extreme outliers relative to what the model expects (i.e. the points represented by the simulated data). Panel B shows that although the DY model recognizes that some high turnover days may not be due to private information events, it still classifies the days with the most extreme turnover as information days. Essentially, the DY model implies that, later in the sample, $CPIE$ is a non-linear function of turnover. Thus, the R^2s in Panel B of Table 4 drop over time. As with the PIN model, the majority of the data are extreme outliers compared to what the DY model expects (i.e. the points representing the simulated data).

Although Fig. 5 is a highly stylized example of the PIN and DY models' failure to describe the data, the problem is widespread and made more severe by the increase in volume observed in the sample. To quantify how often the PIN model fails to fit the data, Panel A of Fig. 6 shows the fraction of days for the median stock-year which the PIN model classifies as "outliers" (likelihoods smaller than 10^{-10}). According to the PIN model, for the median stock about 60% (90%) of the annual observations are classified as outliers in 2005 (2010). Panel B shows the results related to the DY model. According to the DY model, for the median stock about 40% (90%) of the annual observations are classified as outliers in 2005 (2010). Essentially, both the PIN and DY models are very poor descriptions of the order flow data in the most recent years in the sample. The intuition for this is simple, the PIN and DY models assume that the order flow is distributed as a mixture of Poisson random variables. The mean and the variance of a Poisson random variable are equal and as a consequence the Poisson mixtures behind the PIN and the DY model cannot accommodate the high level and volatility of turnover that we observe in the later part of the sample.

Fig. 5 also emphasizes the mechanical nature of the relation between $CPIE_{PIN}$ and turnover. The PIN model essentially classifies all days with higher than average turnover as being days with private information events. Note that this classification does not necessarily relate to the possibility, suggested by Collin-Dufresne and Fos (2014b), that informed traders sometimes choose to trade in days with high liquidity or turnover. Naturally, it is possible that informed traders do in fact trade in some days with high volume. However, the PIN model classifies *all* high volume days as information events. This happens because the relation between $CPIE_{PIN}$ and turnover mechanically results from the model's inability to match both the high level and volatility of the order flow data.

Fig. 5 also gives the intuition for why the median PIN α increases over time in Fig. 3. To see this, recall that α is the unconditional expected value of $CPIE$. Therefore, as we observe more $CPIE_{PIN}$ values approaching one, the estimated PIN α must increase. In fact, the median PIN α becomes close to 50% later in the sample which consistent with the fact that the PIN model assigns a $CPIE_{PIN}$ equal to one (zero) to days with turnover above (below) the average. The same happens to a lesser extent with the DY model, but not with the OWR model.

O'Hara, Yao, and Ye (2014) find that high-frequency trading is associated with an increase in the use of odd lot trades, which do not appear in the TAQ database. Therefore, estimates of the PIN and DY model parameters computed using recent TAQ data may be systematically biased. More broadly, Fig. 6 indicates that even if the DY and PIN models are estimated using a dataset that includes odd lot trades, both models will still be badly misspecified late in the sample.

The OWR model does not suffer from the limitations faced by the PIN and DY models. The OWR model does not rely on assumptions about the number of buys and sells, instead it relies on assumptions about the order imbalance itself (y_e). As a result, the OWR model is not affected by the observed increase in the level of both buys and sells. Moreover, when we compare Panels A and B of Table 5, the OWR model behaves similarly when using both the simulated and the real data, indicating that the variations in $CPIE_{OWR}$ in the real data are consistent with those implied by the model.

Given the strong connection between $CPIE$s and the unconditional probability of information arrival, (α) these results call into question the use of PIN and *Adj.* PIN as proxies for private information. While there are other parameters in the models (i.e. μ, ϵ_b and ϵ_s), these parameters are jointly identified with α. Hence it seems extremely unlikely that in the joint identification of the model parameters, biases in the other parameters 'correct' the biases in α in such a way that PIN and *Adj.* PIN are 'rescued' as reasonable proxies of private information. Thus, while our $CPIE$ results do not speak directly to μ, ϵ_b and ϵ_s, they still call into question PIN and $Adj.\ PIN$ as a measures of private information. Moreover, it seems unlikely that either PIN or $Adj.\ PIN$ can possibly measure private information later in the sample if the models in which they are based cannot even account for the existence 90% of the daily order flow observations for the median stock.

5 Event study evidence

This section examines how well the PIN, DY, and OWR models identify information events around announcements of targets in merger and acquisition (M&A) transactions. Unlike a standard event study, we focus on movements in $CPIE$ rather than price movements. For each model, we examine the period $t \in [-30, 30]$ around the event. To do so, we estimate the parameter vectors $\Theta_{PIN,i}$, $\Theta_{DY,i}$, and $\Theta_{OWR,i}$ in the period $t \in [-312, -60]$ before the event and then compute the daily $CPIE$s for the period $t \in [-30, 30]$ surrounding the announcement. Prior studies such as Aktas, de Bodt, Declerck, and Van Oppens (2007) and Vega (2006) estimate the parameters of the model in various windows around an event in order to compute the PIN. Our procedure is different in that we estimate the parameters of the model one year prior to the event and then employ the estimated parameters as if we were market makers observing the market data (i.e. buys and sells in the PIN and DY models; order flow, overnight and intraday returns in the OWR model) and attempting to infer whether an information event occurred. Table 1 Panel B presents summary statistics for order imbalance, intraday returns, overnight returns, turnover, number of buys, and the number of sells for merger announcement days ($t = 0$). Section 3 describes the maximum likelihood estimation.

5.1 Information event probabilities under the PIN model

Panel A of Fig. 7 shows the average $CPIE_{PIN}$ in event time for our sample of merger targets. The graph shows that, under the PIN model, the probability of an information event increases prior to the event, starting at around 55% 23 days before the announcement and peaking around 79% on the day of the announcement. The rise in the probability of an information event prior to the announcement is consistent with a world where informed traders generate signals about the merger and trade on this information before the merger is announced to the public. $CPIE_{PIN}$ is also higher *after* the identity of the merger target becomes public information. In fact, $CPIE_{PIN}$ remains above the average $CPIE_{PIN}$ observed in the gap period, $[-60, -30]$, for 30 days after the announcement.

Panels B and C of Fig. 7 shed light on the features of the data that produce the observed pattern in the average $CPIE_{PIN}$ in Panel A. Panel B shows the average predictions from OLS regressions of $CPIE_{PIN}$ on order imbalance and absolute order imbalance squared across all of the stocks in the event study sample.[13] The solid line indicates that order imbalance explains only a small fraction of the movement in $CPIE_{PIN}$ during the event window. Panel C shows the average predictions from regressions of $CPIE_{PIN}$ on turnover and turnover squared. The solid line indicates that the variation in $CPIE_{PIN}$ around merger target announcements is explained almost entirely by turnover. The intuition follows directly from the results in Section 4.1, which illustrates that $CPIE_{PIN}$ is extremely sensitive to volume increases. The higher post-event volume levels are enough to keep $CPIE_{PIN}$ above its pre-event mean for a substantial period.

To formalize the intuition behind the figures in Panels B and C of Fig. 7, we run regressions similar to those in Table 3 using our event sample. Specifically, we run regressions of $CPIE_{PIN}$ on absolute value of order imbalance and its squared term during the event window [-30,+30]. The results of these regressions (see Table 6) indicate that absolute order imbalance explains little of the variation in $CPIE_{PIN}$ in the event window while turnover explains most of the variation in $CPIE_{PIN}$. In fact, Panel A of Table 6 shows that for the median stock, adding turnover and turnover squared to these regressions triples the R^2s.

[13] We restrict the predicted $CPIE$ to fall on the interval [0,1].

5.2 Information event probabilities under the DY model

Panel A of Fig. 8 shows the average $CPIE_{DY}$ in event time for our sample of merger announcements. Panels B of Fig. 8 shows the average $CPIE_{DY}$ along with the average predicted $CPIE_{DY}$ with |adj. OIB| and |adj. OIB| squared. Panels C of Fig. 8 shows the average $CPIE_{DY}$ along with the average predicted $CPIE_{DY}$ with turnover and turnover squared. Like $CPIE_{PIN}$, $CPIE_{DY}$ increases from around 55% starting 23 days prior to the announcement and peaks at 73% on the day of the announcement. Also similar to $CPIE_{PIN}$, $CPIE_{DY}$ remains high after the announcement. The reason for this post-event increase mirrors what we observed with the PIN model. Thus, the higher post-event volume levels are enough to keep $CPIE_{DY}$ above its pre-event mean for a substantial period.

As with the PIN model, to formalize the intuition behind the figures in Panels B and C of Fig. 8, we run regressions similar to those in Table 4 using our event sample. Specifically, we run regressions of $CPIE_{DY}$ on the absolute value of adjusted order imbalance and its squared term during the event window [-30,+30]. The results of these regressions (see Table 6) indicate that turnover explains most of the variation in $CPIE_{DY}$. In fact, Panel B of Table 6 shows that for the median stock, the R^2s in these regressions almost double due to turnover and turnover squared.

5.3 Information event probabilities under the OWR model

Panel A of Fig. 9 illustrates the average $CPIE_{OWR}$ in event time for our sample of merger announcements. Similar to the PIN model, the probability of an information event increases from around 45% 6 days before the announcement and peaks on the announcement date at around 49%. In fact, the $CPIE_{OWR}$ is more than two standard deviations from its mean (estimated between $t \in [-60, -29]$) two trading days before the announcement. As with the PIN and DY model results, this pattern is consistent with informed traders acting on private information before the announcement. However, unlike the PIN and DY models, the $CPIE_{OWR}$ drops back to its pre-event mean within a day or two.

Consistent with the simulations and regressions in Table 5, Fig. 9 Panels B–G show that the majority of the variation in measured private information ($CPIE_{OWR}$) comes from intraday (Panel C) and overnight (Panel D) returns as well as the interaction between the two (Panel G). Order imbalance squared (Panel A) provides almost no explanatory power, although the interaction between the order imbalance and returns (Panels E and F) does have some impact.

Panel C of Table 6 presents the results from regressions of $CPIE_{OWR}$ on the squared and interaction terms of order imbalance, intraday, and overnight returns (y_e, r_d, r_o) during the event window [-30,+30]. Panel C of Table 6 shows that for the median stock, the R^2 is 81%. This implies that, as the model suggests, a large proportion of the variation in $CPIE_{OWR}$ is due to the squared and interaction terms of order imbalance, intraday, and overnight returns.

5.4 Discussion of event study results

The event study results suggest that the variation in PIN around events documented in the literature could be due to variation in α that is driven primarily by volume, rather than order imbalance. For instance, Benos and Jochec (2007) show that PIN increases after earnings announcements, while Aktas, de Bodt, Declerck, and Van Oppens (2007) show that PIN increases after M&A target announcements due to increases in both μ and α. Therefore, our evidence suggests that these PIN results are at least partially explained by the fact that the PIN model erroneously attributes increases in volume to private information. Note that this conclusion does not depend on whether private information is higher after announcements or not because, as we show in Panel A of Table 3, under the PIN model, turnover should have little incremental power in identifying the arrival of private information once we control for absolute order imbalance.

Another important implication of our results is that event studies based on measures of private information (e.g. Easley, Engle, O'Hara, and Wu (2008) and Brennan, Huh, and Subrahmanyam (2015)) can be misleading. For instance, it may appear at first glance that the results in Panel A of Fig. 7 suggest that the PIN model identifies private information in a sensible way since $CPIE_{PIN}$ increases dramatically from 55% before the announcement to over 70% on the day of the announcement then falls after the announcement, albeit over a period of weeks. However, the decomposition of the $CPIE$s in Panels B and C of Fig. 7 points to a different interpretation, namely that the dramatic increase in $CPIE$ around the event is actually result of variation in turnover, not order imbalance. Indeed, our results indicate that event study plots of the PIN measure or the $CPIE$ that do not distinguish between variation in these measures related to order imbalance and variation due to turnover can lead to misleading conclusions about the model's ability to properly identify private information. Naturally, turnover may increase due to private information. However, turnover around announcements is also affected by other factors such as portfolio rebalancing and disagreement. Moreover, the fact that $CPIE_{PIN}$ varies with turnover and not absolute order imbalance is not consistent with the PIN model.

The DY model is also prone to identify information events from variations in turnover rather than order imbalance. On the other hand, the OWR model, which uses net order flow to identify private information, exhibits no such tendency.

Under the working hypothesis that there is more informed trade before rather than after merger announcements, our findings suggest that the OWR model identifies private information in a sensible way.[14] Even though the magnitude of the increase in $CPIE_{OWR}$ around the event date is small, it results from variation in the variables the model suggests should be important in identifying private information. This stands in contrast to the results for the PIN and DY models. The small increase in $CPIE_{OWR}$ may simply be indicative of the difficulty of detecting the arrival of private information, even when using variables other than just order imbalance. Moreover, the fact that order imbalance alone explains very little of the variation in $CPIE_{OWR}$ around merger announcements suggests that order flow, however well

[14] There is evidence suggesting that it is unlikely that there is more informed trading activity in the weeks after the news is made public. Indeed, studies such as Meulbroek (1992) show that in 80% of insider trading cases involving mergers, the insiders acted before the announcement of the merger.

modeled, is insufficient to be the sole source of inferences about private information arrival. This result provides empirical support for the proposition in Back, Crotty, and Li (2014) and in Kim and Stoll (2014) that researchers cannot use order flow alone to successfully identify periods of informed trade.

6 Conclusion

The PIN measure, developed in the seminal work of Easley and O'Hara (1987), Easley, Kiefer, O'Hara, and Paperman (1996), and Easley, Kiefer, and O'Hara (1997), is arguably the most widely used measure of information asymmetry in the accounting, corporate finance and asset pricing literature today. Recent work however suggests that PIN fails to capture private information (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), and Collin-Dufresne and Fos (2014a)). This paper analyzes why the model might incorrectly identify informed trade. We perform this analysis by comparing the PIN model with the DY and OWR models. By doing so, we suggest some important insights for future research using, constructing, or testing proxies for informed trade.

We find that both the PIN and DY models tend to identify information events from total volume or turnover, rather than from order flow imbalance. This failure in both models is particularly strong after the increase in volume in the middle of the first decade of the 2000's. For example, after 2006, for the median stock in our sample, the PIN and DY models suggest that about 90% of observed daily order flows have a near-zero probability of occurring. It seems unlikely that models that cannot even account for the existence of 90% of the order flow observations of the median stock can possibly detect private information. The OWR model, on the other hand, does not suffer from this problem, because it relies on assumptions about order flow imbalance rather than about the distribution of the number of buys and sells. These findings suggest that future research concerned with constructing private information measures should focus on modeling order flow imbalance directly instead of modeling the number of buys and sells as the PIN and DY models do.

Our event study results, which do not depend on whether private information arrives before or after the event, suggest that at least part of the variation in PIN around events documented in Aktas, de Bodt, Declerck, and Van Oppens (2007) and Benos and Jochec (2007) are explained by the fact that the PIN model erroneously attributes increases in volume, unrelated to order imbalance, to private information. In addition, we find similar results for the DY model. These findings suggest that future research concerned with event study based tests of private information proxies (e.g. Easley, Engle, O'Hara, and Wu (2008) and Brennan, Huh, and Subrahmanyam (2015)) can be misleading if one overlooks the fact that the pattern in the tested proxies can result from changes in variables that are in principle unrelated to private information. For instance, Brennan, Huh, and Subrahmanyam (2015) interpret the fact that their $CPIE_{PIN}$ measures are higher after earnings announcements than before as evidence of informed trading. We show that $CPIE_{PIN}$ is closely related to volume (controlling for order flow imbalance), therefore their findings may be due to the fact that turnover is higher after earnings announcements.

Under the working hypothesis that there is more informed trade before rather than after merger announcements, our findings suggest that the OWR model identifies private information in a sensible way. Moreover, the fact that order imbalance alone explains very little of the variation in $CPIE_{OWR}$ around merger announcements suggests that order flow, however well modeled, is insufficient to be the sole source of inferences about private information arrival. This result provides empirical support for the proposition in Back, Crotty, and Li (2014) and in Kim and Stoll (2014) that researchers cannot use net order flow alone to successfully identify periods of informed trade. This suggests that future research aimed at building measures of informed trade should focus on the use of variables other than simply net order flow alone.

Our findings also suggest that future research in corporate finance, accounting or asset pricing that uses information asymmetry measures should consider using the OWR α or

$CPIE_{OWR}$ as a measure of private information instead of the PIN or $Adj.\ PIN$.

References

Akins, Brian K., Jeffrey Ng, and Rodrigo S. Verdi, 2012, Investor competition over information and the pricing of information asymmetry, *The Accounting Review* 87, 35–58.

Aktas, Nihat, Eric de Bodt, Fany Declerck, and Herve Van Oppens, 2007, The PIN anomaly around M & A announcements, *Journal of Financial Markets* 10, 169–191.

Andersen, Torben G., and Oleg Bondarenko, 2014, VPIN and the flash crash, *Journal of Financial Markets* 17, 1–46.

Back, Kerry, Kevin Crotty, and Tao Li, 2014, Can information asymmetry be identified from order flows alone?, *Working paper.*

Bakke, Tor-Erik, and Toni. M. Whited, 2010, Which firms follow the market? An analysis of corporate investment decisions, *The Review of Financial Studies* 23, 1941–1980.

Banerjee, Snehal, and Ilan Kremer, 2010, Disagreement and learning: Dynamic patterns of trade, *The Journal of Finance* 65, 1269–1302.

Benos, Evangelos, and Marek Jochec, 2007, Testing the PIN variable, *Working paper.*

Brennan, Michael J., Sahn-Wook Huh, and Avanidhar Subrahmanyam, 2015, High-frequency measures of information risk, *Working paper.*

Chae, Joon, 2005, Trading volume, information asymmetry, and timing information, *The Journal of Finance* 60, 413–442.

Collin-Dufresne, Pierre, and Vyacheslav Fos, 2014a, Do prices reveal the presence of informed trading?, *Journal of Finance* Forthcoming.

——— , 2014b, Insider trading, stochastic liquidity and equilibrium prices, *National Bureau of Economic Research Working paper.*

Da, Zhi, Pengjie Gao, and Ravi Jagannathan, 2011, Impatient trading, liquidity provision, and stock selection by mutual funds, *The Review of Financial Studies* 324, 675–720.

Duarte, Jefferson, Xi Han, Jarrod Harford, and Lance A. Young, 2008, Information asymmetry, information dissemination and the effect of regulation FD on the cost of capital, *Journal of Financial Economics* 87, 24–44.

Duarte, Jefferson, and Lance Young, 2009, Why is PIN priced?, *Journal of Financial Economics* 91, 119–138.

Easley, David, Robert F. Engle, Maureen O'Hara, and Liuren Wu, 2008, Time-varying arrival rates of informed and uninformed trades, *Journal of Financial Econometrics* pp. 171–207.

Easley, David, Soeren S. Hvidkjaer, and Maureen O'Hara, 2002, Is information risk a determinant of asset returns?, *Journal of Finance* 57, 2185–2221.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1997, One day in the life of a very common stock, *Review of Financial Studies* 10, 805–835.

——— , and Joseph B. Paperman, 1996, Liquidity, information, and infrequently traded stocks, *Journal of Finance* 51, 1405–1436.

Easley, David, Marcos Lopez de Prado, and Maureen O'Hara, 2012, Flow toxicity and liquidity in a high-frequency world, *Review of Financial Studies* 25, 1457–1493.

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

Glosten, Lawrence R., and Paul R. Milgrom, 1985, Bid, ask and transaction prices in a specialist market with heterogeneously informed traders, *Journal of Financial Economics* 13, 71–100.

Hendershott, Terrence, Dmitry Livdan, and Norman Schurhoff, 2014, Are institutions informed about news?, *Swiss Finance Institute Research Paper*.

Kandel, Eugene, and Neil D. Pearson, 1995, Differential interpretation of public signals and trade in speculative markets, *Journal of Political Economy* 103, 831–872.

Kim, Sukwon Thomas, and Hans R. Stoll, 2014, Are trading imbalances indicative of private information?, *Journal of Financial Markets* 20, 151–174.

Kyle, Albert S., 1985, Continuous auctions and insider trading, *Econometrica* 53, 1315–1335.

Lai, Sandy, Lilian Ng, and Bohui Zhang, 2014, Does PIN affect equity prices around the world?, *Journal of Financial Economics* 114, 178–195.

Lambert, Richard A., and Christian Leuz, 2012, Information precision, information asymmetry, and the cost of capital, *Review of Finance* 16, 1–29.

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

Meulbroek, Lisa K., 1992, An empirical analysis of illegal insider trading, *Journal of Finance* 47, 1661–1699.

Mohanram, Partha, and Shiva Rajgopal, 2009, Is PIN priced risk?, *Journal of Accounting and Economics* 47, 226–243.

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

O'Hara, Maureen, Chen Yao, and Mao Ye, 2014, What's not there: Odd lots and market data, *Journal of Finance* 69, 2199–2236.

Vega, Clara, 2006, Stock price reaction to public and private information, *Journal of Financial Economics* 82, 103–133.

Table 1: **Summary Statistics.** This table summarizes the full sample and event day (t=0) returns, order imbalance, number of buys and sells, and turnover. We compute intraday and overnight returns as well as daily buys and sells for stocks between 1993 and 2012 using data from the NYSE TAQ database. Following Odders-White and Ready (2008), we compute the intraday return, r_d, at time t as the volume-weighted average price at t (VWAP) minus the opening quote midpoint at t plus dividends at time t, all divided by the opening quote midpoint at time t. We compute the overnight return, r_o, at t as the opening quote midpoint at $t+1$ minus the VWAP at t, all divided by the opening quote midpoint at t. The total return, or sum of the intraday and overnight returns, is the open-to-open return from t to $t+1$. We compute y_e as the daily total number of buys minus total number of sells, divided by the total number of trades. For the PIN and DY models, we use the daily total number of buys and sells. We compute turnover as the number of buys plus sells divided by shares outstanding. We collect M&A dates from Thomson Reuter's SDC database. If the event occurs on a non-trading day, then we use the next available trading day as the event day. We require that target firms have a CRSP `PERMNO`, which we get by matching 6-digit CUSIPs from CRSP and SDC.

(a) Full Sample

	N	Mean	Std	Q1	Median	Q3
y_e	5,286,191	2.766%	31.259%	-10.433%	3.282%	18.996%
r_d	5,286,191	-0.004%	1.500%	-0.707%	-0.024%	0.680%
r_o	5,286,191	0.003%	1.297%	-0.566%	-0.024%	0.525%
turn	5,286,191	2.096%	4.178%	0.159%	0.529%	2.407%
# Buys	5,286,191	1,876	6,917	37	220	1,128
# Sells	5,286,191	1,843	6,894	36	194	1,033

(b) Targets

	N	Mean	Std	Q1	Median	Q3
y_e	5,850	3.140%	28.295%	-9.706%	1.618%	16.082%
r_d	5,859	0.381%	2.074%	-0.657%	0.234%	1.273%
r_o	5,859	0.497%	2.048%	-0.507%	0.250%	1.243%
turn	5,859	4.243%	10.885%	0.212%	0.673%	3.690%
# Buys	5,859	4,311	20,232	57	323	2,294
# Sells	5,859	4,287	20,315	52	287	2,149

Table 2: **Parameter Estimates.** This table summarizes parameter estimates of the PIN, DY, and OWR models for 21,206 PERMNO-Year samples from 1993 to 2012. In all three models, α represents the average unconditional probability of an information event at the daily level. For the PIN and DY models, ϵ_B and ϵ_S represent the expected number of daily buys and sells given no private information or symmetric order flow shocks. μ, μ_b, and μ_s represent the expected additional order flows given an information event, which is good news with probability δ and bad news with probability $1-\delta$. In the DY model, a symmetric order flow shock occurs with probability θ, in which case the expected number of buys and sells increase by Δ_B and Δ_S, respectively. In the OWR model, σ_u represents the standard deviation of the order imbalance due to uninformed traders, which is observed with normally distributed noise with variance σ_z^2. σ_i represents the standard deviation of the informed trader's private signal. σ_{pd} and σ_{po} represent the standard deviation of intraday and overnight returns, respectively. $\overline{CPIE}$ and Std($CPIE$) are the PERMNO-Year mean and standard deviation of $CPIE$.

(a) PIN

	N	Mean	Std	Q1	Median	Q3
α	21,206	0.372	0.122	0.291	0.375	0.445
δ	21,206	0.607	0.209	0.484	0.625	0.762
ϵ_b	21,206	1,625	5,388	33	193	1,039
ϵ_s	21,206	1,596	5,369	35	186	956
μ	21,206	312	593	43	160	314
$\overline{CPIE}$	21,206	0.382	0.135	0.293	0.379	0.449
Std($CPIE$)	21,206	0.451	0.052	0.427	0.470	0.490

(b) DY

	N	Mean	Std	Q1	Median	Q3
α	21,206	0.456	0.092	0.409	0.464	0.509
δ	21,206	0.550	0.192	0.441	0.541	0.680
θ	21,206	0.249	0.137	0.149	0.253	0.344
ϵ_b	21,206	1,418	4,571	26	158	866
ϵ_s	21,206	1,397	4,570	28	148	807
Δ_b	21,206	2,148	10,058	41	190	989
Δ_s	21,206	2,097	9,934	34	160	908
μ_b	21,206	290	575	29	119	310
μ_s	21,206	284	574	27	107	302
$\overline{CPIE}$	21,206	0.455	0.092	0.409	0.461	0.506
Std($CPIE$)	21,206	0.454	0.056	0.431	0.479	0.493

(c) OWR

	N	Mean	Std	Q1	Median	Q3
α	21,206	0.437	0.257	0.214	0.436	0.639
σ_u	21,206	0.075	0.068	0.022	0.062	0.109
σ_z	21,206	0.239	0.143	0.137	0.221	0.332
σ_i	21,206	0.030	0.286	0.013	0.021	0.027
σ_{pd}	21,206	0.010	0.005	0.006	0.009	0.012
σ_{po}	21,206	0.006	0.004	0.004	0.006	0.008
$\overline{CPIE}$	21,206	0.451	0.258	0.227	0.455	0.656
Std($CPIE$)	21,206	0.137	0.047	0.109	0.142	0.171

Table 3: **PIN Model Regressions.** This table reports real and simulated regressions of the $CPIE_{PIN}$ on absolute order imbalance ($|B - S|$), and order imbalance squared ($|B - S|^2$). In Panel A, we simulate 1,000 instances of the PIN model for each PERMNO-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the R^2 attributed to *turn* and $turn^2$ in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 values, and tests of the hypothesis that the observed variation in $CPIE_{PIN}$ is consistent with the PIN model. The p-value of R^2 ($R^2_{inc.}$) is the probability of observing an R^2 at least as small (large) as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

	β		t		R^2			$R^2_{inc.}$		
	$\|B - S\|$	$\|B - S\|^2$	$\|B - S\|$	$\|B - S\|^2$	5%	50%	95%	5%	50%	95%
1993	0.437	-0.079	(10.31)	(-1.80)	71.13%	76.09%	80.38%	7.17%	10.57%	15.25%
1994	0.422	-0.072	(9.63)	(-1.67)	67.49%	73.26%	78.11%	9.39%	13.47%	18.55%
1995	0.410	-0.058	(9.68)	(-1.36)	70.32%	75.39%	79.85%	7.64%	11.39%	16.02%
1996	0.432	-0.085	(9.89)	(-1.90)	69.02%	74.28%	78.87%	8.32%	12.17%	16.97%
1997	0.450	-0.089	(10.30)	(-1.98)	71.99%	76.93%	81.12%	7.36%	10.76%	14.79%
1998	0.482	-0.104	(10.79)	(-2.36)	74.32%	78.71%	82.46%	6.65%	9.53%	13.30%
1999	0.484	-0.112	(11.03)	(-2.47)	75.62%	79.96%	83.46%	6.49%	9.36%	12.92%
2000	0.529	-0.137	(11.88)	(-3.00)	79.78%	83.36%	86.15%	4.98%	7.47%	10.45%
2001	0.638	-0.217	(13.97)	(-4.61)	83.34%	86.13%	88.57%	4.17%	6.00%	8.35%
2002	0.695	-0.260	(14.11)	(-5.30)	82.61%	85.53%	88.06%	4.83%	6.92%	9.54%
2003	0.665	-0.244	(12.38)	(-4.52)	78.88%	82.36%	85.36%	7.90%	10.56%	13.79%
2004	0.650	-0.223	(11.49)	(-4.16)	77.84%	81.38%	84.59%	8.92%	11.67%	15.03%
2005	0.658	-0.220	(12.59)	(-4.46)	80.47%	83.59%	86.45%	7.69%	10.09%	12.95%
2006	0.650	-0.221	(11.96)	(-4.35)	80.31%	83.36%	86.18%	7.76%	10.29%	13.50%
2007	0.632	-0.222	(9.40)	(-4.07)	79.72%	83.35%	86.15%	8.53%	10.93%	14.05%
2008	0.666	-0.235	(12.29)	(-4.83)	82.44%	85.25%	88.00%	6.83%	9.15%	11.78%
2009	0.709	-0.269	(14.37)	(-5.70)	84.29%	86.87%	89.20%	6.22%	8.28%	10.57%
2010	0.704	-0.261	(14.60)	(-5.68)	84.99%	87.41%	89.64%	5.66%	7.55%	9.89%
2011	0.671	-0.234	(14.13)	(-5.21)	85.91%	88.25%	90.21%	5.34%	7.28%	9.39%
2012	0.693	-0.251	(14.92)	(-5.62)	85.68%	87.98%	90.34%	5.22%	7.22%	9.50%

Table 3: **PIN Model Regressions.** Continued.

(b) Real Data

	β		t		R^2			$R^2_{inc.}$		
	$\|B-S\|$	$\|B-S\|^2$	$\|B-S\|$	$\|B-S\|^2$	50%	p-value	% Rej.	50%	p-value	% Rej.
1993	0.300	-0.073	(5.98)	(-1.43)	35.76%	0.39%	98.81%	36.20%	2.57%	94.07%
1994	0.264	-0.047	(5.28)	(-0.92)	32.82%	0.39%	98.02%	40.02%	3.36%	92.17%
1995	0.280	-0.061	(5.77)	(-1.29)	34.20%	0.73%	96.98%	36.97%	5.05%	89.29%
1996	0.277	-0.065	(5.69)	(-1.28)	30.92%	0.51%	98.46%	38.97%	3.85%	92.30%
1997	0.283	-0.073	(5.67)	(-1.36)	30.80%	0.35%	99.05%	38.86%	3.54%	92.99%
1998	0.274	-0.059	(5.26)	(-1.09)	30.12%	0.24%	99.31%	39.58%	3.54%	93.67%
1999	0.280	-0.059	(5.21)	(-1.08)	29.05%	0.18%	99.38%	39.46%	3.29%	94.29%
2000	0.300	-0.079	(5.48)	(-1.39)	29.99%	0.17%	99.73%	39.08%	2.59%	95.63%
2001	0.339	-0.111	(5.67)	(-1.87)	29.44%	0.17%	99.71%	39.39%	3.53%	94.76%
2002	0.279	-0.058	(4.09)	(-0.85)	23.05%	0.10%	99.82%	44.28%	5.59%	91.48%
2003	0.247	-0.032	(3.57)	(-0.47)	21.97%	0.17%	99.73%	41.86%	9.55%	84.87%
2004	0.211	-0.005	(3.14)	(-0.08)	19.55%	0.00%	100.00%	45.22%	8.78%	86.21%
2005	0.254	-0.053	(3.81)	(-0.81)	19.42%	0.38%	99.46%	46.29%	9.21%	85.47%
2006	0.251	-0.066	(3.80)	(-0.96)	16.95%	1.26%	97.86%	48.44%	10.83%	85.30%
2007	0.271	-0.104	(4.01)	(-1.57)	14.30%	2.47%	95.62%	50.32%	14.04%	82.00%
2008	0.268	-0.111	(4.00)	(-1.66)	13.78%	1.93%	96.34%	50.97%	11.49%	86.08%
2009	0.280	-0.117	(4.15)	(-1.74)	14.59%	1.88%	96.79%	49.91%	10.08%	87.58%
2010	0.291	-0.124	(4.39)	(-1.82)	15.96%	1.96%	96.23%	47.64%	10.62%	87.45%
2011	0.295	-0.131	(4.56)	(-2.03)	15.94%	1.01%	97.34%	46.60%	11.14%	86.90%
2012	0.319	-0.145	(4.96)	(-2.23)	17.56%	3.35%	94.75%	45.61%	13.31%	85.12%

Table 4: **DY Model Regressions.** This table reports real and simulated regressions of the $CPIE_{DY}$ on adjusted order imbalance (|adj. OIB|), and adjusted order imbalance squared (|adj. OIB|^2. In Panel A, we simulate 1,000 instances of the DY model for each PERMNO-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the R^2 attributed to *turn* and $turn^2$ in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 values, and tests of the hypothesis that the observed variation in $CPIE_{DY}$ is consistent with the DY model. The p-value of R^2 ($R^2_{inc.}$) is the probability of observing an R^2 at least as small (large) as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

	β		t		R^2			$R^2_{inc.}$		
	|adj. OIB|	|adj. OIB|^2	|adj. OIB|	|adj. OIB|^2	5%	50%	95%	5%	50%	95%
1993	0.518	-0.230	(10.88)	(-4.74)	52.28%	59.44%	66.01%	5.55%	9.86%	15.29%
1994	0.484	-0.214	(10.47)	(-4.42)	50.66%	58.06%	64.97%	5.56%	9.46%	14.95%
1995	0.475	-0.214	(9.96)	(-4.32)	46.81%	54.46%	61.69%	7.01%	11.71%	17.54%
1996	0.516	-0.229	(10.54)	(-4.60)	51.36%	58.62%	65.21%	5.18%	9.09%	14.31%
1997	0.513	-0.221	(10.33)	(-4.40)	50.55%	57.80%	64.50%	4.78%	8.57%	14.03%
1998	0.537	-0.236	(10.60)	(-4.49)	52.85%	60.14%	66.63%	4.00%	7.45%	12.31%
1999	0.607	-0.281	(11.92)	(-5.45)	56.53%	63.49%	69.68%	3.07%	6.11%	10.47%
2000	0.597	-0.272	(11.43)	(-5.09)	55.69%	62.59%	69.09%	2.82%	5.65%	9.73%
2001	0.729	-0.350	(13.81)	(-6.75)	65.81%	71.48%	76.83%	0.62%	1.87%	4.09%
2002	0.769	-0.371	(15.03)	(-7.28)	71.90%	76.37%	80.55%	0.24%	1.04%	2.41%
2003	0.805	-0.394	(16.06)	(-7.99)	74.77%	78.95%	82.78%	0.34%	1.19%	2.71%
2004	0.798	-0.385	(15.94)	(-7.61)	77.39%	81.40%	84.70%	0.23%	0.95%	2.22%
2005	0.787	-0.365	(16.23)	(-7.40)	79.40%	83.08%	86.23%	0.25%	0.97%	2.20%
2006	0.761	-0.332	(15.52)	(-6.74)	79.38%	83.00%	86.15%	0.45%	1.41%	2.88%
2007	0.736	-0.311	(12.97)	(-5.97)	69.81%	74.50%	79.19%	1.23%	2.93%	5.99%
2008	0.755	-0.317	(15.14)	(-6.52)	77.82%	81.67%	85.36%	0.34%	1.21%	2.82%
2009	0.768	-0.331	(16.09)	(-7.01)	79.54%	83.16%	86.38%	0.63%	1.70%	3.51%
2010	0.769	-0.329	(15.95)	(-7.01)	78.65%	82.63%	86.22%	0.56%	1.64%	3.66%
2011	0.754	-0.313	(15.47)	(-6.73)	77.75%	81.79%	85.71%	0.63%	1.87%	4.10%
2012	0.763	-0.328	(15.65)	(-7.01)	77.64%	81.93%	85.61%	0.89%	2.25%	4.69%

Table 4: **DY Model Regressions.** Continued.

(b) Real Data

	β		t		R^2			$R^2_{inc.}$		
	$\lvert$adj. OIB$\rvert$	$\lvert$adj. OIB$\rvert^2$	$\lvert$adj. OIB$\rvert$	$\lvert$adj. OIB$\rvert^2$	50%	p-value	% Rej.	50%	p-value	% Rej.
1993	0.369	-0.170	(7.61)	(-3.48)	34.07%	6.15%	80.56%	15.22%	23.83%	48.21%
1994	0.348	-0.150	(7.51)	(-3.16)	33.55%	7.07%	75.52%	14.53%	23.87%	48.38%
1995	0.342	-0.149	(6.99)	(-3.00)	30.15%	6.52%	77.48%	15.63%	29.41%	43.47%
1996	0.358	-0.164	(7.33)	(-3.42)	31.11%	6.85%	79.27%	14.19%	25.56%	50.64%
1997	0.334	-0.140	(6.49)	(-2.78)	28.00%	5.52%	82.25%	13.92%	26.26%	50.56%
1998	0.329	-0.136	(6.21)	(-2.62)	26.26%	4.50%	84.04%	12.97%	22.18%	57.16%
1999	0.365	-0.166	(6.91)	(-3.16)	27.89%	3.53%	88.72%	12.56%	18.93%	62.38%
2000	0.333	-0.145	(5.75)	(-2.55)	23.49%	3.06%	89.90%	11.88%	20.82%	62.06%
2001	0.374	-0.176	(6.38)	(-3.06)	25.25%	1.43%	94.67%	9.07%	15.71%	74.29%
2002	0.328	-0.130	(4.82)	(-1.90)	21.31%	1.03%	97.16%	9.08%	10.15%	82.14%
2003	0.334	-0.135	(4.84)	(-1.98)	21.55%	0.63%	98.41%	8.58%	10.51%	81.42%
2004	0.295	-0.104	(4.15)	(-1.46)	18.31%	0.39%	99.11%	9.57%	10.09%	83.63%
2005	0.279	-0.103	(4.03)	(-1.51)	16.23%	0.40%	98.83%	10.61%	11.10%	82.60%
2006	0.243	-0.083	(3.40)	(-1.17)	12.46%	1.16%	97.58%	11.15%	16.81%	77.86%
2007	0.219	-0.086	(3.14)	(-1.25)	9.66%	1.94%	95.91%	12.26%	25.72%	65.76%
2008	0.217	-0.086	(3.05)	(-1.23)	8.83%	2.16%	96.34%	11.92%	19.43%	74.90%
2009	0.230	-0.093	(3.24)	(-1.30)	10.04%	2.04%	95.86%	11.43%	19.40%	74.53%
2010	0.241	-0.103	(3.41)	(-1.49)	10.59%	2.45%	95.08%	12.38%	21.74%	71.55%
2011	0.245	-0.102	(3.45)	(-1.50)	10.35%	2.04%	95.95%	13.05%	21.61%	71.57%
2012	0.275	-0.127	(4.04)	(-1.86)	12.22%	2.54%	95.61%	12.20%	23.56%	70.88%

Table 5: **OWR Model Regressions.** This table reports real and simulated regressions of the $CPIE_{OWR}$ on the squared and interaction terms of y_e, r_d, and r_o. In Panel A, we simulate 1,000 instances of the OWR model for each `PERMNO`-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 values. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 values, and tests of the null that the model fits the data. The p-value of R^2 is the probability of observing an R^2 at least as small as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the null at the 5% level.

(a) Simulated Data

	β						t						R^2		
	y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	5%	50%	95%
1993	0.002	0.068	-0.003	0.017	0.016	0.096	(0.42)	(11.52)	(-0.66)	(2.71)	(3.34)	(17.78)	68.29%	79.86%	88.22%
1994	0.002	0.065	-0.003	0.018	0.017	0.093	(0.53)	(12.10)	(-0.67)	(3.14)	(3.80)	(18.95)	70.03%	81.70%	89.67%
1995	0.003	0.065	-0.003	0.019	0.018	0.093	(0.57)	(12.03)	(-0.71)	(3.14)	(4.00)	(18.83)	69.82%	81.98%	89.91%
1996	0.003	0.066	-0.003	0.020	0.019	0.094	(0.68)	(12.73)	(-0.76)	(3.77)	(4.43)	(20.14)	72.12%	83.64%	91.18%
1997	0.003	0.063	-0.003	0.018	0.018	0.092	(0.77)	(14.31)	(-0.80)	(4.05)	(4.73)	(21.45)	73.01%	85.04%	92.43%
1998	0.002	0.070	-0.004	0.018	0.017	0.102	(0.67)	(16.25)	(-1.01)	(4.14)	(4.70)	(24.53)	74.91%	86.68%	93.93%
1999	0.003	0.060	-0.003	0.017	0.018	0.093	(0.74)	(13.90)	(-0.75)	(3.88)	(4.86)	(22.15)	72.82%	84.70%	92.22%
2000	0.003	0.051	-0.002	0.017	0.019	0.085	(0.87)	(13.37)	(-0.58)	(4.20)	(5.64)	(22.86)	73.87%	85.03%	92.21%
2001	0.002	0.066	-0.004	0.014	0.014	0.098	(0.51)	(17.18)	(-1.15)	(3.72)	(4.25)	(26.22)	76.05%	87.58%	94.14%
2002	0.001	0.066	-0.003	0.012	0.013	0.099	(0.44)	(18.37)	(-1.03)	(3.40)	(3.89)	(27.41)	76.47%	87.94%	94.40%
2003	0.002	0.071	-0.005	0.014	0.013	0.105	(0.48)	(19.18)	(-1.53)	(3.50)	(3.84)	(27.86)	77.31%	88.81%	94.93%
2004	0.001	0.068	-0.005	0.012	0.012	0.100	(0.49)	(21.61)	(-1.91)	(4.05)	(4.06)	(30.04)	79.32%	90.05%	95.22%
2005	0.002	0.061	-0.005	0.012	0.012	0.086	(0.60)	(22.68)	(-2.02)	(4.35)	(4.35)	(31.06)	80.89%	90.80%	95.18%
2006	0.001	0.063	-0.004	0.011	0.011	0.089	(0.52)	(22.88)	(-1.91)	(3.95)	(4.14)	(30.37)	80.34%	90.48%	95.19%
2007	0.001	0.051	-0.003	0.002	0.004	0.068	(0.65)	(22.32)	(-1.69)	(0.78)	(1.68)	(28.67)	81.21%	90.63%	95.41%
2008	0.076	0.000	-0.001	0.000	0.004	0.001	(27.51)	(0.07)	(-0.25)	(0.10)	(1.42)	(0.29)	76.59%	88.91%	95.17%
2009	0.002	0.039	-0.002	0.001	0.005	0.060	(1.18)	(18.30)	(-0.73)	(0.35)	(2.36)	(27.24)	80.66%	90.07%	95.06%
2010	0.002	0.038	-0.002	0.000	0.000	0.046	(0.94)	(18.05)	(-1.34)	(0.13)	(0.23)	(22.24)	78.97%	88.62%	94.54%
2011	0.001	0.042	-0.002	0.000	0.000	0.055	(0.79)	(19.58)	(-1.37)	(0.11)	(0.16)	(24.64)	80.82%	90.39%	95.10%
2012	0.001	0.046	-0.003	0.000	0.000	0.055	(0.68)	(19.47)	(-1.55)	(0.11)	(0.22)	(23.02)	79.83%	89.47%	94.62%

Table 5: **OWR Model Regressions.** Continued.

(b) Real Data

	β						t						R^2		
	y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	50%	p-value	% Rej.
1993	-0.000	0.053	-0.000	0.032	0.029	0.055	(-0.03)	(7.24)	(-0.13)	(4.41)	(4.56)	(8.11)	69.97%	15.59%	40.02%
1994	0.000	0.053	-0.001	0.032	0.027	0.060	(0.06)	(8.11)	(-0.17)	(4.69)	(4.68)	(9.44)	72.00%	17.13%	38.49%
1995	0.001	0.052	-0.001	0.033	0.029	0.059	(0.15)	(7.92)	(-0.17)	(4.74)	(4.89)	(9.35)	72.73%	17.59%	41.21%
1996	0.001	0.055	-0.003	0.032	0.028	0.062	(0.28)	(8.61)	(-0.52)	(4.77)	(4.81)	(9.83)	73.65%	16.42%	41.49%
1997	0.002	0.054	-0.002	0.029	0.027	0.061	(0.36)	(8.90)	(-0.53)	(4.85)	(4.84)	(10.17)	74.72%	16.34%	41.04%
1998	0.002	0.069	-0.004	0.025	0.023	0.074	(0.37)	(11.25)	(-0.89)	(4.43)	(4.15)	(12.61)	77.46%	18.52%	35.82%
1999	0.002	0.057	-0.003	0.025	0.025	0.065	(0.56)	(9.59)	(-0.64)	(4.33)	(4.58)	(11.66)	76.48%	19.65%	31.28%
2000	0.003	0.050	-0.003	0.021	0.022	0.066	(0.82)	(10.58)	(-0.98)	(4.50)	(5.15)	(14.37)	79.83%	28.77%	20.29%
2001	0.001	0.068	-0.003	0.018	0.016	0.078	(0.47)	(14.62)	(-0.94)	(4.10)	(3.81)	(16.91)	83.25%	34.94%	20.48%
2002	0.002	0.072	-0.002	0.016	0.014	0.081	(0.47)	(16.83)	(-0.72)	(3.88)	(3.71)	(19.17)	84.71%	36.14%	16.94%
2003	0.002	0.080	-0.003	0.017	0.015	0.080	(0.60)	(20.66)	(-0.94)	(4.38)	(3.93)	(20.51)	87.22%	42.18%	13.72%
2004	0.001	0.077	-0.005	0.016	0.012	0.074	(0.54)	(24.74)	(-1.74)	(4.48)	(3.58)	(21.11)	88.70%	41.66%	14.95%
2005	0.002	0.072	-0.005	0.013	0.010	0.065	(0.83)	(25.08)	(-2.12)	(4.36)	(3.32)	(20.58)	89.54%	44.67%	12.29%
2006	0.002	0.072	-0.005	0.013	0.010	0.066	(0.74)	(25.53)	(-1.61)	(4.12)	(3.36)	(20.42)	89.47%	43.32%	12.28%
2007	0.002	0.058	-0.003	0.004	0.005	0.058	(0.98)	(18.17)	(-0.97)	(1.40)	(1.79)	(17.59)	89.34%	45.41%	10.21%
2008	0.077	0.004	-0.002	0.003	0.006	0.007	(22.41)	(1.10)	(-0.55)	(1.07)	(2.00)	(1.54)	88.02%	42.49%	7.83%
2009	0.003	0.038	-0.002	0.004	0.006	0.053	(1.55)	(15.99)	(-0.87)	(1.85)	(2.42)	(22.33)	89.34%	46.12%	7.45%
2010	0.002	0.035	-0.002	0.002	0.003	0.038	(1.39)	(16.80)	(-0.69)	(1.02)	(1.53)	(15.83)	89.54%	48.77%	7.53%
2011	0.002	0.043	-0.002	0.002	0.003	0.050	(1.27)	(17.71)	(-0.84)	(1.04)	(1.50)	(18.56)	89.84%	49.68%	7.88%
2012	0.002	0.045	-0.003	0.002	0.003	0.039	(1.14)	(20.34)	(-1.05)	(1.20)	(1.54)	(17.30)	90.29%	51.35%	7.71%

Table 6: **M&A Regressions.** This table reports regression results for the PIN, DY, and OWR $CPIE$ around M&A events. For each M&A target firm in our sample, we run regressions of $CPIE$ on order flow and returns data (for the OWR model) from $[-30, +30]$ and report median estimates across all the events. In Panels A and B we compute the incremental $R^2_{\text{inc.}}$ as the increase in R^2 attributed to $turn$ and $turn^2$. For all panels, we report standardized coefficients.

(a) PIN

β		t		R^2	$R^2_{\text{inc.}}$
$\lvert B-S\rvert$	$\lvert B-S\rvert^2$	$\lvert B-S\rvert$	$\lvert B-S\rvert^2$	50%	50%
0.298	-0.117	(5.76)	(-2.29)	20.18%	44.04%

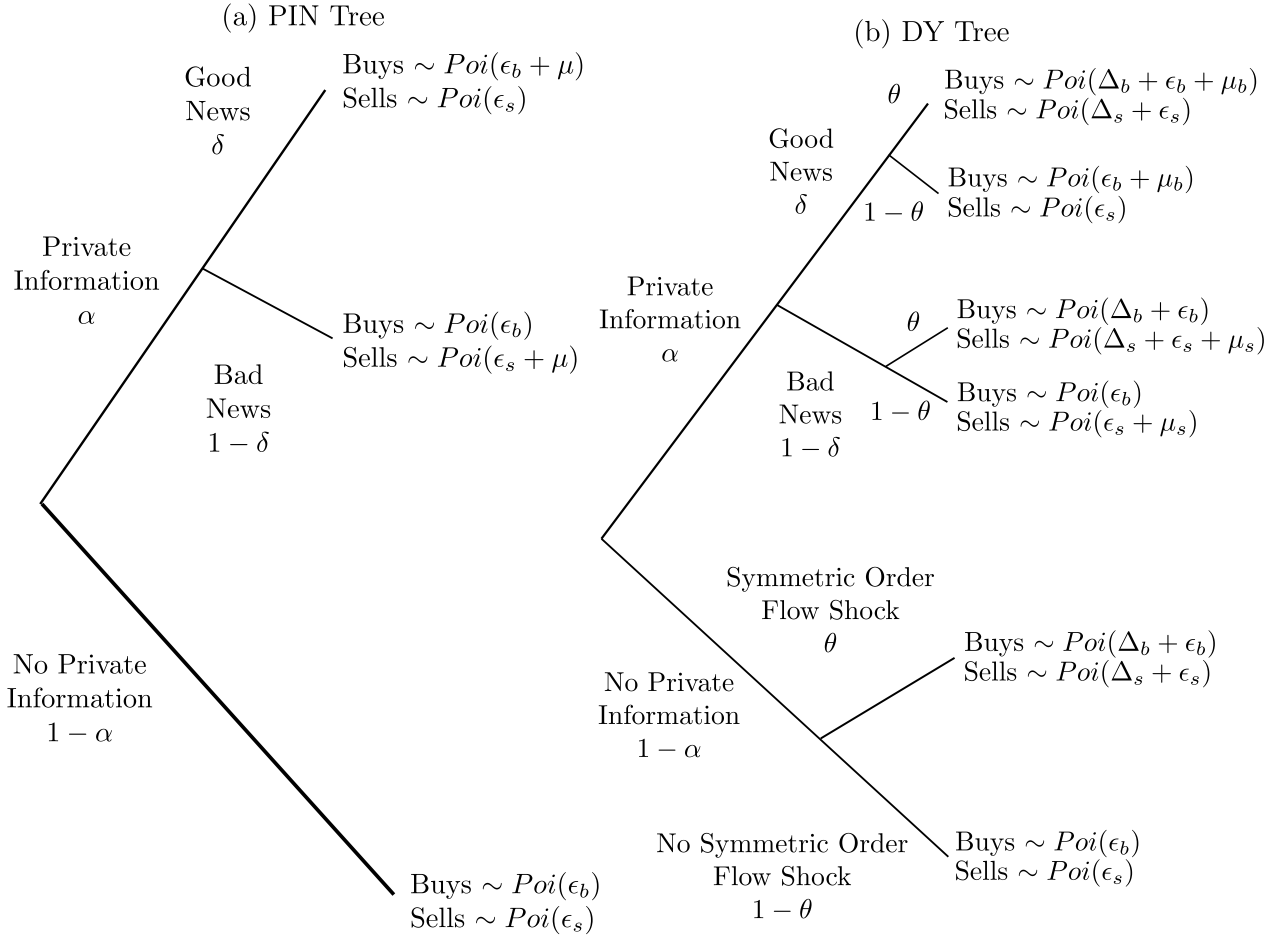
(b) DY

β		t		R^2	$R^2_{\text{inc.}}$
$\lvert$adj. OIB$\rvert$	$\lvert$adj. OIB$\rvert^2$	$\lvert$adj. OIB$\rvert$	$\lvert$adj. OIB$\rvert^2$	50%	50%
0.292	-0.130	(5.43)	(-2.44)	15.83%	11.96%

(c) OWR

β						t						R^2
y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	y_e^2	$y_e \times r_d$	$y_e \times r_o$	r_d^2	$r_d \times r_o$	r_o^2	
0.002	0.057	-0.003	0.016	0.015	0.061	(0.87)	(13.91)	(-0.92)	(4.09)	(3.95)	(15.18)	81.40%

Figure 1: **Model Trees.** Panels A and B represent the PIN and DY models of informed trading. For a given trading day, private information arrives with probability α. When there is no private information, buys and sells are Poisson with intensity ϵ_b and ϵ_s. Private information is good news with probability δ. In the PIN model, the expected number of buys (sells) increases by μ in case of good (bad) news. In the DY model, the expected number of buys (sells) increases by μ_b (μ_s) in case of good (bad) news. The DY model extends the PIN model to include symmetric order flow shocks, which occur with probability θ. In the event of a symmetric order flow shock, buys increase by Δ_b and sells increase by Δ_s.

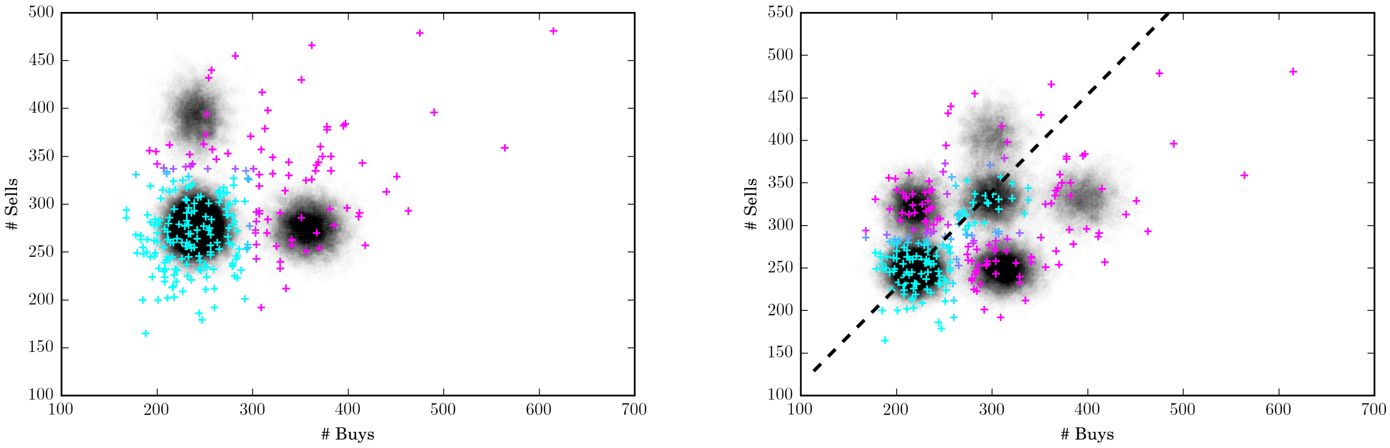


(a) PIN Tree

(b) DY Tree

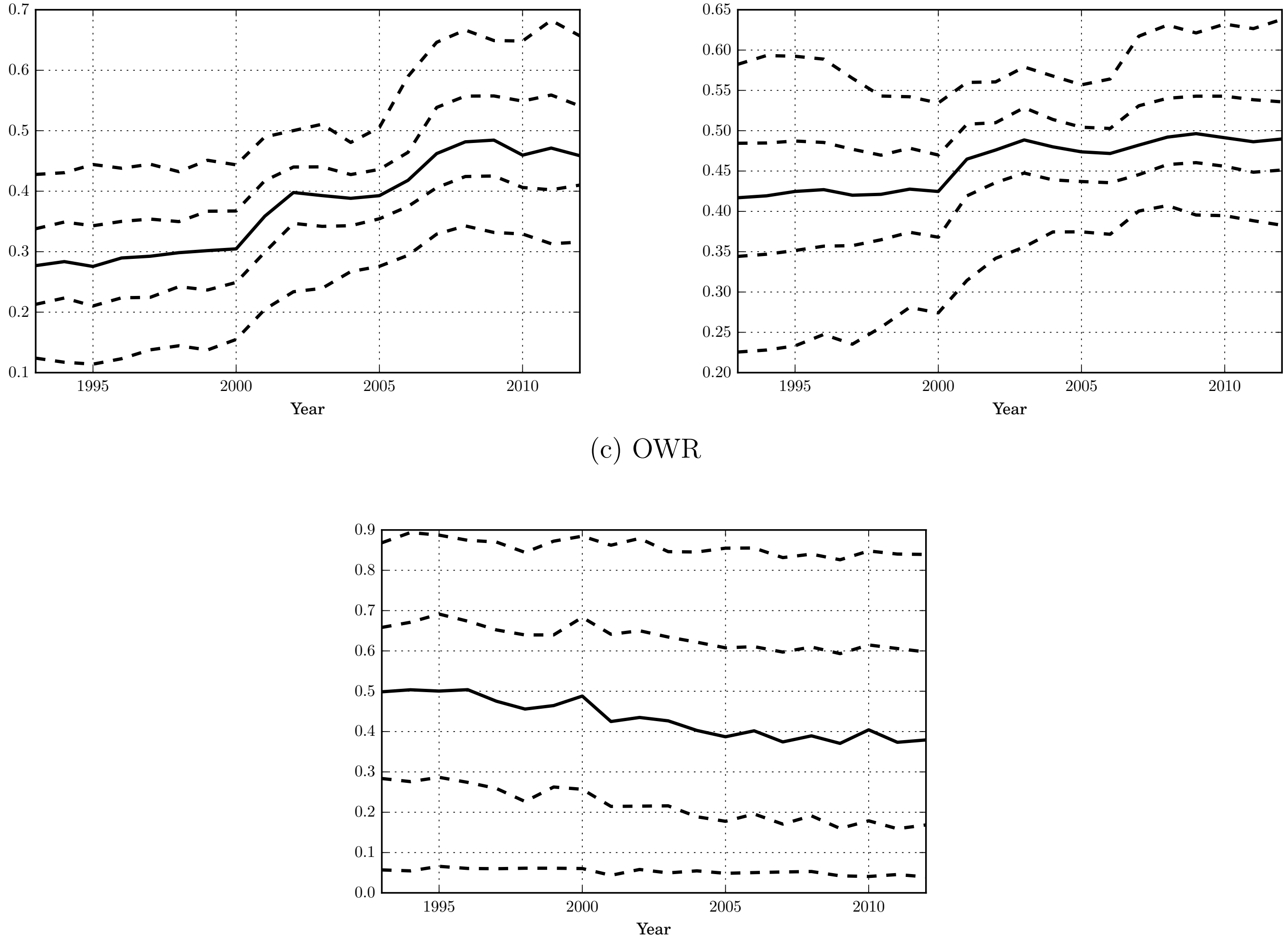
Good News δ
Buys $\sim Poi(\epsilon_b + \mu)$
Sells $\sim Poi(\epsilon_s)$
Private Information α
Buys $\sim Poi(\epsilon_b)$
Sells $\sim Poi(\epsilon_s + \mu)$
Bad News $1 - \delta$
No Private Information $1 - \alpha$
Buys $\sim Poi(\epsilon_b)$
Sells $\sim Poi(\epsilon_s)$
θ
Buys $\sim Poi(\Delta_b + \epsilon_b + \mu_b)$
Sells $\sim Poi(\Delta_s + \epsilon_s)$
Good News δ
Buys $\sim Poi(\epsilon_b + \mu_b)$
Sells $\sim Poi(\epsilon_s)$
$1 - \theta$
Private Information α
θ
Buys $\sim Poi(\Delta_b + \epsilon_b)$
Sells $\sim Poi(\Delta_s + \epsilon_s + \mu_s)$
Bad News $1 - \delta$
$1 - \theta$
Buys $\sim Poi(\epsilon_b)$
Sells $\sim Poi(\epsilon_s + \mu_s)$
Symmetric Order Flow Shock θ
Buys $\sim Poi(\Delta_b + \epsilon_b)$
Sells $\sim Poi(\Delta_s + \epsilon_s)$
No Private Information $1 - \alpha$
No Symmetric Order Flow Shock $1 - \theta$
Buys $\sim Poi(\epsilon_b)$
Sells $\sim Poi(\epsilon_s)$

Figure 2: **XOM 1993.** This figure compares the real and simulated data for XOM in 1993 using the PIN and DY models. In Panels A and B, the real data are marked as +. The real data are shaded according to the model-specific $CPIE$, with lighter markers (+) representing low and darker markers (+) high $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark "cloud" of points, and low probability states appear as a light "cloud" of points. The PIN model has three states: no news, good news, and bad news; the DY model includes three additional states for when there are symmetric order flow shocks.


(a) PIN

(b) DY

Figure 3: **Yearly Alphas.** This figure shows the distribution of yearly α estimates for the PIN, DY, and OWR models, respectively. The solid black line represents the median α, and the dotted lines represent the 5, 25, 75, and 95 percentiles.

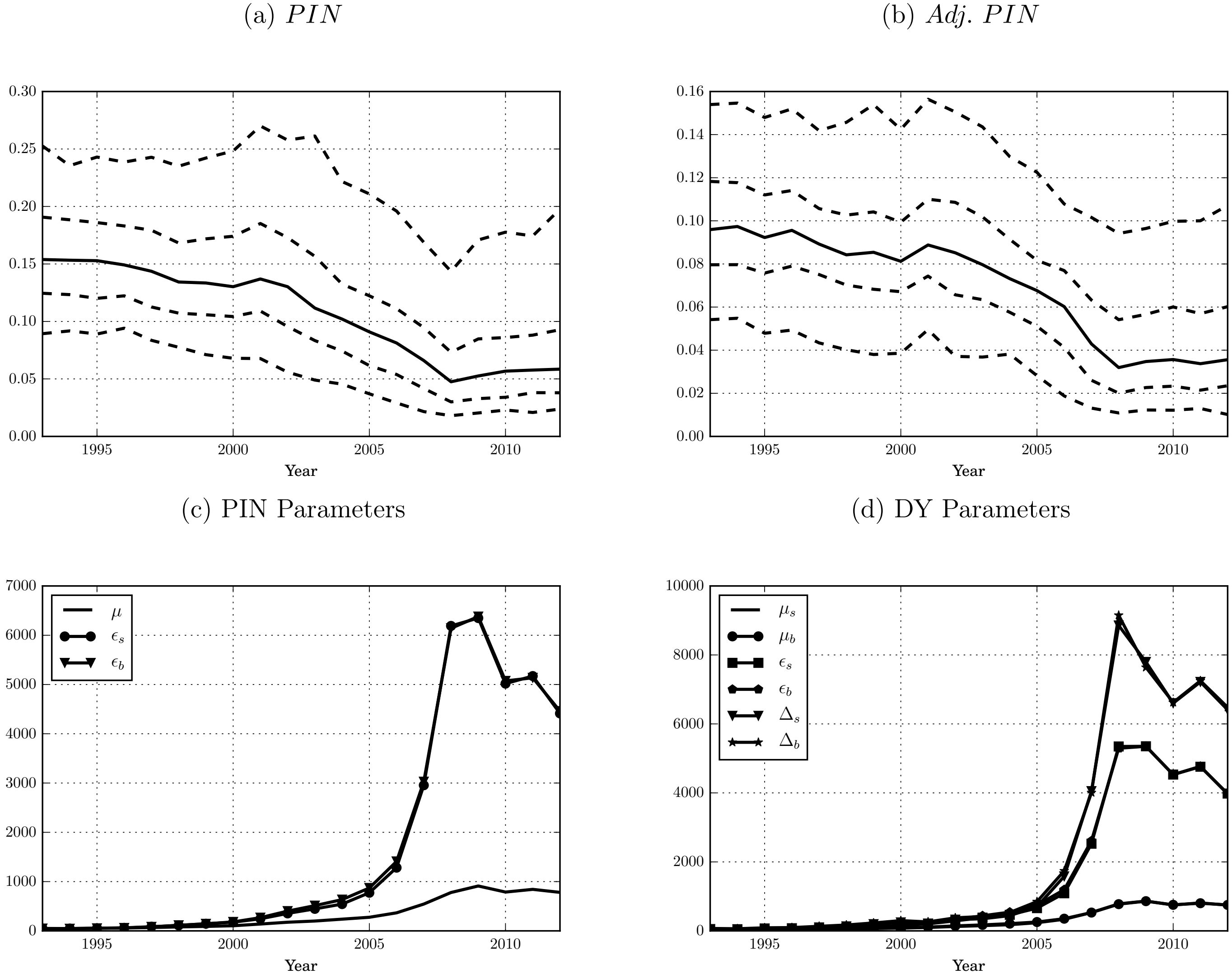


(a) PIN

(b) DY

(c) OWR

Figure 4: **PIN and DY Model Parameters.** This figure shows the distribution of PIN and $Adj.\ PIN$ as well as the mean PIN and DY model parameters for each year in our sample. In Panels A and B the solid black line represents the median stock's probability of informed trading, and the dotted lines represent the 5, 25, 75, and 95 percentiles. In Panels C and D, μ, ϵ, and Δ represent the expected number of trades from informed, uninformed, and symmetric order flow trading shocks, respectively.



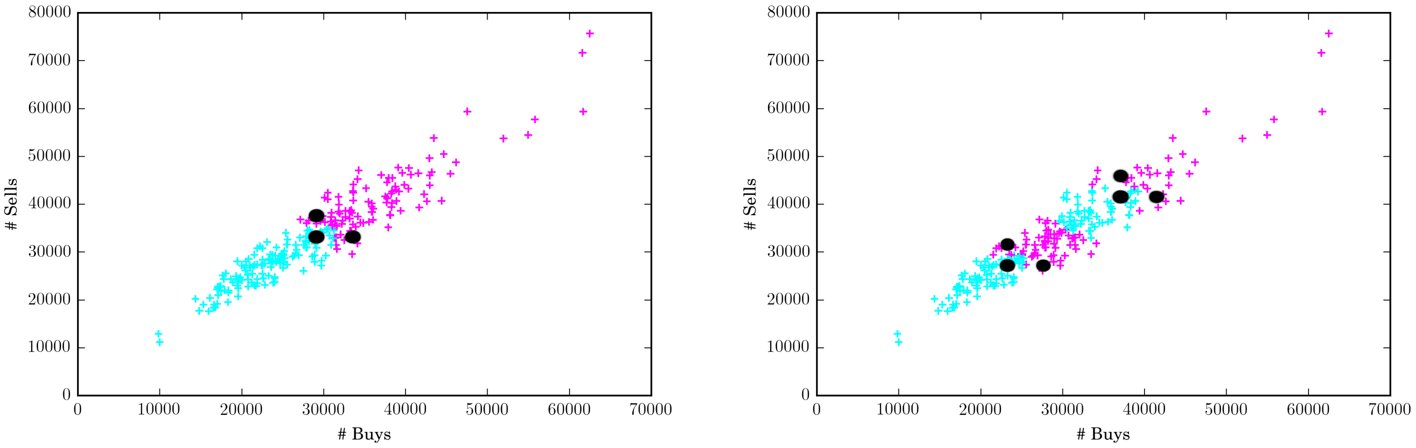
(a) PIN

(b) $Adj.\ PIN$

(c) PIN Parameters

(d) DY Parameters

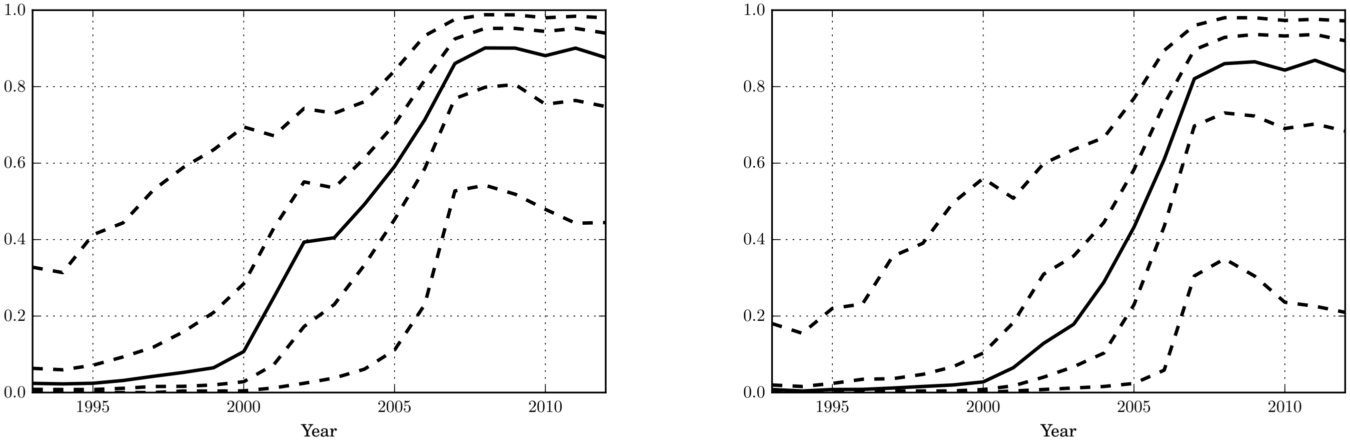
Figure 5: **XOM 2012.** This figure compares the real and simulated data for XOM in 2012 using the PIN and DY models. In Panels A and B, the real data are marked as +. The real data are shaded according to the model-specific $CPIE$, with lighter markers (+) representing low and darker markers (+) high $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark "cloud" of points, and low probability states appear as a light "cloud" of points. The PIN model has three states: no news, good news, and bad news; the DY model includes three additional states for when there are symmetric order flow shocks.



(a) PIN

(b) DY

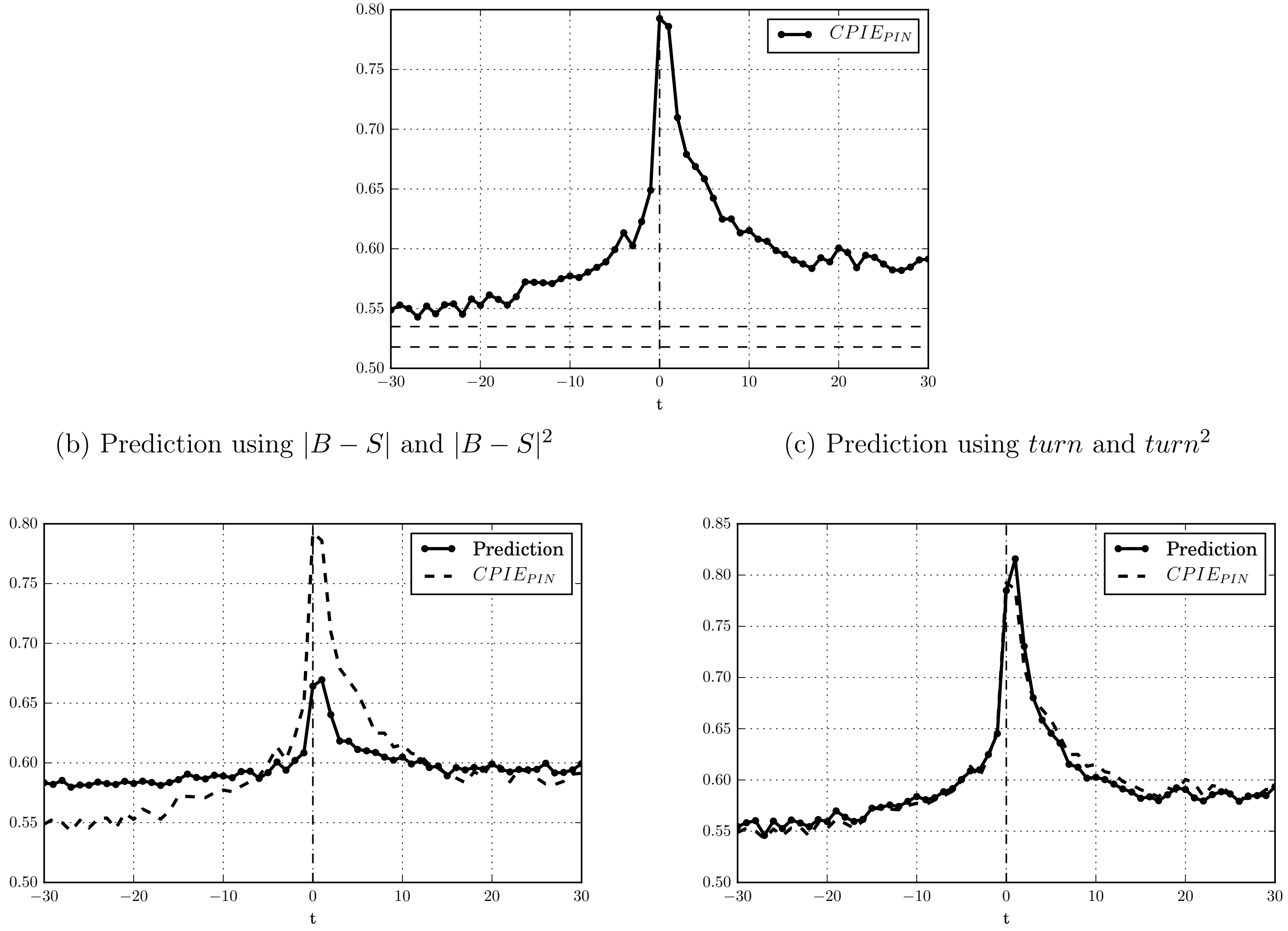
Figure 6: **Days With Near-Zero Probability.** Panels A and B show the distribution of the fraction of days within a `PERMNO`-Year with near-zero probability of occurring under the data-generating processes of the PIN and DY models. These days occur when the total likelihood, given the model parameters and observed order flow data, is less than 10^{-10}. The solid black line represents the median stock, and the dotted lines represent the 5, 25, 75, and 95 percentiles.



(a) PIN

(b) DY

Figure 7: **M&A Target - PIN.** Panel A shows the average $CPIE_{PIN}$ for the PIN model in event time surrounding mergers and acquisitions targets. Panels B and C compare the average with the predicted $CPIE_{PIN}$ using the absolute value of buys minus sells or turnover, respectively. To obtain the predictions, we run regressions of daily $CPIE_{PIN}$ on $|B-S|$ or $turn$, and their respective squared terms.

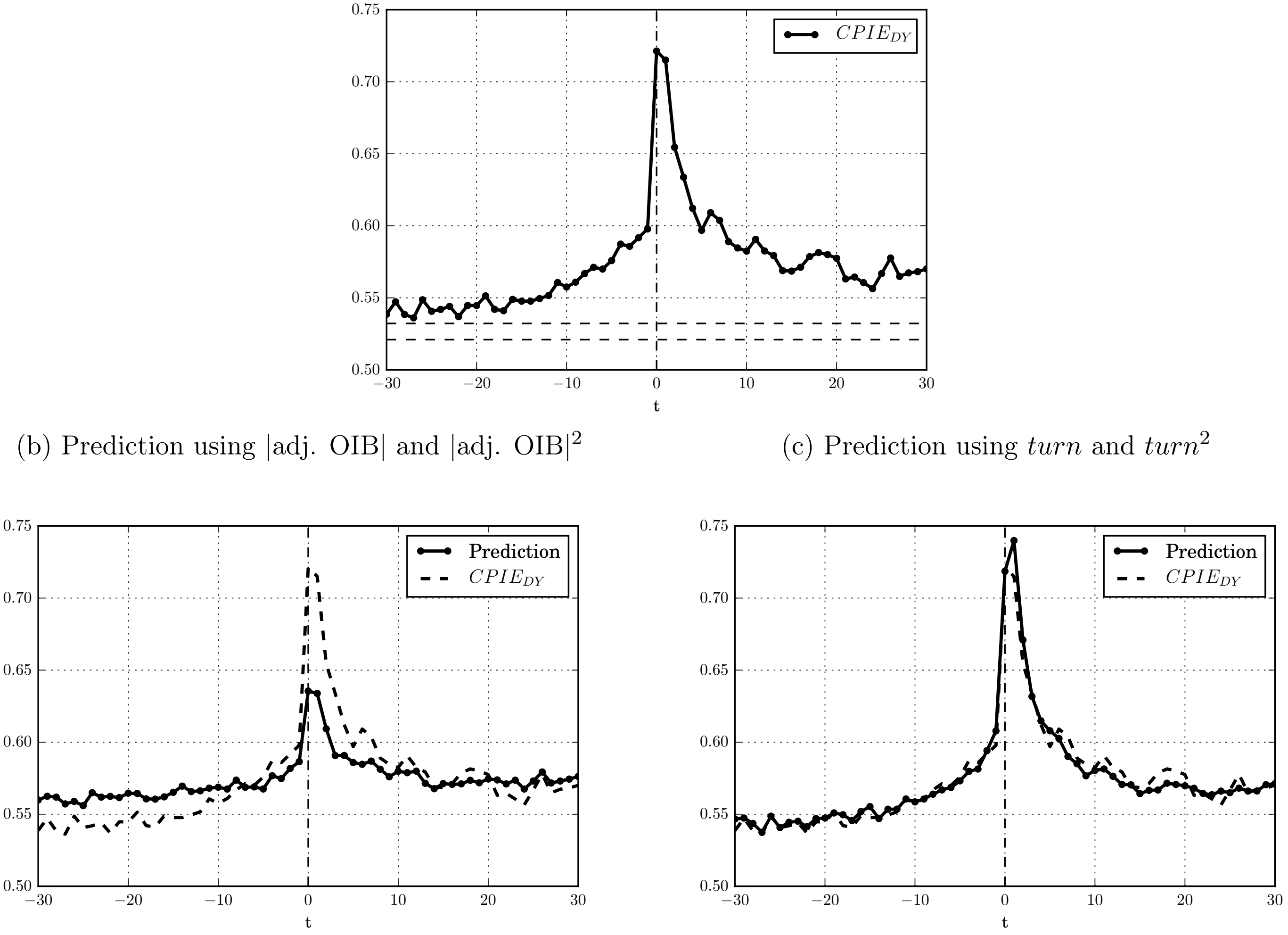


(a) $CPIE_{PIN}$

(b) Prediction using $|B-S|$ and $|B-S|^2$

(c) Prediction using $turn$ and $turn^2$

Figure 8: **M&A Target - DY.** Panel A shows the average $CPIE_{DY}$ for the DY model in event time surrounding mergers and acquisitions targets. Panels B and C compare the average with the predicted $CPIE_{DY}$ using the absolute value of adjusted order imbalance or turnover, respectively. To obtain the predictions, we run regressions of daily $CPIE$ on |adj. OIB| or $turn$, and their respective squared terms.



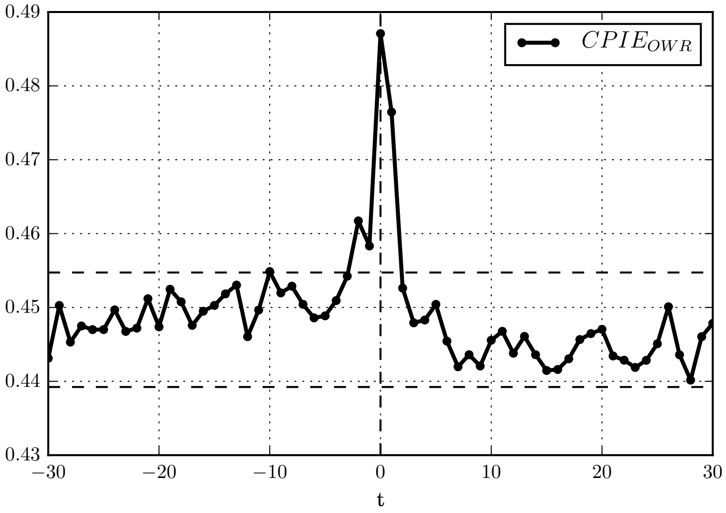
(a) $CPIE_{DY}$

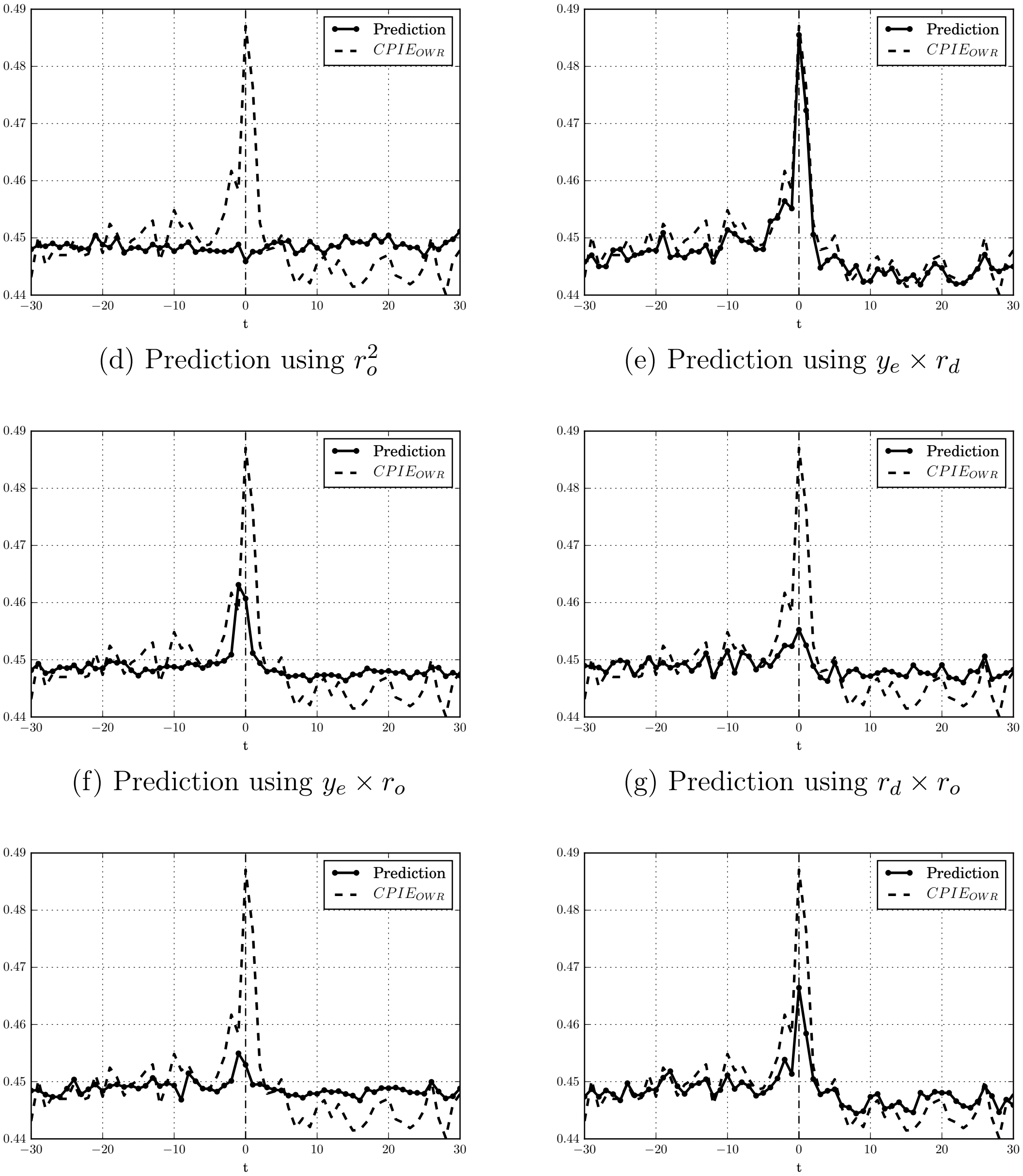
(b) Prediction using |adj. OIB| and |adj. OIB|^2

(c) Prediction using $turn$ and $turn^2$

Figure 9: **M&A Target - OWR.** Panel A shows the average $CPIE_{OWR}$ in event time surrounding mergers and acquisitions targets. Panels B–G compare the average $CPIE_{OWR}$ with the predicted $CPIE_{OWR}$ using the squared and interaction terms of y_e, r_d, and r_o.

(a) $CPIE_{OWR}$





(b) Prediction using y_e^2

(c) Prediction using r_d^2

(d) Prediction using r_o^2

(e) Prediction using $y_e \times r_d$

(f) Prediction using $y_e \times r_o$

(g) Prediction using $r_d \times r_o$

Internet Appendix: What does the PIN model identify as private information?

Jefferson Duarte, Edwin Hu, and Lance Young

May 1st, 2015

Internet Appendix

A. PIN Likelihood

The probabilities of observing $B_{i,t}$ and $S_{i,t}$ on a day without an information event, on a day with positive information event, and on a day with a negative information event are:

$$\begin{aligned}
L_{NI}(D_{PIN,i,t}) &= (1-\alpha_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \quad (1)\\
L_{I^+}(D_{PIN,i,t}) &= \alpha_i\delta_i e^{-(\mu_i+\epsilon_{B_i})}\frac{(\mu_i+\epsilon_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \quad (2)\\
L_{I^-}(D_{PIN,i,t}) &= \alpha_i(1-\delta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_i+\epsilon_{i,S})}\frac{(\mu_i+\epsilon_{i,S})^{S_{i,t}}}{S_{i,t}!} \quad (3)
\end{aligned}$$

B. DY Likelihood

$$\begin{aligned}
L_{NI,NS}(D_{DY,i,t}) &= (1-\alpha_i)(1-\theta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \quad (4)\\
L_{NI,S}(D_{DY,i,t}) &= (1-\alpha_i)\theta_i e^{-(\epsilon_{B_i}+\Delta_{B_i})}\frac{(\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\epsilon_{S_i}+\Delta_{S_i})}\frac{(\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \quad (5)\\
L_{I^-,NS}(D_{DY,i,t}) &= \alpha_i(1-\theta_i)(1-\delta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_{S_i}+\epsilon_{S_i})}\frac{(\mu_{S_i}+\epsilon_{S_i})^{S_{i,t}}}{S_{i,t}!} \quad (6)\\
L_{I^-,S}(D_{DY,i,t}) &= \alpha_i\theta_i(1-\delta_i)e^{-(\epsilon_{B_i}+\Delta_{B_i})}\frac{(\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_{S_i}+\epsilon_{S_i}+\Delta_{S_i})}\frac{(\mu_{S_i}+\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \quad (7)\\
L_{I^+,NS}(D_{DY,i,t}) &= \alpha_i(1-\theta_i)\delta_i e^{-(\mu_{B_i}+\epsilon_{B_i})}\frac{(\mu_{B_i}+\epsilon_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_S}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \quad (8)\\
L_{I^+,S}(D_{DY,i,t}) &= \alpha_i\theta_i\delta_i e^{-(\mu_{B_i}+\epsilon_{B_i}+\Delta_{B_i})}\frac{(\mu_{B_i}+\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\epsilon_{S_i}+\Delta_{S_i})}\frac{(\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \quad (9)
\end{aligned}$$

where $L_{NI,NS}(D_{DY,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information trading or symmetric order flow shock; $L_{NI,S}(D_{DY,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day without private information and with a symmetric order flow shock; $L_{I^-,NS}$ ($L_{I^-,S}$) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information and without (with) symmetric order flow shock; $L_{I^+,NS}$ ($L_{I^+,S}$) is the probability on a day with positive information and without (with) a symmetric order flow shock.

C. OWR Likelihood

Let $\Theta_{OWR,i} = (\alpha_i, \sigma_{u_i}, \sigma_{z_i}, \sigma_{i_i}, \sigma_{p,d_i}, \sigma_{p,o_i})$ be the vector of parameters of this model. The parameter α_i is the probability that there is an information event on a given day. $\sigma^2_{z_i}$ is the variance of the noise of the observed net order flow (y_e); $\sigma^2_{u_i}$ is the variance of the net order flow from noise traders; $\sigma^2_{i_i}$ is the variance of the private signal received by the informed trader; σ^2_{p,d_i} is the variance of the intraday return; σ^2_{p,o_i} is the variance of the overnight return.

The likelihood of observing $D_{OWR,i,t}$ on a day without and with an information event:

$$\begin{aligned} L_{NI} &= (1-\alpha) f_{NI}(D_{OWR,i,t}) \qquad (10) \\ L_I &= \alpha f_I(D_{OWR,i,t}) \qquad (11) \end{aligned}$$

where $f_{NI}(D_{OWR,i,t})$ is the joint probability density of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$ on days without information, $f_I(D_{OWR,i,t})$ is the density of $(y_{e,t}, r_{o,t}, r_{d,t})$ on days with information events. Both $f_{NI}(D_{OWR,i,t})$ and $f_I(D_{OWR,i,t})$ are multivariate normal with zero means and covariance matrices Ω_{NI_i} and Ω_{I_i}. The covariance matrix Ω_{NI_i} has elements:

$$\begin{aligned} Var(y_e) &= \sigma^2_u + \sigma^2_z, \qquad (12) \\ Var(r_d) &= \sigma^2_{pd} + \alpha\sigma^2_i/4, \qquad (13) \\ Var(r_o) &= \sigma^2_{po} + \alpha\sigma^2_i/4, \qquad (14) \\ Cov(r_d, r_o) &= -\alpha\sigma^2_i/4, \qquad (15) \\ Cov(r_d, y_e) &= \alpha^{1/2}\sigma_i\sigma_u/2, \qquad (16) \\ Cov(r_o, y_e) &= -\alpha^{1/2}\sigma_i\sigma_u/2 \qquad (17) \end{aligned}$$

And Ω_{I_i}:

$$\begin{aligned}
Var(y_e) &= (1+1/\alpha)\sigma_u^2 + \sigma_z^2, & (18)\\
Var(r_d) &= \sigma_{pd}^2 + (1+\alpha)\sigma_i^2/4, & (19)\\
Var(r_o) &= \sigma_{po}^2 + (1+\alpha)\sigma_i^2/4, & (20)\\
Cov(r_d, r_o) &= (1-\alpha)\sigma_i^2/4, & (21)\\
Cov(r_d, y_e) &= \alpha^{-1/2}\sigma_i\sigma_u/2 + \alpha^{1/2}\sigma_i\sigma_u/2, & (22)\\
Cov(r_o, y_e) &= \alpha^{-1/2}\sigma_i\sigma_u/2 - \alpha^{1/2}\sigma_i\sigma_u/2 & (23)
\end{aligned}$$

D. Estimating Order Flow, $r_{o,i,t}$ and $r_{d,i,t}$

Wharton Research Data Services (WRDS) provides trades matched to National Best Bid and Offer (NBBO) quotes at 0, 1, 2, and 5 second delay intervals. We use only "regular way" trades, with original time and/or corrected timestamps to avoid incorrect quotes or non-standard settlement terms, for instance, trades that are settled in cash or settled the next business day.[1] Prior to 2000, we match “regular way” trades to quotes delayed for 5 seconds; between 2000 and 2007, we match trades to quotes delayed for 1 second; and after 2007, we match trades to quotes without any delay.

We classify the matched trades as either buys or sells following the Lee and Ready (1991) algorithm, which classifies all trades occurring above (below) the bid-ask mid-point as buyer (seller) initiated. We use a tick test to classify trades that occur at the mid-point of the bid and ask prices. The tick test classifies trades as buyer (seller) initiated if the price was above/(below) that of the previous trade.

To estimate $r_{o,i,t}$ and $r_{d,i,t}$, we run daily cross-sectional regressions of overnight and intraday returns on a constant, historical beta (based on the previous 5 years of monthly CRSP returns), log market cap, log book-to-market (following Fama and French (1992), Fama and French (1993), and Davis, Fama, and French (2000)). We impose min/max values for book

[1] Trade COND of (“@”,“*”, or “ ”) and CORR of (0,1)

equity (before taking logs) of 0.017 and 3.13, respectively. If book equity is negative, we set it to 1 before taking logs, so that it is zero after taking logs. We use the residuals from these daily cross-sectional regressions, winsorized at the 1 and 99% levels as our idiosyncratic intraday and overnight returns.

E. Maximum likelihood procedure

To estimate the PIN likelihood function, we use the maximum of the likelihood maximization with ten different starting points as in Duarte and Young (2009). We note, however, that late in the sample, the likelihood functions of the PIN and of the DY models are very close to zero. This is not surprising given the results in Fig. 6. In fact, after 2006, the PIN model suggests that 90% of the observed daily order flows for the median stock have a near-zero probability (i.e. smaller than 10^{-10}) of occurring. This makes the estimation susceptible to local optima. To get around this problem, we choose one of our ten starting points to be such that the PIN model clusters are close to the observed mean of the number of buys and sells. Specifically, we choose ϵ_B and ϵ_S values equal to the sample means of buys and sells, α equal to 1%, and delta equal to the mean absolute value of order imbalance. The other nine starting points are randomized. We do this in order to ensure that at least one of the starting points is centered properly, as the numerical likelihood estimation using purely random starts often stops at points outside of the central cluster of data.

We use a similar procedure to estimate the DY model. Specifically, we choose (ϵ_B, ϵ_S) values, and $(\epsilon_B + \Delta_B, \epsilon_S + \Delta_S)$ equal to the sample means of buys and sells computed by the k-means algorithm with k=2. The k-means algorithm looks for clusters in the buys and sells such that each observation belongs to the cluster with the nearest mean. Because we know a priori that buys and sells have a strong positive correlation (see Duarte and Young (2009)), we partition the sample into high and low order flow clusters, which correspond to the symmetric order flow shock/no symmetric order flow shock states in the DY model. The other nine starting points are randomized. As with the PIN model, we do this in order to ensure that at least one of the starting points is centered properly, as the numerical likelihood estimation using purely random starts often stops at points outside of the central clusters of data.

F. Computing $CPIE$s

In Section 2 of the paper, we define the $CPIE$ for the PIN, DY, and OWR models as the ratio of the "news" likelihood functions to the sum total of the likelihood functions. In practice, there are many cases in the PIN model for which the data are classified as "impossible" days, meaning $L_{NI}(D_{PIN,i,t})$, $L_{I^+}(D_{PIN,i,t})$, and $L_{I^-}(D_{PIN,i,t})$ are all numerically equivalent to zero (numerically a value less than the computer epsilon). As a result the $CPIE$ ratio results in a divide by zero error.

In order to compute $CPIE$ for these days, we "center" the likelihoods around the state with the highest log-likelihood before computing the $CPIE$. For example, consider the PIN model with:

$$\begin{aligned} L_{\max} &\equiv \max\{L_{NI}, L_{I^+}, L_{I^-}\}, && (24) \\ \ell_{\max} &\equiv \log(L_{\max}) && (25) \end{aligned}$$

where ℓ represents the log of the corresponding likelihood function. We compute the centered versions of each of the likelihood functions:

$$\begin{aligned} \ell'_{NI} &= \ell_{NI} - \ell_{\max}, && (26) \\ \ell'_{I^+} &= \ell_{I^+} - \ell_{\max}, && (27) \\ \ell'_{I^-} &= \ell_{I^-} - \ell_{\max}. && (28) \end{aligned}$$

We compute the $CPIE'$ as:

$$CPIE'_{PIN} = \frac{L'_{I^+} + L'_{I^-}}{L'_{NI} + L'_{I^+} + L'_{I^-}} \qquad (29)$$

such that the most likely state has $L' = 1$. For a high turnover day, it may be the case that $L'_{I^+} = 1$, $L'_{I^-} = 0$ and $L'_{NI} = 0$; hence, the $CPIE$' will be 1. We follow a similar procedure to compute $CPIE_{DY}$.

References

Davis, James L., Eugene F Fama, and Kenneth R French, 2000, Characteristics, covariances, and average returns: 1929 to 1997, *Journal of Finance* 55, 389–406.

Duarte, Jefferson, and Lance Young, 2009, Why is PIN priced?, *Journal of Financial Economics* 91, 119–138.

Fama, Eugene F, and Kenneth R French, 1992, The cross-section of expected stock returns, *Journal of Finance* 47, 427–465.

——— , 1993, Common risk factors in the returns on stock bonds, *Journal of Financial Economics* 33, 3–56.

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