# Does the PIN model mis-identify private information and if so, what is the alternative?*

Jefferson Duarte†, Edwin Hu‡, and Lance Young§

September 7th, 2017

**Abstract**

We show that the PIN model is no more useful in identifying private-information arrival than simply looking at whether turnover is above average or not. This calls into question $PIN$ as a measure of private information since turnover varies for many reasons unrelated to private-information arrival. We also examine an alternative to the PIN model, the Odders-White and Ready (2008) model. Our tests indicate that measures of private information based on the Odders-White and Ready (2008) model are promising alternatives to $PIN$.

*Keywords*: Liquidity; Information Asymmetry

*We thank Torben Andersen, Kerry Back, Pierre Collin-Dufresne, Kevin Crotty, Zhi Da, Bei Dong, Robert Engle, Gustavo Grullon, Terry Hendershott, Sahn-Wook Huh, Jyri Kinnunen, Pete Kyle, Yelena Larkin, Edward X. Li, K. Ramesh, Min Shen, Avi Wohl and seminar participants at the 2015 ITAM Conference, 2015 Annual SoFiE Conference, 2015 Annual MFS Conference, 2015 CICF, 2016 AFA Conference, FGV-EESP, PUC-RJ, Rice University, Texas A&M University, University of Virginia (McIntire), and the University of Washington (Foster) for helpful comments. We thank Elaine Brewer, Frank Gonzalez, Judy Hua, and Edward Martinez for computational support. This paper was previously circulated under the title "What Does the PIN Model Identify as Private Information?".

†Duarte is with the Jesse H. Jones School of Business at Rice University (jefferson.duarte@rice.edu).

‡Hu is with the U.S. Securities and Exchange Commission (hue@sec.gov). The Securities and Exchange Commission, as a matter of policy, disclaims responsibility for any private publication or statement by any of its employees. The views expressed herein are those of the author and do not necessarily reflect the views of the Commission or of the author's colleagues upon the staff of the Commission.

§Young is with the Michael G. Foster School of Business at the University of Washington (youngla@u.washington.edu).

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) is extensively used in the accounting, corporate finance and asset pricing literatures as a measure of information asymmetry.[1] Recently, several papers have documented some perhaps puzzling variation in $PIN$ around events (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), and Collin-Dufresne and Fos (2015)). While these papers suggest potential problems with the PIN model, there remains no definitive test of a model's ability to capture the arrival of private information because such information is inherently unobservable. Therefore, any test of a model's ability to capture private-information arrival is, in effect, a joint hypothesis test. For instance, using earnings announcements, the PIN model has been tested under the working hypothesis that the arrival of private information is more likely before an earnings announcement than after one. This hypothesis, however, is controversial because it is possible that agents convert public information into private signals using superior analysis (e.g. Kim and Verrecchia (1994, 1997)). In such a case, rather than indicating problems with the model, a higher $PIN$ after an earnings announcement would indicate that the PIN model properly captures the arrival of private information. Therefore, the apparently puzzling findings in the literature may be due to incorrect assumptions about the timing of private-information arrival, rather than problems with the PIN model.

Our first research question is whether the PIN model mis-identifies the arrival of private information. Our examination of the PIN model's ability to identify private information uses the working hypothesis that days of high turnover cannot *all* be considered private-information days.[2] Conversely, low turnover days cannot *all* be considered no private-information days. In contrast to the prior literature, our working hypothesis does not make specific assumptions about the arrival of private information. Instead, it is based on two

[1] A Google scholar search reveals that this series of PIN papers has been cited more than 3,500 times as of this writing. Recent examples of papers that use PIN in the finance and accounting literature include Chen, Goldstein, and Jiang (2007), Duarte, Han, Harford, and Young (2008), Bakke and Whited (2010), Da, Gao, and Jagannathan (2011), Ferreira, Ferreira, and Raposo (2011), Akins, Ng, and Verdi (2012), Brennan, Huh, and Subrahmanyam (2015), and Bennett, Garvey, Milbourn, and Wang (2017).

[2] In what follows, we refer to buyer initiated trades as 'buys', seller initiated trades as 'sells', turnover as the number of buys plus sells, order flow as either buys or sells, and absolute order flow imbalance as the absolute value of the difference between buys and sells.

uncontroversial, but closely related, principles. First, although turnover may be related to the arrival of private information, it also varies for myriad reasons unrelated to private information. For instance, turnover can increase due to disagreement (e.g. Kandel and Pearson (1995), and Banerjee and Kremer (2010)). Turnover is subject to calendar effects because traders coordinate trade on certain days to reduce trading costs (Admati and Pfleiderer (1988)). Furthermore, turnover can vary due to portfolio rebalancing (Lo and Wang (2000)) and taxation reasons (Lakonishok and Smidt (1986)). Therefore, a model that identifies the arrival of private information from turnover alone treats all the reasons for which turnover might vary as private information related. Second, even if one were to attempt to infer private-information arrival from turnover, reliable inferences cannot be gleaned from a simple heuristic based on whether turnover is low or high. This notion is so uncontroversial that we are unaware of any paper in the vast market microstructure literature that proposes identifying private-information arrival in this way.[3] Indeed, the literature has not even reached a consensus about the *sign* of the relation between turnover and information asymmetry.[4]

To address our first research question, we employ a variable called the Conditional Probability of an Information Event ($CPIE$). $CPIE$ is the probability that a model assigns to the arrival of private information on a particular day, given the data on that day. For instance, $CPIE_{PIN}$ is the probability of private-information arrival on a given day, conditional on the PIN model parameters and the observed daily order flow. By examining day-to-day variation in $CPIE$, an econometrician can infer how the model identifies private information.

We compare variation in $CPIE_{PIN}$ to variation in the $CPIE$ of a model that mechanically identifies the arrival of private information from turnover, hereafter called the Mechanical model. The idea behind this comparison is that if the PIN model identifies private-information arrival mechanically from turnover, it should be of no more use in identifying the arrival of such information than the Mechanical model. Our Mechanical model sets the probability of private-information arrival to one whenever turnover is higher than average and zero otherwise.[5] Formally, $CPIE_{Mech,j,t}$ is a dummy variable with value one when turnover

[3]Stickel and Verrecchia (1994) propose identifying information arrival in general based on whether volume is above or below the mean, but not private information in particular.

[4]See O'Hara (1997) for a review.

[5]We use the term Mechanical 'model' for convenience. The Mechanical model is intentionally not a structural model, instead it is just a heuristic.

on day $t$ for stock $j$ is above the annual mean of daily turnover for stock $j$ and zero otherwise. As such, we also refer to $CPIE_{Mech}$ as the 'Mechanical dummy.' Essentially, the Mechanical model amounts to the economically implausible statement that private information is sure to arrive on any day when turnover is 'high' (above the mean) and no private information ever arrives on days when turnover is 'low' (below the mean). As a result, the Mechanical model cannot produce reliable inferences about the arrival of private information.

To make our comparison, we estimate time-series regressions of $CPIE_{PIN}$ on $CPIE_{Mech}$ for each stock $j$. We find that the Mechanical dummy alone explains around 65% of the variation in $CPIE_{PIN}$. A natural question is whether this high average $R^2$ is the result of a potentially complex, non-linear functional relation between turnover and $CPIE_{PIN}$ instead of a relation captured by a simplistic dummy variable (i.e. $CPIE_{Mech}$). To address this possibility, we show that turnover and its square add only 8% to the explanatory power of $CPIE_{Mech}$. More significantly, while turnover varies for many reasons unrelated to the arrival of private information, turnover *may* vary with the arrival of private information. Thus, it is possible that the relation between $CPIE_{PIN}$ and turnover obtains because the PIN model properly captures that portion of variation in turnover that results from the arrival of private information. To address this concern, we control for a series of variables that the literature suggests are related to both private-information arrival and turnover. Our results indicate that these controls add only 4% to the explanatory power of turnover, its square, and $CPIE_{Mech}$. In summary, our regressions indicate that the PIN model mechanically identifies the arrival of private information from turnover.

Two limitations of the PIN model combine to cause this conflation of variation in turnover with the arrival of private information. First, under the PIN model, increases in expected turnover can only come about through the arrival of private information.[6] Second, the PIN model cannot simultaneously match both the mean and the variance of turnover due to its restrictive distributional assumptions. As a result of these limitations, when confronted with actual data, the model mechanically interprets periods of 'high' ('low') turnover as periods of private-information (no private-information) arrival.

[6]In the PIN model, realized turnover varies even without the arrival of private information. *Expected* turnover, however, varies only with the arrival of private information.

To further demonstrate that the PIN model's inferences are unreliable due to the conflation of turnover with private-information arrival, we compare $CPIE_{PIN}$ with $CPIE_{Mech}$ in two settings from the literature on private information. First, Cohen, Malloy, and Pomorski (2012) propose a method to identify opportunistic insider trades. Their results suggest that opportunistic insider trades reveal private information. Using this idea, we compare variation in $CPIE_{PIN}$ and $CPIE_{Mech}$ around opportunistic insider trades. Second, Hasbrouck (1988, 1991a,b) point out that non-information related price changes (e.g. dealer inventory control) should be subsequently reversed, while price moves from information related trades should not. Hence, we analyze whether the relation between price reversals and $CPIE_{PIN}$ is robust to controlling for $CPIE_{Mech}$.

We find that the event time correlation between $CPIE_{PIN}$ and $CPIE_{Mech}$ around opportunistic insider trades is above 99%. Moreover, the relation between $CPIE_{PIN}$ and price reversals disappears once we control for $CPIE_{Mech}$. Thus, the most popular model of private information in the literature yields inferences about private-information arrival that are no more reliable than simply looking at whether daily turnover is above or below its mean.[7] These results call into question the use of proxies based on the PIN model.

Despite the PIN model's problems, all is not lost in the quest for intuitive measures of information asymmetry based on structural models. Our second research question is whether an alternative to the PIN model, developed by Odders-White and Ready (2008) (hereafter OWR), is a viable substitute. We analyze the OWR model as an alternative to the PIN model because, unlike the PIN model which uses only order flow, the OWR model identifies the arrival of private information using intra-day and overnight returns along with order imbalance.[8] The intuition behind the OWR model is that the market maker only partially updates prices during the day in response to an order flow shock. This comes about because she is uncertain whether the abnormal order flow reflects the arrival of bona fide private information or a noise trade shock. However, the subsequent overnight price pattern is

[7]The PIN model is based on the theoretical notion, originally developed by Glosten and Milgrom (1985), that periods of informed trade can be identified by abnormally large absolute order flow imbalances. However, we show that empirically the PIN model identifies private-information arrival from turnover and not from absolute order flow imbalance.

[8]There are other structural models of private information that are based on order flow alone (e.g. Duarte and Young (2009)). However, Back, Crotty, and Li (2014) and Kim and Stoll (2014) show evidence consistent with the idea that order flow imbalance alone does not reveal private information.

different depending on whether private information arrives or not. If a private-information event occurs during the day, the overnight price response is a continuation of the intra-day reaction as uncertainty about the arrival of private information is resolved overnight and prices adjust to completely impound the now-public information. If no private information arrives, the overnight price response is a reversal of the intra-day reaction. As a result, the OWR model identifies private-information arrival from the interactions between order imbalance, intra-day, and overnight returns, as well as their variances.[9]

Similar to our analysis with the PIN model, we analyze whether the OWR model mechanically conflates the arrival of private information with turnover using regressions of $CPIE_{OWR}$ on the Mechanical dummy. The average $R^2$ in these regressions is around 3%. This stands in contrast to the 65% $R^2$ with $CPIE_{PIN}$. Thus, unlike the PIN model, the OWR does not identify private information mechanically from turnover. Naturally, this does not mean that the OWR is necessarily a good model. Rather, it simply means that the OWR model meets the 'low bar' of our uncontroversial working hypothesis that days of high (low) turnover cannot *all* be considered private-information (no private-information) days.

To gain further insight, we examine the OWR model under two additional working hypotheses. The first of these additional hypothesis is that opportunistic insiders trade up to the point that prices reveal their private information. Consistent with this hypothesis, we find that $CPIE_{OWR}$ rises before opportunistic insider trades then drops immediately afterward. The second working hypothesis is that private-information arrival is associated with weaker future price reversals. Consistent with this hypothesis, we find that $CPIE_{OWR}$ is related to smaller future return reversals.[10] With the caveat that our examination of the OWR model relies on two potentially controversial working hypotheses, our interpretation of these results is that the OWR model is a promising alternative to the PIN model.

Our paper contributes to an emerging literature that uses daily measures of private information. In a contemporaneous paper, Brennan, Huh, and Subrahmanyam (2015) examine high-frequency measures of good and bad news in event study settings. In contrast, we use

[9]Many potential proxies for private information that are not based on structural models (e.g. bid-ask spreads, impulse responses from structural VARs, and $VPIN$) have been analyzed in the literature (e.g. Andersen and Bondarenko (2014)). We focus on structural models (the OWR and PIN models).

[10]Even though the calculation of the $CPIE_{OWR}$ uses returns, our return continuation tests are constructed to avoid a mechanical relation between $CPIE_{OWR}$ and future returns. See further discussion in Section 3.

$CPIE_{PIN}$ to shed light on how the PIN model identifies private information. A related literature shows that the PIN model does not fit the order flow data well. For instance, Gan, Wei, and Johnstone (2014) show that the PIN model poorly describes the empirical distribution of order flow, while Duarte and Young (2009) argue that $PIN$ is a biased measure of private information because the PIN model does not match the positive covariance of buys and sells. While these results are suggestive of problems with the PIN model, the fact that it does not match some of the moments of the order flow distribution does not imply that $PIN$ fails to capture the variable of economic interest, namely private-information arrival. We contribute to this literature because our tests focus directly on *how* the model identifies private-information arrival by using $CPIE_{PIN}$. Finally, we contribute to the extensive and growing literature that employs measures of private information by showing that proxies based on the OWR model can potentially replace the widely used $PIN$ metric.

The remainder of the paper is as follows. Section 1 outlines the data we use for our empirical results. Section 2 shows that the PIN model mechanically associates variation in turnover with the arrival of private information. Section 3 evaluates the OWR model as an alternative to the PIN model. Section 4 concludes.

# 1 Data

To estimate the PIN and OWR models, we collect trade and quote data for all NYSE stocks between 1993 and 2012 from the NYSE TAQ database. We require that the firms in our sample have only one type of common stock (i.e. a single `PERMNO` and 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, of which about 36% (25%) are in the top (bottom) three Fama-French size deciles. For each stock in the sample, we classify each trade as either a buy or a sell, following the Lee and Ready (1991) algorithm. Following the literature, we estimate the PIN model for each stock $j$ using a sample consisting of the number of buys and sells for each day ($B_{j,t}$ and $S_{j,t}$). In our regression analysis, we also use the daily absolute order flow imbalance ($|B_{j,t} - S_{j,t}|$), and turnover ($turn_{j,t} = B_{j,t} + S_{j,t}$).

The OWR model requires intra-day and overnight returns as well as order imbalance.

Following OWR we compute the intra-day return on day $t$ as the volume-weighted average price (VWAP) during the trading day $t$ minus the opening quote midpoint on day $t$ plus dividends issued on day $t$, all divided by the opening quote midpoint on day $t$. We compute the overnight return on day $t$ as the opening quote midpoint on day $t+1$ minus the VWAP on day $t$, all divided by the opening quote midpoint on day $t$. Thus, the open-to-open return from day $t$ to day $t+1$ is the sum of the intra-day and overnight returns. We follow OWR by removing systematic effects from returns to obtain measures of idiosyncratic overnight and intra-day returns ($r_{o,j,t}$ and $r_{d,j,t}$). We compute order imbalance ($y_{e,j,t}$) as the daily share volume of buys minus the share volume of sells, divided by the total share volume. Like OWR, we remove days around unusual distributions or large dividends, as well as CUSIP or ticker changes. We also drop days for which there are missing overnight returns, intra-day returns, order imbalance, buys, or sells. See the Internet Appendix for further details.

There are two differences between our empirical procedures and those of OWR. First, OWR estimate $y_e$ as the idiosyncratic component of order flow imbalance divided by shares outstanding. We do not follow this procedure in defining $y_e$ because we find that it produces noisy estimates. Specifically, we find that $y_e$ defined as shares bought minus shares sold divided by shares outstanding, as in OWR, suffers from scale effects late in the sample, when order flow is several orders of magnitude larger than shares outstanding. Second, OWR remove a whole trading year of data surrounding distribution events, but we remove only one trading week [-2,+2] around these events.

We also examine a sample of opportunistic insider trades, as defined in Cohen, Malloy, and Pomorski (2012), from the Thomson Reuters' database of insider trades. In order to classify a trader as opportunistic or routine, we require three years of consecutive insider trades. We classify a trader as routine if she places a trade in the same calendar month for at least three years. All non-routine insiders' trades are classified as opportunistic. Our event sample includes 32,676 opportunistic insider trades.

Table 1 contains summary statistics for all the variables used to estimate the models. Panel A gives summary statistics for our entire sample and Panel B displays the summary statistics for opportunistic insider trading days.

# 2 Does PIN mis-identify private information?

Section 2.1 describes the PIN model. Section 2.2 shows that the PIN model mechanically identifies the arrival of private information from turnover. Section 2.3 demonstrates that the PIN model yields inferences that are no more reliable than simply looking at whether daily turnover is above or below its mean to identify private-information arrival.

## 2.1 Description and estimation of 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 noise traders and informed traders. Fig. 1 shows a tree diagram of the model. The intuition behind Fig. 1 is that at the beginning of each day, informed traders receive a private signal with probability $\alpha$. If the informed traders receive no signal, they do not trade. Therefore, buy and sell orders arrive at the normal rate of noise trade ($\epsilon_B$ for buys and $\epsilon_S$ for sells). If the informed receive a signal (positive with probability $\delta$ and negative with probability $1-\delta$), they join the noise traders and place orders with the market maker at the rate $\mu$. These orders lead to larger expected absolute order flow imbalances on days when the informed traders receive a signal.

Formally, let $B_{j,t}$ ($S_{j,t}$) represent the number of buys (sells) for stock $j$ on day $t$, $\Theta_{PIN,j} = (\alpha_j, \mu_j, \epsilon_{B_j}, \epsilon_{S_j}, \delta_j)$ be the vector of PIN model parameters for stock $j$, and $D_{PIN,j,t} = [\Theta_{PIN,j}, B_{j,t}, S_{j,t}]$ be the vector of PIN model parameters together with the daily number of buys and sells. The likelihood of observing a given number of buys and sells on day $t$ ($L(D_{PIN,j,t})$) is equal to the likelihood of observing $B_{j,t}$ and $S_{j,t}$ on a day without private information ($L_{NI}(D_{PIN,j,t})$), added to the likelihood of $B_{j,t}$ and $S_{j,t}$ on a day with positive ($L_{I^+}(D_{PIN,j,t})$) as well as negative ($L_{I^-}(D_{PIN,j,t})$) information. Conditional on the information event, $B_{j,t}$ and $S_{j,t}$ are independent Poisson random variables, hence:

$$L_{NI}(D_{PIN,j,t}) = (1-\alpha_j)e^{-\epsilon_{B_j}}\frac{\epsilon_{B_j}^{B_{j,t}}}{B_{j,t}!}e^{-\epsilon_{S_j}}\frac{\epsilon_{S_j}^{S_{j,t}}}{S_{j,t}!} \quad (1)$$

$$L_{I^+}(D_{PIN,j,t}) = \alpha_j\delta_j e^{-(\mu_j+\epsilon_{B_j})}\frac{(\mu_j+\epsilon_{B_j})^{B_{j,t}}}{B_{j,t}!}e^{-\epsilon_{S_j}}\frac{\epsilon_{S_j}^{S_{j,t}}}{S_{j,t}!} \quad (2)$$

$$L_{I^-}(D_{PIN,j,t}) = \alpha_j(1-\delta_j)e^{-\epsilon_{B_j}}\frac{\epsilon_{B_j}^{B_{j,t}}}{B_{j,t}!}e^{-(\mu_j+\epsilon_{S_j})}\frac{(\mu_j+\epsilon_{j,S})^{S_{j,t}}}{S_{j,t}!} \quad (3)$$

Let $I_{j,t}$ be a dummy equal to one if the informed receive a private signal about stock $j$ on day $t$ and zero otherwise. $CPIE_{PIN,j,t}$ is the econometrician's posterior probability of private-information arrival given the data observed on day $t$, and the PIN model parameters. That is, $CPIE_{PIN,j,t} = P\left[I_{j,t} = 1 | D_{PIN,j,t}\right]$. According to Bayes' theorem:

$$CPIE_{PIN,j,t} = \frac{L_{I^-}(D_{PIN,j,t}) + L_{I^+}(D_{PIN,j,t})}{L_{I^-}(D_{PIN,j,t}) + L_{I^+}(D_{PIN,j,t}) + L_{NI}(D_{PIN,j,t})} \tag{4}$$

In the absence of buy and sell data, an econometrician would assign probability $\alpha_j = E[CPIE_{PIN,j,t}]$ to the arrival of private information for stock $j$ on day $t$, where the expectation is taken with respect to the joint distribution of $B_{j,t}$ and $S_{j,t}$.

We estimate the PIN model numerically via maximum likelihood for every firm-year in our sample. Specifically, we maximize $\prod_{t=1}^{T} L(D_{PIN,j,t})$. Maximization of this likelihood function is prone to numerical issues because of two features of the data. First, days with thousands of buys and sells are common. As a result, attempting to directly compute the exponentials and factorials in Equations 1 to 3 often generates values that are too large to be represented by a typical computer. To address this problem we follow Duarte and Young (2009) and compute $L_{NI}(D_{PIN,j,t})$, $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$ by first computing their logarithms. For instance, consider the computation of $L_{NI}(D_{PIN,j,t})$. Letting $\ell_{NI}$=$ln[L_{NI}(D_{PIN,j,t})]$, according to Equation 1 we have:

$$\ell_{NI} = ln(1-\alpha_j) - \epsilon_{B_j} + ln(\epsilon_{B_j}) \times B_{j,t} - \sum_{k=1}^{B_{j,t}} \ln(k) - \epsilon_{S_j} + ln(\epsilon_{S_j}) \times S_{j,t} - \sum_{k=1}^{S_{j,t}} \ln(k) \tag{5}$$

The computation of $\ell_{NI}$ as above does not result in numerical overflow problems even for very large numbers of trades because $B_{j,t}$ and $S_{j,t}$ enter Equation 5 multiplicatively instead of as exponents in Equation 1. Moreover, the negative terms in Equation 5 net out with the positive terms, resulting in values of $\ell_{NI}$ that can be readily exponentiated to compute $L_{NI}(D_{PIN,j,t})$. We compute $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$ as the exponential of $\ell_{I^+} = ln[L_{I^+}(D_{PIN,j,t})]$ and $\ell_{I^-} = ln[L_{I^-}(D_{PIN,j,t})]$.

Second, the PIN model cannot match both the high level and volatility of order flow late in the sample. As a result, the likelihood functions are very close to zero, which makes the estimation susceptible to local optima. To get around this problem, we follow Duarte and Young (2009) by maximizing the likelihood using ten different sets of starting points and

choosing the parameter estimates associated with the largest final likelihood value. Moreover, for our first set of starting points, we choose $\epsilon_B$ and $\epsilon_S$ values equal to the sample means of buys and sells, $\alpha$ equal to 1%, $\delta$ equal to 50% and $\mu$ equal to the mean absolute value of order flow imbalance. We do this in order to ensure that at least one of the starting points is centered properly. The other nine starting points are randomized.

These same two features of the data also plague direct computation of $CPIE_{PIN}$ in Equation 4 with numerical overflow and underflow problems. To address this problem we first define $\ell_{\max} = \max\{\ell_{NI}, \ell_{I^+}, \ell_{I^-}\}$. We then compute $CPIE_{PIN}$ as:

$$CPIE_{PIN,j,t} = \frac{e^{(\ell_{I^+} - \ell_{\max})} + e^{(\ell_{I^-} - \ell_{\max})}}{e^{(\ell_{NI} - \ell_{\max})} + e^{(\ell_{I^+} - \ell_{\max})} + e^{(\ell_{I^-} - \ell_{\max})}} \tag{6}$$

The equation above handles days with thousands of buys and sells because it replaces direct computation of the likelihoods ($L_{NI}(D_{PIN,j,t})$, $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$) with their logs ($\ell_{NI}, \ell_{I^+}, \ell_{I^-}$). It also handles days when the likelihood in the denominator of Equation 4 is such a small positive number that typical computer systems cannot distinguish it from zero. The computation of $CPIE_{PIN}$ using Equation 6 avoids this problem because the denominator of Equation 6 has a lower bound of one.

It is important to note that Equation 6 addresses a *computational* problem, not a mathematical problem. Equation 6 is not an approximation or an arbitrary normalization of Equation 4. In fact, a simple algebraic manipulation shows that these expressions are equivalent. Thus, Equation 6 is a mathematically-sound way to rewrite Equation 4 in order to avoid computational problems that would lead to a large number of missing $CPIE_{PIN}$ observations. Indeed, direct computation of Equation 4 would result in the complete loss of all $CPIE_{PIN}$ observations for the median stock by 2004.

Fig. 2 shows how the distribution of $\alpha$ changes over time. The PIN model $\alpha$ increases over time, rising from about 30% in 1993 to 50% in 2012.[11] Table 2 contains summary statistics for the parameter estimates as well as the cross-sectional sample means and standard deviations of $CPIE_{PIN}$. These statistics show that, as expected, the mean $CPIE_{PIN}$ behaves like $\alpha$.

[11]The increase in our PIN model $\alpha$ parameters is somewhat larger than that in Brennan, Huh, and Subrahmanyam (2015). This small difference arises because we have a smaller number of stocks since we apply sample filters similar to those in OWR. Without these filters, the increase in our PIN model $\alpha$ parameters from 1993 to 2012 is comparable to that in Brennan, Huh, and Subrahmanyam (2015).

We also estimate the parameter vector $\Theta_{PIN,j}$ in the period $t \in [-312, -60]$ before opportunistic insider trades. These parameter estimates are used to compute the $CPIE$s used in our opportunistic insider trading event study. The summary statistics of the parameter estimates for the event studies are similar to those in Table 2 and in Fig. 2.

## 2.2 How does the PIN model identify private information?

Section 2.2.1 uses a single stock, Exxon-Mobil, as an example of how the PIN model identifies private information. Section 2.2.2 shows that the PIN model's conflation of turnover with private-information arrival is widespread in the cross section of stocks.

### 2.2.1 A single stock example

To illustrate how the PIN model conflates turnover with the arrival of private information, Fig. 3 presents scatter plots of real and simulated buy and sell data for Exxon-Mobil. The simulated data are generated from the PIN model using Exxon-Mobil's estimated PIN model parameters for 1993 and 2012. Panels A and B plot the simulated and real order flow for Exxon-Mobil in 1993 and 2012 respectively, 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 value of $CPIE$, with darker points (+ **magenta**) representing high $CPIEs$, and lighter points signifying (+ cyan) low $CPIE$s. Panels C and D plot Exxon-Mobil's $\text{CPIE}_{PIN}$ as function of turnover. The vertical lines in these panels represent the annual mean of daily turnover.

The simulated data in Panels A and B of Fig. 3 illustrate the central intuition behind the PIN model. The simulated data fall into three categories corresponding to the nodes of the tree in Fig. 1. The data in these three categories create the distinct dark clusters in Panels A and B. In each panel, two of the clusters are made up of days characterized by relatively large absolute order flow imbalance, with a large number of sells (buys) and relatively few buys (sells). These are private-information days. The third group of days has relatively low numbers of buys and sells because there is no private-information arrival.

The real data, on the other hand, show no such distinct clusters. In Panel B of Fig. 3 the three simulated clusters from the PIN model rarely overlap with the data. Note that

the diameter of the clusters reflects the amount of variation in buys and sells that the model anticipates. Any day that falls outside the clusters appears *to the model* to be an extreme event. Far from being restricted to Exxon-Mobil, this problem affects nearly all of the stocks in our sample. Indeed, according to the PIN model, for the median stock about 60% of the annual observations can reasonably be classified as outliers ($L(D_{PIN,j,t}) < 10^{-10}$) in 2005.[12]

Panels A and B also plot a dotted line representing the annual mean of daily turnover. These lines, along with the $CPIE$ color scheme for the observed data, suggest that the PIN model mechanically identifies private information from turnover. To clarify this mechanical identification, Panels C and D plot $CPIE_{PIN}$ as function of turnover. Panels C and D show that the PIN model is essentially 'sure' that any day with turnover even slightly above a particular threshold (near the mean) is a private-information day (i.e. $CPIE_{PIN} = 1$). On the other hand, any day with turnover below this threshold is classified as a day with no private information (i.e. $CPIE_{PIN} = 0$). Note that this mechanical identification of private information does not necessarily relate to the possibility that informed traders may sometimes choose to trade on days with high liquidity or turnover (see Collin-Dufresne and Fos (2014)). Naturally, it is possible that informed traders do trade on some days with high turnover. However, our point is that the PIN model mechanically identifies almost *all* days with above average turnover as definitely private-information days and all days with below average turnover as definitely no private-information days.

Fig. 3 also highlights the intuition of why the PIN model mechanically associates turnover with private-information arrival. This conflation happens because of two limitations of the model. Note from Fig. 1 that when the informed traders receive no signal, they do not trade. Therefore, buy and sell orders arrive at the normal rate of noise trade and turnover is distributed as Poisson with intensity $\epsilon_B + \epsilon_S$. If the informed receive a signal, they join the noise traders in placing orders. Hence, turnover is distributed as Poisson with intensity $\epsilon_B + \epsilon_S + \mu$. Thus, under the PIN model, private-information arrival is necessarily the cause of any increase in expected daily turnover. Second, the size of the simulated clusters relative to the amount of variation in actual buys and sells indicates that the model's assumption of a

[12]The consequences of this problem for the likelihood maximization and $CPIE_{PIN}$ calculation are discussed in Section 2.1.

mixture of Poisson distributions, which have equal mean and variance, cannot accommodate the large variance of turnover that we see in the data. Thus, turnover on any given day tends to appear, to the model, to be either extremely high or extremely low (i.e. outside the simulated clusters in Panels A and B of Fig. 3). As a result of these two limitations, the model treats the vast majority of days when turnover is larger than the model expects as private-information days and most of the days when turnover is smaller than the model expects as no private-information days. That is, $CPIE_{PIN}$ mimics a dummy variable that is equal to one when turnover is above some threshold (near the mean) and zero otherwise.

### 2.2.2 $CPIE_{PIN}$ as a function of turnover

Fig. 3 shows the intuition behind the PIN model's mechanical identification of private-information events for one stock. In this section, we show that the problem is widespread. To do so, we first introduce the Mechanical model. The Mechanical model treats any day with above (below) average turnover as a private-information (no private-information) day:

$$CPIE_{Mech,j,t} = \begin{cases} 0, & \text{if } turn_{j,t} < \overline{turn}_j \\ 1, & \text{if } turn_{j,t} \geq \overline{turn}_j, \end{cases} \tag{7}$$

where $\overline{turn}_j$ is the average daily turnover computed over the same sample period as we used to compute the PIN model parameters.

To compare time series variation in $CPIE_{PIN}$ with variation in $CPIE_{Mech}$, we run the regression $CPIE_{PIN,j,t} = \beta_{0,j} + \beta_{1,j} \times CPIE_{Mech,j,t} + \varepsilon_{j,t}$ for each stock-year $j$ in the sample. For each stock $j$ and day $t$, we calculate $CPIE_{PIN,j,t}$ and $CPIE_{Mech,j,t}$ using data and estimates of the PIN model parameters for the entire calendar year containing day $t$. Naturally, market makers and traders do not have all of this information on day $t$. Therefore $CPIE_{PIN,j,t}$ and $CPIE_{Mech,j,t}$ cannot be used to set prices or conduct trading strategies. However, they are useful to gauge the similarity between the PIN model and a mechanical model of private-information arrival. Such an assessment is important to researchers who do observe order flow, PIN model parameters, and turnover over their entire sample period and thus can construct both measures for use in their work.

The results in Table 3 show that $CPIE_{PIN}$ is very closely approximated by the Mechanical dummy. Note that since $CPIE_{Mech}$ is a dummy variable, the intercept ($\beta_{0,j}$) in the

regression is the expected value of $CPIE_{PIN}$ when turnover is below the mean. Similarly, the sum of the coefficients $(\beta_{0,j} + \beta_{1,j})$ is the expected value of $CPIE_{PIN}$ when turnover is above the mean. The coefficient estimates reveal that for days with turnover below the mean ($CPIE_{Mech} = 0$), the median stock's $CPIE_{PIN}$ is close to zero, around 0.02 in 1993 and 0.12 in 2012. In contrast, for days with turnover above the mean ($CPIE_{Mech} = 1$), $CPIE_{PIN}$ for the median stock is 0.66 (0.64 +0.02) in 1993, rising to 0.95 (0.12+0.83) in 2012. Furthermore, the median $R^2$ is 58% in 1993, rising to nearly 70% in 2012. The average median $R^2$ across years is close to 65%.

The $R^2$s in Table 3 also allow us to examine how pervasive the mechanical conflation of private-information arrival with turnover is in the cross section. Stocks with the lowest (highest) $R^2$s are those for which variation in $CPIE_{Mech}$ explains the least (most) variation in $CPIE_{PIN}$. To assess how this conflation varies in the cross section, we select six stocks whose $R^2$s are at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles in 1993 and 2012. The values of these $R^2$s are indicated in Table 3. The six stocks are BXG (Bluegreen Corp.), EDBR (Edison Brothers Stores Inc.), TEK (Tektronix.), JWN (Nordstrom Inc.), MLM (Martin Marietta Materials Inc.), and VZ (Verizon Communications Inc.).

Fig. 4 presents plots of $CPIE_{PIN}$ as a function of turnover for all six stocks. Panel A plots $CPIE_{PIN}$ as function of turnover for the stock at the $5^{th}$ percentile in 1993 (BXG). BXG is among the stocks for which $CPIE_{PIN}$ is least well described by the Mechanical dummy. Even so, the PIN model assigns a probability larger than 99% to the arrival of private information if turnover is above 47 trades and assigns a probability smaller than 2 basis points to any day with turnover less than 24 trades. This covers about 85% of the trading days on which BXG traded in 1993. Panel B plots $CPIE_{PIN}$ as function of turnover for the stock at the $5^{th}$ percentile in 2012 (JWN). For JWN, any day with turnover below 8,751 trades is assigned a $CPIE_{PIN}$ of zero and any day with turnover above 10,088 trades is assigned a $CPIE_{PIN}$ of one. This covers 85% of JWN's trading days in 2012. The plots for the stocks at the $50^{th}$ and $95^{th}$ percentiles are even more striking, particularly in 2012.

Table 3 and Fig. 4 indicate that $CPIE_{PIN}$ is very well approximated by $CPIE_{Mech}$, not only for Exxon-Mobil, but also throughout the cross section. The approximation, however, is not perfect. Therefore, a natural question is whether, despite the high $R^2$s in Table

3, $CPIE_{Mech}$ oversimplifies the relation between $CPIE_{PIN}$ and turnover. To address the possibility of a more complicated, non-linear relation between $CPIE_{PIN}$ and $turn$, we regress $CPIE_{PIN}$ on $turn$, $turn^2$, and $CPIE_{Mech}$.

Panel A of Table 4 displays the results of these regressions. The coefficients on $CPIE_{Mech}$, $turn$ and $turn^2$ for the median stock are in general significant at the 1% level. However, it is important to note that the interpretation of the coefficients ($\beta_0$ and $\beta_1$) from Table 3 does not carry over to Table 4 because $CPIE_{Mech}$ is, by construction, mechanically related to $turn$ and $turn^2$. That is, $\beta_0$ is no longer the expected value of $CPIE_{PIN}$ when turnover is less than its mean and the sum of the coefficients $\beta_0 + \beta_1$ is no longer the expected value of $CPIE_{PIN}$ when turnover is greater than its mean. Moreover, the standardization of the variables in Table 4 forces the intercept to zero. As such, we focus on the difference in the $R^2$s across Tables 3 and 4, which tells us the contribution of $turn$ and $turn^2$ relative to $CPIE_{Mech}$ in explaining variation in $CPIE_{PIN}$. Specifically, the average median $R^2$ across all of the stock-years in Table 3 is close to 65%, while the average $R^2$ in Panel A of Table 4 is 73%. This small difference of 8% in the average $R^2$s indicates that $turn$ and $turn^2$ add little to the explanatory power of $CPIE_{Mech}$, a simple dummy variable based on turnover.

Our interpretation for the high $R^2$s in Panel A of Table 4 is that the PIN model mistakenly identifies all variation in turnover due to disagreement, calendar effects, portfolio rebalancing, and taxation as private-information arrival. However, one objection to this interpretation is that while turnover varies for many reasons unrelated to the arrival of private information, turnover *can* vary with the arrival of private information.

To address this objection, Panel B of Table 4 shows the results from regressions including a series of control variables that are correlated with turnover and plausibly related to the arrival of private information. To come up with a list of such variables, we look to the OWR and PIN models for guidance. Specifically, the PIN model suggests that the daily absolute order flow imbalance ($|B - S|$) is related to private-information arrival.[13] Moreover, recall that the OWR model identifies private-information arrival from the interactions between order imbalance, intra-day, and overnight returns, as well as their variances. That is, the

[13]We also control for ($|B - S|^2$) to address any potential non-linearities in the relation between $|B - S|$ and $CPIE_{PIN}$.

OWR model suggests that the squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$ and $r_o \times y_e$) vary with private-information arrival. All of these variables are also plausibly related to turnover. Therefore, if the relation between $CPIE_{PIN}$ and turnover is simply due to the fact that the PIN model captures the portion of variation in turnover that happens to be due to the arrival of private information, then including these controls in the regressions should attenuate the coefficient estimates and increase the $R^2$s from those in Panel A of Table 4. The results in Panel B indicate that this is not the case. In fact these controls increase the average $R^2$ for the median stock by only 4% over the 73% average $R^2$ in Panel A.

Overall, our results strongly support the conclusion that the PIN model mechanically identifies the arrival of private information from turnover. Even though the PIN model is based on the theoretical implication that periods of informed trade can be identified by abnormally large absolute order flow imbalance, empirically the PIN model violates this notion. In fact, the simple Mechanical dummy explains most of the variation in $CPIE_{PIN}$.

## 2.3 Does the PIN model produce reliable inferences?

To demonstrate that the PIN model's inferences are unreliable due to the conflation of turnover with private information, we compare $CPIE_{PIN}$ with $CPIE_{Mech}$ in two settings from the literature on private information: opportunistic insider trades and price reversals. First, we show that $CPIE_{PIN}$ identifies opportunistic insider trades in the same way as $CPIE_{Mech}$ using the insider trade classification scheme developed in Cohen, Malloy, and Pomorski (2012).[14] There is a large literature that suggests that insiders may have private information and may trade on that information.[15] Cohen, Malloy, and Pomorski (2012) show that a long-short portfolio that exploits the trades of opportunistic traders (opportunistic buys minus opportunistic sells) earns value-weighted abnormal returns of 82 basis points per month (9.8 percent annualized, t-statistic=2.15). They also show that opportunistic insiders' trades show significant predictive power for future news about the firm, and that

[14]See Section 1 for a further discussion of the classification of insider trades as opportunistic.

[15]See for instance Jaffe (1974), Seyhun (1986, 1998), Rozeff and Zaman (1988), Lin and Howe (1990), Bettis, Vickery, and Vickery (1997), Lakonishok and Lee (2001), Kahle (2000), Ke, Huddart, and Petroni (2003), Piotroski and Roulstone (2005), Jagolinzer (2009).

the fraction of opportunistic insiders in a given month is negatively related to the number of recent news releases by the SEC regarding illegal insider trading cases. Opportunistic insider trades therefore provide a convenient laboratory to show the consequences of the PIN model's conflation of turnover with private-information arrival.

Unlike a standard event study, we focus on variation in $CPIE$ rather than price movements. We estimate the $PIN$ parameter vector, $\Theta_{PIN,j}$, in the period $t \in [-312, -60]$ before the event and then compute daily $CPIE$s based on market data for the period $t \in [-20, 20]$ surrounding the event. Prior studies (e.g Benos and Jochec (2007)) estimate the parameters of the model in various windows around an event in order to compute $PIN$. Our procedure is different in that we estimate the parameters of the PIN and the Mechanical model one year prior to the event and then employ the estimated parameters as econometricians observing the daily market data (i.e. buys and sells) and attempting to infer whether a private-information event occurred.

Fig. 5 shows the average $CPIE_{PIN}$ and $CPIE_{Mech}$ in event time for our sample of opportunistic insider trades. The graph shows that the pattern in $CPIE_{PIN}$ around insider trades is nearly identical to that of $CPIE_{Mech}$. With both models, the probability of private-information arrival increases prior to the event, starting below 59% 20 days before the announcement, peaking around 68% on the day the insider executes the trade and slowly decreasing to 60% 10 days after the event. Indeed, the correlation between the average $CPIE_{PIN}$ and average $CPIE_{Mech}$ in event time is over 99%.

A researcher unaware that the PIN model conflates turnover with private-information arrival might view the pattern in $CPIE_{PIN}$ in Fig. 5 as indicative of the actual pattern of private-information arrival around insider trades. However, the *identical* pattern in $CPIE_{Mech}$ suggests otherwise. The average $CPIE_{Mech}$ on a particular day in event time is, by construction, the fraction of stocks with turnover above the estimation period mean daily turnover. Consequently, $CPIE_{Mech}$ says nothing about the arrival of private information *per se*. For instance, $CPIE_{Mech}$ slowly decreases after insider trades simply because daily turnover tends to slowly decline after opportunistic insiders trade. Therefore, the same pattern in $CPIE_{PIN}$ around the event simply reflects the PIN model's mechanical identification of private-information arrival from turnover. Thus, far from suggesting actual variation

in the probability of informed trade, the results in Fig. 5 indicate that the PIN model produces inferences that are no more reliable than those from a simple heuristic that assigns probability one (zero) to the arrival of private information when turnover is high (low).

Second, we examine the relation between $CPIE_{PIN}$, $CPIE_{Mech}$ and price reversals. The market microstructure literature has long held that price changes related to informed trades should be permanent, while non-information related price changes (e.g. those related to dealer inventory control, price pressure, price discreteness, etc.) should be transient (e.g. Hasbrouck (1988, 1991a,b)). Therefore, we examine the relation between $CPIE_{PIN}$ and return autocorrelations. Specifically, we consider the following regression:

$$r_{j,t+1} = \alpha + \beta_1 r_{j,t} + \beta_2 CPIE_{j,t} + \beta_3 (r_{j,t} \times CPIE_{j,t}) + \beta_4 (r_{j,t} \times X_{j,t}) + \beta_5 X_{j,t} + \upsilon_{j,t+1}. \quad (8)$$

In the above, $r_{j,t}$ is the open-to-open, risk-adjusted return ($r_{j,d,t} + r_{j,o,t}$) on day $t$ for stock $j$, $X_{j,t}$ is a vector of variables related to turnover ($CPIE_{Mech}$, $turn$, and $turn^2$), and $CPIE_{j,t}$ is either $CPIE_{PIN,j,t}$ or $CPIE_{Mech,j,t}$. These $CPIE$s are estimated using stock $j$ data for each calendar year. We estimate the regressions above using a panel regression approach.

Table 5 reports the results of these regressions. The negative coefficients on $r_{j,t}$ in each of the regressions in Table 5 shows a tendency of daily returns to reverse. The coefficient $\beta_3$ measures the effect of the model's $CPIE$ on the correlation between the return on day $t$ and the return on the next trading day. The first two columns of Table 5 show that the estimates for $\beta_3$ are positive and significant for both $CPIE_{PIN}$ and $CPIE_{Mech}$. This suggests that both $CPIE$s are associated with smaller future return reversals.

A researcher unaware of the conflation of turnover and the arrival of private information in the PIN model could interpret the fact that $\beta_3$ is positive and significant in Column 1 as evidence that the PIN model captures the arrival of private information that has a persistent impact on prices. However, the third and forth columns of Table 5 show that the damping effect of $CPIE_{PIN}$ on return reversals disappears once we include $CPIE_{Mech}$, $turn$, and $turn^2$. This is problematic for many reasons, but particularly because turnover increases upon public news and, unlike liquidity shocks, public news events should not be associated with return reversals. Thus, as with insider trades, the PIN model performs no better in the return reversals context than a simplistic heuristic.

In sum, the results in this section indicate that any proxy for private information based on the PIN model, including $PIN$, necessarily produces misleading results. To see this, note that the $PIN$ of a stock is $\alpha\mu/(\alpha\mu + \epsilon_B + \epsilon_S)$. Recall from Section 2.1 that $\alpha = E[CPIE_{PIN,t}]$. Hence, $CPIE_{PIN}$ and $PIN$ are linked via the unconditional probability of private-information arrival, $\alpha$. Consequently, given our results that $CPIE_{PIN}$ is no more reliable than a simplistic heuristic, any proxy for private information based on the PIN model (e.g. $PIN$ or $\alpha$) is equally unreliable.

# 3 An alternative to the PIN model

This section analyzes the OWR model. Section 3.1 describes the model and Section 3.2 uses the methods in Section 2.2 to analyze the OWR model.

## 3.1 Description and estimation of the OWR model

OWR extend Kyle (1985) to allow for days with and without private-information arrival. Fig. 6 shows a time line for the events in the model. Under the OWR model, private information arrives before the opening of the trading day with probability $\alpha$. On days when private information arrives, the model assumes that the information is publicly revealed after the close of trade. Econometricians can make inferences about the probability of private-information arrival under the OWR model because the covariance matrix of the three variables ($y_e$, $r_d$, $r_o$) differs between days with and without private-information arrival.[16]

To see how the covariance matrix of ($y_e$, $r_d$, $r_o$) differs between private-information and no private-information days, consider the covariance of the intra-day and overnight returns. This covariance is positive on days with private-information arrival, reflecting the fact that the information event is not completely captured in prices during the day. Thus, the revelation of the private information after the close causes the overnight return to continue the partial intra-day price reaction. In contrast, the covariance of the intra-day and overnight returns is negative in the absence of private-information arrival since the market marker's reaction to the noise trade during the day is reversed overnight when she learns that there was no

[16]Unlike the market maker who must update prices before observing the overnight revelation of information, econometricians using the OWR model can make inferences about the arrival of private information after viewing the overnight price response.

private signal. The intuition for why the other elements of the covariance matrix of ($y_e$, $r_d$, $r_o$) differ between private-information and no private-information days is similar.

Formally, let $\Theta_{OWR,j} = (\alpha_j, \sigma_{z,j}, \sigma_{u,j}, \sigma_{i,j}, \sigma_{p,d,j}, \sigma_{p,o,j})$ be the vector of OWR parameters for stock $j$. The parameter $\alpha_j$ is the unconditional probability of private-information arrival on any given day for stock $j$; $\sigma^2_{z,j}$ is the variance of the noise in the observed order imbalance ($y_{e,j}$); $\sigma^2_{u,j}$ is the variance of the order imbalance from noise traders; $\sigma^2_{i,j}$ is the variance of the private signal received by the informed traders; $\sigma^2_{p,d,j}$ is the variance of the public news component of the intra-day return; $\sigma^2_{p,o,j}$ is the variance of the public news component of the overnight return. Let $D_{OWR,j,t} = [\Theta_{OWR,j}, y_{e,j,t}, r_{d,j,t}, r_{o,j,t}]$ be the vector of model parameters along with the order imbalance and the intra-day as well as overnight returns. The likelihood function on a day without private-information arrival is $L_{NI}(D_{OWR,j,t}) = (1-\alpha_j)N(0, \Sigma_{NI,j})$, where $N(0, \Sigma_{NI,j})$ is the normal density with mean zero and covariance matrix:

$$\Sigma_{NI,j} = \begin{bmatrix} \sigma^2_{u,j} + \sigma^2_{z,j} & \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2} & \frac{-\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2} \\ \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2} & \sigma^2_{p,d,j} + \frac{\alpha_j\sigma^2_{i,j}}{4} & \frac{-\alpha_j\sigma^2_{i,j}}{4} \\ \frac{-\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2} & \frac{-\alpha_j\sigma^2_{i,j}}{4} & \sigma^2_{p,o,j} + \frac{\alpha_j\sigma^2_{i,j}}{4} \end{bmatrix} \tag{9}$$

On the other hand, the likelihood function on a day with private-information arrival is $L_I(D_{OWR,j,t}) = \alpha_j N(0, \Sigma_{I,j})$, where the covariance matrix $\Sigma_{I,j}$ is:

$$\Sigma_{I,j} = \begin{bmatrix} (1+\frac{1}{\alpha_j})\sigma^2_{u,j} + \sigma^2_{z,j} & \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2}(1+\frac{1}{\alpha_j}) & \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2}(\frac{1}{\alpha_j}-1) \\ \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2}(1+\frac{1}{\alpha_j}) & \sigma^2_{p,d,j} + \frac{(1+\alpha_j)\sigma^2_{i,j}}{4} & \frac{(1-\alpha_j)\sigma^2_{i,j}}{4} \\ \frac{\alpha_j^{1/2}\sigma_{i,j}\sigma_{u,j}}{2}(\frac{1}{\alpha_j}-1) & \frac{(1-\alpha_j)\sigma^2_{i,j}}{4} & \sigma^2_{p,o,j} + \frac{(1+\alpha_j)\sigma^2_{i,j}}{4} \end{bmatrix} \tag{10}$$

Let $I_{j,t}$ be an indicator function with value one when private information arrives on day $t$ for stock $j$. As was the case for the PIN model, $CPIE_{OWR,j,t}$ is the econometrician's posterior probability of private-information arrival given the model parameters and the data observed on that day ($y_{e,j,t}, r_{d,j,t}, r_{o,j,t}$). That is, $CPIE_{OWR,j,t} = P[I_{j,t} = 1 | D_{OWR,j,t}]$. Bayes' theorem implies that $CPIE_{OWR,j,t}$ is given by:

$$CPIE_{OWR,j,t} = \frac{L_I(D_{OWR,j,t})}{L_I(D_{OWR,j,t}) + L_{NI}(D_{OWR,j,t})} \tag{11}$$

In the absence of order flow and return data, an econometrician would assign a probability $\alpha_j = E[CPIE_{OWR,j,t}]$ to the arrival of private information for stock $j$ on day $t$, where the expectation is taken with respect to the joint distribution of the data vector ($y_{e,j,t}, r_{o,j,t}, r_{d,j,t}$).

As with the PIN, we estimate the OWR model numerically via maximum likelihood. Specifically, we maximize $\prod_{t=1}^{T} L(D_{OWR,j,t})$, where $L(D_{OWR,j,t})$ is the sum of $L_{NI}(D_{OWR,j,t})$ and $L_I(D_{OWR,j,t})$. In contrast to the PIN model, we do not encounter any numerical issues in directly computing either $L(D_{OWR,j,t})$ or $CPIE_{OWR}$ with Equation 11.

Table 6 contains summary statistics for the OWR parameter estimates and $CPIE_{OWR}$. As with the PIN model, we see from Table 6 that the mean $CPIE_{OWR}$ behaves like $\alpha$ in the OWR model. Fig. 7 plots the time series of the estimated OWR $\alpha$.[17] We also estimate the OWR model for each stock $j$ in the period $t \in [-312, -60]$ before opportunistic insider trades. These parameter estimates are used to compute $CPIE_{OWR}$ in our event study using insider trades. The summary statistics for the parameter estimates used in the insider trade event study are similar to those in Table 6.

## 3.2 Assessing the OWR model

Table 7 presents results from time-series regressions of $CPIE_{OWR}$ on $CPIE_{Mech}$. In contrast to Table 3, the results in Table 7 show that $CPIE_{OWR}$ is very poorly approximated by the Mechanical dummy. Recall that since $CPIE_{Mech}$ is a dummy variable, the intercept ($\beta_0$) in this regression is the expected value of $CPIE_{OWR}$ when turnover is below its mean. Similarly, the sum of the coefficients ($\beta_0 + \beta_1$) is the expected value of $CPIE_{OWR}$ when turnover is above its mean. For most years, $\beta_0$ varies between 0.2 and 0.6, while $\beta_1$ varies between 0.01 and 0.07. Thus, on days with high turnover, the average $CPIE_{OWR}$ is *not* substantially higher than on days with low turnover. Furthermore, the median $R^2$s are low, around 3% on average. These results indicate that the OWR model does not mechanically conflate turnover and private-information arrival.

As with the PIN model, we also sort stocks based on their $R^2$s in Table 7. The stocks with the lowest (highest) $R^2$s are those for which variation in $CPIE_{Mech}$ explains the least (most) of variation in $CPIE_{OWR}$. Fig. 8 presents plots of $CPIE_{OWR}$ as function of turnover for six stocks with $R^2$s at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles in 1993 and 2012. The values of these $R^2$s are indicated in Table 7. The six stocks are Cilcorp Inc (CER), Alexander

[17] Note that the estimated OWR $\alpha$ parameters are in general higher than those in OWR. This is due to the fact that our definition of $y_e$ is different from that in OWR (see discussion in Section 1 above). In fact, we get $\alpha$ estimates close to those reported in OWR if we define $y_e$ in the same way that they do.

& Alexander Services (AAL), Alza Corp (AZA), Pioneer Natural Resources (PXD), Eagle Materials (EXP), and T N S Inc (TNS). None of the plots in Fig. 8 reveal any apparent mechanical relation between turnover and $CPIE_{OWR}$.

Panel A of Table 8 shows the results from regressions including $turn$ and $turn^2$. In contrast to the results in Table 4, $CPIE_{Mech}$, $turn$ and $turn^2$ explain little of the variation in $CPIE_{OWR}$. Indeed, the average $R^2$ in Panel A is around 7%. Panel B presents the results of regressions controlling for the squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order flow imbalance ($y_e^2$), three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$ and $r_o \times y_e$), $|B - S|$ and its square. In contrast to the results in Table 4, the inclusion of the control variables dramatically increases the $R^2$s. Indeed, the $R^2$s in Panel B rise to around 85%, on average, after including the controls. In contrast to the PIN model, this suggests that turnover plays little role in identifying private-information arrival under the OWR model. Instead, the identification comes from the variables that the OWR model suggests are related to private-information arrival.

In sum, the results in Tables 7 and 8 yield no evidence that inferences from the OWR model are necessarily unreliable. To gain further insight into the OWR model's performance, we consider two additional working hypotheses in the context of opportunistic insider trades and of return reversals.

Consider first the relation between $CPIE_{OWR}$ and opportunistic insider trades. Under the working hypothesis that opportunistic insiders trade up to the point that prices reveal their information, $CPIE$s should be higher coincident with opportunistic trades and decline after the trades. Therefore, we examine $CPIE_{OWR}$ around opportunistic insider trades ($t \in [-20, 20]$). As in Section 2.3, these are based on estimates of $\Theta_{OWR,j}$ computed in the period $t \in [-312, -60]$ before the event.

Fig. 9 presents the average $CPIE_{OWR}$ and $CPIE_{Mech}$ in event time for our sample of opportunistic insider trades. Interestingly, $CPIE_{OWR}$ rises a few days before the insider trades, suggesting that whatever private signal the insider is responding to is also received and acted upon by others. Most importantly, unlike $CPIE_{Mech}$ (and thus $CPIE_{PIN}$), $CPIE_{OWR}$ drops dramatically immediately after the trade. Hence, variation in $CPIE_{OWR}$ is consistent with the idea that opportunistic insiders trade up to the point that prices fully

reveal private information.

Next we examine the relation of $CPIE_{OWR}$ with future return reversals. Specifically, under the working hypothesis that private-information arrival is associated with weaker price reversals, a credible model of private-information arrival should have a $CPIE$ that is associated with smaller future price reversals. Therefore, we run Regression 8, using $CPIE_{OWR}$. Before continuing, however, there are two issues worth clarifying. First, recall that the independent variable in this regression is the open-to-open, risk-adjusted return ($r_{j,t} = r_{d,j,t}+r_{o,j,t}$) on day $t$. Thus, there is no overlap between the intra-day and overnight returns that are used to compute $CPIE_{OWR,j,t}$ on day $t$ and the return on day $t+1$. This is important because if $CPIE_{OWR,j,t}$ and $r_{j,t+1}$ were computed using overlapping data, then the relation between them would be mechanical. Second, while the OWR model relies on $r_{d,j,t} \times r_{o,j,t}$ to identify private-information arrival, it is a one period model and has no predictions about the relation between $CPIE_{OWR,j,t}$ and the correlation between $r_{j,t}$ and $r_{j,t+1}$. That is, the OWR model has implications about the relation between private-information arrival and the covariance between intra-day and subsequent overnight return ($r_{d,j,t}$ and $r_{o,j,t}$), but it has no implication whatsoever about the covariance between the daily returns $r_{j,t}$ and $r_{j,t+1}$. Thus, for the regressions in this section we rely on our working hypothesis to yield implications for the effect of private-information arrival on the covariance between the daily returns $r_{j,t}$ and $r_{j,t+1}$, not on the OWR model *per se*.

Table 9 reports the coefficient estimates and t-statistics for these regressions. Most importantly, the results in Table 9 show that the estimate for $\beta_3$ in the OWR model is positive and significant, indicating that $CPIE_{OWR}$ is associated with smaller future return reversals. Indeed, a one standard deviation shock to $CPIE_{OWR}$ is associated with a 27% (2.416/8.881) decline in the subsequent reversal. For completeness, we also report the results using $CPIE_{Mech}$. As expected, the $\beta_3$ estimates for $CPIE_{Mech}$ are very different from that for $CPIE_{OWR}$. Furthermore, controlling for $CPIE_{Mech}$ as well as $turn$ and $turn^2$ does not significantly change $\beta_3$ estimates for $CPIE_{OWR}$. This, along with the results in Table 8, suggests that $CPIE_{OWR}$ does not simply capture the effect of turnover (say due to public news, for instance) on return reversals. On the contrary, the OWR model appears to capture the arrival of private information with persistent impact on prices.

# 4 Conclusion

Our findings indicate that the PIN model mechanically groups *all* sources of variation in turnover (e.g. disagreement, calendar effects, portfolio rebalancing, taxation, etc.) under the umbrella of private-information arrival. Indeed, our results indicate that the PIN model is no more useful in identifying private information than a mechanical model that assigns probability one to the arrival of private information on days when turnover is above average and zero to the arrival of private information on any other day. These findings indicate that the most widely used measure of information asymmetry, $PIN$, is not reliable.

We also examine an alternative to the PIN model, the OWR model, which infers the arrival of private information from returns and order flow. We find that the OWR model does not mechanically identify private information from turnover. Furthermore, we present results that suggest that the OWR model is able to capture the arrival of private information in the context of opportunistic insider trades, and of return reversals. These results indicate that the OWR model is a promising alternative to the PIN model. That being said, our OWR results come with the caveat that they depend on potentially controversial working hypotheses about the timing of the arrival of private information. Therefore, future research examining the OWR model in different contexts is needed to make definitive conclusions about the OWR model's ability to identify private information.

# References


Admati, Anat R., and Paul Pfleiderer, 1988, A theory of intraday patterns: Volume and price variability, *Review of Financial Studies* 1, 3–40.

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.

Bennett, Benjamin, Gerald Garvey, Todd Milbourn, and Zexi Wang, 2017, Managerial compensation and stock price informativeness, *Working paper*.

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

Bettis, Carr, Don Vickery, and D.W. Vickery, 1997, Mimickers of corporate insiders who make large-volume trades, *Financial Analysts Journal* 53, 57–66.

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

Chen, Qi, Itay Goldstein, and Wei Jiang, 2007, Price informativeness and investment sensitivity to stock price, *Review of Financial Studies* 20, 619–650.

Cohen, Lauren, Christopher Malloy, and Lukasz Pomorski, 2012, Decoding inside information, *Journal of Finance* 67, 1009–1043.

Collin-Dufresne, Pierre, and Vyacheslav Fos, 2014, Insider trading, stochastic liquidity and equilibrium prices, *National Bureau of Economic Research Working paper*.

——— , 2015, Do prices reveal the presence of informed trading?, *Journal of Finance* 70, 1555–1582.

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

Ferreira, Daniel, Miguel A. Ferreira, and Carla C. Raposo, 2011, Board structure and price informativeness, *Journal of Financial Economics* 99, 523–545.

Gan, Quan, Wang C. Wei, and David J. Johnstone, 2014, Does the probability of informed trading model fit empirical data?, *FIRN Research Paper*.

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.

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

——— , 1991a, Measuring the information content of stock trades, *Journal of Finance* 46, 179–207.

——— , 1991b, The summary informativeness of stock trades, *Review of Financial Studies* 4, 571–594.

Jaffe, Jeffrey F., 1974, Special information and insider trading, *The Journal of Business* 47, 410–428.

Jagolinzer, Alan D., 2009, SEC rule 10b5-1 and insiders' strategic trade, *Management Science* 55, 224–239.

Kahle, Kathleen M., 2000, Insider trading and the long-run performance of new security issues, *Journal of Corporate Finance* 6, 25–53.

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

Ke, Bin, Steven Huddart, and Kathy Petroni, 2003, What insiders know about future earnings and how they use it: Evidence from insider trades, *Journal of Accounting and Economics* 35, 315–346.

Kim, Oliver, and Robert E. Verrecchia, 1994, Market liquidity and volume around earnings announcements, *Journal of Accounting and Economics* 17, 41–67.

——— , 1997, Pre-announcement and event-period private information, *Journal of Accounting and Economics* 24, 395–419.

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.

Lakonishok, Josef, and Inmoo Lee, 2001, Are insiders' trades more informative?, *Review of Financial Markets* 14, 79–111.

Lakonishok, Josef, and Seymour Smidt, 1986, Volume for winners and losers: Taxation and other motives for stock trading, *The Journal of Finance* 41, 951–973.

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

Lin, Ji-Chai, and John S. Howe, 1990, Insider trading in the OTC market, *Journal of Finance* 45, 1273–1284.

Lo, Andrew W., and Jiang Wang, 2000, Trading volume: Definitions, data analysis, and implications of portfolio theory, *Review of Financial Studies* 13, 257–300.

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, 1997, *Market Microstructure Theory* (Blackwell Publishing).

Piotroski, Joseph D., and Darren. T. Roulstone, 2005, Do insider trades reflect contrarian beliefs and superior knowledge about cash flow realizations?, *Journal of Accounting and Economics* 39, 55–81.

Rozeff, Michael S., and Mar A. Zaman, 1988, Market efficiency and insider trading: New evidence, *Journal of Business* 61, 25–44.

Seyhun, H. Nejat, 1986, Insiders' profits, costs of trading, and market efficiency, *Journal of Financial Economics* 16, 189–212.

——— , 1998, *Investment Intelligence from Insider Trading* (MIT Press).

Stickel, Scott E., and Robert E. Verrecchia, 1994, Evidence that trading volume sustains stock price changes, *Journal of Finance* 50, 57–67.

Table 1: **Summary Statistics.** This table summarizes the full sample and opportunistic insider trading day returns, order imbalance as well as the number of buys (B) and sells (S). 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, CRSP and COMPUSTAT. Following OWR, we compute the intraday return 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 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 intra-day ($r_d$) and overnight ($r_o$) returns are risk adjusted using daily cross-sectional regressions of each return measure on a constant, historical beta (based on the previous five years of monthly returns), the natural logarithm of market capitalization, and the natural logarithm of the book-to-market ratio. We compute the order imbalance ($y_e$) as the daily total volume of buys minus total volume of sells, divided by the total volume. Our sample of opportunistic insider trades is constructed using the method detailed in Cohen, Malloy, and Pomorski (2011).

(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% |
| $B$ | 5,286,191 | 1,876 | 6,917 | 37 | 220 | 1,128 |
| $S$ | 5,286,191 | 1,843 | 6,894 | 36 | 194 | 1,033 |

(b) Opportunistic Insider Trades

| | N | Mean | Std | Q1 | Median | Q3 |
|---|---|---|---|---|---|---|
| $y_e$ | 32,676 | 4.980% | 20.425% | -5.106% | 3.874% | 15.353% |
| $r_d$ | 32,676 | 0.151% | 1.566% | -0.632% | 0.086% | 0.865% |
| $r_o$ | 32,676 | 0.056% | 1.247% | -0.467% | 0.020% | 0.528% |
| $B$ | 32,676 | 3,852 | 10,645 | 354 | 1,129 | 3,478 |
| $S$ | 32,676 | 3,787 | 10,554 | 300 | 996 | 3,303 |

Table 2: **PIN Model Parameter Estimates.** This table summarizes the parameter estimates for the PIN model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\delta$ represents the probability of good news, and $1 - \delta$ represents the probability of bad news. The parameters $\epsilon_B$ and $\epsilon_S$ represent the expected number of daily buys and sells given no private information, and $\mu$ is the expected increase in the number of trades given the arrival of private information. $CPIE_{PIN}$ is the probability of private-information arrival on a particular day, conditional on the PIN model parameters and the observed buys and sells. $\overline{CPIE}$ and Std($CPIE$) are the mean and standard deviation of $CPIE_{PIN}$ computed for each firm-year. In the table below, we report the mean, standard deviation, first, second, and third quartiles for each parameter, $\overline{CPIE}$ and Std($CPIE$) across all firm-years.

| | N | Mean | Std | Q1 | Median | Q3 |
|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.372 | 0.122 | 0.291 | 0.375 | 0.445 |
| $\delta$ | 21,206 | 0.607 | 0.209 | 0.484 | 0.625 | 0.762 |
| $\epsilon_B$ | 21,206 | 1,625 | 5,388 | 33 | 193 | 1,039 |
| $\epsilon_S$ | 21,206 | 1,596 | 5,369 | 35 | 186 | 956 |
| $\mu$ | 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 |

Table 3: **Regressions of $CPIE_{PIN}$ on the Mechanical Dummy.** This table reports results from the regression: $CPIE_{PIN,j,t} = \beta_0 + \beta_1 \times CPIE_{Mech,j,t} + \varepsilon_{j,t}$, where $CPIE_{Mech,j,t}$ is a dummy variable which is one if stock $j$'s turnover on day $t$ is greater than the mean daily turnover of stock $j$ during the calendar year, and zero otherwise. We report median coefficient and $t$-statistic estimates (in parentheses) as well as the $5^{th}$, $50^{th}$, $95^{th}$ percentiles of $R^2$ for each year in our sample. We compute Newey-West standard errors with a lag length selected according to the Akaike Information Criterion (AIC) from a regression of $CPIE_{PIN}$ on a constant, trend, and quadratic trend.

| | $\beta_0$ | | $\beta_1$ | | $R^2$ | | |
|---|---|---|---|---|---|---|---|
| | | | | | $5^{th}$ | $50^{th}$ | $95^{th}$ |
| 1993 | 0.019 | (2.67) | 0.638 | (14.66) | 29.18% | 57.99% | 75.20% |
| 1994 | 0.022 | (2.93) | 0.644 | (15.32) | 28.87% | 58.74% | 74.66% |
| 1995 | 0.017 | (2.58) | 0.635 | (14.56) | 26.08% | 57.00% | 74.21% |
| 1996 | 0.021 | (2.68) | 0.656 | (15.74) | 29.64% | 59.07% | 73.73% |
| 1997 | 0.019 | (2.51) | 0.668 | (15.65) | 31.49% | 59.41% | 77.26% |
| 1998 | 0.018 | (2.36) | 0.685 | (16.38) | 35.37% | 61.47% | 79.79% |
| 1999 | 0.022 | (2.44) | 0.693 | (16.50) | 32.63% | 61.33% | 78.41% |
| 2000 | 0.018 | (2.14) | 0.721 | (17.06) | 34.27% | 63.44% | 82.20% |
| 2001 | 0.033 | (2.42) | 0.776 | (19.77) | 49.31% | 67.02% | 81.75% |
| 2002 | 0.040 | (2.59) | 0.812 | (22.38) | 50.69% | 70.17% | 83.84% |
| 2003 | 0.045 | (2.77) | 0.802 | (21.53) | 51.96% | 68.41% | 81.38% |
| 2004 | 0.041 | (2.63) | 0.818 | (22.71) | 54.63% | 70.34% | 83.12% |
| 2005 | 0.050 | (2.88) | 0.841 | (24.67) | 53.07% | 72.42% | 85.66% |
| 2006 | 0.065 | (3.27) | 0.856 | (26.99) | 34.54% | 74.08% | 88.21% |
| 2007 | 0.191 | (5.54) | 0.792 | (21.74) | 23.17% | 61.03% | 89.14% |
| 2008 | 0.166 | (5.20) | 0.808 | (23.30) | 19.65% | 64.15% | 92.00% |
| 2009 | 0.127 | (4.35) | 0.823 | (23.93) | 24.57% | 68.11% | 91.98% |
| 2010 | 0.122 | (4.53) | 0.842 | (25.64) | 23.67% | 69.35% | 89.89% |
| 2011 | 0.173 | (5.33) | 0.797 | (21.75) | 19.37% | 62.00% | 89.76% |
| 2012 | 0.119 | (4.45) | 0.831 | (25.06) | 24.49% | 69.09% | 88.81% |

Table 4: **Regressions of $CPIE_{PIN}$ on $CPIE_{Mech}$ Including Other Variables.** Panel A reports results from regressions of $CPIE_{PIN}$ on $CPIE_{Mech}$, $turn$, and $turn^2$. The variables $turn$ and $turn^2$ represent the daily sum of buys and sells and its square, respectively. Panel B reports results from regressions of $CPIE_{PIN}$ on $CPIE_{Mech}$, $turn$, $turn^2$, and additional controls: $|B-S|$, $|B-S|^2$, $y_e^2$, $r_d^2$, $r_o^2$, $y_e \times r_d$, $y_e \times r_o$, and $r_d \times r_o$. The variables $B$ and $S$ represent the daily number of buys and sells respectively. The variables $r_d$ and $r_o$ are the intra-day and overnight return respectively. The variable $y_e$ is the order imbalance. We report median estimates, $t$-statistics (in parentheses), and $R^2$ values for each year in our sample. All coefficients are multiplied by 100. All variables are standardized. We compute Newey-West standard errors with a lag length selected according to the AIC from a regression of $CPIE_{PIN}$ on a constant, trend, and quadratic trend.

| | A. No Controls | | | | | | | B. With Controls | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $CPIE_{Mech}$ | | $turn$ | | $turn^2$ | | $R^2$ | $CPIE_{Mech}$ | | $turn$ | | $turn^2$ | | $R^2$ |
| 1993 | 31.7 | (3.96) | 83.7 | (4.89) | -35.2 | (-2.59) | 70.46% | 31.0 | (4.58) | 64.1 | (3.71) | -22.6 | (-1.55) | 76.68% |
| 1994 | 32.1 | (4.10) | 82.7 | (4.90) | -31.9 | (-2.30) | 71.36% | 31.7 | (4.79) | 63.0 | (3.75) | -20.3 | (-1.34) | 77.29% |
| 1995 | 31.3 | (4.01) | 82.4 | (4.90) | -34.8 | (-2.46) | 70.03% | 30.5 | (4.49) | 61.6 | (3.63) | -21.0 | (-1.36) | 75.39% |
| 1996 | 35.0 | (4.49) | 79.9 | (4.83) | -32.4 | (-2.49) | 70.19% | 34.4 | (5.11) | 60.8 | (3.71) | -20.9 | (-1.40) | 75.69% |
| 1997 | 35.5 | (4.39) | 79.2 | (4.74) | -33.6 | (-2.66) | 70.33% | 35.3 | (4.93) | 61.7 | (3.75) | -22.2 | (-1.56) | 75.74% |
| 1998 | 38.4 | (4.76) | 76.1 | (4.43) | -31.6 | (-2.44) | 71.25% | 37.4 | (5.37) | 58.4 | (3.53) | -21.9 | (-1.54) | 76.56% |
| 1999 | 38.4 | (4.80) | 79.6 | (4.82) | -38.0 | (-2.99) | 70.83% | 38.1 | (5.35) | 64.8 | (3.92) | -29.3 | (-2.10) | 75.75% |
| 2000 | 41.8 | (5.09) | 77.9 | (4.72) | -36.8 | (-3.09) | 71.95% | 40.9 | (5.66) | 60.5 | (3.78) | -26.1 | (-2.11) | 76.41% |
| 2001 | 50.0 | (5.98) | 66.0 | (4.14) | -30.7 | (-2.70) | 73.04% | 49.4 | (6.81) | 48.4 | (3.32) | -21.5 | (-1.76) | 77.97% |
| 2002 | 54.8 | (6.73) | 62.2 | (3.90) | -30.7 | (-2.70) | 74.89% | 54.4 | (7.51) | 48.3 | (3.33) | -22.8 | (-1.85) | 78.84% |
| 2003 | 55.6 | (7.09) | 59.8 | (3.91) | -31.6 | (-2.87) | 72.94% | 54.8 | (7.65) | 48.8 | (3.40) | -25.4 | (-2.19) | 76.65% |
| 2004 | 57.7 | (7.27) | 61.0 | (3.86) | -33.6 | (-3.02) | 74.38% | 57.2 | (7.85) | 51.9 | (3.46) | -28.8 | (-2.41) | 77.82% |
| 2005 | 59.9 | (7.66) | 64.7 | (3.96) | -38.6 | (-3.41) | 76.15% | 59.5 | (8.18) | 55.4 | (3.68) | -35.0 | (-2.95) | 79.18% |
| 2006 | 61.1 | (7.85) | 64.8 | (4.08) | -40.1 | (-3.59) | 77.38% | 61.0 | (8.36) | 58.1 | (3.93) | -37.8 | (-3.35) | 79.80% |
| 2007 | 36.5 | (4.51) | 120.1 | (5.22) | -87.4 | (-4.23) | 70.79% | 36.0 | (4.47) | 119.6 | (5.37) | -89.5 | (-4.47) | 73.05% |
| 2008 | 42.7 | (5.25) | 107.1 | (5.16) | -76.4 | (-4.11) | 72.85% | 42.3 | (5.33) | 100.5 | (5.12) | -73.7 | (-4.24) | 74.65% |
| 2009 | 51.0 | (6.20) | 81.7 | (4.61) | -56.4 | (-3.91) | 74.11% | 50.1 | (6.48) | 75.6 | (4.51) | -53.9 | (-3.87) | 76.56% |
| 2010 | 54.1 | (6.90) | 80.0 | (4.60) | -54.3 | (-3.87) | 74.19% | 53.8 | (7.28) | 73.8 | (4.47) | -54.8 | (-3.87) | 77.06% |
| 2011 | 42.6 | (5.17) | 104.4 | (5.05) | -76.4 | (-4.09) | 70.66% | 42.0 | (5.36) | 101.4 | (5.07) | -76.8 | (-4.27) | 73.45% |
| 2012 | 52.9 | (6.60) | 78.1 | (4.53) | -53.2 | (-3.85) | 74.07% | 52.3 | (6.73) | 70.7 | (4.37) | -50.3 | (-3.80) | 76.82% |

Table 5: **Return Reversal Regressions with $CPIE_{PIN}$.** This table reports regressions of the daily return at time $t+1$ on the return at time $t$, $CPIE$ ($CPIE_{PIN}$ or $CPIE_{Mech}$), and their interaction. Returns are measured from open to open as the sum of the intraday ($r_d$) and overnight returns ($r_o$). $CPIE$s are standardized. We include stock and year fixed effects and cluster standard errors by stock and year. Coefficients are multiplied by 100, and $t$-statistics are reported in parentheses. Stars indicate the statistical significance of the coefficient estimates at the 10, 5, and 1% levels respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| $r_t$ | $-7.235^{***}$ | $-7.330^{***}$ | $-7.327^{***}$ | $-7.328^{***}$ |
| | $(-6.757)$ | $(-6.914)$ | $(-6.884)$ | $(-6.885)$ |
| $CPIE_{PIN}$ | $0.021^{***}$ | | $0.011^{***}$ | $0.011^{***}$ |
| | (4.960) | | (3.944) | (4.228) |
| $CPIE_{Mech}$ | | $0.021^{***}$ | $0.013^{***}$ | $0.014^{***}$ |
| | | (5.110) | (3.753) | (4.040) |
| $CPIE_{PIN} \times r_t$ | $0.610^{**}$ | | $-0.162$ | $-0.248$ |
| | (2.277) | | $(-0.478)$ | $(-0.822)$ |
| $CPIE_{Mech} \times r_t$ | | $0.882^{***}$ | $1.007^{***}$ | $1.032^{***}$ |
| | | (3.218) | (2.811) | (2.904) |
| $turn$ | | | | $-0.018^{***}$ |
| | | | | $(-3.069)$ |
| $turn^2$ | | | | $0.0002^{*}$ |
| | | | | (1.877) |
| $turn \times r_t$ | | | | $0.386^{*}$ |
| | | | | (1.928) |
| $turn^2 \times r_t$ | | | | $-0.006^{**}$ |
| | | | | $(-2.078)$ |
| $R^2$ | 0.55% | 0.55% | 0.55% | 0.56% |
| Observations | 5,283,617 | 5,283,617 | 5,283,617 | 5,283,617 |

Table 6: **OWR Parameter Estimates.** This table summarizes the parameter estimates for the OWR model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\sigma_u$ represents the standard deviation of the order imbalance due to uninformed traders, which is observed with normally distributed noise with variance $\sigma_z^2$. The parameter $\sigma_i$ is the standard deviation of the informed trader's private signal, while $\sigma_{pd}$ and $\sigma_{po}$ are the standard deviations of the public news component of the idiosyncratic intraday and overnight returns, respectively. $CPIE_{OWR}$ is the probability of private-information arrival on a particular day, conditional on the OWR model parameters and the observed market data. $\overline{CPIE}$ and Std($CPIE$) represent the mean and standard deviation of $CPIE_{OWR}$ computed for each firm-year. In the table below, we report the mean, standard deviation, first, second, and third quartiles for each parameter, $\overline{CPIE}$ and Std($CPIE$) across all firm-years.

| | N | Mean | Std | Q1 | Median | Q3 |
|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.437 | 0.257 | 0.214 | 0.436 | 0.639 |
| $\sigma_u$ | 21,206 | 0.075 | 0.068 | 0.022 | 0.062 | 0.109 |
| $\sigma_z$ | 21,206 | 0.239 | 0.143 | 0.137 | 0.221 | 0.332 |
| $\sigma_i$ | 21,206 | 0.030 | 0.286 | 0.013 | 0.021 | 0.027 |
| $\sigma_{pd}$ | 21,206 | 0.010 | 0.005 | 0.006 | 0.009 | 0.012 |
| $\sigma_{po}$ | 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 7: **Regressions of $CPIE_{OWR}$ on the Mechanical Dummy.** This table reports results from the regression: $CPIE_{OWR,j,t} = \beta_0 + \beta_1 \times CPIE_{Mech,j,t} + \nu_{j,t}$, where $CPIE_{Mech,j,t}$ is a dummy variable which is one if stock $j$'s turnover on day $t$ is greater than the mean daily turnover of stock $j$ during the calendar year, and zero otherwise. We report median coefficient and $t$-statistic estimates (in parentheses) as well as the $5^{th}$, $50^{th}$, $95^{th}$ percentiles of $R^2$ for each year in our sample. We compute Newey-West standard errors with a lag length selected according to the AIC from a regression of $CPIE_{OWR}$ on a constant, trend, and quadratic trend.

| | | | | | $R^2$ | | |
|---|---|---|---|---|---|---|---|
| | $\beta_0$ | | $\beta_1$ | | $5^{th}$ | $50^{th}$ | $95^{th}$ |
| 1993 | 0.612 | (53.67) | 0.063 | (3.17) | 0.39% | 4.44% | 12.90% |
| 1994 | 0.608 | (53.90) | 0.056 | (2.89) | 0.28% | 3.58% | 12.17% |
| 1995 | 0.601 | (52.36) | 0.061 | (3.09) | 0.25% | 4.17% | 12.24% |
| 1996 | 0.581 | (51.68) | 0.061 | (3.03) | 0.31% | 4.02% | 11.85% |
| 1997 | 0.570 | (51.22) | 0.060 | (3.08) | 0.19% | 4.16% | 12.07% |
| 1998 | 0.501 | (46.07) | 0.076 | (3.56) | 0.74% | 5.85% | 13.88% |
| 1999 | 0.571 | (53.40) | 0.068 | (3.43) | 0.73% | 5.35% | 14.08% |
| 2000 | 0.610 | (62.82) | 0.054 | (3.08) | 0.18% | 4.36% | 13.27% |
| 2001 | 0.467 | (39.97) | 0.047 | (2.57) | 0.08% | 3.02% | 10.87% |
| 2002 | 0.480 | (46.02) | 0.050 | (2.69) | 0.11% | 3.14% | 10.00% |
| 2003 | 0.387 | (36.94) | 0.041 | (2.30) | 0.05% | 2.28% | 9.20% |
| 2004 | 0.295 | (33.13) | 0.042 | (2.42) | 0.05% | 2.72% | 8.91% |
| 2005 | 0.209 | (32.31) | 0.040 | (2.42) | 0.11% | 2.85% | 8.54% |
| 2006 | 0.248 | (31.57) | 0.034 | (2.06) | 0.02% | 1.99% | 7.06% |
| 2007 | 0.199 | (33.23) | 0.039 | (2.39) | 0.11% | 3.69% | 11.77% |
| 2008 | 0.294 | (28.69) | 0.009 | (0.57) | 0.02% | 1.68% | 9.56% |
| 2009 | 0.281 | (39.51) | 0.024 | (2.05) | 0.03% | 2.64% | 9.84% |
| 2010 | 0.201 | (34.60) | 0.022 | (1.92) | 0.03% | 2.30% | 8.92% |
| 2011 | 0.222 | (37.15) | 0.031 | (2.13) | 0.02% | 2.73% | 8.64% |
| 2012 | 0.162 | (31.39) | 0.024 | (1.86) | 0.02% | 1.90% | 8.12% |

Table 8: **Regressions of $CPIE_{OWR}$ on $CPIE_{Mech}$ Including Other Variables.** Panel A reports results from regressions of $CPIE_{OWR}$ on $CPIE_{Mech}$, $turn$, and $turn^2$. The variables $turn$ and $turn^2$ represent the daily sum of buys and sells and its square, respectively. Panel B reports results from regressions of $CPIE_{OWR}$ on $CPIE_{Mech}$, $turn$, $turn^2$, and additional controls: $|B-S|$, $|B-S|^2$, $y_e^2$, $r_d^2$, $r_o^2$, $y_e \times r_d$, $y_e \times r_o$, and $r_d \times r_o$. The variables $B$ and $S$ represent the daily number of buys and sells respectively. The variables $r_d$ and $r_o$ are the intra-day and overnight return respectively. The variable $y_e$ is the order imbalance. We report median coefficient, $t$-statistic (in parentheses), and $R^2$ values for each year of our sample. Coefficients are multiplied by 100. All variables are standardized. We compute Newey-West standard errors with a lag length selected according to the AIC from a regression of $CPIE_{OWR}$ on a constant, trend, and quadratic trend. Estimates less than 0.01 are indicated with '-'.

| | A. No Controls | | | | | | | B. With Controls | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $CPIE_{Mech}$ | | $turn$ | | $turn^2$ | | $R^2$ | $CPIE_{Mech}$ | | $turn$ | | $turn^2$ | | $R^2$ |
| 1993 | 1.79 | (0.19) | 32.93 | (1.29) | -10.08 | (-0.45) | 8.57% | 0.77 | (0.15) | 17.47 | (1.12) | -13.91 | (-0.93) | 72.56% |
| 1994 | 0.30 | (0.04) | 31.16 | (1.17) | -5.04 | (-0.27) | 7.29% | 0.68 | (0.15) | 13.81 | (0.96) | -10.53 | (-0.75) | 74.31% |
| 1995 | 1.61 | (0.17) | 27.85 | (1.08) | -4.96 | (-0.22) | 8.12% | 1.37 | (0.30) | 13.41 | (0.86) | -9.75 | (-0.65) | 74.83% |
| 1996 | 2.23 | (0.22) | 27.10 | (1.08) | -4.89 | (-0.26) | 7.94% | 0.81 | (0.19) | 15.71 | (1.07) | -12.27 | (-0.87) | 75.85% |
| 1997 | 0.79 | (0.09) | 27.66 | (1.08) | -2.61 | (-0.15) | 8.83% | 0.77 | (0.18) | 14.14 | (1.00) | -10.85 | (-0.81) | 77.11% |
| 1998 | 2.73 | (0.27) | 30.08 | (1.16) | -4.17 | (-0.18) | 10.66% | 1.14 | (0.30) | 13.27 | (1.00) | -11.18 | (-0.84) | 79.51% |
| 1999 | 1.57 | (0.16) | 38.36 | (1.45) | -11.34 | (-0.53) | 10.55% | 1.17 | (0.26) | 12.56 | (0.95) | -10.93 | (-0.87) | 78.61% |
| 2000 | 0.74 | (0.09) | 33.30 | (1.22) | -8.52 | (-0.42) | 8.55% | 0.53 | (0.17) | 9.66 | (0.70) | -8.72 | (-0.74) | 81.64% |
| 2001 | - | (-0.01) | 22.31 | (0.82) | -0.39 | (-0.03) | 6.82% | 0.07 | (0.04) | 6.09 | (0.63) | -5.73 | (-0.59) | 84.86% |
| 2002 | -0.23 | (-0.04) | 20.82 | (0.77) | 1.52 | (0.09) | 7.15% | - | (0.02) | 7.11 | (0.73) | -5.47 | (-0.63) | 85.92% |
| 2003 | -0.97 | (-0.12) | 13.78 | (0.55) | 7.78 | (0.35) | 6.65% | 0.19 | (0.08) | 4.81 | (0.57) | -3.28 | (-0.40) | 88.42% |
| 2004 | -1.46 | (-0.17) | 9.06 | (0.33) | 10.93 | (0.41) | 7.80% | - | (-0.01) | 4.92 | (0.56) | -2.65 | (-0.33) | 89.86% |
| 2005 | -1.89 | (-0.22) | 12.13 | (0.43) | 12.43 | (0.44) | 8.48% | -0.07 | (-0.05) | 2.48 | (0.30) | -1.54 | (-0.20) | 90.71% |
| 2006 | -1.75 | (-0.19) | 9.62 | (0.39) | 6.85 | (0.28) | 6.68% | 0.02 | (0.02) | 3.39 | (0.46) | -2.35 | (-0.34) | 90.51% |
| 2007 | -0.99 | (-0.10) | 6.75 | (0.26) | 9.91 | (0.40) | 8.86% | 0.27 | (0.11) | 0.53 | (0.07) | -0.40 | (-0.07) | 90.33% |
| 2008 | 1.71 | (0.17) | -0.06 | (-0.01) | 4.77 | (0.24) | 3.52% | 0.20 | (0.08) | -1.39 | (-0.19) | 0.56 | (0.08) | 89.70% |
| 2009 | -0.03 | (-0.01) | 9.12 | (0.38) | 5.31 | (0.26) | 5.87% | 0.12 | (0.06) | -1.43 | (-0.20) | 0.41 | (0.09) | 91.03% |
| 2010 | -0.94 | (-0.10) | 8.68 | (0.36) | 4.79 | (0.23) | 5.81% | -0.06 | (-0.04) | -1.54 | (-0.25) | 0.79 | (0.13) | 91.08% |
| 2011 | 0.07 | (0.01) | 13.40 | (0.46) | 0.36 | (0.02) | 5.98% | -0.05 | (-0.02) | -1.35 | (-0.19) | 0.58 | (0.08) | 91.60% |
| 2012 | -0.86 | (-0.09) | 11.56 | (0.45) | 1.53 | (0.07) | 5.38% | -0.14 | (-0.08) | -0.74 | (-0.12) | 0.21 | (0.05) | 91.68% |

Table 9: **Return Reversal Regressions with $CPIE_{OWR}$.** This table reports regressions of the daily return at time $t+1$ on the return at time $t$, $CPIE$ ($CPIE_{OWR}$ or $CPIE_{Mech}$), and their interaction. Returns are measured from open to open as the sum of the intraday ($r_d$) and overnight returns ($r_o$). $CPIE$s are standardized. We include stock and year fixed effects and cluster standard errors by stock and year. Coefficients are multiplied by 100, and $t$-statistics are reported in parentheses. Stars indicate the statistical significance of the coefficient estimates at the 10, 5, and 1% levels respectively.

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| $r_t$ | $-8.881^{***}$ | $-7.330^{***}$ | $-9.083^{***}$ | $-9.246^{***}$ |
| | $(-6.864)$ | $(-6.914)$ | $(-6.892)$ | $(-7.112)$ |
| $CPIE_{OWR}$ | $0.014^{***}$ | | $0.011^{***}$ | $0.011^{***}$ |
| | (4.368) | | (3.842) | (3.851) |
| $CPIE_{Mech}$ | | $0.021^{***}$ | $0.019^{***}$ | $0.021^{***}$ |
| | | (5.110) | (4.982) | (5.348) |
| $CPIE_{OWR} \times r_t$ | $2.416^{***}$ | | $2.347^{***}$ | $2.549^{***}$ |
| | (4.157) | | (4.163) | (4.575) |
| $CPIE_{Mech} \times r_t$ | | $0.882^{***}$ | $0.550^{**}$ | $0.415^{**}$ |
| | | (3.218) | (2.464) | (1.995) |
| $turn$ | | | | $-0.016^{***}$ |
| | | | | $(-2.748)$ |
| $turn^2$ | | | | 0.0001 |
| | | | | (0.992) |
| $turn \times r_t$ | | | | $0.936^{***}$ |
| | | | | (3.544) |
| $turn^2 \times r_t$ | | | | $-0.014^{***}$ |
| | | | | $(-3.788)$ |
| $R^2$ | 0.61% | 0.55% | 0.62% | 0.63% |
| Observations | 5,283,617 | 5,283,617 | 5,283,617 | 5,283,617 |

Figure 1: **PIN Model Tree.** For a given trading day, private information arrives with probability $\alpha$. When there is no private information, buys and sells are distributed as Poisson random variables with intensity $\epsilon_B$ and $\epsilon_S$. Private information is good (bad) news with probability $\delta$ $(1-\delta)$. The expected number of buys (sells) increases by $\mu$ in case of good (bad) news arrival.

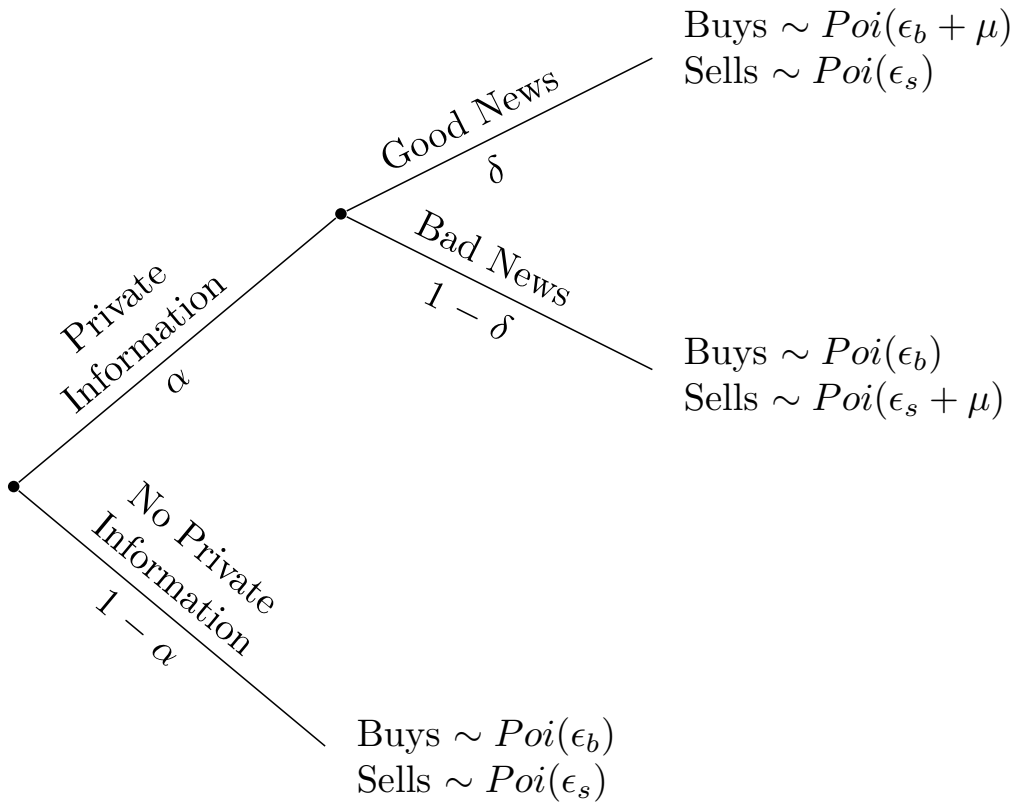


Figure 2: **Yearly $\alpha$ Parameter Estimates for the PIN Model.** The solid black line represents the median value, and the dashed lines represent the $5^{th}$, $25^{th}$, $75^{th}$, and $95^{th}$ percentiles.

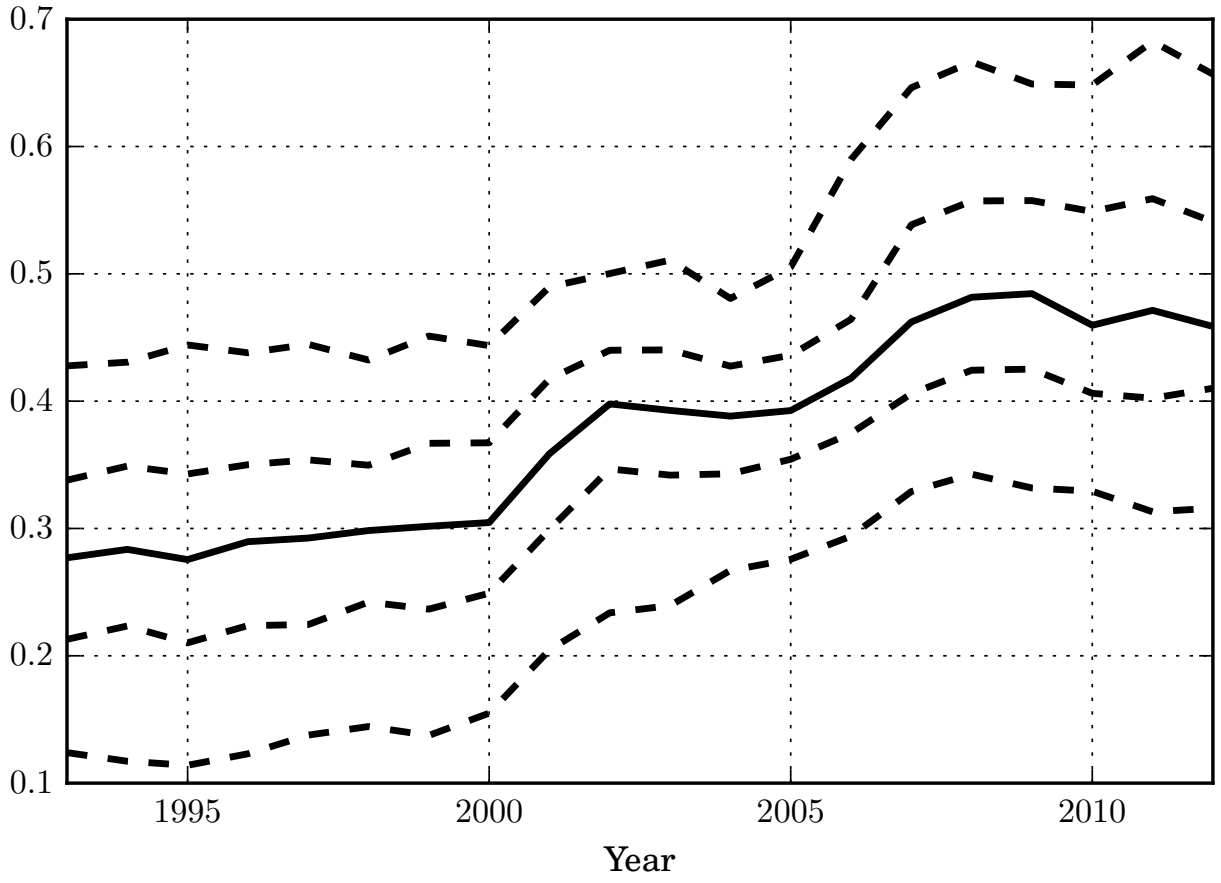
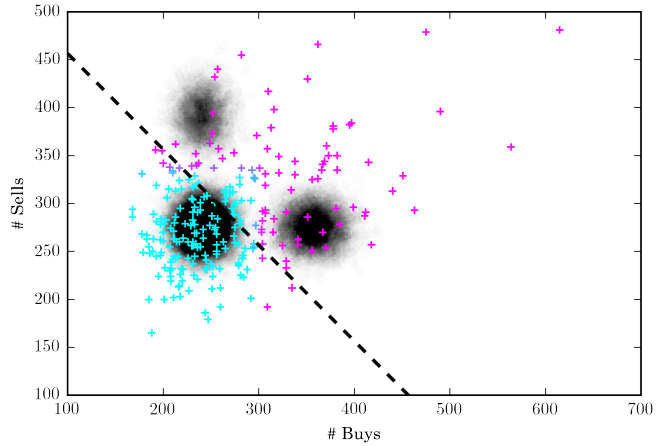
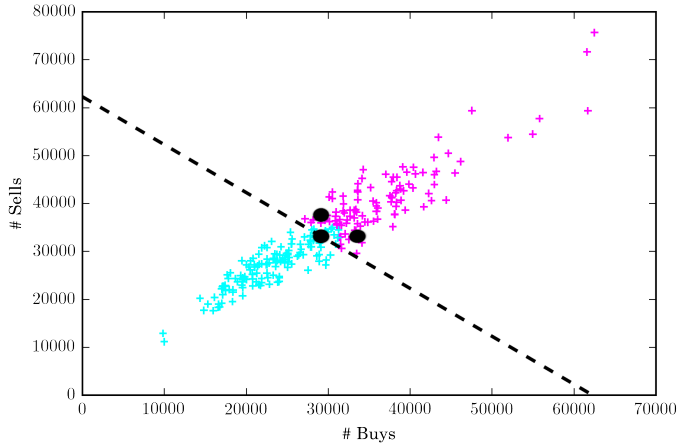


Figure 3: **PIN Model Example.** This figure compares real and simulated data for Exxon-Mobil (XOM) in 1993 and 2012 from the PIN model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{PIN}$, with darker markers (+ magenta) representing high and lighter markers (+ cyan) low $CPIE$s. All the observations below (above) the dashed lines have turnover below (above) the annual mean of daily turnover. High (low) probability states in the simulated data appear as a dark (light) "cloud" of points. The PIN model has three states: no news, good news, and bad news. Panels C and D plot the $CPIE$s for the real data as a function of turnover along with a dashed line indicating the mean turnover.

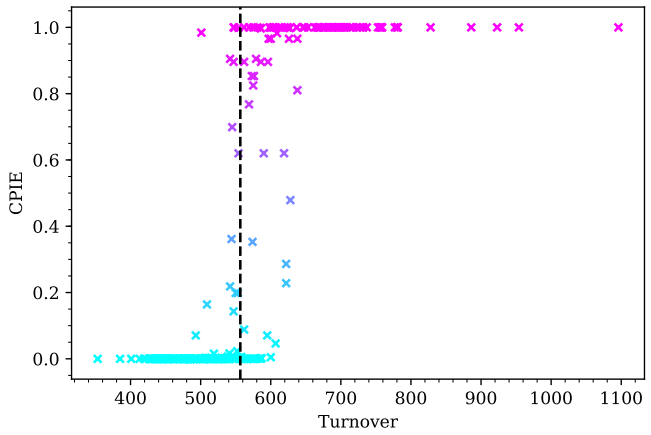
(a) XOM 1993




(b) XOM 2012




(c) XOM 1993




(d) XOM 2012

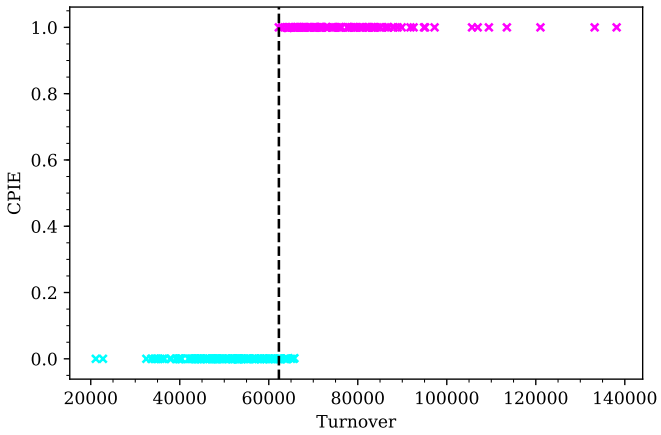
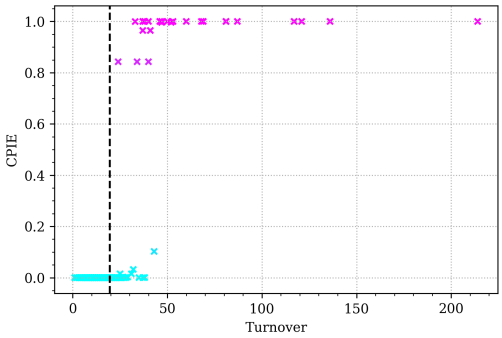
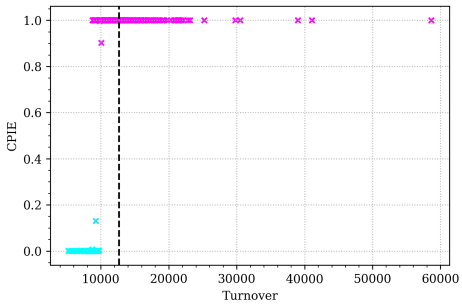


Figure 4: ***CPIE*$_{PIN}$ as Function of Turnover.** Panels (a), (c), and (e) report data from 1993 for the stocks at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles (BXG, EDBR, and TEK respectively) of $R^2$ from a regression of $CPIE_{PIN}$ on $CPIE_{Mech}$ (see Table 3). Panels (b), (d), and (f) show the same plots for the stocks at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles (JWN, MLM, and VZ respectively) of $R^2$ in 2012. The dotted lines represent the yearly mean of daily turnover.

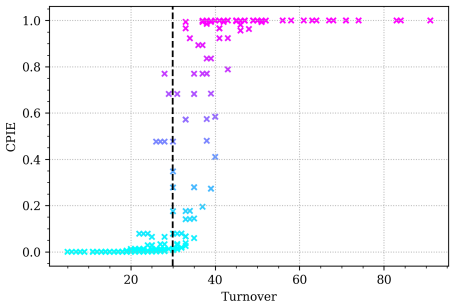
(a) BXG 1993 5%



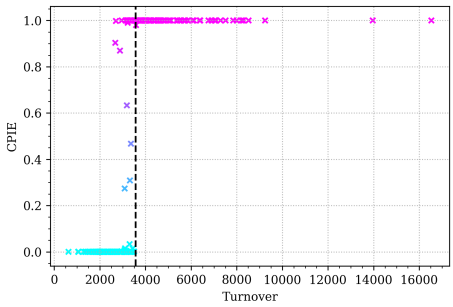

(b) JWN 2012 5%



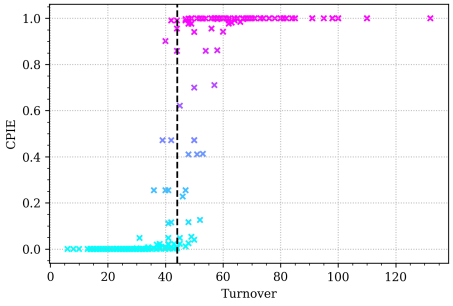

(c) EDBR 1993 50%




(d) MLM 2012 50%




(e) TEK 1993 95%




(f) VZ 2012 95%

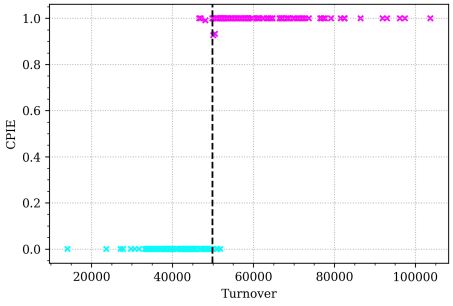


Figure 5: ***CPIE*$_{PIN}$ around Insider Trades.** The solid line is the average $CPIE_{PIN}$ in event time surrounding opportunistic insider trades. The dashed line is the average $CPIE_{Mech}$.

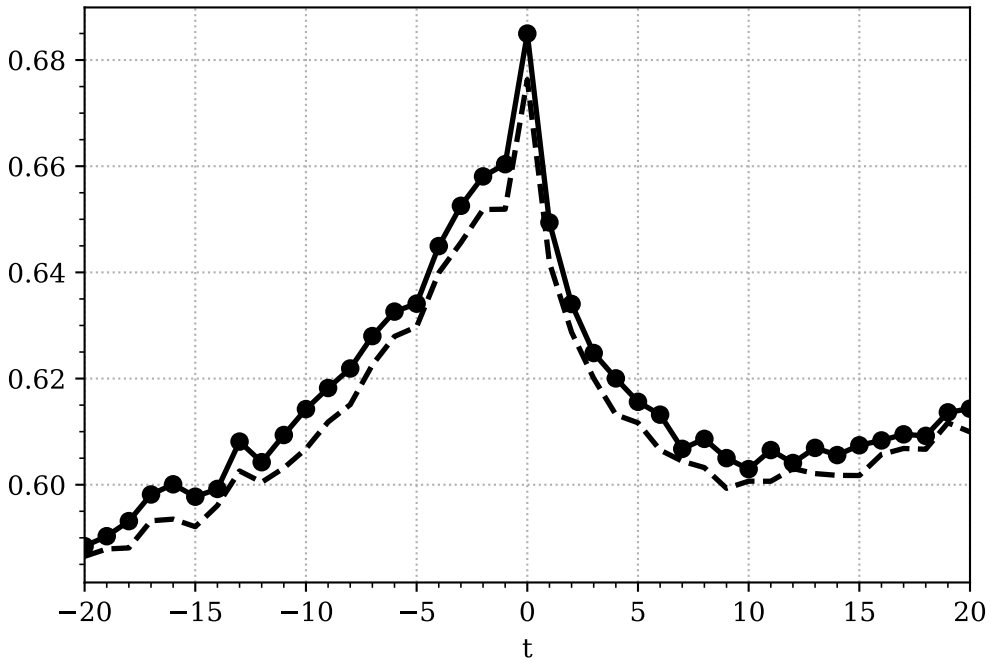


Figure 6: **OWR Model Tree.** In the OWR model, prior to markets opening, private information arrives with probability $\alpha$. Once markets open, investors submit their trades generating order imbalance ($y_e$), and the intraday return ($r_d$). After markets close, private information becomes public and is reflected in the overnight return ($r_o$). The variables ($y_e$, $r_d$, $r_o$) are normally distributed with mean zero. The covariance differs between days with private-information arrival, $\Sigma_I$, and days without the arrival of private information, $\Sigma_{NI}$. When there is no private-information arrival, there is a price reversal in the overnight return ($cov(r_d, r_o) < 0$) and when there is private-information arrival there is a continuation in the returns ($cov(r_d, r_o) > 0$).

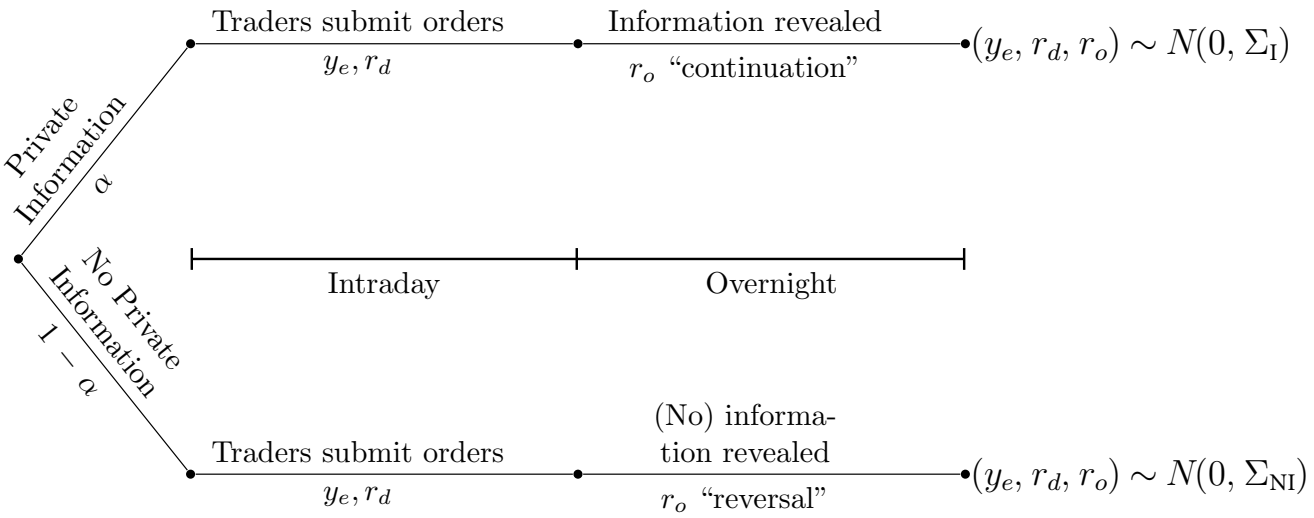


Figure 7: **Yearly $\alpha$ Parameter Estimates for the OWR Model.** The solid black line represents the median value, and the dashed lines represent the $5^{th}$, $25^{th}$, $75^{th}$, and $95^{th}$ percentiles.

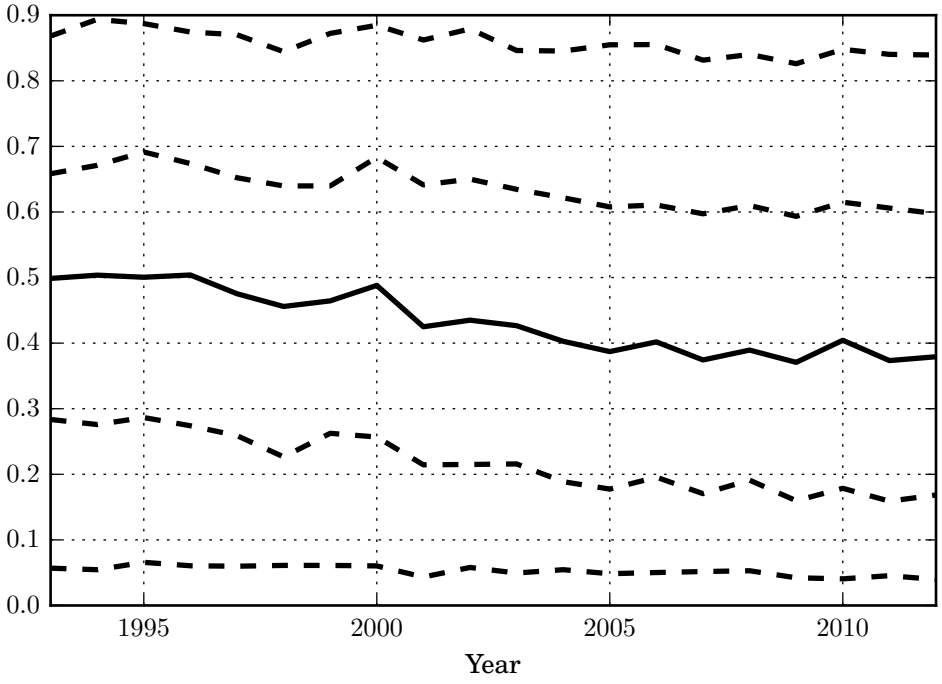
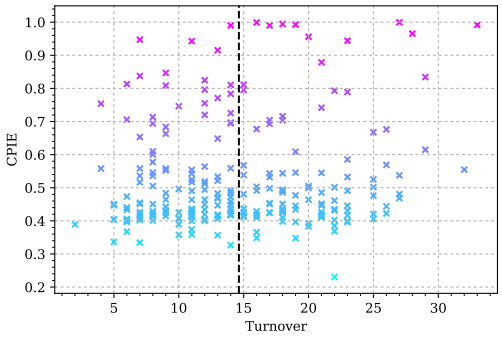
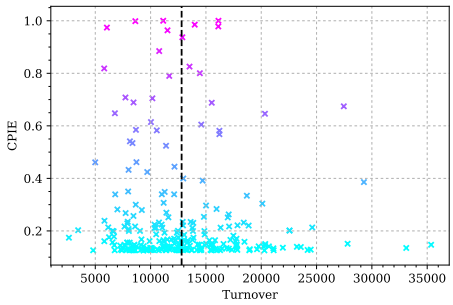


Figure 8: $\boldsymbol{CPIE_{OWR}}$ **as Function of Turnover.** Panels (a), (c), and (e) report data from 1993 for the stocks at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles (CER, AAL, and AZA respectively) of $R^2$ from a regression of $CPIE_{OWR}$ on $CPIE_{Mech}$ (see Table 7). Panels (b), (d), and (f) show the same plots for the stocks at the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles (PXD, EXP, and TNS respectively) of $R^2$ in 2012. The dotted lines represent yearly mean of daily turnover.

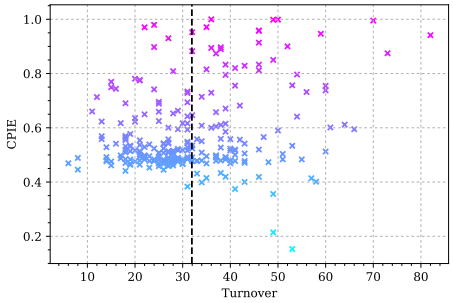
(a) CER 1993 5%



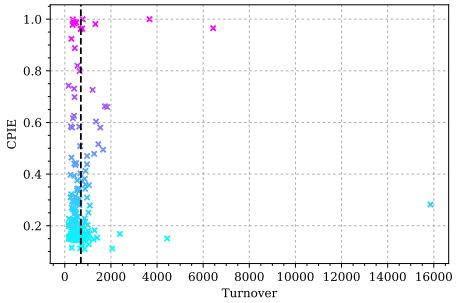

(b) PXD 2012 5%



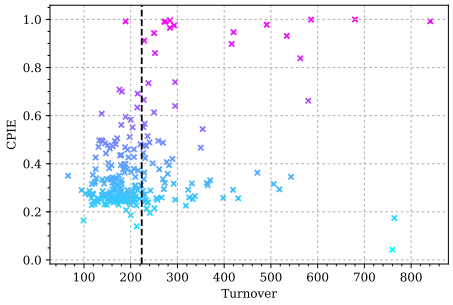

(c) AAL 1993 50%




(d) EXP 2012 50%




(e) AZA 1993 95%




(f) TNS 2012 95%

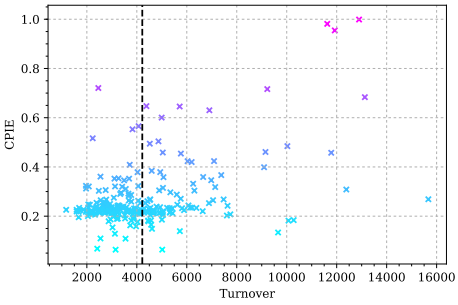
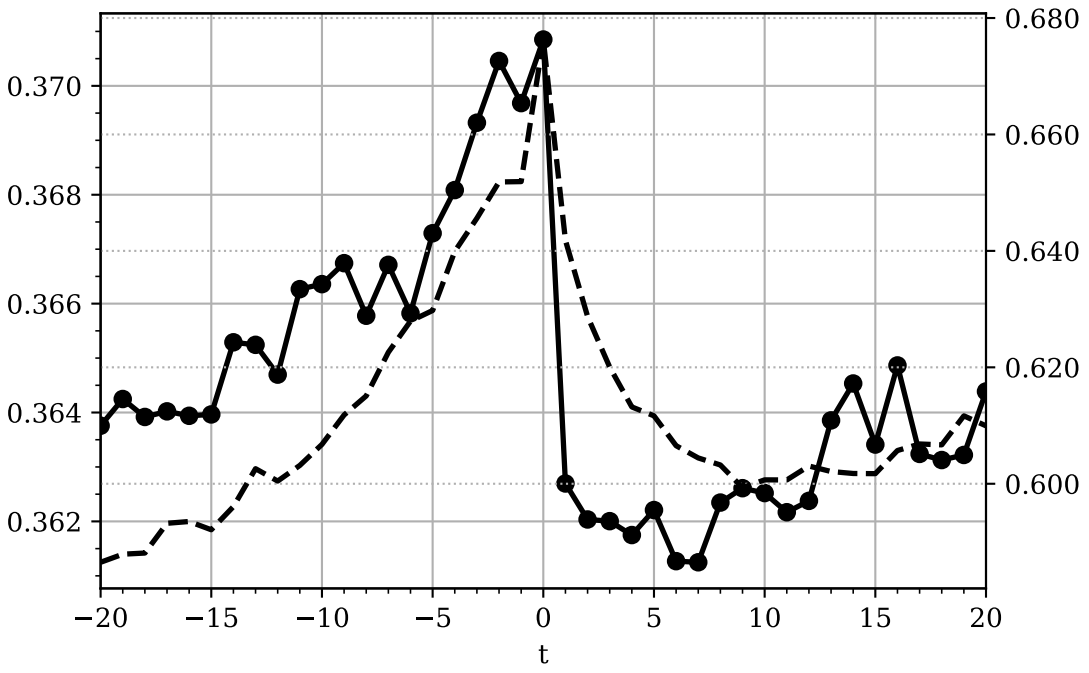


Figure 9: ***CPIE*$_{OWR}$ around Insider Trades.** The solid line is the average $CPIE_{OWR}$ in event time surrounding opportunistic insider trades. The dashed line is the average $CPIE_{Mech}$, plotted on the right-hand y-axis.



# Internet Appendix: Does the PIN model mis-identify private information and if so, what is the alternative?

September 7th, 2017

# A Estimating order flow, $r_{o,j,t}$ and $r_{d,j,t}$

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.

The OWR model requires intra-day and overnight returns. Following OWR we compute the intra-day return on day $t$ as the volume-weighted average price (VWAP) during the trading day $t$ minus the opening quote midpoint on day $t$ plus dividends issued on day $t$, all divided by the opening quote midpoint on day $t$. We compute the overnight return on day $t$ as the opening quote midpoint on day $t+1$ minus the VWAP on day $t$, all divided by the opening quote midpoint on day $t$. The opening quote midpoint is not available in TAQ in 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.

We follow OWR by removing systematic effects from returns to obtain measures of idiosyncratic overnight and intra-day returns ($r_{o,j,t}$ and $r_{d,j,t}$). To estimate $r_{o,j,t}$ and $r_{d,j,t}$, 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 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

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

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 ($r_{d,j,t}$) and overnight ($r_{o,j,t}$) returns.

# 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.

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.