Does the PIN model mis-identify private information and if so, what are our alternatives?[*]

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

March 10th, 2017

Abstract

We investigate whether the Easley and O'Hara (1987) PIN model mis-identifies private information from variation in turnover. We find 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 the PIN as a measure of private information since turnover varies for many reasons unrelated to private information arrival. We also examine two alternatives to the PIN model, the Generalized PIN model (GPIN) and the Odders-White and Ready (2008) model (OWR). Our tests do not reveal any problems with these two models' ability to identify private information, but indicate that the OWR model performs somewhat better.

Keywords: Liquidity; Information Asymmetry

[*] We thank Torben Andersen, Kerry Back, Pierre Collin-Dufresne, Kevin Crotty, Zhi Da, Robert Engle, Gustavo Grullon, Terry Hendershott, Sahn-Wook Huh, Jyri Kinnunen, Pete Kyle, Yelena Larkin, Avi Wohl and seminar participants at the 2015 ITAM Conference, 2015 Annual SoFiE Conference, 2015 Annual MFS Conference, 2015 CICF, 2016 AFA Conference, Rice University, Texas A&M University, University of Virginia (McIntire), and the University of Washington (Foster) for helpful comments. We thank Bei Dong, Edward X. Li, K. Ramesh, and Min Shen for earnings announcements time stamp data. 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 authors 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 accounting, corporate finance and asset pricing literatures as a measure of information asymmetry.[1] The PIN model is based on the notion, originally developed by Glosten and Milgrom (1985), that periods of informed trade can be identified by abnormally large absolute order flow imbalances.[2] Recently, however, several papers have documented the perhaps puzzling fact that PIN is higher after public news announcements than before (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), and Collin-Dufresne and Fos (2015)). All of these papers are informative in that they suggest potential problems with the PIN model. As yet, however, there remains no definitive test of the PIN model's ability to capture private information arrival because the arrival of private information is inherently unobservable. Thus, any test of a model of private information arrival is, in effect, a joint hypothesis test. For instance, in the context of earnings announcements, the PIN model has been tested using the working hypothesis that the arrival of private information is more likely before an earnings announcement than after it. It is possible however that agents convert public information into private signals using superior analysis (e.g. Kim and Verrecchia (1994, 1997)). In such a case, a higher PIN after an earnings announcement would indicate that PIN is properly capturing private information. Therefore, this a joint test of the efficacy of the model along with a test of the working hypothesis about the timing of the arrival of private information.

Our first research question is whether PIN mis-identifies the arrival private information. This is an important question because of the widespread use of PIN throughout the financial economics and accounting literature. To address this research question, we create a variable called the Conditional Probability of an Information Event ($CPIE$). To compute the $CPIE$

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

implied by the PIN model ($CPIE_{PIN}$), we estimate the PIN model's parameters using an entire year of data, and then use the observed market data (i.e. buys and sells) to estimate the posterior or model-implied probability of an information event for each day in our sample.

We then test the PIN model by examining whether $CPIE_{PIN}$ is mechanically driven by turnover, using the working hypothesis that turnover varies for myriad reasons unrelated to private information.[3] Our test is therefore also a joint hypothesis test. However, unlike the working hypotheses about the timing of the arrival of private information that have been used in the extant literature, the idea that turnover varies for reasons unrelated to private information arrival is uncontroversial. For instance, turnover can increase due to disagreement (e.g. Kandel and Pearson (1995), and Banerjee and Kremer (2010)). Turnover is also 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)).

We find that the PIN model *primarily* identifies information events based on turnover, controlling for absolute order flow imbalance. In regressions of $CPIE_{PIN}$ on absolute order imbalance, turnover, and their squared terms, turnover and turnover squared account for, on average, around 65% of the overall R^2. Two limitations of the PIN model combine to create this problem. First, under the PIN model, increases in expected turnover can only come about through the arrival of private information.[4] Second, the PIN model cannot 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 above average turnover as periods of private information arrival. The identification of information events from turnover becomes more pronounced late in the sample with the increase in both the level and variance of turnover. For example, after 2002, our results indicate that the most popular model of private information in the literature, the PIN model, yields inferences that are no more useful than simply looking at whether daily volume is above or below the mean to identify information events.

[3] Theoretically, the PIN model identifies periods of informed trade with abnormally large absolute order flow imbalances. Empirically, however, the PIN model may actually identify private information from turnover and not from order imbalance.

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

To demonstrate how the conflation of turnover with private information is consequential to the broader finance and accounting literature, we consider a setting from the literature that uses the PIN model in an event study context. Specifically, Benos and Jochec (2007) find that PIN is higher after earnings announcements than before. They interpret their findings as evidence that PIN fails to identify private information. In contrast, in a contemporaneous paper, Brennan, Huh, and Subrahmanyam (2015) find that a measure similar to $CPIE_{PIN}$ is higher after earnings announcements, and interpret their results as indicating that private information arrival is indeed higher after earnings announcements (and merger announcements).[5] Our event study findings resolve the impasse between the different interpretations of Benos and Jochec (2007) and Brennan, Huh, and Subrahmanyam (2015) because they indicate that higher PIN or $CPIE_{PIN}$ after earnings announcements can be simply attributed to the fact that turnover is typically much higher after earnings announcements.[6] This example is emblematic of a pervasive issue in the literature since the PIN model ultimately yields unclear inferences about private information arrival.

Despite the PIN model's mechanical conflation of turnover with information arrival, all is not lost in the quest for intuitive measures of information asymmetry based on structural models. The second main contribution of this paper is to analyze two alternatives to the PIN model that break the link between volume and private information. The first is a highly tractable generalization of the PIN model (the GPIN model) that we develop, which relies only on order flow to identify private information. The second is a model developed by Odders-White and Ready (2008) (the OWR model), which uses price impacts along with order flow to identify private information.

Even though there are many measures of private information in the literature, there are at least two reasons why it is interesting to focus on the OWR and the GPIN models.[7] First,

[5] While they do not consider earnings announcements, Aktas, de Bodt, Declerck, and Van Oppens (2007) show that PIN is higher after merger announcements. They interpret this finding as a failure of PIN to capture private information. In contrast to Benos and Jochec (2007) and Aktas, de Bodt, Declerck, and Van Oppens (2007) we use $CPIE_{PIN}$ to conduct this event study instead of PIN.

[6] The literature suggests that turnover remains high after earnings announcements for many reasons unrelated to the arrival of private information (e.g Bamber, Barron, and Stevens (2011)).

[7] There are many measures of private information that are not based on structural models (e.g. bid-ask spreads, impulse responses from structural VARs, and $VPIN$). For reasons of space, we focus on measures of private information that are based on structural models. Moreover, some of these measures have been carefully analyzed in the literature (e.g. Andersen and Bondarenko (2014)).

we focus on two structural models that do not conflate turnover and private information.[8] Second, our choices of alternatives to the PIN model include one that is based on order flow alone (the GPIN model) and another which uses order flows and returns (the OWR model). On one hand, Easley, Kiefer, and O'Hara (1997) emphasize the need for private information proxies that are computed using order flow alone.[9] On the other hand Back, Crotty, and Li (2014) and Kim and Stoll (2014) show evidence consistent with the idea that order imbalance alone does not reveal private information. Therefore our choices of alternatives to the PIN model span both branches of this literature.

As neither the GPIN and OWR models suffer from the mechanical conflation of turnover and private information arrival, to examine their performance we cannot rely on the working hypothesis that turnover varies for many reasons unrelated to private information arrival. We therefore use the GPIN and OWR $CPIE$s ($CPIE_{GPIN}$ and $CPIE_{OWR}$) to diagnose potential problems with the GPIN and OWR models' ability to identify private information arrival in the context of insider trades and price continuation. Cohen, Malloy, and Pomorski (2012) propose a method to identify instances of opportunistic insider trades. Their results show that these trades are profitable, suggesting they reveal private information. Therefore, one criterion to see if a model correctly identifies informed trade, is to examine the variation of its $CPIE$ around opportunistic trades. Specifically, under the working hypothesis that opportunistic insiders will trade up to the point that prices reveal their information, $CPIE$s should be higher coincident with opportunistic trades and decline after the trades. Furthermore, Hasbrouck (1988, 1991a,b) point out that non-information related price changes (e.g. dealer inventory control) should be subsequently reversed, while information related trades should not. Therefore, under the working hypothesis that private information arrival is associated with weaker price reversals, $CPIE$s should be associated with smaller future price reversals if the model properly identifies private information arrival.[10] Both of these tests rely on working hypotheses that are not as strongly established in the literature as the

[8] Easley, Engle, O'Hara, and Wu (2008) and Duarte and Young (2009) develop measures of private information based on structural models. We show in Internet Appendices A and B that these two models also identify private information from turnover in the later part of our sample period.

[9] Easley, Kiefer, and O'Hara (1997) regress prices on measures analogous to $CPIE_{PIN}$. They point out that their results are not mechanical because the PIN model is based only on order flow.

[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.

hypothesis that turnover varies for reasons unrelated to private information. Thus, these tests cannot be considered definitive. However, they are informative because they can at least suggest problems with the models' ability to identify private information.

Our results suggest that the OWR model performs in our tests somewhat better than the GPIN model. The superior performance of the OWR model is perhaps not surprising given that it uses returns and order flow data. In their totality, however, our results suggest that the GPIN model is a promising alternative to the PIN model that relies on order flow alone. On the other hand, if relying on order flow alone is not a requirement, then measures of private information based on the OWR model are promising alternatives to PIN.

Collin-Dufresne and Fos (2015) show that PIN and other measures of adverse selection are lower when Schedule 13D fillers trade and conclude that these measures may fail to capture informed trading when informed traders can select when and how to trade. The failure of the PIN model that we document is more general than that in Collin-Dufresne and Fos (2015) because we show that the PIN model suffers from a pervasive problem that extends well beyond failing to identify informed trade under certain circumstances. Another group of papers in this literature have shown that the PIN model does not fit the order flow data well. For instance, Gan, Wei, and Johnstone (2014) show that the distribution of order flow used in 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 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 – private information arrival. We contribute to this group of papers because our tests focus on a direct measure of *how* the model identifies private information arrival ($CPIE_{PIN}$). Moreover, our results also add to the debate (e.g. Back, Crotty, and Li (2014)) of whether models of private information based on order flow alone perform as well as those based on order flow and returns. In particular, we provide preliminary evidence that models based on returns in addition to order flow (OWR) may perform better than models based on order flow alone (GPIN). Finally, we also contribute to the literature that uses measures of private information by showing that proxies based on the OWR and GPIN models 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 generalizes the PIN model to deal with this shortcoming and evaluates a model based on order flow imbalance alone (GPIN) alongside another model that identifies private information from both returns and order flow (OWR). Section 4 concludes.

1 Data

To estimate the PIN, GPIN, and OWR models, we collect trades and quotes data for all NYSE stocks between 1993 and 2012 from the NYSE TAQ database. We require that the stocks in our sample have only one issue (i.e. one `PERMNO`), are common stocks (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. Despite our sample selection criteria, about 36% (25%) of the stocks in our sample are in the top (bottom) three Fama-French size deciles. For each stock in the sample, we classify each day's trades as either buys or sells, following the Lee and Ready (1991) algorithm. Internet Appendix C describes the computation of the number of buys and sells.

We estimate both the PIN and GPIN models using only the daily number of buys and sells ($B_{i,t}$ and $S_{i,t}$). The OWR model, however, also requires intraday and overnight returns as well as order imbalances. Following Odders-White and Ready (2008) we compute the intraday return at day t as the volume-weighted average price (VWAP) at t minus the opening quote midpoint at t plus dividends at time t, all divided by the opening quote midpoint at time t.[11] We compute the overnight return at t as the opening quote midpoint at $t+1$ minus the VWAP at t, all divided by the opening quote midpoint at t. The total return, or sum of the intraday and overnight returns is the open-to-open return from t to $t+1$. We compute order imbalance (y_e) as the daily share volume of buys minus the share

[11] 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.

volume of sells, divided by the total share volume. We follow Odders-White and Ready and remove systematic effects from returns to obtain measures of unexpected overnight and intraday returns ($r_{o,i,t}$ and $r_{d,i,t}$). See Internet Appendix C for details.

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

For the event study portion of our analysis, we examine earnings announcements. Our sample of earnings announcements includes all CRSP/COMPUSTAT firms listed in NYSE between 1995–2009 for which we have exact timestamps collected from press releases in Factiva which fall within a [-1,0] window relative to COMPUSTAT earnings announcement dates following Dong, Li, Ramesh, and Shen (2015). Because we have exact timestamps for the earnings announcements, we can cleanly separate between the pre and post event periods, thus avoiding ambiguity about when exactly the information becomes public. We use only earnings that are announced after the market is closed. We remove all announcements occurring on non-trading days. Our final sample includes 21,979 earnings announcements.

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 traders' trades are classified as opportunistic. Cohen, Malloy, and Pomorski (2012) show that opportunistic insider trades predict abnormal returns, information events, and regulator actions, which is consistent with the presence of private information. Our event sample includes 32,676 opportunistic insider trades.

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

2 Does PIN mis-identify private information?

This section analyzes whether PIN mis-identifies private information because the underlying model mechanically identifies the arrival of private information from turnover. Section 2.1 briefly describes the PIN model and $CPIE_{PIN}$. Section 2.2 shows that the PIN model identifies the arrival of private information from increases in turnover. Section 2.3 gives an example of how the conflation of private information arrival and turnover in the PIN model is consequential to the literature.

2.1 Description 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 informed traders and uninformed traders. At the beginning of each day, the informed traders receive a private signal with probability α. If the private signal is positive (which occurs with probability δ), buy orders from informed and uninformed traders arrive following a Poisson distribution with intensity $\mu + \epsilon_B$, while sell orders come only from the uninformed traders and arrive with intensity ϵ_S. If the private signal is negative (with probability $1 - \delta$), sell orders from informed and uninformed traders arrive following a Poisson distribution with intensity $\mu + \epsilon_S$, while buy orders come only from the uninformed traders and arrive with intensity ϵ_B. If the informed traders receive no private signal, they do not trade; thus, all buy and sell orders come from the uninformed traders and arrive with intensity ϵ_B and ϵ_S, respectively. Fig. 1 shows a tree diagram of this model. The difference in arrival rates captures the intuition that on days with positive private information, the arrival rate of buy orders increases over and above the normal rate of noise trading because informed traders enter the market to place buy orders.

Similarly, the arrival rate of sell orders rises when the informed traders seek to sell based on their negative private signals. Therefore, in theory, the PIN model identifies the arrival of private information through increases in the absolute value of the order imbalance.

To formalize the concept of $CPIE_{PIN}$, let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock i on day t and $\Theta_{PIN,i} = (\alpha_i, \mu_i, \epsilon_{B_i}, \epsilon_{S_i}, \delta_i)$ represent the vector of the PIN model parameters for stock i. Let $D_{PIN,i,t} = [\Theta_{PIN,i}, B_{i,t}, S_{i,t}]$. The likelihood function of the Easley, Kiefer, O'Hara, and Paperman (1996) model is $\prod_{t=1}^{T} L(D_{PIN,i,t})$, where $L(D_{PIN,i,t})$ is equal to the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information ($L_{NI}(D_{PIN,i,t})$) added to the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with positive information ($L_{I^+}(D_{PIN,i,t})$) and to the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information ($L_{I^-}(D_{PIN,i,t})$). Each of the likelihood functions ($L_{NI}(D_{PIN,i,t})$, $L_{I^+}(D_{PIN,i,t})$ and $L_{I^-}(D_{PIN,i,t})$) corresponds to a node of the tree in Fig. 1. See Internet Appendix D for details.

Using the PIN model, for each stock-day, we compute the probability of an information event conditional both on the model parameters and on the observed total number of buys and sells. Let the indicator $I_{i,t}$ take the value of one if an information event occurs for stock i on day t, and zero otherwise. For the PIN model, we compute $CPIE_{PIN,i,t} = P\left[I_{i,t} = 1 | D_{PIN,i,t}\right]$. This probability is given by $(L_{I^-}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t}))/L(D_{PIN,i,t})$. $CPIE_{PIN,i,t}$ represents the econometrician's posterior probability of an information event given the data observed on that day, and the underlying model parameters.

Note that if we condition down with respect to the data, $CPIE_{PIN,i,t}$ reduces to the model's unconditional probability of information events (α_i). The unconditional probability represents the econometrician's beliefs about the likelihood of an information event before seeing any actual orders or trades. In the absence of buy and sell data, an econometrician would assign a probability $\alpha_i = E[CPIE_{PIN,i,t}]$ to an information event for stock i on day t, where the expectation is taken with respect to the joint distribution of $B_{i,t}$ and $S_{i,t}$.

We estimate the PIN model numerically via maximum likelihood for every firm-year in our sample. The estimation procedure is similar to that used in Duarte and Young (2009). The parameter estimates are used for computing $CPIE_{PIN}$ in Section 2.2. Table 2 contains summary statistics for the parameter estimates of the PIN model. Table 2 also contains summary statistics of the cross-sectional sample means and standard deviations of

$CPIE_{PIN}$. The results in Table 2 show that the mean $CPIE_{PIN}$ behaves exactly like α. Hence, changes in $CPIE_{PIN}$ and changes in the estimated α are analogous. Fig. 2 Panel A shows how the distribution of α changes over time. Interestingly, the PIN model α increases over time, with the median PIN α rising from about 30% in 1993 to 50% in 2012.[12] Panel B of Fig. 2 plots the time series of PIN. Note that PIN decreases over time in spite of the fact that α increases. This happens because, according to the PIN model, the intensity of noise trading is increasing over time while the intensity of informed trading remains relatively flat as shown in Panel C of Fig. 2. It is important to note, however, that the time series patterns of the model parameters in Fig. 2 have no implications for how the PIN model identifies private information.

We also estimate the parameter vectors $\Theta_{PIN,i}$ in the period $t \in [-312, -60]$ before an earnings announcement. These parameter estimates are used to compute the $CPIE$s in Section 2.3. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 2 and in Figure 2.

2.2 How does the PIN model identify private information?

To show that the PIN model conflates turnover with the arrival of private information, we start with a scatter plot of real and simulated order flow data for Exxon-Mobil in Fig. 3. Panels A and B plot 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 $CPIE$, with darker points (+ **magenta**) representing low and lighter points (+ cyan) high $CPIE$s. Panels C and D plot the CPIE$_{PIN}$ as function of turnover. The vertical lines in these panels represent the annual mean of daily turnover.

Panel A of Fig. 3 illustrates the central intuition behind the PIN model. The simulated data comprise three types of days, which create three distinct clusters. Two of the clusters are made up of days characterized by relatively large absolute order flow imbalance, with

[12] The increase in our estimated PIN model α parameters is somewhat larger than that in Brennan, Huh, and Subrahmanyam (2015). This small difference arises because Brennan, Huh, and Subrahmanyam (2015) have a larger number of stocks per year due to the fact that we apply sample filters similar to those in Odders-White and Ready (2008). In fact, without these filters, the increase in our estimated PIN model α parameters from 1993 to 2012 is comparable to that in Brennan, Huh, and Subrahmanyam (2015).

a large number of sells (buys) and relatively few buys (sells). The third group of days has relatively low numbers of buys and sells because there is no private information arrival. Generalizing from this figure, days with large absolute order flow imbalances correspond to informed traders entering the market in the PIN model.

The real data, on the other hand, show no distinct clusters in Panel A, and in Panel B of Fig. 3 the PIN model's three clusters barely overlap with even a small portion of the data. This implies that the model cannot account for existence of the majority of the daily observations of order flow for Exxon-Mobil in 2012. In essence, the model classifies almost all daily observations as extreme outliers. The intuition for this is that the PIN model assumes that order flow is distributed as a mixture of three bivariate Poisson random variables (i.e. the three clusters in Panels A and B). The mean and the variance of a Poisson random variable are equal and, as a consequence, the Poisson mixtures behind the PIN model cannot accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample.

Panels A and B also plot a line that separates the scatter plots in two regions. All the observations below (above) these lines have turnover below (above) the annual mean of daily turnover. These lines along with the $CPIE$ color scheme for the observed data suggest that the PIN model is mechanically identifying 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 essentially classifies days with above average turnover as private information days (i.e. $CPIE_{PIN}$ equal to one) and days with below average turnover as days without private information (i.e. $CPIE_{PIN}$ equal to zero). Panels C and D emphasize the mechanical nature of the relation between $CPIE_{PIN}$ and turnover. Note that this identification does not necessarily relate to the possibility, suggested by Collin-Dufresne and Fos (2014), that informed traders sometimes choose to trade on days with high liquidity or turnover. Naturally, it is possible that informed traders do in fact trade on some days with high turnover. However, the point here is that the PIN model identifies essentially *all* days with above average turnover as information events.

Fig. 3 also delivers the intuition of why the PIN model mechanically conflates turnover with private information arrival. Essentially this conflation happens because of two limitations of the PIN model. First, under the PIN model, increases in expected turnover can only come about through the arrival of private information. Specifically, recall that $I_{i,t}$ indicates an information event. Note that under the model the number of buys plus sells (turnover) is distributed as a Poisson random variable with intensity:

$$\lambda(I_{i,t}) = \begin{cases} \epsilon_B + \epsilon_S & \text{when } I_{i,t} = 0 \\ \epsilon_B + \epsilon_S + \mu & \text{when } I_{i,t} = 1 \end{cases} \tag{1}$$

Thus, under the PIN model, private information is *necessarily* the cause of any variation in expected daily turnover. Second, the PIN model assumes that order flow is distributed as a mixture of three bivariate Poisson random variables (i.e. the three clusters in Panels A and B of Fig. 3). This assumption is too restrictive to accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample. Hence the poor fit to the turnover data along with the connection between turnover and arrival of private information in the PIN model causes the model to mechanically identify shocks to turnover as due to the arrival of private information.

Fig. 3 shows the PIN model's naive identification of private information events for one stock, however this is not an isolated example.[13] In fact, the problem is widespread. To quantify how often the PIN model classifies information events as simple function of turnover we define

$$CPIE_{Naive,i,t} = \begin{cases} 0, & \text{if } turn_{i,t} < \overline{turn_i} \\ 1, & \text{if } turn_{i,t} \geq \overline{turn_i} \end{cases} \tag{2}$$

That is, $CPIE_{Naive,i,t}$ is a dummy variable equal to one when turnover for stock i on day t ($turn_{i,t}$) is larger than or equal to the annual average of daily turnover of stock i ($\overline{turn_i}$) and zero otherwise. To our knowledge there is no paper in the literature that proposes identifying private information in similar manner.[14] It is clear, however, from Panel D of Fig. 3 that the PIN model essentially identifies the arrival of private information for Exxon-Mobil in 2012 according to this rule. We use $CPIE_{Naive}$ to gauge the extent to which the PIN model conflates the arrival of private information with turnover. Specifically, Panel A of Fig. 4 shows the distribution of the fraction of days for which $CPIE_{PIN}$ is identical to $CPIE_{Naive}$

[13] Here we use the word naive in a technical, statistical sense not in a pejorative sense.

[14] Stickel and Verrecchia (1994) propose identifying information arrival in general with a similar measure, but not private information in particular.

($|CPIE_{PIN} - CPIE_{Naive}| < 10^{-10}$). $CPIE_{PIN}$ and $CPIE_{Naive}$ are identical for about 85% of the annual observations for the median stock since 2002.

Another way to gauge the extent to which the PIN model breaks down later in our sample period is to count the number of days that the PIN model classifies as outliers. Panel B of Fig. 4 shows the fraction of days for the median stock-year which the PIN model classifies as "outliers" (likelihoods smaller than 10^{-10}). According to the PIN model, for the median stock about 60% (90%) of the annual observations are classified as outliers in 2005 (2010).[15]

Figs. 3 and 4 also give the intuition for why the median PIN α increases over time in Fig. 2. To see this, recall that α is the unconditional expected value of $CPIE_{PIN}$. Therefore, as we observe more $CPIE_{PIN}$ values approaching one, the estimated PIN α must increase. In fact, the median PIN α becomes close to 50% later in the sample which consistent with the fact that the PIN model assigns a $CPIE_{PIN}$ equal to one (zero) to days with turnover above (below) the average.

Given the strong connection between $CPIE$s and the unconditional probability of information arrival (α), Figs. 3 and 4 call into question the use of PIN as proxy for private information. The PIN of a stock, defined as $\frac{\alpha\mu}{\alpha\mu+\epsilon_B+\epsilon_S}$, is the unconditional probability that any given trade is initiated by an informed trader. Therefore $CPIE_{PIN}$ and PIN are linked via the unconditional probability of an information event, α. While the parameters μ, ϵ_B and ϵ_S also affect PIN, these parameters are jointly identified with α. Hence it seems extremely unlikely that in the joint identification of the model parameters, biases in the other parameters 'correct' the biases in α in such a way that PIN is 'rescued' as a reasonable proxy for private information. Thus, while our $CPIE$ results do not speak directly to μ, ϵ_B and ϵ_S, they still call into question PIN as a measure of private information.

To formally show that the PIN model identifies private information from turnover instead of order flow, we use daily data to estimate the following regression for every stock-year i in our sample: $CPIE_{PIN,i,t} = \alpha_i + \beta_{0,i}|B-S|_{i,t} + \beta_{1,i}|B-S|^2_{i,t} + \beta_{2,i}turn_{i,t} + \beta_{3,i}turn^2_{i,t} + \varepsilon_{i,t}$. We then compare the results from regressions with data created by simulating the PIN

[15] O'Hara, Yao, and Ye (2014) find that high-frequency trading is associated with an increase in the use of odd lot trades, which do not appear in the TAQ database. Therefore, estimates of the PIN model parameters computed using recent TAQ data may be systematically biased. More broadly, Fig. 4 indicates that even if the PIN model are estimated using data that includes odd lot trades, the model will still be badly misspecified late in the sample.

model to results from regressions with real data. To create the simulated data, we first estimate the parameters of the PIN model for each firm-year in our sample. Then, for each firm-year, we generate 1,000 artificial firm-years' worth of data (i.e. $B_{i,t}$ and $S_{i,t}$) using the estimated parameters. We then compute the $CPIE_{PIN,i,t}$ for each trading day in a simulated trading year and regress these $CPIE$s on absolute order flow imbalance and turnover. The results of the regressions using simulated data are useful because they reveal how the PIN model is intended to identify private information arrival and also allow us to build empirical distributions of the R^2s of the regressions of $CPIE$s on absolute order imbalance and turnover under the null hypothesis that the PIN model correctly describes the order flow data.[16]

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

The columns of Table 3 labeled as '$R^2_{inc.}$' include statistics on the increase in the R^2 that is due to the inclusion of turnover ($turn$) and turnover squared ($turn^2$) in the regressions. Specifically, $R^2_{inc.}$ is equal to the difference between the R^2 of the extended regression model with turnover terms and the R^2 of a regression that includes only order imbalance terms. We report the average of the median, the 5^{th}, and the 95^{th} percentiles of the $R^2_{inc.}$s of these regressions across the 1,000 simulations. The incremental increase in R^2s are relatively low,

[16] Since there are many moments that the PIN model can fail to match, there are many tests that might reject the PIN model (e.g. Duarte and Young (2009)). Our regression tests are not designed to analyze whether the PIN model matches particular moments in the data but instead are focused on how the PIN model identifies the fundamental variable of interest because $CPIE_{PIN}$ is a direct measure of private information according to the PIN model.

with an average value of around 10%, which implies that, under the model's data generating process, turnover has only modest incremental power in explaining $CPIE_{PIN}$. The picture that emerges from these regressions is that if the PIN model were a perfectly accurate representation of trading activity, $CPIE_{PIN}$ would be determined solely by the absolute order flow imbalance on each day.

Panel B of Table 3 reports regression results for the real rather than simulated data. With the real data, the picture is very different. The R^2s of the regressions of $CPIE_{PIN}$ on $|B - S|$ and $|B - S|^2$ are much smaller than those in the simulations. On the other hand, the incremental R^2s from turnover are much higher than those in Panel A. The incremental R^2 also increases over time with a value of about 36% in 1993, to nearly 46% in 2012. This implies that turnover and turnover squared explain a much larger degree of variation in $CPIE_{PIN}$ than absolute order imbalance. In fact, the average ratio of the median R^2s, $R^2_{inc.}/(R^2 + R^2_{inc.})$, is about 65%. The difference arises because, in the real data, absolute order flow imbalance and turnover are only weakly correlated. For instance, large absolute order flow imbalances are possible when turnover is below average, and vice versa. Under the PIN model, however, the two are highly correlated.

We test the hypothesis that $R^2_{inc.}$s in the actual data are consistent with those generated under the PIN model. Panel B reports the average p-value (the probability of observing an $R^2_{inc.}$ in the simulations at least as large as what we observe in the data) across all stocks, and the frequency that we reject the null at the 5% level implied by the distribution of simulated $R^2_{inc.}$s. The PIN model is rejected in about 89% of the stock-years in our sample, and there is on average less than a 7% chance of the PIN model generating $R^2_{inc.}$s as high as what we see in the data.

The results in Table 3 indicate that the PIN model identifies private information from increases in turnover, as opposed to changes in absolute order imbalances for the majority of the sample. These findings are inconsistent with the microstructure assumptions of the PIN model—controlling for absolute order imbalance there should be no room for turnover in explaining private information arrival.

2.3 Is the conflation of private information arrival and turnover consequential?

The previous section shows that the PIN model primarily identifies private information from turnover. The question remains, however, whether this is merely an inconsequential specification issue or whether this changes the interpretation of results in the broader finance and accounting literature. We address this question with an example that uses the PIN model in an event study context. Specifically, we examine how well the PIN model identifies information events around earnings announcements.

Unlike a standard event study, we focus on movements in $CPIE$ rather than price movements. We examine the period $t \in [-20, 20]$ around the event. To do so, we estimate the parameter vector $\Theta_{PIN,i}$ in the period $t \in [-312, -60]$ before the event and then compute the daily $CPIE$s for the period $t \in [-20, 20]$ surrounding the announcement. Prior studies estimate the parameters of the model in various windows around an event in order to compute the PIN. Our procedure is different in that we estimate the parameters of the model one year prior to the event and then employ the estimated parameters as if we were an econometrician observing the market data (i.e. buys and sells) and attempting to infer whether an information event occurred. Table 1 Panel B presents summary statistics for order imbalance, intraday returns, overnight returns, number of buys, and the number of sells for earnings announcement days ($t = 0$).

Panel A of Fig. 5 shows the average $CPIE_{PIN}$ in event time for our sample of earnings announcements. The graph shows that, under the PIN model, the probability of an information event increases prior to the event, starting below 55% 20 days before the announcement and peaking above 80% on the day after the announcement and remains high for 20 trading days after the actual earnings become public information. Panels B and C of Fig. 5 shed light on the features of the data that produce the observed pattern in the average $CPIE_{PIN}$ in Panel A. Panel B shows the average predictions from OLS regressions of $CPIE_{PIN}$ on absolute order imbalance and absolute order imbalance squared across all of the stocks in the event study sample. The solid line indicates that absolute order imbalance explains only a small fraction of the variation in $CPIE_{PIN}$ within the event window. Panel C shows the average predictions from regressions of $CPIE_{PIN}$ on turnover and turnover squared.

The solid line indicates that the variation in $CPIE_{PIN}$ around earnings announcements is explained almost entirely by turnover. The intuition follows directly from the results in Section 2.2, which shows that $CPIE_{PIN}$ is mechanically driven by turnover increases. The higher post-event turnover levels are enough to keep $CPIE_{PIN}$ above its pre-event mean for a substantial period.

To formalize the intuition behind Panels B and C of Fig. 5, we run regressions similar to those in Table 3 using our event sample. Specifically, we run regressions of $CPIE_{PIN}$ on absolute value of order imbalance and its squared term during the event window [-20,+20]. The results of these regressions (see Table 4) indicate that absolute order imbalance explains little of the variation in $CPIE_{PIN}$ in the event window while turnover explains most of the variation in $CPIE_{PIN}$. In fact, Table 4 shows that for the median stock, adding turnover and turnover squared to these regressions nearly quadruples the R^2s.

To see how the conflation of turnover with private information arrival is consequential to researchers using PIN, consider the results in Panel A of Fig. 5. It may appear to a researcher unaware that the PIN model conflates turnover with private information arrival that the results in Panel A of Fig. 5 suggest that the PIN model identifies private information in a sensible way. After all, $CPIE_{PIN}$ increases dramatically from 55% before the announcement to over 75% on the day of the announcement then falls after the announcement, albeit over a period of weeks. However, the decomposition of the $CPIE$s in Panels B and C of Fig. 5 points that the dramatic increase in $CPIE$ around the event is actually result of variation in turnover. Therefore, the results in Panels B and C of Fig. 5 lead to a different interpretation of the findings in Panel A because turnover around earnings announcement can vary for many reasons unrelated to the arrival of private information. Traditionally the literature has attributed high turnover around announcements to disagreement (e.g. Kandel and Pearson (1995)). Karpoff (1986) suggests that high turnover after earnings announcements may also be due to divergent prior expectations, while Frazzini and Lamont (2007) attributes high turnover to small investors' lack of attention. None of these studies suggest that the higher turnover around announcements is necessarily the result of increased informed trade, per se. Indeed, even the PIN model suggests that once we control for absolute order imbalance, turnover should have little power to identify informed trade.

3 Two alternatives to the PIN model

This section analyzes two models that do not conflate the arrival of private information with turnover. The first model, is an generalization of the PIN model (the GPIN model) based on order flow alone. The other model is the OWR model, which infers the arrival of private information from returns and order flow imbalance. Section 3.1 presents the GPIN model. Section 3.2 describes the OWR model and Section 3.3 presents two analyses of how these models identify private information.

3.1 Generalizing the PIN model

As we discuss in Section 2.2 and specifically in Equation 1, the conflation of turnover with the arrival of private information in the PIN model happens because of two limitations of the PIN model. First, under the PIN model, expected daily turnover (λ) is a deterministic and increasing function of private information arrival (Equation 1). Second, the PIN model assumption about the distribution of order flow is is too restrictive to accommodate the high level and volatility of turnover that we observe, especially in the later part of the sample.

In this section, we present a generalization of the PIN model that addresses these two limitations of the PIN model. In a nutshell, the GPIN model allows expected daily turnover (λ) to be drawn independently of the arrival of private information while keeping the same information structure of the PIN model. Essentially, the GPIN model relaxes the restrictive feature of the PIN model that the liquidity provider can infer the arrival of private information from turnover, while keeping the assumption that the liquidity provider infers the arrival of private information from the relative number of buy versus sell orders. This is a natural generalization since there is no theoretical reason why turnover should be deterministically associated with the arrival of private information. Expected turnover may vary for many reasons unrelated to the arrival of private information. For instance, turnover is expected to vary for calendar reasons (e.g. trading days close to holidays have lower turnover). Moreover, there is no theoretical justification for why turnover should be positively associated with the arrival of private information. On one hand, trading by informed traders may increase turnover. For instance,Collin-Dufresne and Fos (2015, 2014) show that in some contexts, informed traders may disguise their trades in periods of high liquidity such that market movements conceal the nature of their information. On the other hand, liquidity traders may postpone trading when the arrival of private information is likely leading to a negative relation between turnover and private information (e.g. Chae (2005)).

To generalize the PIN model, we first reparameterize it to focus on the intuition of its failure. Panel A of Fig. 6 displays a reparameterization of the PIN model in terms of three new parameters. First, the ratio of the intensity of uninformed buyer initiated trades to the intensity of the total number of uninformed trades ($\theta = \epsilon_B/(\epsilon_B + \epsilon_S)$). Second, the ratio of the expected number of informed to uninformed trades on days where there is private information ($\eta = \mu/(\epsilon_B + \epsilon_S)$). Third, the overall intensity of the number of buys plus sells on days without private information arrival ($\lambda(0) = \epsilon_B + \epsilon_S$). Panel A of Fig. 6 is a re-parameterization of the PIN model in Fig. 1 using the parameters $\lambda(0)$, η, and θ instead of ϵ_B, ϵ_S, and μ.

The GPIN model generalizes the PIN model because it draws the expected turnover (λ_t) independently of the arrival of private information instead of assuming, as the PIN model does, that (λ_t) is a deterministic function of turnover. Panel B of Fig. 6 presents the tree structure for the Generalized PIN model (GPIN). Specifically, the GPIN model in Panel B of Fig. 6 draws λ_t from a $Gamma(r, p/(1-p))$ distribution with shape parameter r and scale parameter $p/(1-p)$. Naturally, we could generalize the PIN model by drawing λ_t from another distribution, however the fact that λ_t is drawn from a *Gamma* distribution makes the model particularly tractable because, in this case, turnover $(B + S)$ is distributed as *Negative Binomial* (see Appendix E for proof), which dramatically simplifies the numerical estimation of the model.[17] In the maximum likelihood estimation the order flow intensity (λ) parameters r and p can be estimated in a first stage from the distribution of $(B + S)$, independently of the remaining information structure parameters which can be estimated in a second stage. $CPIE_{GPIN}$ is calculated in the same way as in the PIN model. Moreover, if we condition down with respect to the data, $CPIE_{GPIN}$ reduces to the model's unconditional probability of information events (α). See Appendix E for a detailed discussion of the model,

[17] The mixture of the *Poisson* and *Gamma* distributions is the well-known *Negative Binomial* distribution (see Casella and Berger (2002)).

the associated $GPIN$ measure, the likelihood function, and the $CPIE_{GPIN}$ calculation.

To illustrate how the GPIN model works, we present a stylized example of the GPIN in Fig. 7. Analogous to the PIN model plot in Fig. 3, we plot simulated and real order flow data for Exxon-Mobil in 1993 and 2012, with buys on the horizontal axis and sells on the vertical axis. Panels A and B of Fig. 7 illustrate the central intuition behind the GPIN model. The simulated data comprise three types of days, which create three distinct clusters. Two of the clusters are made up of days characterized by a high proportion of imbalanced trades (large $\frac{|B-S|}{B+S}$), with a large number of sells (buys) and relatively few buys (sells). The third group of days has a low proportion of imbalanced trades, no private information arrival, and is clustered around the dashed line in the center of the scatter plots.

The GPIN model implies that days with information events are the ones in which the proportion of imbalanced trades is large. An econometrician using the GPIN model, moving along the dashed line in Panels A and B, would observe that days with above average turnover–days the PIN model classifies as information events–are no longer classified as such, because higher turnover is driven by a large draw of the parameter λ_t under the GPIN model. Instead, the GPIN model identifies private information when moving away from the dashed line–when the proportion of imbalanced trades is high.

Panels C and D plot $CPIE_{GPIN}$ as function of turnover. As opposed to the analogous plot of the PIN model in Fig. 3, Panels C and D do not indicate any relation between turnover and $CPIE_{GPIN}$.[18] Although the GPIN model is not a perfect description of the order flow data, it manages to fix the problem of the PIN model which mechanically identifies private information arrival from turnover.

Table 5 contains summary statistics for the parameter estimates of the GPIN model. Table 5 also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{GPIN}$. We see that the mean $CPIE_{GPIN}$ behaves exactly like α. We also estimate the GPIN model for every stock in our sample in the period $t \in [-312, -60]$ before opportunistic insider trades. These parameter estimates are used to compute the

[18] Internet Appendix E shows the results of regressions of $CPIE_{GPIN}$ on the proportion of imbalanced trades and turnover. These regressions are analogous to those that we performed with the PIN model in Table 3. The results of these regressions indicate that the GPIN model does not conflate turnover with the arrival of private information.

$CPIE_{GPIN}$ in Section 3.3. The summary statistics of the parameter estimates for the event studies are qualitatively similar to those in Table 5.

3.2 The OWR model

Odders-White and Ready (2008) extend Kyle (1985) by allowing for days with information events and days without information events. Private information arrives before the opening of the trading day with probability α. On days when private information arrives, the model assumes that the information is publically revealed after the close of trade. The OWR model identifies the arrival of private information through order flow imbalance, y_e, the intraday price response to order imbalance, r_d, and through subsequent overnight price changes, r_o.[19] The vector (y_e, r_d, r_o) is assumed to be multivariate normal with mean zero and a covariance matrix that differs between information days and non-information days.[20]

Fig. 8 shows the time line of the model. The intuition behind the OWR model is that the market maker updates prices in response to order flow because the order flow could reflect an information event. However, the subsequent price pattern is different depending on whether there actually was an information event or not. If an information event occurs, the overnight price response reflects a continuation of the market makers' intraday reaction. If no information event occurs, the overnight price response reverses the market makers' initial price reaction. Therefore, an econometrician can make inferences about the probability of an information event in the OWR model because the covariance matrix of the three variables (y_e, r_d, r_o) differs between days when private information arrives and days when only public information is available.[21]

To see how the covariance matrix of (y_e, r_d, r_o) differs between information and non-information days, consider first the covariance of the intraday and overnight returns. This covariance is positive for information events, reflecting the fact that the information event is not completely captured in prices during the day and the revelation of the private information

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

[20] We follow Odders-White and Ready and remove systematic effects from returns to obtain measures of unexpected overnight and intraday returns (r_o and r_d). See Section 1 and Internet Appendix C for a detailed description of how we compute y_e, r_o and r_d.

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

means that the overnight return continues the partial intraday price reaction. In contrast, this covariance is negative in the absence of an information event since the market marker's reaction to the noise trade during the day is reversed when she learns that there was no private signal.

The other moments in the covariance matrix of (y_e, r_d, r_o) are also affected by the arrival of private information. If no information event occurs, then $Var(y_e)$ is composed of only the variances of the uninformed order flow and the noise in the data. However, if an event occurs, $Var(y_e)$ increases because the order flow reflects at least some informed trading. Similarly, $Var(r_d)$ is higher for an information event, because it reflects the market maker's partial reaction to the day's increased order flow. Since the private signal is revealed after trading closes, $Var(r_o)$ also increases in the wake of an information event, as it reflects the remainder of the market maker's partial reaction to the informed trade component in order flow. Likewise, information events make $cov(y_e, r_d)$ and $cov(y_e, r_o)$ rise. The higher covariance between order flow and intraday returns occurs because, in an information event, both order flow and the intraday return (partially) reflect the impact of informed trading. Along these same lines, because the market maker cannot separate the informed from the uninformed order flow, she is unable to fully adjust the price during the day to reflect the informed trader's private signal. However, since the private signal is publically revealed and fully reflected in prices after the close, $cov(y_e,\ r_o)$ is higher during an information event.

In contrast to the PIN and GPIN models, the OWR model does not contain a direct analog to the probability of informed trading (PIN). To understand this result, note that the probability of informed trade in the PIN and GPIN models is given by the ratio of the expected number of informed trades to the expected total number of trades on a given day. Since the OWR model employs only the difference between buys and sells, it does not make assumptions about the distribution of number of trades. Thus, the OWR is mute regarding the ratio of the expected number of informed trades to expected number of trades. This may appear to be a limitation of the OWR model, but this is actually an advantage because it allows the OWR model to disentangle variations in turnover from the arrival of informed trading, much like the GPIN model.

Even though the OWR model does not have a measure analogous to the PIN measure, the OWR model admits other useful measures of private information. For instance, the OWR model has a $CPIE_{OWR}$ which reduces to the model's unconditional probability of information events (α) if we condition down with respect to the data. Moreover, Odders-White and Ready (2008) motivate their model as a tool to separate the expected liquidity provider losses due to trading with informed traders into the frequency of private information arrival and the expected magnitude of the private information. Hence, the OWR allows for the construction of private information measures that are based on both dimensions. The PIN and GPIN models, on the other hand, focus only on the frequency of information arrival and are silent with respect to the expected magnitude of the private information. Hence, our comparison of the GPIN and OWR models with $CPIE_{GPIN}$ and $CPIE_{OWR}$ focuses on the dimension of private information that both models have in common, namely the frequency of information arrival. The fact that we are using $CPIE$s to compare the models does not imply that we are taking the position that frequency measures are the only private information metrics that are worthy of consideration.

As with the PIN and GPIN models, we estimate the OWR model numerically via maximum likelihood. Table 6 contains summary statistics for the parameter estimates of the OWR model. Table 6 also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{OWR}$. As in the PIN and GPIN models, we see that the mean $CPIE_{OWR}$ behaves exactly like α in the OWR model. The estimated OWR α parameters are in general higher than those in Odders-White and Ready (2008). This is due to the fact that our definition of y_e is different from that in Odders-White and Ready (2008) (see the discussion in Section 1 above).[22] Fig. 9 plots the time series of the estimated OWR α. In contrast to the PIN α, the OWR α is decreasing over time. This pattern may indicate that private information arrival is less likely later in our sample. While interesting, understanding this pattern is outside the scope of this paper and we leave this investigation for future research. We also estimate the OWR model for each stock i in the period $t \in [-312, -60]$ before opportunistic insider trades. These parameter estimates are used to compute the $CPIE$s in Section 3.3.1. The summary statistics of the parameter estimates for the event

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

studies are qualitatively similar to those in Table 6. Internet Appendix F has a detailed description of model, its likelihood function, and the $CPIE_{OWR}$ calculation. Appendix F also displays the results of regressions of $CPIE_{OWR}$ similar to those that we perform with $CPIE_{PIN}$ in Section 2.2. These regressions indicate that the OWR model identifies the arrival of private information in a way consistent with its theory.

3.3 Assessing the GPIN and OWR models

In this section, we use two different approaches to diagnose potential problems with OWR and GPIN models' ability to detect private information.

3.3.1 $CPIE_{GPIN}$ and $CPIE_{OWR}$ around insider trading

In this section we investigate whether the OWR and GPIN models are capable of identifying opportunistic insider trades using the insider trade classification scheme developed in Cohen, Malloy, and Pomorski (2012).[23] There is a large literature that suggests that insiders may have private information and may trade on that information.[24] Recently, 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 the trades of opportunistic insiders show significant predictive power for future news about the firm, and that the fraction of traders who are opportunistic 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 examine the models' ability to detect the arrival of actionable private information.

Panel A (B) of Fig. 13 presents the average $CPIE_{GPIN}$ ($CPIE_{OWR}$) in event time for our sample of opportunistic insider trades. Both models show a statistically significant spike in $CPIE$s at $t = 0$, consistent with the arrival of private information on the day that insiders trade. Specifically, at $t = 0$, the $CPIE$s are more than two standard deviations higher than the mean estimated between $t \in [-40, 21]$.

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

[24] 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).

While $CPIE_{GPIN}$ rises on the day that insider actually trades, counterintutitively it also spikes on several days after the insider trade. This suggests that the GPIN model may be yielding 'false positives' in the sense that it appears to identify the arrival of private information when we have no a priori economic reason to suspect any such information arrival (e.g. day t+5 and day t+16 after the insider trade). On the other hand, the $CPIE_{OWR}$ rises a few days before the insider trades and clearly drops after the trade. The fact that $CPIE_{OWR}$ increases a few days before the insider trades suggests that whatever private signal the insider is responding to is also received by others that attempt to act on it as well.

3.3.2 Are $CPIE_{GPIN}$ and $CPIE_{OWR}$ related to return continuation?

The market microstructure literature has long held that price changes related to informed trades should not be subsequently reversed while non-information related price changes (e.g. dealer inventory control, price pressure, price discreteness etc.) are transient (e.g. Hasbrouck (1988, 1991a,b)). In this section, we investigate whether $CPIE_{GPIN}$ and $CPIE_{OWR}$ are associated with subsequent return reversals. In particular, we examine the relation between $CPIE$s and return autocorrelations. The intuition is that if a model's $CPIE$ on day t actually reflects a high probability of informed trade then we expect that the return on day t will be continued over the subsequent day. To capture this idea we model return autocorrelations as linear functions of $CPIE$. Specifically, we consider the following regressions: $r_{i,t+1} = \alpha + \beta_{OWR,1} r_{i,t} + \beta_{OWR,2} CPIE_{OWR,t} + \beta_{OWR,3}(r_{i,t} \times CPIE_{OWR,t}) + \varepsilon_{i,t+1}$, and $r_{i,t+1} = \alpha + \beta_{GPIN,1} r_{i,t} + \beta_{GPIN,2} CPIE_{GPIN,t} + \beta_{GPIN,3}(r_{i,t} \times CPIE_{GPIN,t}) + \varepsilon_{i,t+1}$.

In the above regressions, $r_{i,t}$ is the open-to-open, risk-adjusted return $(r_{i,d,t} + r_{i,o,t})$ on day t. Thus, there is no overlap between the intraday and overnight returns that are used to compute $CPIE_{OWR,i,t}$ on day t and the return on day $t+1$. This is important because if there were overlap between $CPIE_{OWR,i,t}$ and $r_{i,t+1}$, then the relation would be mechanical. The coefficients $\beta_{OWR,3}$ and $\beta_{GPIN,3}$ reflect the impact of the model's $CPIE$ on the correlation between the return on day t and the return the next trading day. We estimate the regressions above using a panel regression approach including firm and year fixed effects with standard errors clustered by firm and year. Table 7 reports the coefficient estimates and t-statistics for these regressions. We standardize both $CPIE_{OWR}$ and $CPIE_{GPIN}$ in these regressions to have a standard deviation of one. The results in Table 7 show that the estimates for both $\beta_{OWR,3}$ and $\beta_{GPIN,3}$ are positive and significant, indicating that both $CPIE_{GPIN}$ and $CPIE_{OWR}$ are associated with smaller future return reversals. To see this note that both regressions show a tendency of daily returns to reverse because the coefficients on lagged returns in both regressions are negative. In fact, a one standard deviation shock to $CPIE_{OWR}$ is associated with a 27% (2.417/8.883) decline in the subsequent reversal, while a one standard deviation shock to $CPIE_{GPIN}$ is associated with a 4% (0.271/6.955) drop in the subsequent reversal. Consistent with Hasbrouck (1991a,b), both the OWR and GPIN models appear to capture the arrival of information that has a persistent impact on prices. However, the impact of $CPIE_{OWR}$ on return reversals is much larger than that of $CPIE_{GPIN}$, which is consistent with the OWR model doing a better job capturing informed trade than the GPIN model.

4 Conclusion

Recent work suggests that the PIN measure, developed in the seminal work of Easley and O'Hara (1987), Easley, Kiefer, O'Hara, and Paperman (1996), and Easley, Kiefer, and O'Hara (1997), may fail to capture private information (e.g. Aktas, de Bodt, Declerck, and Van Oppens (2007), Benos and Jochec (2007), and Collin-Dufresne and Fos (2015)). However, PIN remains the most widely used measure of information asymmetry in the accounting, corporate finance and asset pricing literature today. This may be because there is no definitive test of any model's ability to capture private information arrival since arrival of private information is by definition unobservable.

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. This is at odds with a vast literature that suggests turnover varies for many reasons unrelated to the arrival of private information. This failure of the PIN model is particularly strong after the increase in turnover in the early 2000s. In fact, after 2002 for the median stock in our sample, the PIN model is essentially equivalent to a naïve model that assigns a probability of one to the arrival of private information on any day where turnover is above average and zero probability to the arrival of private information on any other day. These findings indicate that the the PIN model does not deliver reliable proxies for private information.

We highlight how consequential the PIN model's conflation of turnover with volume is to the literature with an event study around earnings announcements. Our results suggest the findings in Brennan, Huh, and Subrahmanyam (2015) and in Benos and Jochec (2007) can simply be attributed to the fact that turnover (instead of the probability of private information arrival) is typically much higher after earnings announcements. This example is emblematic of a pervasive issue in the literature since, in many, if not all, contexts the PIN model ultimately yields misleading inferences about private information arrival.

We also examine two alternatives to the PIN model that do not conflate turnover with the arrival of private information. One is the Odders-White and Ready (2008) model (OWR), which infers the arrival of private information from returns and order flow, the other is an extension of the PIN model (the GPIN model), which is solely based on order flow but corrects the PIN model's mechanical association of private information arrival with variation in turnover. Our results do not suggest any problems with mis-identifying private information for either the GPIN or the OWR models. This being said, we believe the OWR model performs somewhat better than the GPIN model for two reasons. First, both $CPIE_{OWR}$ and $CPIE_{GPIN}$ increase in periods of opportunistic insider trading. However, $CPIE_{OWR}$ encouragingly decreases dramatically immediately following the insider trades, while $CPIE_{GPIN}$ displays suspicious spikes in the post event period with no apparent economic interpretation. Second, the relation between $CPIE_{OWR}$ and future return continuation is about seven times larger than that of the $CPIE_{GPIN}$.

As we pointed out previously, there is no perfect identification strategy for private information arrival. Thus, our insider trading and return continuation tests of the OWR and GPIN models are not definitive. However, these two tests are informative because they do not reveal problems with the the OWR and GPIN models' ability to sensibly identify private information arrival. Furthermore, in their totality, our results suggest that the GPIN model is a promising alternative to the PIN model for research that requires a model based on order flow alone. On the other hand, if relying on order flow alone is not a requirement, then the OWR model is the more promising alternative to the PIN model.

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.

Bamber, Linda S., Orire E. Barron, and Douglas E. Stevens, 2011, Trading volume around earnings announcements and other financial reports: Theory, research design, empirical evidence, and directions for future research, *Contemporary Accounting Research* 28, 431–471.

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.*

Casella, George, and Roger Berger, 2002, *Statistical Inference* (Thomson Learning).

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

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.

Dong, Bei, Edward Xuejun Li, K. Ramesh, and Min Shen, 2015, Priority dissemination of public disclosures, *The Accounting Review* 90, 2235–2266.

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

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

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

Easley, David, 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 structureandprice-informativeness, *Journal of Financial Economics* 99, 523–545.

Frazzini, Andrea, and Owen Lamont, 2007, The earnings announcement premium and trading volume, *working paper*.

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.

Karpoff, Jonathan M., 1986, A theory of trading volume, *The Journal of Finance* 41, 1069–1087.

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

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 event day (t=0) returns, order imbalance, and number of buys and sells. We compute intraday and overnight returns as well as daily buys and sells for stocks between 1993 and 2012 using data from the NYSE TAQ database. Following Odders-White and Ready (2008), we compute the intraday return, r_d, at time t as the volume-weighted average price at t (VWAP) minus the opening quote midpoint at t plus dividends at time t, all divided by the opening quote midpoint at time t. We compute the overnight return, r_o, at t as the opening quote midpoint at $t+1$ minus the VWAP at t, all divided by the opening quote midpoint at t. We compute y_e as the daily total volume of buys minus total volume of sells, divided by the total volume. For the PIN and GPIN models, we use the daily total number of buys and sells. Our sample of earnings announcements includes all CRSP/COMPUSTAT firms listed in NYSE between 1995–2009 for which we have exact timestamps collected from press releases in Factiva which fall within a [-1,0] window relative to COMPUSTAT earnings announcement dates. Opportunistic insider trades are defined as 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%
# Buys	5,286,191	1,876	6,917	37	220	1,128
# Sells	5,286,191	1,843	6,894	36	194	1,033

(b) Earnings Announcements

	N	Mean	Std	Q1	Median	Q3
y_e	21,979	5.099%	22.122%	-4.787%	4.373%	16.400%
r_d	21,979	0.002%	2.424%	-1.252%	-0.004%	1.271%
r_o	21,979	0.075%	2.313%	-1.042%	0.013%	1.153%
# Buys	21,979	4,572	13,491	223	956	3,421
# Sells	21,979	4,465	13,546	191	831	3,165

(c) 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%
# Buys	32,676	3,852	10,645	354	1,129	3,478
# Sells	32,676	3,787	10,554	300	996	3,303

Table 2: **PIN Parameter Estimates.** This table summarizes parameter estimates of the PIN model for 21,206 `PERMNO`-Year samples from 1993 to 2012. α represents the average unconditional probability of an information event at the daily level. δ represents the probability of good news, and $1-\delta$ represents the probability of bad news. ϵ_B and ϵ_S represent the expected number of daily buys and sells given no private information. μ represents the expected additional order flows given an information event. $\overline{CPIE}$ and Std($CPIE$) are the `PERMNO`-Year mean and standard deviation of $CPIE_{PIN}$.

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

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

(a) Simulated Data

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

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

(b) Real Data

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

Table 4: **PIN Regressions Around Earnings Announcements.** This table reports regression results for $CPIE_{PIN}$ around Earnings Announcements. For each announcing firm in our sample we run regressions of $CPIE_{PIN}$ on absolute order imbalance ($|B-S|$) and absolute order imbalance squared ($|B-S|^2$) from $[-20,+20]$ and report median estimates across all the events. We compute the incremental $R^2_{\text{inc.}}$ as the increase in R^2 attributed to *turn* and $turn^2$ in an extended regression model. We report standardized coefficients.

β		t		R^2	$R^2_{\text{inc.}}$
$\|B-S\|$	$\|B-S\|^2$	$\|B-S\|$	$\|B-S\|^2$	50%	50%
0.143	-0.032	(1.07)	(-0.35)	15.42%	44.44%

Table 5: **GPIN Parameter Estimates.** This table summarizes parameter estimates of the GPIN model for 21,206 `PERMNO`-Year samples from 1993 to 2012. α represents the average unconditional probability of an information event at the daily level. δ represents the probability of good news, and $1-\delta$ represents the probability of bad news. The total number of trades in any given day (t) is drawn from a Poisson distribution with intensity λ_t, where λ_t is draw from a Gamma distribution with shape parameter r and scale parameter $p/(1-p)$. The number of buys on a day with no private information is draw from a Poisson distribution with intensity $\theta \times \lambda_t$. On days with negative news, the number of buys is drawn from a Poisson with intensity $\theta/(1+\eta) \times \lambda_t$. $\overline{CPIE}$ and Std($CPIE$) are the `PERMNO`-Year mean and standard deviation of $CPIE_{GPIN}$.

	N	Mean	Std	Q1	Median	Q3
α	21,206	0.493	0.088	0.448	0.498	0.543
δ	21,206	0.495	0.184	0.372	0.492	0.616
r	21,206	7.210	4.724	4.056	5.976	8.960
p	21,206	0.948	0.080	0.932	0.984	0.997
θ	21,206	0.515	0.049	0.493	0.514	0.546
η	21,206	0.316	0.242	0.152	0.240	0.413
$\overline{CPIE}$	21,206	0.494	0.087	0.449	0.499	0.543
Std($CPIE$)	21,206	0.414	0.082	0.367	0.445	0.478

Table 6: **OWR Parameter Estimates.** This table summarizes parameter estimates of the OWR model for 21,206 PERMNO-Year samples from 1993 to 2012. α represents the average unconditional probability of an information event at the daily level. σ_u represents the standard deviation of the order imbalance due to uninformed traders, which is observed with normally distributed noise with variance σ_z^2. σ_i represents the standard deviation of the informed trader's private signal. σ_{pd} and σ_{po} represent the standard deviation of intraday and overnight returns, respectively. $\overline{CPIE}$ and Std$(CPIE)$ are the PERMNO-Year mean and standard deviation of $CPIE_{OWR}$.

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

Table 7: **Return Reversals.** This table reports regressions of the daily return at time $t+1$ on the return, $CPIE$ ($CPIE_{GPIN}$ or $CPIE_{OWR}$), and the interaction at time t. Returns are measured from open to open and they are computed as the sum of the intraday (r_d) and overnight returns (r_o). We standardize both $CPIE$ measures to have a standard deviation of one. We include stock and year fixed effects and cluster standard errors by stock and year. * indicates statistical significance at the 10% level, ** at the 5%, and *** at the 1% level.

	r_{t+1}	
	OWR	GPIN
r_t	-8.883***	-6.955***
	(-6.88)	(-6.91)
CPIE$_t$	0.0136***	
	(4.36)	
CPIE$_t$ × r_t	2.417***	
	(4.16)	
CPIE$_t$		0.00704***
		(4.03)
CPIE$_t$ × r_t		0.271**
		(2.58)
R^2(%)	0.61	0.54
Obs.	5,284,078	5,284,078

Figure 1: **PIN Tree.** For a given trading day, private information arrives with probability α. When there is no private information, buys and sells are Poisson with intensity ϵ_B and ϵ_S. Private information is good news with probability δ. The expected number of buys (sells) increases by μ in case of good (bad) news.

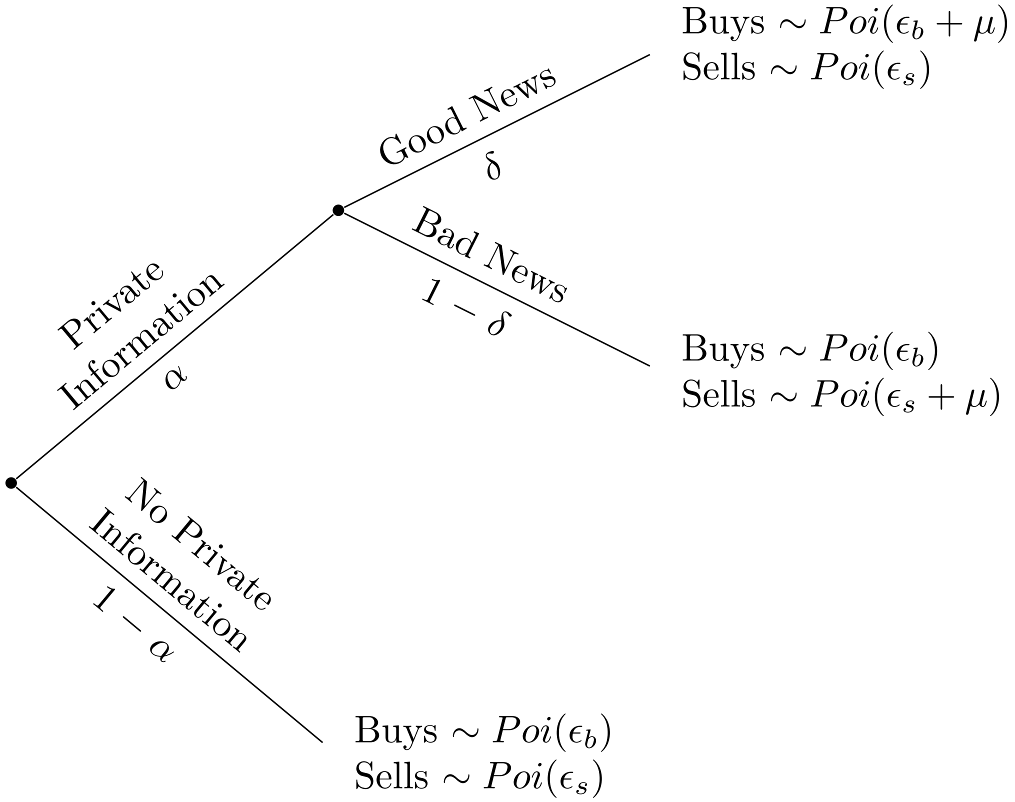
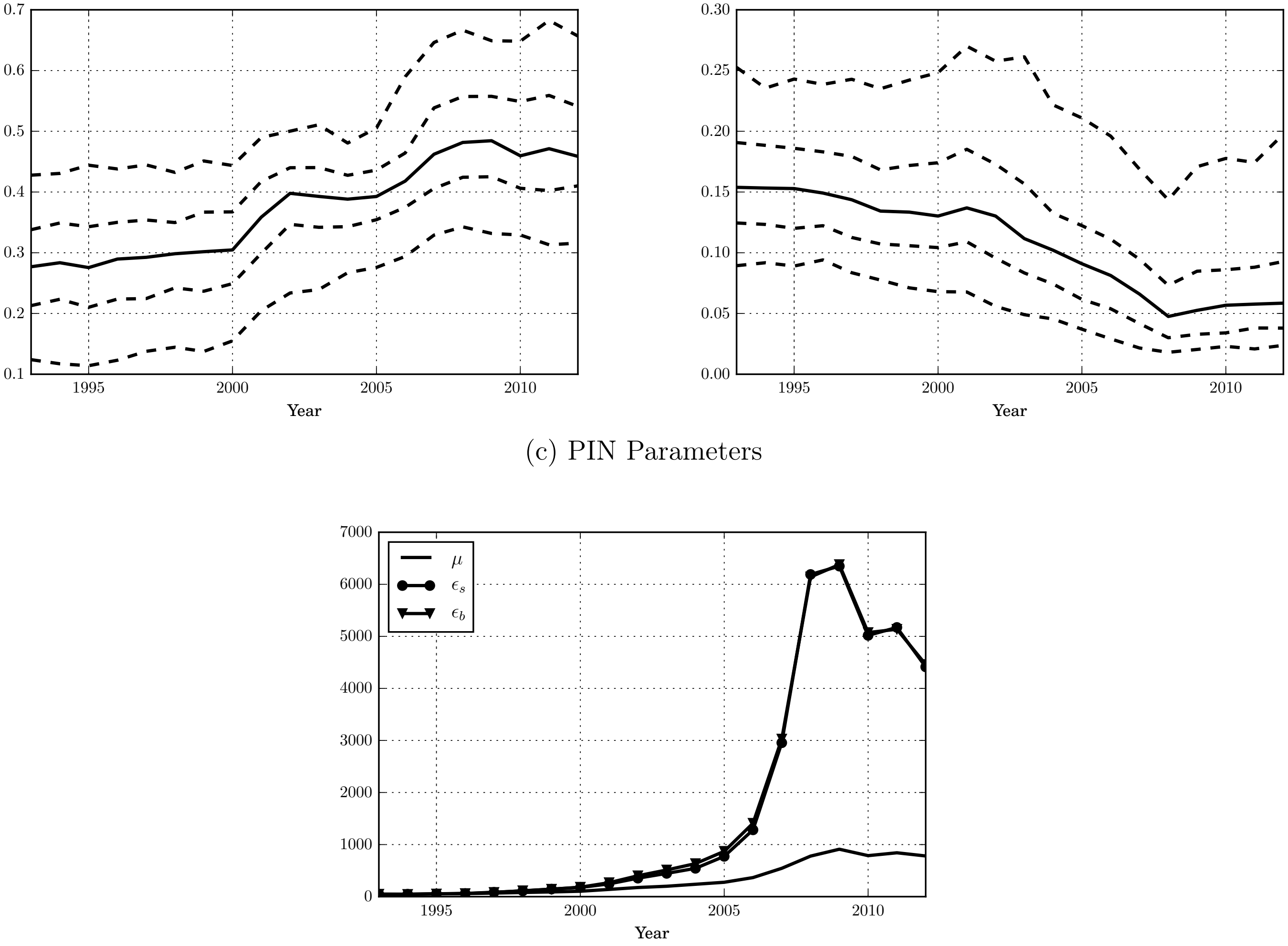


Figure 2: PIN **Parameters.** This figure shows the distribution of yearly α, PIN, and $\mu, \epsilon_B, \epsilon_S$ parameter estimates for the PIN model. The solid black line represents the median value, and the dotted lines represent the 5, 25, 75, and 95 percentiles.

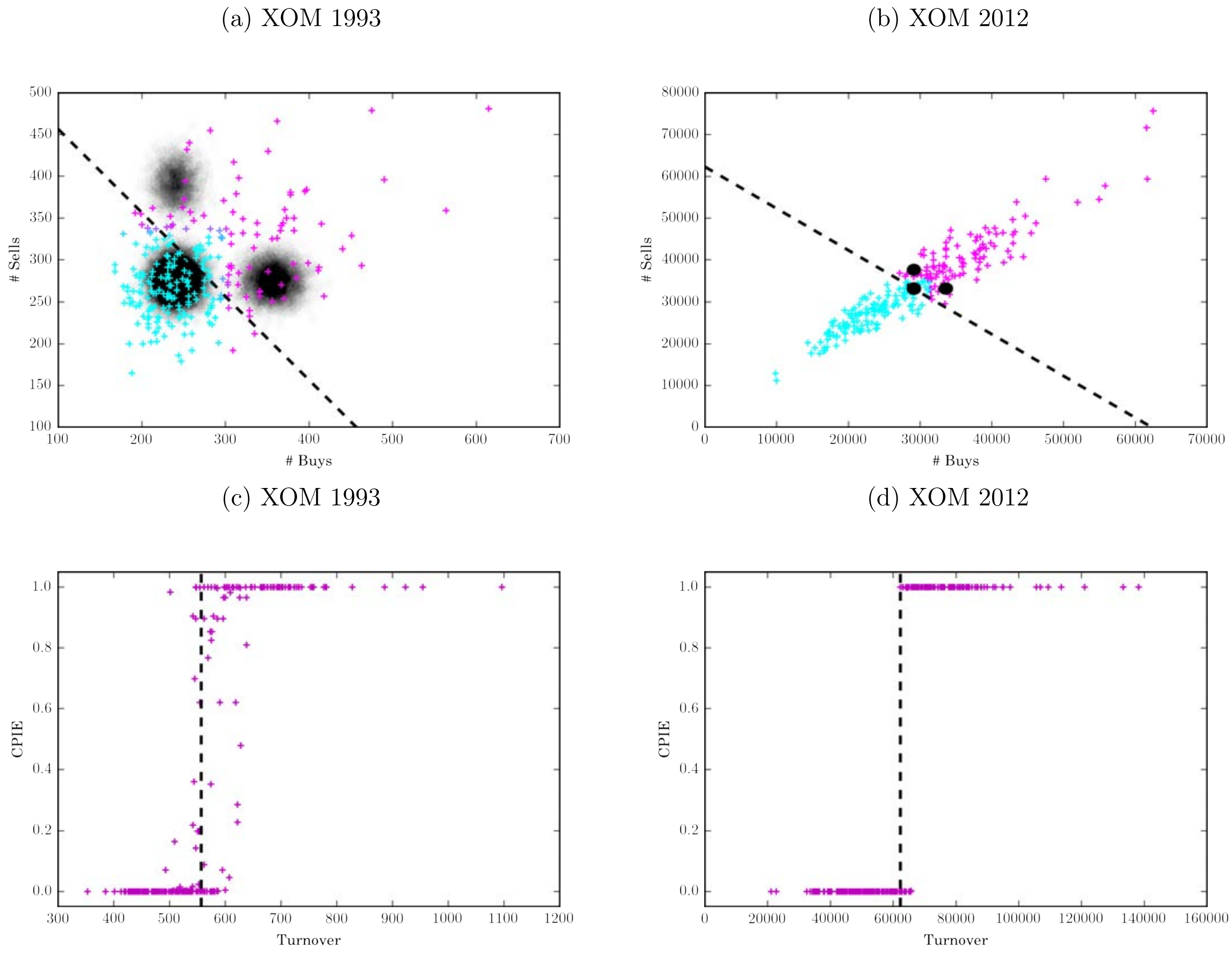


(a) PIN α

(b) PIN

(c) PIN Parameters

Figure 3: **XOM EO.** This figure compares the real and simulated data for XOM in 1993 and 2012 using 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. 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. All the observations below (above) the dashed lines in Panels A and B have turnover below (above) the annual mean of daily turnover. Panels C and D plot the CPIEs for the real data as a function of turnover along with a dashed line indicating the mean turnover.

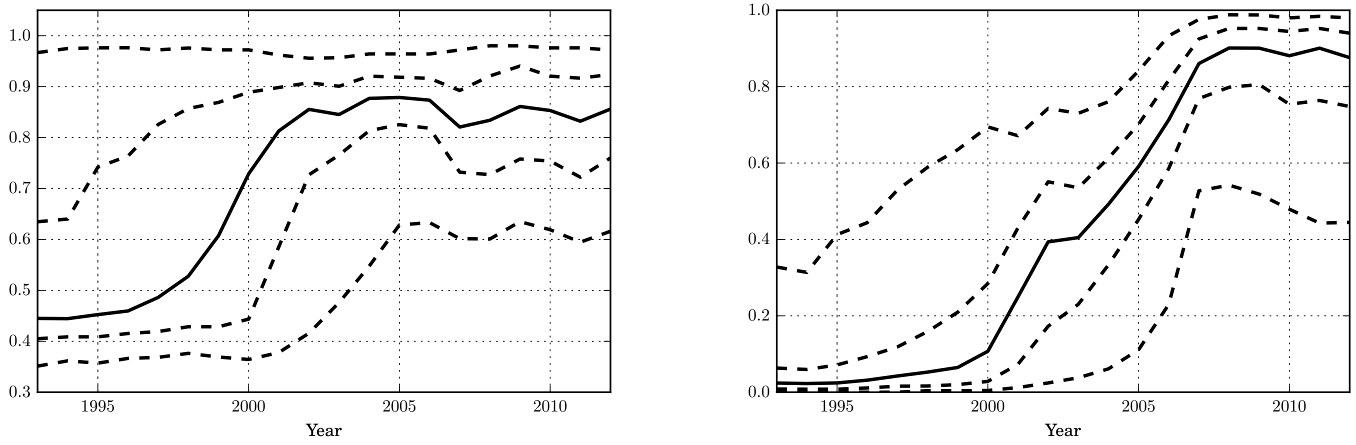

(a) XOM 1993

(b) XOM 2012

(c) XOM 1993

(d) XOM 2012

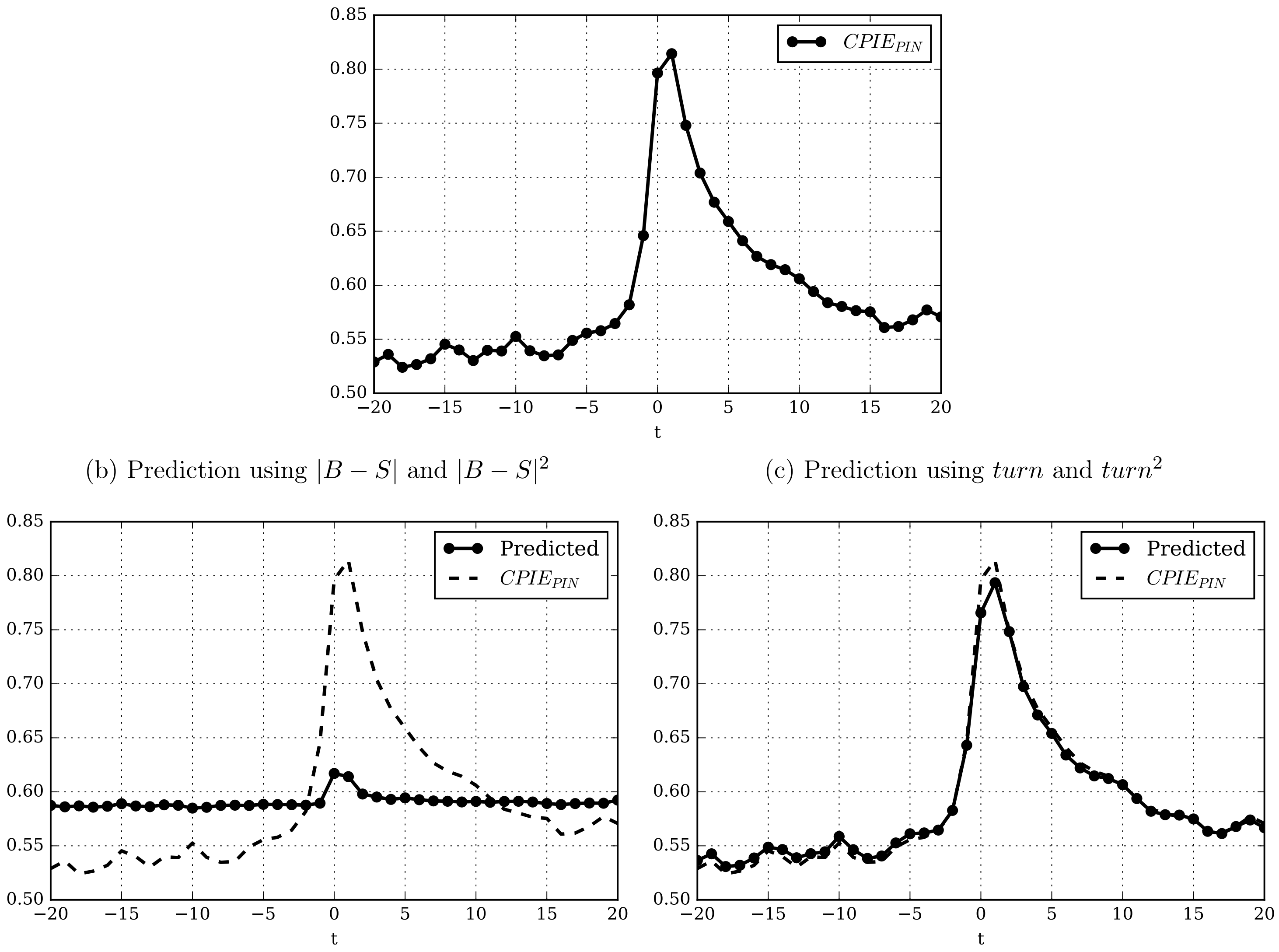
Figure 4: **Breakdown of the PIN Model.** Panel A shows the distribution of the percent of trading days in a year in which the PIN model identifies private information essentially in the same way as the naive identification scheme. That is, Panel A plots the percentage of days where the $|CPIE_{PIN} - CPIE_{Naive}| < 10^{-10}$. $CPIE_{Naive}$ is one for a given stock-day if turnover is higher than the annual mean of daily turnover, and is zero otherwise. Panel B shows the distribution of the percent of days where the likelihood, given the model parameters and observed order flow data is less than 10^{-10} — days, according to the model, with near-zero probability of occurring. The solid black line represents the median stock, and the dashed lines represent the 5, 25, 75, and 95 percentiles.



(a) Days with $CPIE_{PIN} \approx CPIE_{Naive}$

(b) Days with Near-Zero Probability

Figure 5: **Earnings Announcements - PIN.** Panel A shows the average $CPIE_{PIN}$ for the PIN model in event time surrounding earnings announcements. Panels B and C compare the average $CPIE_{PIN}$ with the $CPIE_{PIN}$ predicted with either the absolute order imbalance ($|B - S|$) or turnover ($turn$), respectively. To obtain the predictions, we run regressions of daily $CPIE_{PIN}$ on $|B - S|$ or $turn$, and their respective squared terms.

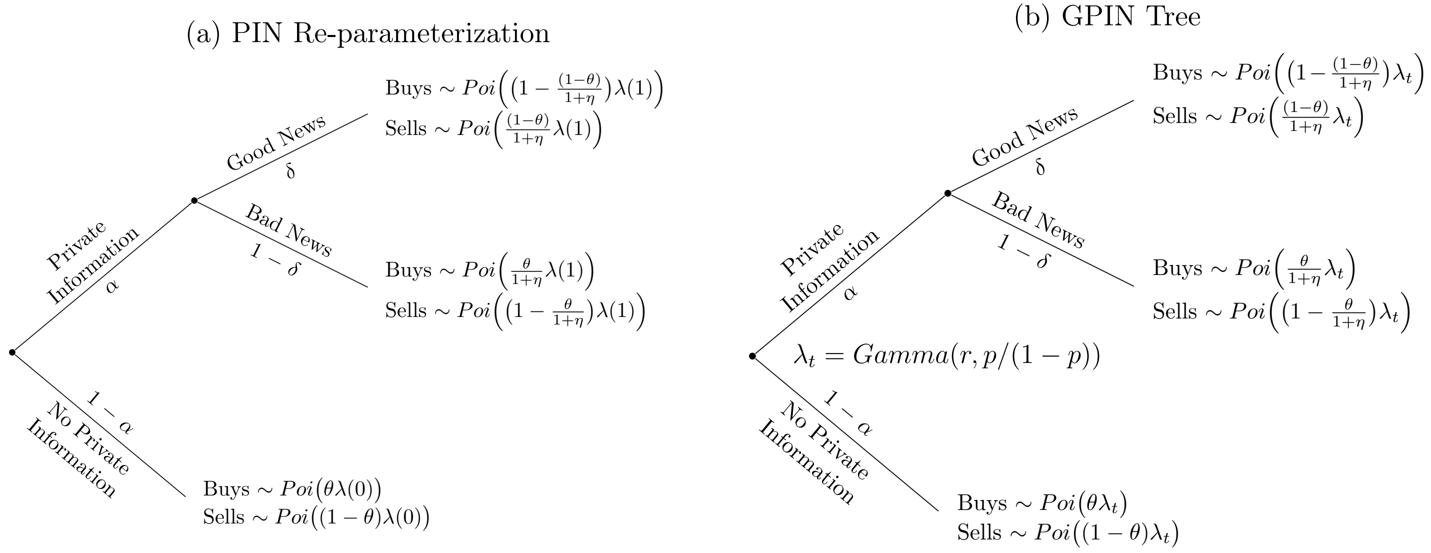


(a) $CPIE_{PIN}$

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

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

Figure 6: **GPIN Tree.** Panel A presents a re-parameterization of the PIN model in terms of ratio of the intensity of uninformed buyer initiated trades to the intensity of the total number of uninformed trades ($\theta = \epsilon_B/(\epsilon_B + \epsilon_S)$), the ratio of the expected number of informed to uninformed trades on days where there is private information ($\eta = \mu/(\epsilon_B + \epsilon_S)$), and the overall intensity of the number of buys plus sells as a function of the arrival of private information ($\lambda(I_{i,t})$). Panel B presents the GPIN model. The GPIN model extends the PIN model by allowing the intensity of the number of trades on a given day t (λ_t) to be drawn from a Gamma distribution with location and scale parameters r and $p/(1-p)$, respectively. The information structure remains the same as the one in the PIN model. For a given trading day, private information arrives with probability α. When there is no private information, the number of buys (sells) is distributed as a Poisson with intensity $\theta \times \lambda_t$ $\big((1-\theta) \times \lambda_t\big)$. Private information is good (bad) news with probability δ $(1-\delta)$. When there is good news, the number of sells (buys) is Poisson with intensity $\frac{(1-\theta)}{1+\eta}\lambda_t$ $\big((1-\frac{(1-\theta)}{1+\eta})\lambda_t\big)$. When there is bad news, the number of buys (sells) is Poisson with intensity $\frac{\theta}{1+\eta}\lambda_t$ $\big((1-\frac{\theta}{1+\eta})\lambda_t)\big)$.

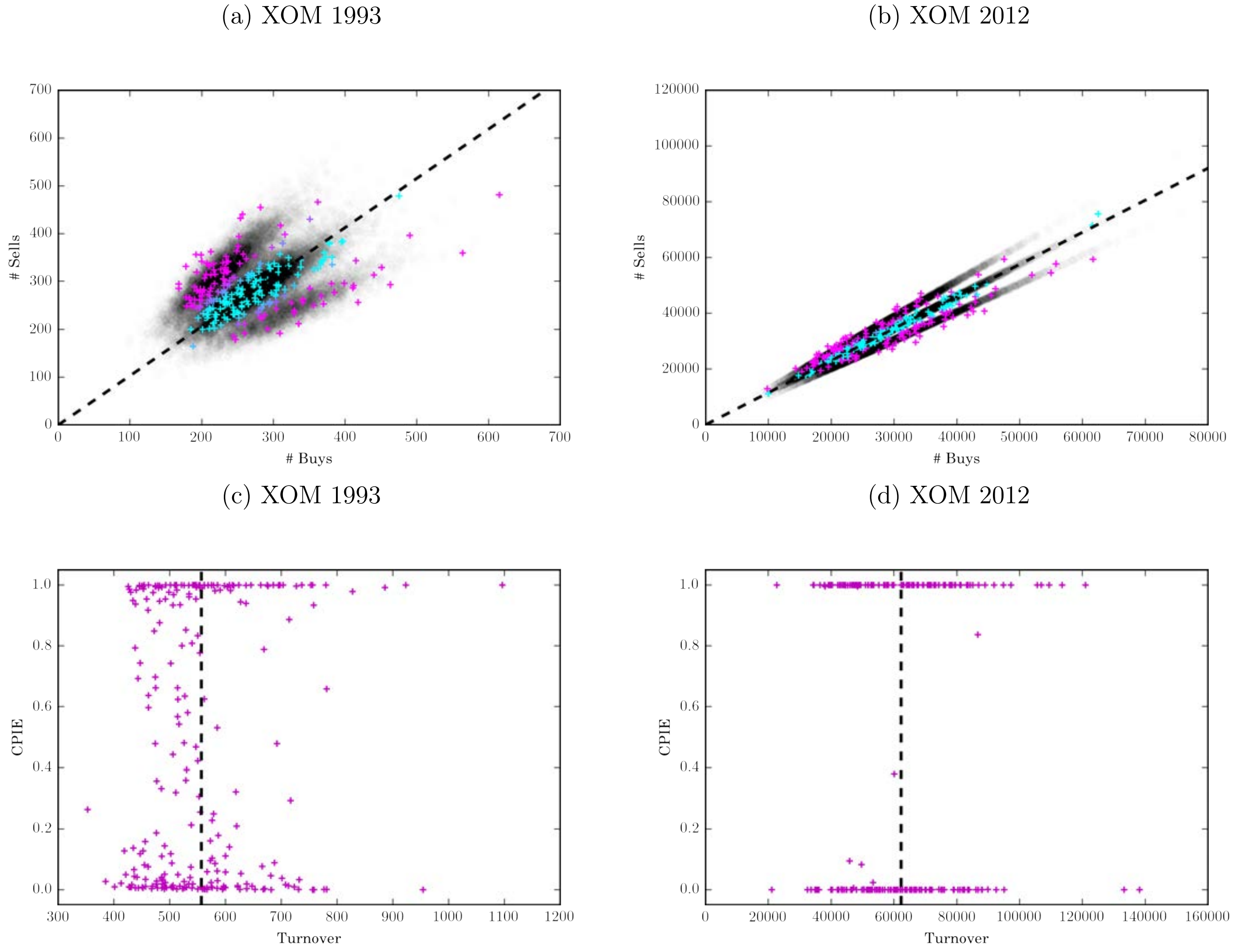


(a) PIN Re-parameterization

(b) GPIN Tree

Private Information
α
Good News
δ
Bad News
$1-\delta$
$1-\alpha$
No Private Information
Buys $\sim Poi\Big((1-\frac{(1-\theta)}{1+\eta})\lambda(1)\Big)$
Sells $\sim Poi\Big(\frac{(1-\theta)}{1+\eta}\lambda(1)\Big)$
Buys $\sim Poi\Big(\frac{\theta}{1+\eta}\lambda(1)\Big)$
Sells $\sim Poi\Big((1-\frac{\theta}{1+\eta})\lambda(1)\Big)$
Buys $\sim Poi\big(\theta\lambda(0)\big)$
Sells $\sim Poi\big((1-\theta)\lambda(0)\big)$
Private Information
α
Good News
δ
Bad News
$1-\delta$
$\lambda_t = Gamma(r, p/(1-p))$
$1-\alpha$
No Private Information
Buys $\sim Poi\Big((1-\frac{(1-\theta)}{1+\eta})\lambda_t\Big)$
Sells $\sim Poi\Big(\frac{(1-\theta)}{1+\eta}\lambda_t\Big)$
Buys $\sim Poi\Big(\frac{\theta}{1+\eta}\lambda_t\Big)$
Sells $\sim Poi\Big((1-\frac{\theta}{1+\eta})\lambda_t\Big)$
Buys $\sim Poi\big(\theta\lambda_t\big)$
Sells $\sim Poi\big((1-\theta)\lambda_t\big)$

Figure 7: **XOM GPIN.** This figure compares the real and simulated data for XOM in 1993 using the GPIN model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{GPIN}$, with darker markers (+ magenta) representing high and lighter markers (+ cyan) low $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark "cloud" of points, and low probability states appear as a light "cloud" of points. The GPIN model has three states: no news, good news, and bad news. Panels C and D plot the CPIE values for the real data as a function of turnover along with a dashed vertical line indicating the annual mean of daily turnover.


(a) XOM 1993

(b) XOM 2012

(c) XOM 1993

(d) XOM 2012

Figure 8: **OWR Tree.** In the OWR model, prior to markets opening, private information arrives with probability α. 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 and covariance Σ, where Σ is function of the information arrival indicator (I). For instance, when there is no private information, there is a reversal in the returns ($cov(r_d, r_o) < 0$) and when there is private information there is a continuation in the returns ($cov(r_d, r_o) > 0$).

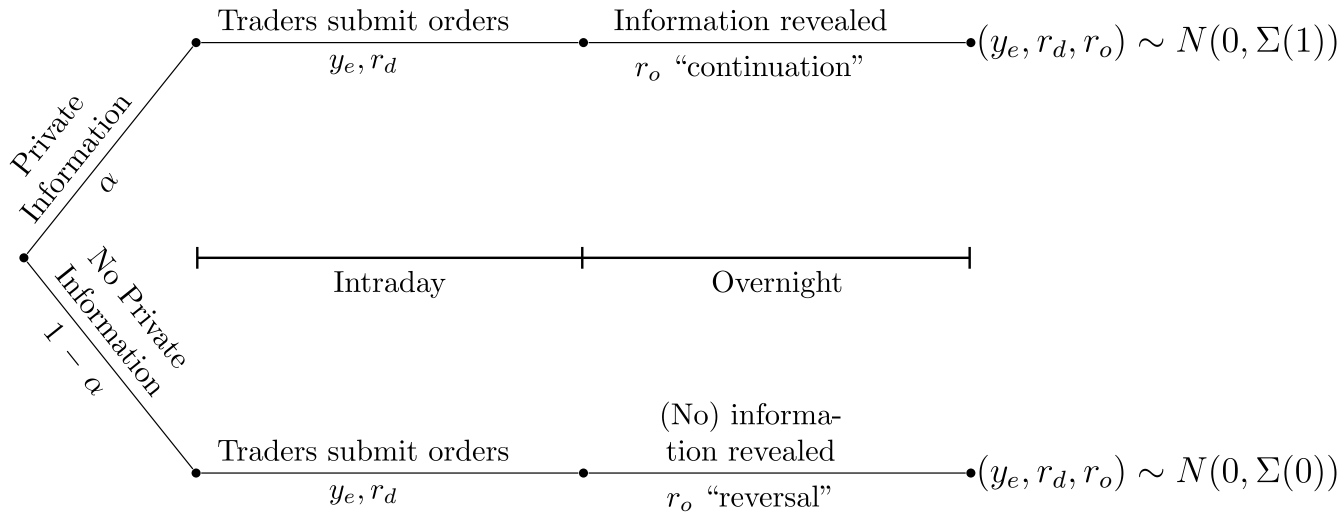


Figure 9: **OWR** α**.** This figure shows the distribution of yearly α parameter estimates for the OWR model. The solid black line represents the median value, and the dashed lines represent the 5, 25, 75, and 95 percentiles.

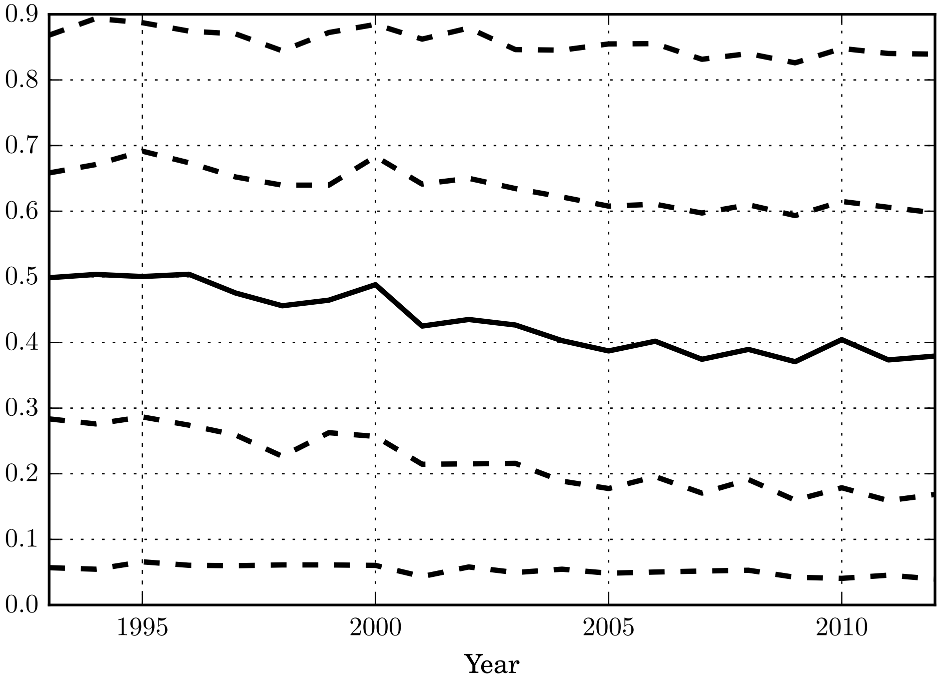
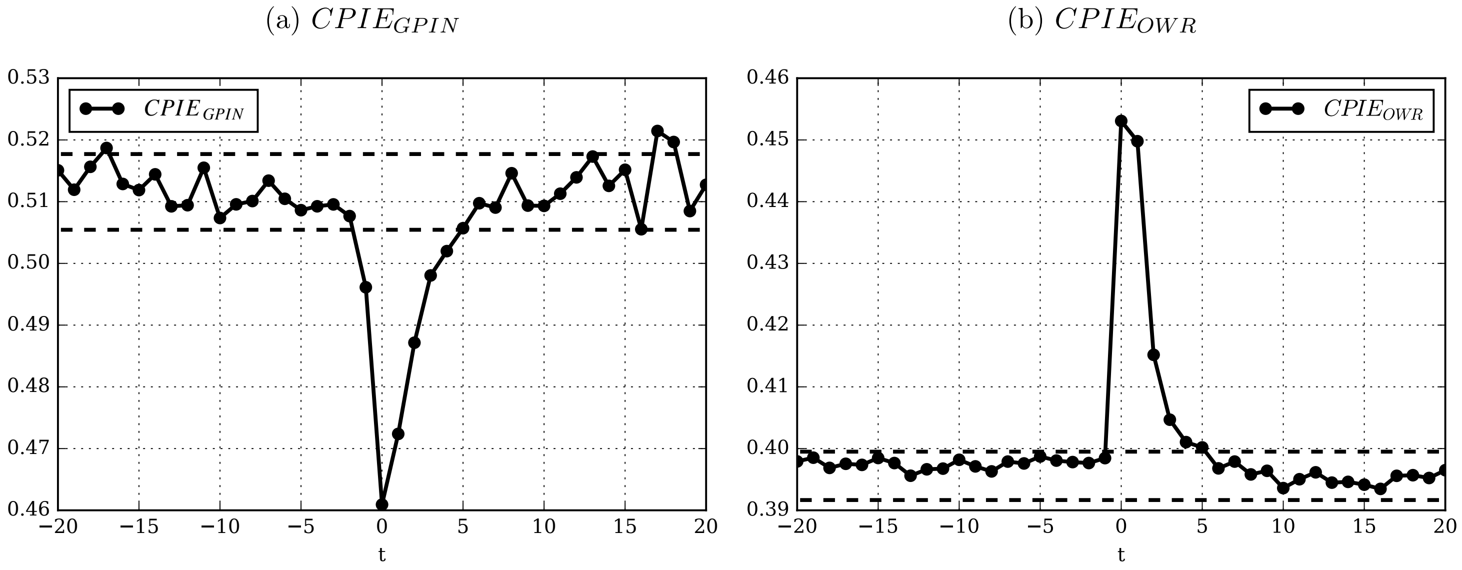
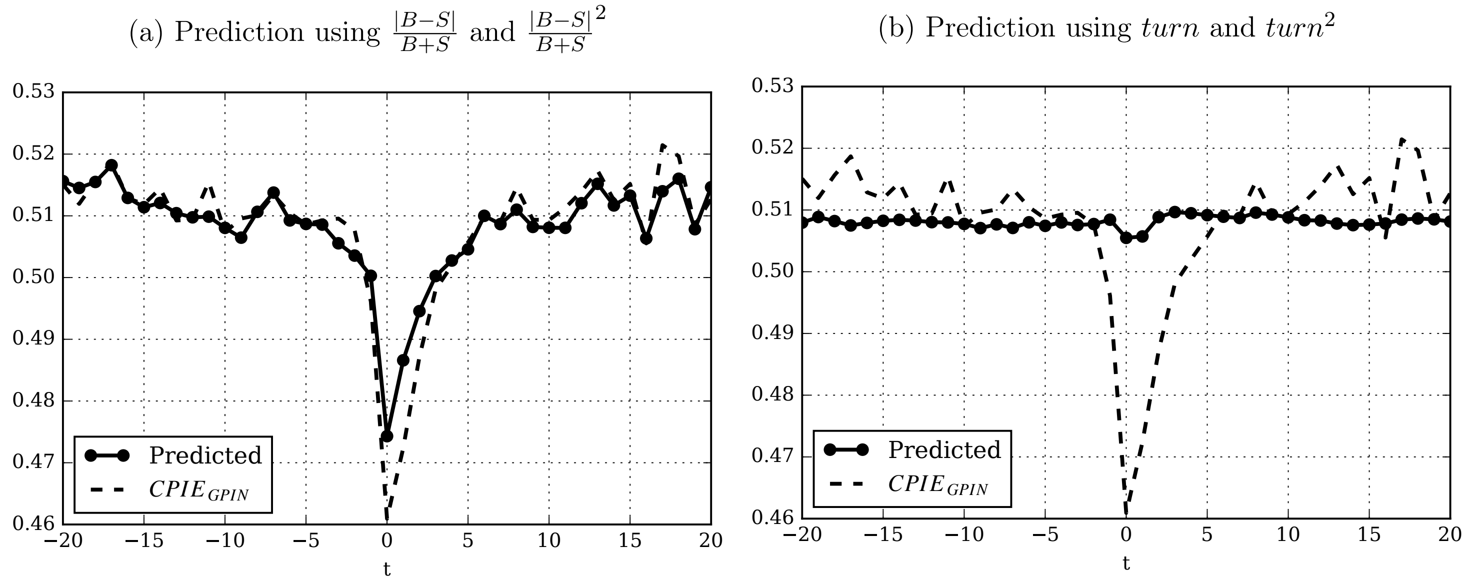


Figure 10: **Earnings Announcements.** Panel A (B) shows the average $CPIE_{GPIN}$ ($CPIE_{OWR}$) for the GPIN (OWR) model in event time surrounding earnings announcements.


(a) $CPIE_{GPIN}$

(b) $CPIE_{OWR}$

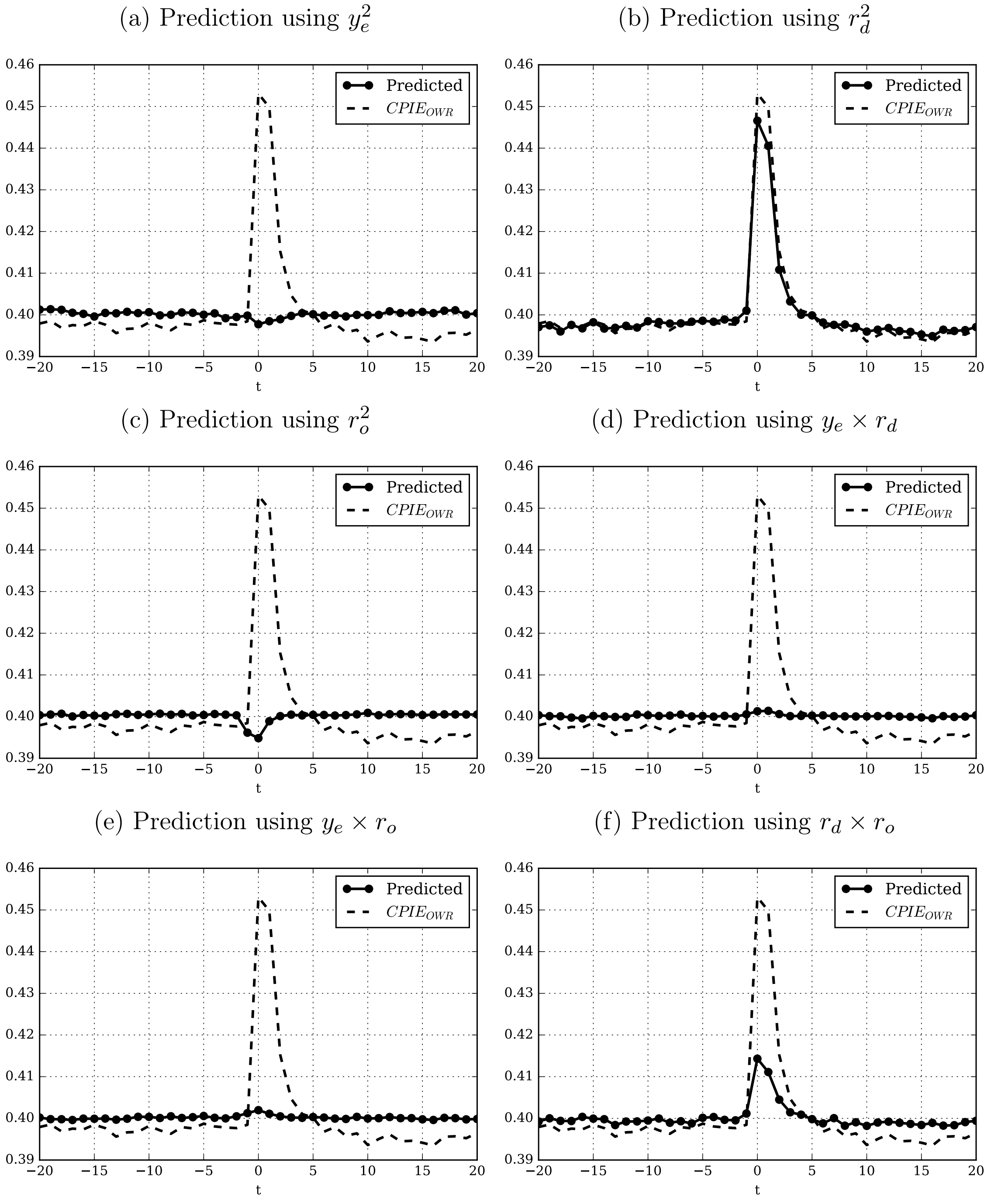
Figure 11: **Earnings Announcements - GPIN Decomposition.** Panels A and B compare the average $CPIE_{GPIN}$ with the $CPIE_{GPIN}$ predicted using either $\frac{|B-S|}{B+S}$ or turnover ($turn$), respectively. To obtain the predictions, we run regressions of daily $CPIE_{GPIN}$ on $\frac{|B-S|}{B+S}$ or $turn$, and their respective squared terms.



(a) Prediction using $\frac{|B-S|}{B+S}$ and $\frac{|B-S|}{B+S}^2$

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

Figure 12: **Earnings Announcements - OWR Decomposition.** Panels A–F compare the average $CPIE_{OWR}$ with the $CPIE_{OWR}$ predicted using the squared and interaction terms of y_e, r_d, and r_o.



(a) Prediction using y_e^2

(b) Prediction using r_d^2

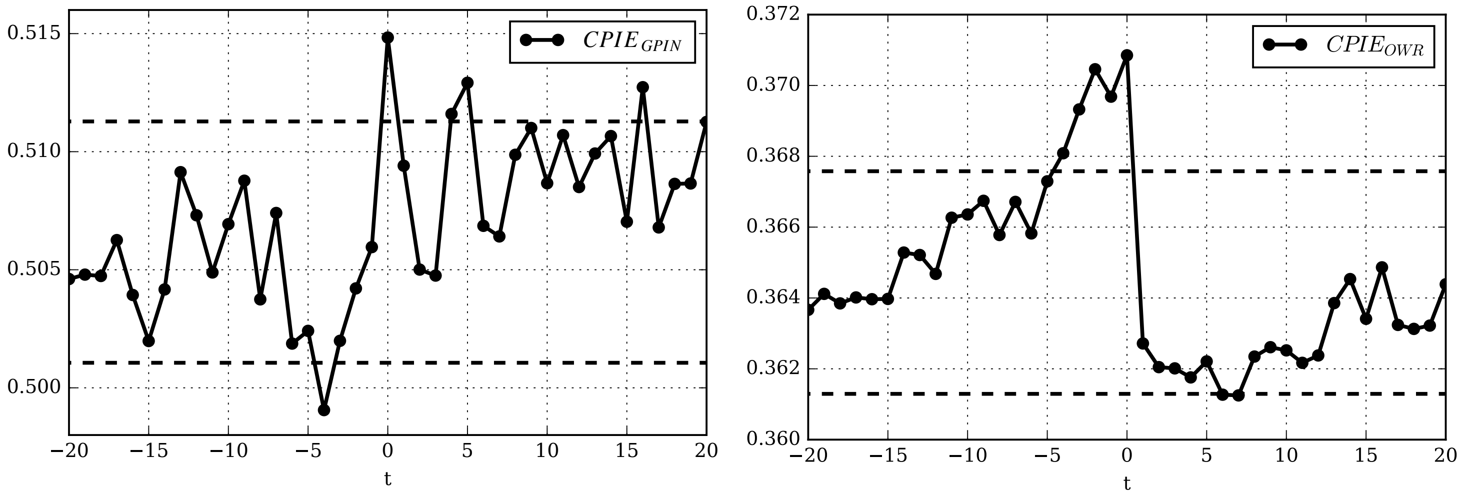
(c) Prediction using r_o^2

(d) Prediction using $y_e \times r_d$

(e) Prediction using $y_e \times r_o$

(f) Prediction using $r_d \times r_o$

Figure 13: **Opportunistic Insider Trades.** Panel A (B) shows the average $CPIE_{GPIN}$ ($CPIE_{OWR}$) for the GPIN (OWR) model in event time surrounding opportunistic insider trades.



(a) $CPIE_{GPIN}$

(b) $CPIE_{OWR}$

Internet Appendix: Does the PIN model mis-identify private information and if so, what are our alternatives?

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

March 8th, 2017

[*] 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 authors colleagues upon the staff of the Commission.

A The DY model

Duarte and Young (2009) propose an extension of the PIN model that accounts for the positive correlation between buys and sells. We show in this Appendix that the Duarte and Young (2009) model also performs poorly late in our sample from 1993–2012.

A.1 The DY model

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

In contrast to the PIN model, the DY model allows for symmetric order flow shocks. These shocks increase both the number of buyer- and seller-initiated trades but are unrelated to private information events. Symmetric order flow shocks can happen for a variety of reasons, such as disagreement among traders about the interpretation of public news. Alternatively, liquidity shocks may occur that cause investors holding different collections of assets to simultaneously rebalance their portfolios, resulting in increases to both buys and sells. Regardless of the mechanism, symmetric order flow shocks arrive on any given day with probability θ. On days with symmetric order flow shocks, both the number of buyer- and seller-initiated trades increase by amounts drawn from independent Poisson distributions with intensity Δ_B or Δ_S, respectively. Buy and sell orders from uninformed traders arrive according to a Poisson distribution with intensities ϵ_B ($\epsilon_B + \Delta_B$) and ϵ_S ($\epsilon_S + \Delta_S$) on days without (with) symmetric order flow shocks. Fig. A1 shows the structure of the DY model.

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

A.2 Estimation of the DY model

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

$$\begin{aligned} L(D_{DY,i,t}) &= L_{NI,NS}(D_{DY,i,t}) + L_{NI,S}(D_{DY,i,t}) + L_{I^-,NS}(D_{DY,i,t}) \\ &\quad + L_{I^-,S}(D_{DY,i,t}) + L_{I^+,NS}(D_{DY,i,t}) + L_{I^+,S}(D_{DY,i,t}) \end{aligned} \tag{1}$$

where $L_{NI,NS}(D_{DY,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information or a symmetric order flow shock; $L_{NI,S}(D_{DY,i,t})$ is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day without private information but with a symmetric order flow shock; $L_{I^-,NS}$ ($L_{I^-,S}$) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative information and without (with) a symmetric order flow shock; and $L_{I^+,NS}$ ($L_{I^+,S}$) is the probability on a day with positive information and without (with) a symmetric order flow shock. Analogous to the original PIN model, each term in the likelihood function corresponds to a branch in the tree in Fig. A1 and each term is given by:

$$L_{NI,NS}(D_{DY,i,t}) = (1-\alpha_i)(1-\theta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \tag{2}$$

$$L_{NI,S}(D_{DY,i,t}) = (1-\alpha_i)\theta_i e^{-(\epsilon_{B_i}+\Delta_{B_i})}\frac{(\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\epsilon_{S_i}+\Delta_{S_i})}\frac{(\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \tag{3}$$

$$L_{I^-,NS}(D_{DY,i,t}) = \alpha_i(1-\theta_i)(1-\delta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_{S_i}+\epsilon_{S_i})}\frac{(\mu_{S_i}+\epsilon_{S_i})^{S_{i,t}}}{S_{i,t}!} \tag{4}$$

$$L_{I^-,S}(D_{DY,i,t}) = \alpha_i\theta_i(1-\delta_i)e^{-(\epsilon_{B_i}+\Delta_{B_i})}\frac{(\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_{S_i}+\epsilon_{S_i}+\Delta_{S_i})}\frac{(\mu_{S_i}+\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \tag{5}$$

$$L_{I^+,NS}(D_{DY,i,t}) = \alpha_i(1-\theta_i)\delta_i e^{-(\mu_{B_i}+\epsilon_{B_i})}\frac{(\mu_{B_i}+\epsilon_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_S}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \tag{6}$$

$$L_{I^+,S}(D_{DY,i,t}) = \alpha_i\theta_i\delta_i e^{-(\mu_{B_i}+\epsilon_{B_i}+\Delta_{B_i})}\frac{(\mu_{B_i}+\epsilon_{B_i}+\Delta_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-(\epsilon_{S_i}+\Delta_{S_i})}\frac{(\epsilon_{S_i}+\Delta_{S_i})^{S_{i,t}}}{S_{i,t}!} \tag{7}$$

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

A.3 $CPIE_{DY}$

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

$$CPIE_{DY,i,t} = \frac{L_{I^+,NS}(D_{DY,i,t}) + L_{I^+,S}(D_{DY,i,t}) + L_{I^-,S}(D_{DY,i,t}) + L_{I^-,NS}(D_{DY,i,t})}{L(D_{DY,i,t})} \tag{8}$$

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

Table A1 contains summary statistics for the parameter estimates for the DY model as well as summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{DY}$. We see that the mean $CPIE$ behaves exactly like α. Hence, changes in $CPIE_{DY}$ and changes in the estimated alphas are analogous.

A.4 How does the DY model identify private information?

To illustrate how the $CPIE_{DY}$ works, we present a stylized example of the DY model in Fig. A2. In Panel A we plot simulated and real order flow data for Exxon-Mobil during 1993, with buys on the horizontal axis and sells on the vertical axis. Real data are marked as +, and simulated data as transparent dots. The real data are shaded according to the $CPIE$, with lighter points (+ cyan) representing low and darker points (+ magenta) high $CPIE$s.

The DY model generates six data clusters, greatly improving upon the PIN model's coverage of the data in 1993. The two clusters on the dotted line are not related to private information, but the other four clusters are. An econometrician using the DY model, moving along the dotted line, would observe that high turnover days–considered information days under the PIN model–are no longer classified as such, because higher turnover may be driven by symmetric order flow shocks under the DY model. Instead, the DY model identifies private information when moving away from the dotted line; when buys are greater than sells and vice versa.

Unfortunately, late in the sample the DY model breaks down. Panel B of Fig. A2 shows that the DY model, like the PIN model, fails to fit the majority of the order flow data for Exxon-Mobil in 2012. The problem of fitting the data is not limited to our stylized example. Fig. A3 shows that after 2005 the DY model estimates that the total likelihood for 80% of the order flow data of the median stock is less than 10^{-10}.

As a more formal test of the DY model, Table A2 presents regressions of $CPIE_{DY}$ based on simulated and real data. The right-hand side variables are the absolute order imbalance adjusted for buy/sell correlations ($|adj.OIB|$), turnover and its squared term. We define the adjusted absolute order imbalance as the absolute value of the residual from a regression of buys on sells. We use this measure to analyze the DY model because, as Fig. A2 suggests, the DY model implies that days with information events are far from the dashed line in this figure.[1] Turnover, as before, is defined as the sum of buys and sells. We report median coefficient estimates and $t-$statistics across all firms within a particular year. The coefficients are standardized as above. We report the average of the median, the 5^{th}, and the 95^{th} percentiles of the R^2s and R^2_{inc}s.

As with the $CPIE_{PIN}$, in theory, turnover has little additional power in explaining $CPIE_{DY}$. The incremental R^2s in Table A2 Panel A are low with an average value close to 4%. This is smaller than the average incremental R^2s of the PIN model. The intuition for this result is that the DY model disentangles turnover and order flow shocks by including the possibility of symmetric order flow shocks. Buying and selling activity can simultaneously be higher than average, but this is not indicative of private information unless there is a large order flow imbalance.

Panel B of Table A2 reports regression results for the real, rather than simulated, data. The DY model behaves very differently when using real data as opposed to data generated from the model. The R^2s for the real data are much lower than those in the simulated data, declining from 35% in 1993 to 12% in 2012. The incremental R^2 indicates that turnover and turnover squared explain a large degree of variation in $CPIE_{DY}$. Indeed, the average ratio of the median R^2s, $R^2_{inc.}/(R^2 + R^2_{inc.})$, is about 40%. The p-values are the average probability (under the DY model) of observing an incremental R^2 larger or equal to the observed in the real data and $\%Rej.$ is the frequency that we reject the null hypothesis that the incremental R^2 is consistent with the DY model at 5% significance. In 1993, our hypothesis test based rejects the model at 5% significance for 48% of the stocks, while in 2012 this percentage increases to around 70%.

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

B The EEOW model

Easley, Engle, O'Hara, and Wu (2008) propose a model in which PIN is time-varying. They estimate this model for a sample of 16 stocks. We show in this Appendix that the Easley, Engle, O'Hara, and Wu (2008) model also performs poorly late in our sample from 1993–2012.

B.1 The EEOW model

Easley, Engle, O'Hara, and Wu (2008) develop a model in which PIN is time-varying. They do so by allowing the expected number of informed and non-informed trades to be time-varying. Specifically, let the vector $\psi_t = [\alpha\mu_t, 2\epsilon_t]'$ represent the expected number of informed and non-informed trades on day t, and $\widetilde{\psi}_{i,t} = \psi_{i,t}e^{-g_i t}$, $i = 1, 2$, be the detrended arrival rates. In the Easley, Engle, O'Hara, and Wu (2008) model, $\widetilde{\psi}_t$ follows the generalized autoregressive process

$$\widetilde{\psi}_t = \omega + \Phi\widetilde{\psi}_{t-1} + \Gamma\widetilde{Z}_t \tag{9}$$

where ω, Φ, and Γ are matrices of constants and $\widetilde{Z}$ is a vector composed by the detrended absolute order imbalance and the detrended total number of trades minus the absolute order imbalance. Conditional on ψ_t, the arrival rates of buyer and seller initiated trades are like those in the PIN model. That is, the uninformed buys and sells are distributed as Poisson with intensity ϵ_t, the number of trades initiated by informed investors are distributed as Poisson with intensity μ_t, and private information arrives in the beginning of the day with probability α.

B.2 Estimation of the EEOW model

As with the PIN model, we estimate the EEOW model numerically via maximum likelihood. We estimate the EEOW for the same 16 stocks as in Easley, Engle, O'Hara, and Wu (2008) for the sample period between 1993 and 2012. Let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock i on day t and $\Theta_{EEOW,i,t} = (\alpha_i, \mu_{i,t}, \epsilon_{i,t}, \delta_i)$ represent a vector with some the EEOW model parameters for stock i. Let $D_{EEOW,i,t} = [\Theta_{EEOW,i,t}, B_{i,t}, S_{i,t}]$. Conditional on $\epsilon_{i,t}$ and $\mu_{i,t}$, the EEOW model behaves as the PIN model. Therefore, conditional on $\epsilon_{i,t}$ and

$\mu_{i,t}$, the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without an information event, on a day with positive information event, and on a day with a negative information event are:

$$L_{NI}(D_{EEOW,i,t}) = (1-\alpha_i)e^{-\epsilon_{i,t}}\frac{\epsilon_{i,t}^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{i,t}}\frac{\epsilon_{i,t}^{S_{i,t}}}{S_{i,t}!} \tag{10}$$

$$L_{I^+}(D_{EEOW,i,t}) = \alpha_i\delta_i e^{-(\mu_{i,t}+\epsilon_{i,t})}\frac{(\mu_{i,t}+\epsilon_{i,t})^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{i,t}}\frac{\epsilon_{i,t}^{S_{i,t}}}{S_{i,t}!} \tag{11}$$

$$L_{I^-}(D_{EEOW,i,t}) = \alpha_i(1-\delta_i)e^{-\epsilon_{i,t}}\frac{\epsilon_{i,t}^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_{i,t}+\epsilon_{i,t})}\frac{(\mu_{i,t}+\epsilon_{i,t})^{S_{i,t}}}{S_{i,t}!} \tag{12}$$

where $L_{NI}(D_{EEOW,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information trading; L_{I^-} (L_{I^+}) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative (positive) information. The likelihood function of the EEOW model is $\prod_{t=1}^{T} L(D_{EEOW,i,t})$, where $L(D_{EEOW,i,t}) = L_{NI}(D_{EEOW,i,t}) + L_{I^+}(D_{EEOW,i,t}) + L_{I^-}(D_{EEOW,i,t})$. The parameters $\epsilon_{i,t}$ and $\mu_{i,t}$ are obtained from equation 9 and the time trend parameter are set equal to the mean of the log-growth of Z.

B.3 $CPIE_{EEOW}$

$CPIE_{EEOW}$ is defined in the same way as $CPIE_{PIN}$. That is, $CPIE_{EEOW}$ is the ratio $(L_{I^+}(D_{EEOW,i,t})+L_{I^-}(D_{EEOW,i,t}))/L(D_{EEOW,i,t})$. We compute $CPIE_{EEOW}$ in the same way that we compute $CPIE_{PIN}$(see Appendix Section D.3 for details).

B.4 How does the EEOW model identify private information?

To illustrate how the $CPIE_{EEOW}$ works, we present a stylized example of the EEOW model in Fig. A4. In Panel A we plot simulated and real order flow data for Exxon-Mobil during 1993, with buys on the horizontal axis and sells on the vertical axis. Real data are marked as +, and simulated data as transparent dots. The real data are shaded according to the $CPIE$, with lighter points (+ cyan) representing low and darker points (+ magenta) high $CPIE$s.

The EEOW model generates three data clusters, but generates increased variation in turnover relative to the PIN, due to serial correlation in arrival rates with a deterministic trend. The clusters are not as distinct as those generated by the PIN model, but the intuition is the same.

Unfortunately, late in the sample the EEOW model breaks down. Panel B of Fig. A4 shows that the EEOW model, like the PIN and DY models, fails to fit the majority of the order flow data for Exxon-Mobil in 2012. Even though the EEOW model is a much richer description of the order flow data than the PIN model, it also breaks down due to the increase in turnover that we see in the late period of our sample. The problem of fitting the data is not limited to our stylized example, and is reflected in our formal regression tests.

Table A3 presents regressions of $CPIE_{EEOW}$ based on simulated and real data. The right-hand side variables are the absolute order imbalance ($|OIB|$), turnover and their squared terms. We report median coefficient estimates and $t-$statistics across the 16 symbols from Easley, Engle, O'Hara, and Wu (2008) within a particular year. The coefficients are standardized as above. We report the average of the median, the 5^{th}, and the 95^{th} percentiles of the R^2s and R^2_{inc}s.

As with the $CPIE_{PIN}$, in theory, turnover has little additional power in explaining $CPIE_{EEOW}$. The incremental R^2s in Table A3 Panel A are low with an average value close to 4%. This is smaller than the average incremental R^2s of the PIN model. The intuition for this result is that the EEOW model disentangles turnover and order flow shocks by modelling their arrival rates separately in the GARCH model.

Panel B of Table A3 reports regression results for the real, rather than simulated, data. The EEOW model behaves very differently when using real data as opposed to data generated from the model. The R^2s for the real data are much lower than those in the simulated data. The incremental R^2 indicates that turnover and turnover squared explain a large degree of variation in $CPIE_{EEOW}$. The p-values are the average probability (under the EEOW model) of observing an incremental R^2 larger or equal to the observed in the real data and $\%Rej.$ is the frequency that we reject the null hypothesis that the incremental R^2 is consistent with the EEOW model at 5% significance. In the first half of our sample, our hypothesis test based rejects the model at 5% significance for 60% of the stocks, while in the latter half this percentage increases to around 63%.

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

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

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

To estimate $r_{o,i,t}$ and $r_{d,i,t}$, we run daily cross-sectional regressions of overnight and intraday returns on a constant, historical β (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 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 ($r_{d,i,t}$) and overnight ($r_{o,i,t}$) returns.

D Details of the PIN model

D.1 PIN Likelihood

Let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock i on day t and $\Theta_{PIN,i} = (\alpha_i, \mu_i, \epsilon_{B_i}, \epsilon_{S_i}, \delta_i)$ represent the vector of the PIN model parameters for stock i. Let $D_{PIN,i,t} = [\Theta_{PIN,i}, B_{i,t}, S_{i,t}]$. The likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without an information event, on a day with positive information event, and on a day with a negative

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

information event are:

$$L_{NI}(D_{PIN,i,t}) = (1-\alpha_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \tag{13}$$

$$L_{I^+}(D_{PIN,i,t}) = \alpha_i\delta_i e^{-(\mu_i+\epsilon_{B_i})}\frac{(\mu_i+\epsilon_{B_i})^{B_{i,t}}}{B_{i,t}!}e^{-\epsilon_{S_i}}\frac{\epsilon_{S_i}^{S_{i,t}}}{S_{i,t}!} \tag{14}$$

$$L_{I^-}(D_{PIN,i,t}) = \alpha_i(1-\delta_i)e^{-\epsilon_{B_i}}\frac{\epsilon_{B_i}^{B_{i,t}}}{B_{i,t}!}e^{-(\mu_i+\epsilon_{i,S})}\frac{(\mu_i+\epsilon_{i,S})^{S_{i,t}}}{S_{i,t}!} \tag{15}$$

where $L_{NI}(D_{PIN,i,t})$ is the likelihood of observing $B_{i,t}$ and $S_{i,t}$ on a day without private information trading; L_{I^-} (L_{I^+}) is the likelihood of $B_{i,t}$ and $S_{i,t}$ on a day with negative (positive) information.

D.2 Maximum likelihood procedure

To estimate the PIN likelihood function, we use the maximum of the likelihood maximization with ten different starting points as in Duarte and Young (2009). We note, however, that late in the sample, the likelihood functions of the PIN are very close to zero. After 2006, the PIN model suggests that 90% of the observed daily order flows for the median stock have a near-zero probability (i.e. smaller than 10^{-10}) of occurring. This makes the estimation susceptible to local optima. To get around this problem, we choose one of our ten starting points to be such that the PIN model clusters are close to the observed mean of the number of buys and sells. Specifically, we choose ϵ_B and ϵ_S values equal to the sample means of buys and sells, α equal to 1%, and delta equal to the mean absolute value of order imbalance. The other nine starting points are randomized. We do this in order to ensure that at least one of the starting points is centered properly, as the numerical likelihood estimation using purely random starts often stops at points outside of the central cluster of data.

D.3 Computing $CPIE_{PIN}$

In Section 2 of the paper, we define the $CPIE$ as the ratio of the "news" likelihood functions to the sum total of the likelihood functions. In practice, there are many cases in the PIN model for which the data a near-zero probability of occurring, meaning $L(D_{PIN,i,t}) = L_{NI}(D_{PIN,i,t}) + L_{I^+}(D_{PIN,i,t}) + L_{I^-}(D_{PIN,i,t})$ is smaller than 10^{-10}. As a result the $CPIE$ ratio frequently results in a divide by zero error.

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

$$L_{\max} \equiv \max\{L_{NI}, L_{I^+}, L_{I^-}\}, \tag{16}$$

$$\ell_{\max} \equiv \log(L_{\max}) \tag{17}$$

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

$$\ell'_{NI} = \ell_{NI} - \ell_{\max}, \tag{18}$$

$$\ell'_{I+} = \ell_{I^+} - \ell_{\max}, \tag{19}$$

$$\ell'_{I-} = \ell_{I^-} - \ell_{\max}. \tag{20}$$

We compute the $CPIE'$ as:

$$CPIE'_{PIN} = \frac{L'_{I+} + L'_{I-}}{L'_{NI} + L'_{I+} + L'_{I-}} \tag{21}$$

such that the most likely state has $L' = 1$. For a high turnover day, it may be the case that $L'_{I+} = 1$, $L'_{I-} = 0$ and $L'_{NI} = 0$; hence, the $CPIE$’ will be 1. This computational procedure is equivalent to taking the limit of $CPIE_{PIN}$ as $L(D_{PIN,i,t})$ goes to zero. We follow a similar procedure to compute $CPIE_{DY}$.

E Details of the GPIN model

The GPIN model extends the PIN model to allow for continuous variation in turnover unrelated to private information arrival.

E.1 The microstructure of the GPIN model

The market maker knows that the number of trades (i.e. $B + S$) on day t is distributed as a Poisson random variable with intensity λ_t. The trade intensity, λ_t, is drawn from a Gamma distribution with parameters r and p. In what follows, in the interest of clarity, we suppress the t subscript on λ. The market maker does not observe λ directly, she only sees the buy and sell orders as they arrive. The market maker also knows that at the beginning of every day the probability that informed traders receive a private signal is α. If the informed receive a private signal, then the market maker knows that some fraction of the day's total number of trades will be informed. If the informed traders receive no private signal, then all trades are uninformed. If there is no information in the market, then conditional on λ, the sum of buys and sells is drawn from a Poisson distribution with arrival rate λ. If informed traders do receive a private signal, η represents the ratio of the expected number of informed to uninformed trades. Thus, if informed traders receive a private signal then the fraction of informed trade to total trade is $\frac{\eta}{1+\eta}$. The corresponding fraction of uninformed trade is equal to $1 - \frac{\eta}{1+\eta} = \frac{1}{1+\eta}$. Thus, if informed traders receive a private signal, then conditional on λ, the total arrival rate of orders remains equal to $(\frac{1}{1+\eta} + \frac{\eta}{1+\eta})\lambda{=}\lambda$. It is immediately clear from this intuition that the probability of informed trade under the GPIN model is simply the unconditional expected fraction of informed trade to total trade, $PIN_{GPIN} = \frac{\alpha\eta}{1+\eta}$. The PIN_{GPIN} does not involve λ because λ determines the overall intensity of trade, but not the split between informed and uninformed trade.

Formally, the probability that any given trade is informed is equal to the expected number of informed trades divided by the expected number of trades. This ratio is:

$$\frac{\text{E[Inf. Trades]}}{\text{E[Trades]}} = \frac{\text{E[E[Inf. Trades}|\lambda]]}{\text{E[E[Trades}|\lambda]]}. \tag{22}$$

The numerator for the $GPIN$ is $\text{E}[\alpha\delta\big(\frac{\eta}{1+\eta}\big)\lambda + \alpha(1-\delta)\big(\frac{\eta}{1+\eta}\big)\lambda]$, and the denominator is simply $\text{E}[\lambda]$. Simplifying we get that $PIN_{GPIN} = \frac{\alpha\eta}{1+\eta}$.

To see the connection between the PIN_{PIN} and PIN_{GPIN}, first note that we can write the formula for PIN_{PIN} using Equation 22. Using the reparameterization of the PIN model presented in Section 3.1, the numerator is $\alpha \times E[\text{Inf. Trades}|\lambda = \lambda(1)] + (1-\alpha) \times E[\text{Inf. Trades}|\lambda = \lambda(0)]$. The expected number of informed trades on days with private information ($\lambda = \lambda(1)$) in the PIN model is μ and zero otherwise, hence the numerator of Equation 22 reduces to $\alpha \times \mu$. Under the PIN model, the denominator of Equation 22 is $\alpha \times E[\text{Trades}|\lambda = \lambda(1)] + (1-\alpha) \times E[\text{Trades}|\lambda = \lambda(0)]$. The expected number of trades on days with private information ($\lambda = \lambda(1)$) in the PIN model is $\epsilon_B + \epsilon_S + \mu$ and $\epsilon_B + \epsilon_S$ otherwise. Hence the denominator of Equation 22 reduces to $\epsilon_B + \epsilon_S + \alpha \times \mu$, which leads to the formula $PIN_{PIN} = \frac{\alpha\mu}{\alpha\mu+\varepsilon_B+\varepsilon_S}$. Note that unlike the PIN_{PIN}, α does not appear in the denominator of the PIN_{GPIN}. This difference occurs because, in the PIN model, everything else equal, stocks with higher α have higher expected turnover. This relation has a direct impact on the denominator of Equation 22 and comes about because of the conflation of expected turnover and the arrival of private information in the PIN model (see Equation 1 in the paper). In the GPIN model, on the other hand, expected turnover (λ) is drawn independently of private information arrival. Hence, α has no effect on expected turnover and thus no place in the denominator of Equation 22.

Finally, to verify that the GPIN model captures the same microstructure intuition as the PIN model, consider the bid-ask spread under the GPIN model and the PIN model. Following similar logic to that in Easley, Keifer, O'Hara and Paperman (1996), the expression for the opening bid-ask spread under the GPIN model is the same as that under the PIN model:

$$\frac{\alpha\eta}{1+\eta} \times (\overline{V} - \underline{V}) = PIN_{GPIN} \times (\overline{V} - \underline{V}) \tag{23}$$

where $\overline{V}$ is the value of the firm conditional on good news and $\underline{V}$ represents the value of the firm conditional on bad news.

E.2 Negative binomial distribution in GPIN model

In the GPIN model, conditional on λ_t the distribution of turnover ($B + S$) is *Poisson* with intensity λ_t. Moreover, λ_t is drawn from $Gamma(r, p/(1-p))$ distribution. Hence, the probability that $B + S$ is equal to x in a given day is:

$$f(x; r, p) = \int_0^\infty \frac{\lambda^x}{x!} \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{(\frac{p}{1-p})^r \Gamma(r)} d\lambda = \frac{(1-p)^r p^{-r}}{\Gamma(r)} p^{r+x} \Gamma(r+x) \tag{24}$$

which is the well known $Negative\ Binomial(r, p)$ (see Casella and Berger (2002)).

E.3 GPIN maximum likelihood estimation

Let $\Theta_{GPIN} = (\alpha, \delta, \eta, \theta, r, p)$ be the vector of parameters of the GPIN model. Let $B_{i,t}$ ($S_{i,t}$) represent the number of buys (sells) for stock i on day t and $D_{GPIN,i,t} = [\Theta_{GPIN,i}, B_{i,t}, S_{i,t}]$. The likelihood function of the extended PIN model is $\prod_{t=1}^{T} L(D_{GPIN,i,t})$, where

$$L(D_{PIN,i,t}) = L_{NI}(D_{GPIN,i,t}) + L_{I^+}(D_{GPIN,i,t}) + L_{I^-}(D_{GPIN,i,t}). \tag{25}$$

Define the function:

$$f(B,S;r,p,\theta) = \frac{\theta^B(1-\theta)^S}{B!S!}\frac{(1-p)^r p^{-r}}{\Gamma(r)} p^{r+B+S}\Gamma(r+B+S) \tag{26}$$

And the parameters $\theta_{I^+} = (\eta+\theta)/(1+\eta)$, $\theta_{I^-} = \theta/(1+\eta)$

$$\begin{aligned} L_{NI}(D_{GPIN,i,t}) = && (1-\alpha)f(B,S;r,p,\theta) \\ L_{I^+}(D_{GPIN,i,t}) = && \alpha\delta f(B,S;r,p,\theta_{I^+}) \\ L_{I^+}(D_{GPIN,i,t}) = && \alpha(1-\delta)f(B,S;r,p,\theta_{I^-}) \end{aligned} \tag{27}$$

Conditional on λ_t and analogous to the original PIN model, each term in the likelihood function corresponds to a branch in the GPIN tree in the paper. We maximize the GPIN likelihood function in two steps. First we estimate the parameters r and p to fit the $Negative\ Binomial(r,p)$ distribution to the turnover data. We then maximize the GPIN likelihood with fixed r and p to obtain estimates of α, δ, η and θ. Analogous to the estimation of the PIN likelihood, in each step we use the maximum likelihood based on ten random starting points to avoid picking up local maxima.

E.4 Computing $CPIE_{GPIN}$

As with the PIN model, for each stock-day, we compute the probability of an information event conditional on both the model parameters and on the number of buys and sells observed that day. We compute $CPIE_{GPIN,i,t} = P\left[I_{i,t} = 1 | D_{GPIN,i,t}\right]$, which is equal to $(L_{I^-}(D_{GPIN,i,t}) + L_{I^+}(D_{GPIN,i,t}))/L(D_{GPIN,i,t})$. We compute $CPIE_{GPIN}$ in the same way as we compute $CPIE_{PIN}$, see Section D.3 for details. The analytical formula for $CPIE_{GPIN}$ is:

$$CPIE_{GPIN} = \frac{\alpha\delta\theta_{I^+}^B(1-\theta_{I^+})^S + \alpha(1-\delta)\theta_{I^-}^B(1-\theta_{I^-})^S}{(1-\alpha)\theta^B(1-\theta)^S + \alpha\delta\theta_{I^+}^B(1-\theta_{I^+})^S + \alpha(1-\delta)\theta_{I^-}^B(1-\theta_{I^-})^S} \tag{28}$$

E.5 The GPIN model does not conflate turnover with private information

As a formal test of the GPIN model we run regressions of $CPIE_{GPIN}$ on the proportion of imbalanced trades $(\frac{|B-S|}{B+S})$ and a squared term $((\frac{|B-S|}{B+S})^2)$.[3] We use $\frac{|B-S|}{B+S}$ to analyze the GPIN model because, as we discuss in the paper, the GPIN model implies that days with information events are the ones in which the proportion of imbalanced trades is large.

Panel A of Table A4 presents the results of regressions based on simulated data. As in the case of the regressions for the PIN model in the paper, we report the median coefficient estimates and t-statistics. The coefficients are standardized so they represent the increase in $CPIE_{GPIN}$ due to a one standard deviation increase in the corresponding independent variable. We also report the average of the median, the 5^{th}, and the 95^{th} percentiles of the empirical distribution of R^2s of these regressions generated by the 1,000 simulations. In general the GPIN model identifies private information from the proportion of imbalanced trades. The median R^2 values are high, ranging from 61%-92%, while the incremental R^2 from turnover is small-typically below 4%.

Panel B of Table A4 reports regression results for the real rather than simulated data. In contrast to the PIN model, in the real data the GPIN model identifies private information from the proportion of imbalanced trades and not turnover. The median R^2 values are high, ranging from 38%–72%, while the incremental R^2 from turnover is small—typically below 1%. Naturally, the GPIN model is not a perfect description of the order flow data. This can be seen from the fact that R^2 values using the real data are on average lower than those in the simulated data. However, the GPIN model fixes the conflation of arrival of private information with turnover, namely in the majority of stock-year observations in the real data the incremental R^2 due to turnover is at least as large as the incremental R^2 in the simulated data. Therefore, the GPIN model, while not a perfect description of the order flow data, fixes the problem of the PIN model which mechanically identifies private information from higher turnover.

[3] We do not directly compare the simulations of the GPIN model to those of the PIN model. Instead we compare the real data for each model to the simulated data under the null hypothesis that each model identifies information consistent with the theory.

F Details about the OWR model

F.1 OWR Likelihood

Let $\Theta_{OWR,i} = (\alpha_i, \sigma_{u_i}, \sigma_{z_i}, \sigma_{i_i}, \sigma_{p,d_i}, \sigma_{p,o_i})$ be the vector of parameters of this model. The parameter α_i is the probability that there is an information event on a given day. $\sigma^2_{z_i}$ is the variance of the noise of the observed net order flow (y_e); $\sigma^2_{u_i}$ is the variance of the net order flow from noise traders; $\sigma^2_{i_i}$ is the variance of the private signal received by the informed trader; σ^2_{p,d_i} is the variance of the intraday return; σ^2_{p,o_i} is the variance of the overnight return. Let $r_{d,i,t}$, $(r_{o,i,t})$ represent the intraday and overnight returns for stock i on day t, and $(y_{e,i,t})$ represent the order flow imbalance for stock i on day t. Let $D_{OWR,i,t} = [\Theta_{OWR,i}, r_{d,i,t}, r_{o,i,t}, y_{e,i,t}]$. The likelihood of observing $D_{OWR,i,t}$ on a day without and with an information event is:

$$
\begin{aligned}
L_{NI} &= (1-\alpha) f_{NI}(D_{OWR,i,t}) && (29)\\
L_I &= \alpha f_I(D_{OWR,i,t}) && (30)
\end{aligned}
$$

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

$$
\begin{aligned}
Var(y_e) &= \sigma_u^2 + \sigma_z^2, && (31)\\
Var(r_d) &= \sigma_{pd}^2 + \alpha\sigma_i^2/4, && (32)\\
Var(r_o) &= \sigma_{po}^2 + \alpha\sigma_i^2/4, && (33)\\
Cov(r_d, r_o) &= -\alpha\sigma_i^2/4, && (34)\\
Cov(r_d, y_e) &= \alpha^{1/2}\sigma_i\sigma_u/2, && (35)\\
Cov(r_o, y_e) &= -\alpha^{1/2}\sigma_i\sigma_u/2 && (36)
\end{aligned}
$$

And Ω_{I_i}:

$$\begin{aligned}
Var(y_e) &= (1+1/\alpha)\sigma_u^2 + \sigma_z^2, & (37)\\
Var(r_d) &= \sigma_{pd}^2 + (1+\alpha)\sigma_i^2/4, & (38)\\
Var(r_o) &= \sigma_{po}^2 + (1+\alpha)\sigma_i^2/4, & (39)\\
Cov(r_d, r_o) &= (1-\alpha)\sigma_i^2/4, & (40)\\
Cov(r_d, y_e) &= \alpha^{-1/2}\sigma_i\sigma_u/2 + \alpha^{1/2}\sigma_i\sigma_u/2, & (41)\\
Cov(r_o, y_e) &= \alpha^{-1/2}\sigma_i\sigma_u/2 - \alpha^{1/2}\sigma_i\sigma_u/2 & (42)
\end{aligned}$$

F.2 How does the OWR model identify private information?

In theory, the OWR model identifies private information from the covariance matrix of the three variables in the model $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$. To analyze the model, we run the regression of $CPIE_{OWR}$ on the squared and interaction terms of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$:

$$CPIE_{OWR,i,t} = \beta_0 + \beta_1 y_{e,i,t}^2 + \beta_2 r_{d,i,t}^2 + \beta_3 r_{o,i,t}^2 + \beta_4 y_{e,i,t} r_{d,i,t} + \beta_5 y_{e.i,t} r_{o,i,t} + \beta_6 r_{d,i,t} r_{o,i,t} + u_{i,t}. \quad (43)$$

Panel A of Table A5 presents median coefficient estimates, t-statistics, and three percentiles of R^2s across all firms within a particular year using simulated data. The results highlight the intuition behind the model. The probability of an information event on any given day is increasing in the square of intraday returns, the interaction between imbalance and intraday (or overnight) returns, and the interaction between intraday and overnight returns. The coefficient estimates on the square of the order imbalance and on the square of overnight returns are too small to be precisely measured. The high R^2s indicate that, practically speaking, the square of intraday returns, the interaction between intraday and overnight returns and the interaction between intraday returns and order flow imbalance are sufficient to explain a large part of the variation in $CPIE_{OWR}$.

Panel B of Table A5 shows the median coefficient estimates, t-statistics, and the results of the hypothesis tests based on R^2s across all firms within a particular year using real data. Unlike the PIN, DY, and EEOW models, the coefficient estimates are consistent across the simulated and real data. For instance in simulated data regressions in Panel A, 2008 is the only year in which y_e^2 is the most important term. In the real data regressions in Panel B, 2008 is also the only year in which y_e^2 is the most important term, indicating that the model matches the features of the data quite well, even for clear outliers like 2008. Furthermore, as with the simulated data regressions, the high median R^2s indicate that a large part of the variation in $CPIE_{OWR}$ is explained by the squared and interaction terms of $(y_{e,i,t}, r_{o,i,t}, r_{d,i,t})$ as implied by the model. The average across years of the R^2s in Panel B is about 83% and these R^2s increase over time, reaching 90% in 2012. Moreover, we reject the null hypothesis that the R^2s observed in the real data are consistent with the OWR model at 5% level for about 40% of the sample in 1993 and for about 8% of the sample in 2012.

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

Table A1: **DY Estimates.** This table summarizes parameter estimates of the DY model for 21,206 PERMNO-Year samples from 1993–2012. α represents the average unconditional probability of an information event at the daily level. ϵ_B and ϵ_S represent the expected number of daily buys and sells given no private information or symmetric order flow shocks. μ_b, and μ_s represent the expected additional order flows given an information event, which is good news with probability δ and bad news with probability $1 - \delta$. A symmetric order flow shock occurs with probability θ, in which case the expected number of buys and sells increase by Δ_B and Δ_S, respectively. $\overline{CPIE}$ and Std$(CPIE)$ are the PERMNO-Year mean and standard deviation of $CPIE_{DY}$.

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

Table A2: **DY Model Regressions.** This table reports real and simulated regressions of the $CPIE_{DY}$ on absolute adjusted order imbalance (|adj. OIB|), and absolute adjusted order imbalance squared (|adj. OIB|^2). In Panel A, we simulate 1,000 instances of the DY model for each `PERMNO`-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the R^2 attributed to *turn* and *turn*^{2} in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 and $R^2_{inc.}$ values, and tests of the null hypothesis that the observed relation between $CPIE_{DY}$ and *turn* is consistent with the DY model. The p-value is the average probability of observing an $R^2_{inc.}$ at least as large as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

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

Table A2: **DY Model Regressions.** Continued.

(b) Real Data

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

Table A3: **EEOW Model Regressions.** This table reports real and simulated regressions of the $CPIE_{EEOW}$ on absolute order imbalance (|OIB|), and absolute adjusted order imbalance squared (|OIB|^2). In Panel A, we simulate 1,000 instances of the EEOW model for each symbol in Easley, Engle, O'Hara, and Wu (2008) found in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the R^2 attributed to *turn* and *turn*^{2} in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 and $R^2_{inc.}$ values, and tests of the null hypothesis that the observed relation between $CPIE_{EEOW}$ and *turn* is consistent with the EEOW model. The p-value is the average probability of observing an $R^2_{inc.}$ at least as large as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

	β		t		R^2			$R^2_{inc.}$		
	|OIB|	|OIB|^2	|OIB|	|OIB|^2	5%	50%	95%	5%	50%	95%
1993	0.090	-0.062	(10.10)	(-7.01)	21.43%	43.47%	58.48%	0.10%	1.49%	6.83%
1994	0.130	-0.109	(8.66)	(-6.41)	33.40%	49.39%	58.85%	2.70%	8.03%	15.24%
1995	0.116	-0.077	(7.73)	(-6.02)	25.81%	37.36%	47.11%	0.77%	4.64%	16.22%
1996	0.088	-0.065	(7.23)	(-5.64)	20.76%	36.89%	55.70%	0.27%	2.44%	8.56%
1997	0.136	-0.113	(7.20)	(-6.02)	18.11%	30.81%	47.95%	0.44%	2.54%	9.36%
1998	0.186	-0.161	(8.41)	(-7.26)	22.91%	37.23%	49.63%	0.23%	2.39%	6.76%
1999	0.215	-0.154	(7.33)	(-5.96)	23.48%	29.98%	42.98%	0.21%	1.66%	6.05%
2000	0.208	-0.132	(7.09)	(-5.75)	34.32%	45.42%	53.80%	0.89%	6.04%	19.04%
2001	0.207	-0.174	(9.12)	(-7.27)	19.84%	32.75%	47.72%	0.09%	1.32%	6.55%
2002	0.297	-0.188	(9.78)	(-7.24)	36.97%	47.89%	53.86%	0.43%	4.08%	10.76%
2003	0.243	-0.179	(7.86)	(-5.79)	27.14%	36.72%	47.99%	2.52%	7.45%	16.81%
2004	0.108	-0.097	(7.14)	(-6.24)	11.43%	28.25%	40.21%	0.31%	2.78%	6.98%
2005	0.438	-0.322	(9.86)	(-7.73)	36.17%	42.20%	48.99%	1.08%	5.71%	9.84%
2006	0.409	-0.241	(7.86)	(-5.40)	41.73%	49.05%	57.73%	0.53%	1.88%	3.69%
2007	0.393	-0.268	(9.56)	(-5.68)	37.00%	48.45%	58.58%	0.25%	7.15%	12.66%
2008	0.321	-0.226	(10.41)	(-6.02)	46.49%	51.57%	64.73%	0.05%	0.59%	2.07%
2009	0.360	-0.275	(8.02)	(-4.77)	33.72%	42.80%	46.99%	0.46%	5.36%	14.52%
2010	0.443	-0.311	(11.06)	(-7.85)	43.37%	48.64%	53.34%	0.20%	0.80%	3.40%
2011	0.532	-0.322	(15.92)	(-10.48)	51.05%	55.02%	67.94%	0.16%	3.34%	6.74%
2012	0.241	-0.199	(7.99)	(-5.29)	27.52%	38.67%	51.73%	2.46%	4.10%	9.54%

Table A3: **EEOW Model Regressions.** Continued.

(b) Real Data

	β		t		R^2	$R^2_{inc.}$		
	|OIB|	|OIB|^2	|OIB|	|OIB|^2	50%	50%	p-value	% Rej.
1993	0.166	-0.057	(3.44)	(-2.17)	24.68%	7.74%	1.50%	64.29%
1994	0.238	-0.138	(6.68)	(-3.32)	34.99%	15.81%	0.50%	57.14%
1995	0.153	-0.114	(5.38)	(-3.95)	27.72%	11.75%	8.00%	46.67%
1996	0.156	-0.078	(3.96)	(-2.04)	27.53%	13.00%	3.00%	53.33%
1997	0.221	-0.134	(4.63)	(-2.67)	26.21%	16.24%	1.00%	56.25%
1998	0.190	-0.073	(3.46)	(-1.64)	28.01%	17.19%	0.00%	66.67%
1999	0.186	-0.094	(3.34)	(-1.76)	25.15%	7.69%	4.00%	57.14%
2000	0.297	-0.148	(6.11)	(-3.34)	35.84%	15.06%	3.50%	57.14%
2001	0.270	-0.162	(5.00)	(-2.95)	30.90%	15.10%	0.00%	83.33%
2002	0.330	-0.181	(6.16)	(-3.50)	39.71%	11.78%	12.50%	50.00%
2003	0.189	-0.096	(3.51)	(-2.15)	18.92%	9.01%	48.00%	28.57%
2004	0.193	-0.093	(3.80)	(-1.97)	26.30%	12.90%	0.00%	54.55%
2005	0.146	-0.059	(2.74)	(-1.18)	32.17%	15.22%	36.50%	41.67%
2006	0.262	-0.127	(4.91)	(-2.34)	29.26%	15.61%	0.00%	60.00%
2007	0.109	0.002	(1.63)	(0.03)	36.89%	29.67%	0.00%	70.00%
2008	0.123	-0.018	(1.88)	(-0.28)	37.87%	32.38%	0.00%	72.73%
2009	0.118	-0.032	(2.01)	(-0.53)	27.83%	16.28%	0.00%	57.14%
2010	0.050	0.030	(0.67)	(0.45)	19.31%	14.07%	0.00%	90.00%
2011	0.095	-0.001	(1.28)	(0.03)	33.20%	28.73%	0.00%	88.89%
2012	0.163	-0.030	(2.18)	(-0.45)	26.24%	19.81%	0.00%	66.67%

Table A4: **GPIN Model Regressions.** This table reports real and simulated regressions of the $CPIE_{GPIN}$ on the proportion of imbalanced trades $\left(\frac{|B-S|}{B+S}\right)$ and its square. In Panel A, we simulate 1,000 instances of the GPIN model for each `PERMNO`-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 ($R^2_{inc.}$) values. We compute the incremental $R^2_{inc.}$ as the R^2 attributed to *turn* and $turn^2$ in an extended regression model. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 and $R^2_{inc.}$ values, and tests of the null hypothesis that the observed relation between $CPIE_{GPIN}$ and *turn* is consistent with the GPIN model. The *p*-value is the average probability of observing an $R^2_{inc.}$ at least as large as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the hypothesis at the 5% level.

(a) Simulated Data

	β		t		R^2			$R^2_{inc.}$		
	$\frac{\lvert B-S\rvert}{B+S}$	$\left(\frac{\lvert B-S\rvert}{B+S}\right)^2$	$\frac{\lvert B-S\rvert}{B+S}$	$\left(\frac{\lvert B-S\rvert}{B+S}\right)^2$	5%	50%	95%	5%	50%	95%
1993	0.382	-0.134	(8.22)	(-3.04)	57.61%	63.37%	68.65%	1.79%	4.07%	7.31%
1994	0.355	-0.119	(8.19)	(-2.83)	56.90%	62.64%	67.90%	1.76%	4.23%	7.74%
1995	0.350	-0.113	(7.86)	(-2.59)	59.18%	64.87%	69.82%	1.68%	3.86%	7.24%
1996	0.364	-0.122	(8.31)	(-2.90)	60.59%	65.85%	70.81%	1.60%	3.84%	6.94%
1997	0.369	-0.126	(8.03)	(-2.84)	58.63%	64.01%	69.13%	1.29%	3.34%	6.34%
1998	0.388	-0.131	(8.93)	(-3.03)	60.99%	66.95%	71.74%	1.02%	2.81%	5.69%
1999	0.465	-0.190	(10.90)	(-4.26)	64.29%	69.23%	73.64%	1.01%	2.71%	5.10%
2000	0.447	-0.171	(9.34)	(-3.60)	60.81%	65.74%	70.43%	0.82%	2.42%	4.95%
2001	0.425	-0.123	(6.77)	(-2.08)	59.82%	65.02%	70.21%	0.71%	2.13%	4.40%
2002	0.243	0.007	(2.86)	(0.08)	55.43%	61.22%	66.58%	0.52%	1.87%	3.97%
2003	0.033	0.202	(0.30)	(1.95)	56.10%	62.06%	67.76%	0.51%	1.78%	4.05%
2004	-0.477	0.679	(-4.25)	(6.10)	56.37%	62.52%	68.15%	0.38%	1.47%	3.43%
2005	0.343	-0.062	(3.38)	(-0.67)	64.83%	70.03%	74.47%	0.16%	0.86%	2.23%
2006	0.294	-0.018	(3.16)	(-0.21)	72.38%	77.14%	80.90%	0.06%	0.42%	1.30%
2007	0.778	-0.338	(17.81)	(-7.59)	86.47%	88.49%	90.35%	0.02%	0.17%	0.54%
2008	0.784	-0.335	(18.60)	(-7.90)	90.29%	91.75%	93.13%	0.01%	0.12%	0.42%
2009	0.774	-0.321	(19.72)	(-8.04)	91.13%	92.47%	93.73%	0.01%	0.12%	0.40%
2010	0.773	-0.318	(19.47)	(-7.97)	90.93%	92.27%	93.57%	0.01%	0.13%	0.45%
2011	0.783	-0.335	(19.80)	(-8.16)	91.08%	92.48%	93.67%	0.01%	0.11%	0.40%
2012	0.781	-0.332	(19.89)	(-8.23)	90.82%	92.27%	93.54%	0.01%	0.12%	0.41%

Table A4: **GPIN Model Regressions.** Continued.

(b) Real Data

	β		t		R^2	$R^2_{inc.}$		
	$\frac{\lvert B-S \rvert}{B+S}$	$\left(\frac{\lvert B-S \rvert}{B+S}\right)^2$	$\frac{\lvert B-S \rvert}{B+S}$	$\left(\frac{\lvert B-S \rvert}{B+S}\right)^2$	50%	50%	p-value	% Rej.
1993	0.336	-0.113	(8.20)	(-2.93)	57.90%	1.00%	87.77%	3.26%
1994	0.321	-0.108	(8.12)	(-2.92)	56.55%	1.11%	84.63%	3.30%
1995	0.317	-0.098	(7.99)	(-2.62)	58.03%	1.08%	82.66%	4.03%
1996	0.339	-0.117	(8.73)	(-3.06)	59.28%	0.99%	84.95%	3.08%
1997	0.339	-0.117	(8.38)	(-2.98)	57.53%	1.03%	82.25%	4.16%
1998	0.362	-0.132	(9.59)	(-3.34)	61.34%	0.88%	82.57%	3.73%
1999	0.433	-0.183	(11.55)	(-4.88)	62.95%	0.80%	81.82%	5.18%
2000	0.419	-0.168	(9.74)	(-3.95)	58.88%	0.75%	81.03%	4.00%
2001	0.402	-0.143	(7.32)	(-2.62)	50.55%	0.48%	84.33%	3.52%
2002	0.255	-0.020	(3.57)	(-0.27)	42.07%	0.47%	80.50%	3.75%
2003	0.126	0.101	(1.70)	(1.36)	40.55%	0.46%	80.20%	3.19%
2004	-0.067	0.280	(-0.88)	(3.54)	38.32%	0.42%	75.32%	4.72%
2005	0.249	-0.015	(3.29)	(-0.20)	41.68%	0.41%	70.49%	6.64%
2006	0.264	-0.021	(3.81)	(-0.34)	43.41%	0.36%	59.43%	13.40%
2007	0.762	-0.447	(16.12)	(-9.57)	66.36%	0.31%	40.73%	25.49%
2008	0.800	-0.480	(18.60)	(-11.20)	70.98%	0.23%	39.63%	26.42%
2009	0.813	-0.492	(19.08)	(-11.49)	71.79%	0.23%	39.13%	31.68%
2010	0.814	-0.488	(18.94)	(-11.44)	72.77%	0.21%	41.33%	28.77%
2011	0.809	-0.480	(18.79)	(-11.21)	71.67%	0.22%	39.58%	29.71%
2012	0.804	-0.475	(18.83)	(-11.14)	72.72%	0.20%	42.12%	26.87%

Table A5: **OWR Model Regressions.** This table reports real and simulated regressions of the $CPIE_{OWR}$ on the squared and interaction terms of y_e, r_d, and r_o. In Panel A, we simulate 1,000 instances of the OWR model for each `PERMNO`-Year in our sample (1993–2012) and report mean standardized estimates for the median stock, along with 5%, 50%, and 95% values of the R^2 values. In Panel B, we report standardized estimates for the median stock using real data, along with the median R^2 values, and tests of the null that the model fits the data. The p-value is the average probability of observing an R^2 at least as small as what is observed in the real data. The % Rej. is the fraction of stocks for which we reject the null at the 5% level.

(a) Simulated Data

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

Table A5: **OWR Model Regressions.** Continued.

(b) Real Data

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

Figure A1: **DY Tree.** For a given trading day, private information arrives with probability α. When there is no private information, buys and sells are Poisson with intensity ϵ_B and ϵ_S. Private information is good news with probability δ. The expected number of buys (sells) increases by μ in case of good (bad) news. Non-information related order flow shocks arrive with probability θ. In the event of an order flow shock, buys and sells increase by δ_b and δ_s respectively.

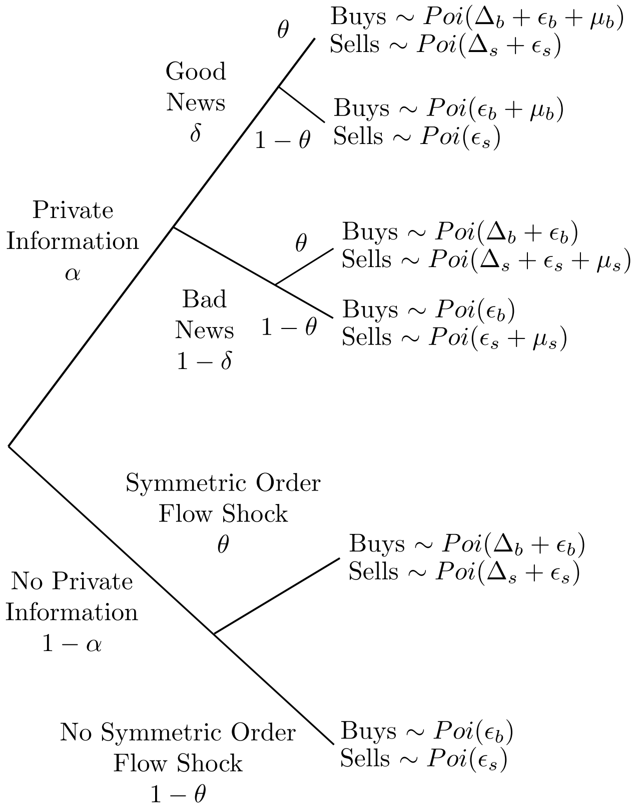
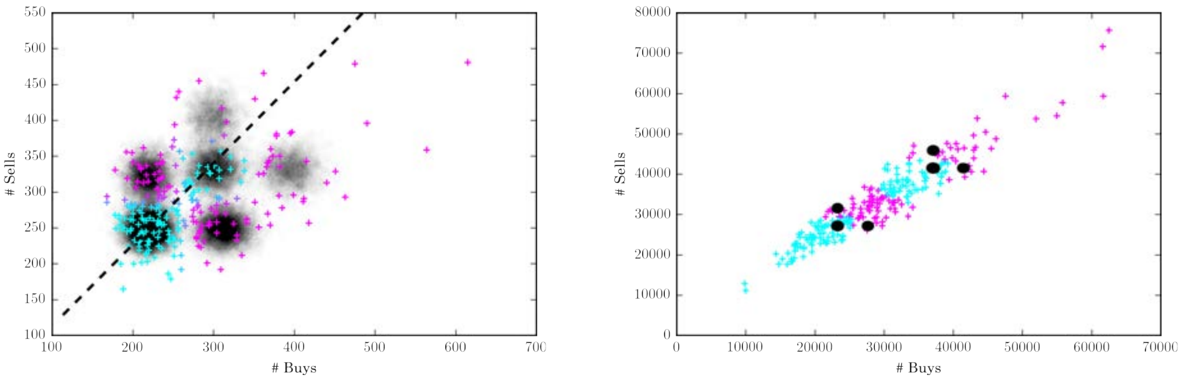

θ
Buys $\sim Poi(\Delta_b + \epsilon_b + \mu_b)$
Sells $\sim Poi(\Delta_s + \epsilon_s)$
Good News δ
Buys $\sim Poi(\epsilon_b + \mu_b)$
Sells $\sim Poi(\epsilon_s)$
$1 - \theta$
Private Information α
θ
Buys $\sim Poi(\Delta_b + \epsilon_b)$
Sells $\sim Poi(\Delta_s + \epsilon_s + \mu_s)$
Bad News $1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

Figure A2: **XOM DY.** This figure compares the real and simulated data for XOM in 1993 and in 2012 using the DY model. In Panels A and B, the real data are marked as +. The real data are shaded according to the $CPIE_{DY}$, with darker markers (+ magenta) representing high and lighter markers (+ cyan) low $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark "cloud" of points, and low probability states appear as a light "cloud" of points. The DY model extends the three states of the PIN model corresponding to no news, good news, and bad news with three additional states with higher order flows due to non-information symmetric order flow shocks.



(a) XOM 1993

(b) XOM 2012

Figure A3: **Breakdown of the DY Model.** This figure shows the distribution of the percent of days where the total likelihood, given the model parameters and observed order flow data is less than 10^{-10}—days, according to the model, with near-zero probability of occurring. The solid black line represents the median stock, and the dotted lines represent the 5, 25, 75, and 95 percentiles.

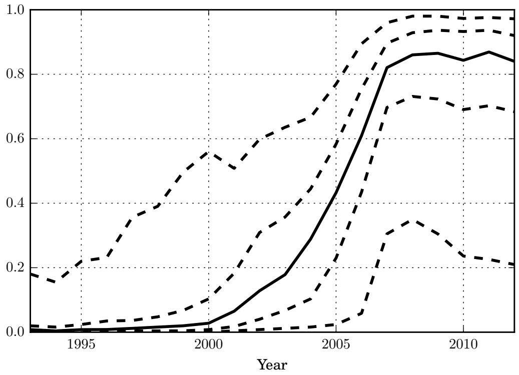
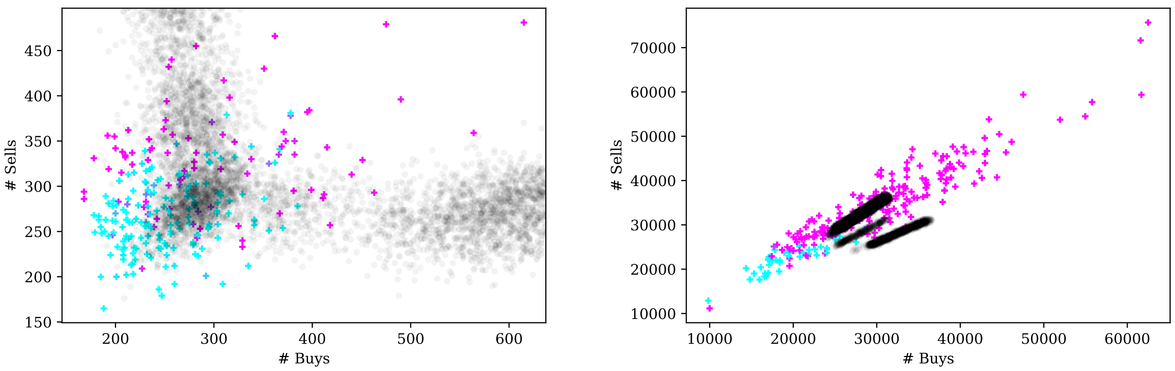


Figure A4: **XOM EEOW.** This figure compares the real and simulated data for XOM in 1993 and in 2012 using the EEOW model. In Panels A and B, the real data are marked as **+**. The real data are shaded according to the $CPIE_{DY}$, with darker markers (+ magenta) representing high and lighter markers (+ cyan) low $CPIE$s. The simulated data points are represented by transparent dots, such that high probability states appear as a dense, dark "cloud" of points, and low probability states appear as a light "cloud" of points.



(a) XOM 1993

(b) XOM 2012

References

Casella, George, and Roger Berger, 2002, *Statistical Inference* (Thomson Learning).

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

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

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

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.