# A comparison of some structural models of private information arrival*

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

December 17th, 2018

**Abstract**

We show that the PIN and the Duarte and Young (2009) (APIN) models do not match the variability of noise trade in the data and that this limitation has severe implications for how these models identify private information. We examine two alternatives to these models, the Generalized PIN model (GPIN) and the Odders-White and Ready (2008) model (OWR). Our tests indicate that measures of private information based on the OWR and GPIN models are promising alternatives to the APIN's $Adj.PIN$ and $PIN$.

*Keywords*: Liquidity; Information Asymmetry

*We thank Torben Andersen, Kerry Back, Pierre Collin-Dufresne, Kevin Crotty, Zhi Da, Bei Dong, Robert Engle, Gustavo Grullon, Terry Hendershott, Sahn-Wook Huh, Jyri Kinnunen, Pete Kyle, Yelena Larkin, Edward X. Li, K. Ramesh, Min Shen, Avi Wohl and seminar participants at the 2015 ITAM Conference, 2015 Annual SoFiE Conference, 2015 Annual MFS Conference, 2015 CICF, 2016 AFA Conference, 2018 CFRIC, FGV-EESP, FIU, PUC-RJ, Rice University, Southern Methodist University, Texas A&M University, University of Virginia (McIntire), and the University of Washington (Foster) for helpful comments. We thank Elaine Brewer, Frank Gonzalez, Judy Hua, and Edward Martinez for computational support. This paper was previously circulated under the title "Does the PIN model mis-identify private information and if so, what are the alternatives?". Data for the paper are available at: `https://edwinhu.github.io/pin`.

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

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

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

# A comparison of some structural models of private information arrival

December 17th, 2018

**Abstract**

We show that the PIN and the Duarte and Young (2009) (APIN) models do not match the variability of noise trade in the data and that this limitation has severe implications for how these models identify private information. We examine two alternatives to these models, the Generalized PIN model (GPIN) and the Odders-White and Ready (2008) model (OWR). Our tests indicate that measures of private information based on the OWR and GPIN models are promising alternatives to the APIN's $Adj.PIN$ and $PIN$.

*Keywords*: Liquidity; Information Asymmetry

One seminal contribution of market microstructure to the broader finance and accounting literature is the development of adverse selection measures based on structural models. Using different settings Glosten and Milgrom (1985) and Kyle (1985) model how liquidity providers account for adverse selection when responding to the trades of both informed and uninformed agents. Easley and O'Hara (1987) add the notion that the quantity of private information varies over time, with private information randomly arriving on some days, but not others. Using this insight, Easley, Kiefer, O'Hara, and Paperman (1996) estimate the PIN Model, a structural model that extracts the amount of adverse selection in a given stock from order flow data.[1] The PIN model, and to a lesser extent the APIN model, have attracted considerable attention in the accounting, corporate finance, and asset pricing literatures because they produce much needed proxies for information asymmetry.[2]

This paper comprehensively examines several alternative structural microstructure models including the PIN model, the APIN model, a new extension of the PIN model—the GPIN model, and the OWR model. We first examine the extent to which the models can match the observed moments of order flow and perform statistical tests on the nested models. Second, we examine whether the models yield better inferences than mechanical heuristics based on turnover that, by construction, misidentify private information arrival. These placebo 'tests', while somewhat *ad-hoc*, reveal how the statistical limitations of the models impact the economic viability of the models' inferences. To see this note that for any model to be considered successful at identifying private information arrival, it cannot yield the same inferences as a simple mechanical heuristic. Lastly, we examine the performance of the models that survive the first two tests by assessing their ability to identify the arrival of opportunistic insider trades and whether the models' signals about the arrival of private information are associated with smaller future price reversals.

To perform our second and third tests, we employ a variable called the Conditional

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

[2] A Google scholar search reveals that the PIN papers cited above alone have been cited more than 3,500 times as of this writing. Recent examples of papers that use $PIN$ and $Adj.PIN$ 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 (2018), and Bennett, Garvey, Milbourn, and Wang (2017).

Probability of an Information Event ($CPIE$). $CPIE$ is the conditional probability that a model assigns to the arrival of private information on a particular day, given the model parameters and data on that day. For instance, $CPIE_{PIN}$ is the probability of private-information arrival on a given day, conditional on the PIN model parameters and the observed daily order flow. In our second test, we compare a model's $CPIE$ with the $CPIE$ of a mechanical heuristic, or placebo, based solely on the level of turnover. In our third test, we examine variation in models' $CPIE$s around insider trades as well as the association of their $CPIE$s with future return reversals.

We first examine the PIN and APIN models. Specifically, we consider these models' ability to match the observed means, variances, and covariance between buys and sells as well as the means and variances of turnover. This analysis reveals that the PIN model cannot match the large amount of variability of trade that we see in the data. Indeed, the model-implied variances of buys and sells are around 550 times smaller than the variances of actual buys and sells. The APIN model improves the fit to the data over the PIN model by mixing between two PIN models, one with high noise trade intensity and one with low noise trade intensity. Indeed, 99% of the firm years in the cross section have likelihood ratio (LR) tests that reject the PIN in favor of the APIN at the 6% level or less. However, in spite of the improved fit, our results show that the APIN model-implied variances for buys and sells are less than half of what we see in the actual data.

To show how the statistical limitations of the PIN and APIN models affect their inferences, in our second test we estimate time-series regressions of each model's $CPIE$ on the $CPIE$ from a mechanical heuristic ($CPIE_{Mech}$). If the model mechanically identifies private-information arrival from turnover, we expect that its $CPIE_{Mech}$ will explain most of the variation in the model's $CPIE$. For instance, we compare variation in $CPIE_{PIN}$ to a purely mechanical heuristic that 'identifies' the arrival of private information from turnover. We call this the PIN-Mechanical Heuristic, and its $CPIE$ is $CPIE_{Mech,PIN}$. Formally, $CPIE_{Mech,PIN,j,t}$ is an indicator variable with value one when turnover on day $t$ for stock $j$ is above the annual mean of daily turnover for stock $j$ and zero otherwise. The PIN-Mechanical Heuristic amounts to the economically implausible statement that private information is sure to arrive on any day when turnover is above the mean and no private

information ever arrives on days when turnover is below the mean.[3]

We find that both the PIN and APIN models yield inferences about the arrival of private information that closely track mechanical heuristics based on turnover. In particular, we find that the PIN-Mechanical Heuristic alone explains around 59% of the variation in $CPIE_{PIN}$ for the median stock in our sample. Furthermore, this effect is widespread throughout the cross-section. The APIN model is not much better. For the median stock in the sample, the $R^2$ in a regression of $CPIE_{APIN}$ on $CPIE_{Mech,APIN}$ is 54%.[4] Moreover, we find that controlling for a long list of variables (including order flow imbalance as well as intra-day and overnight squared returns) that could reasonably proxy for private-information arrival does nothing to increase the $R^2$ in the regressions in any material way. This indicates that the while most of the variation in $CPIE_{PIN}$ and $CPIE_{APIN}$ is mechanically related to turnover, even the remaining variation has little to do with the arrival of private information.

Despite the PIN and APIN model's problems, all is not lost in the quest for intuitive measures of information asymmetry based on structural models. We also analyze two alternatives to the PIN and APIN models. First, we introduce a highly tractable generalization of the PIN model (the GPIN model) that, like the PIN and APIN models, relies only on order flow to identify private information arrival. As in the APIN model, the GPIN model allows expected daily turnover from noise trading to be random, while keeping the same information structure as the PIN model. However, in contrast to the APIN model, the GPIN model does not rely on mixing only two discrete PIN models. Instead, it allows for a continuum of PIN models. Second, we consider the OWR model, which uses returns along with order flow to identify private information.

[3]That is, when examining the PIN model our working hypothesis is that we cannot infer private information arrival from just looking at whether turnover is above or below the mean. This working hypothesis is based on two uncontroversial, but closely related, principles. First, although turnover may be related to the arrival of private information, it also varies for myriad reasons unrelated to private information. For instance, turnover can increase due to disagreement (e.g., Kandel and Pearson, 1995; Banerjee and Kremer, 2010). Turnover is subject to calendar effects because traders coordinate trade on certain days to reduce trading costs (Admati and Pfleiderer, 1988). Furthermore, turnover can vary due to portfolio rebalancing (Lo and Wang, 2000) and taxation reasons (Lakonishok and Smidt, 1986). Thus, any model that identifies the arrival of private information purely from turnover effectively classifies all variation in turnover as private information related. Second, even if one were to attempt to infer private-information arrival from turnover alone, reliable inferences cannot be gleaned from a simple heuristic based on whether turnover is low or high.

[4]We show in the appendix that the PIN and APIN model's mechanical conflation of private-information arrival with turnover is much more pronounced later in our sample. Indeed, the $R^2$s of the regression of $CPIE$ on $CPIE_{Mech}$ from both models are consistently upwards of 70% after 2006.

Our choice of alternatives to the PIN and APIN models include one that is based on order flow alone (the GPIN model) and another which uses order flow and returns (the OWR model). The distinction is important because, even though the PIN and APIN models are based on order flow alone, Kim and Stoll (2014) show evidence that order imbalance alone does not reveal private information.[5] Moreover, in contemporaneous work, Back, Crotty, and Li (2018) develop and empirically examine a model that uses both order flow and returns. Using their model, they make the theoretical point that order flow may not reveal private information if liquidity providers provide less liquidity for stocks with high degrees of information asymmetry and informed traders trade less in illiquid stocks. Thus, the degree to which a model that uses returns as well as order imbalance can detect private-information arrival better than a model based on order flow alone is an important empirical question.

Our analysis reveals that by allowing for a continuum of PIN models, the GPIN model can produce variation in buys, sells and turnover closer to that in the actual data. Indeed, the mean GPIN model-implied variance is around 68% of the empirical variance for the mean firm year in the sample. As the GPIN model's improved performance in matching the moments of the data suggests, the LR test rejects the PIN model at the 1% level in favor of the GPIN model for 99% of the firm years in the cross section.[6]

Unlike the PIN and APIN models, we find that the mechanical heuristics along with turnover and turnover squared explain only between 4% and 10% of the variation in $CPIE_{GPIN}$ and $CPIE_{OWR}$ for the median stock. This stands in contrast to the 64% $R^2$ for the PIN model and 56% for the APIN model. Furthermore, adding variables such as order flow imbalance, intra-day, and overnight squared returns dramatically increases the $R^2$ in the regressions to 35% for the GPIN model and 43% for the OWR model. This indicates that variation in $CPIE_{GPIN}$ and $CPIE_{OWR}$ is related to variables that are plausibly connected

[5] Both in Glosten and Milgrom (1985) and Kyle (1985) models, prices must adjust to reflect to the arrival of private information. However, Easley, Kiefer, O'Hara, and Paperman (1996) do not use this price-response mechanism when developing the PIN model, instead they rely only on the implications of Glosten and Milgrom (1985) to the relation between order imbalance and private information arrival.

[6] It is important to note that the OWR model differs from the PIN, APIN, and GPIN models in that it does not attempt to model the number of buys and sells. Instead, the OWR model focuses on the net imbalance between buys and sells (i.e., $y_e$). This implies that moments such as the variance of turnover and the covariance between buys and sells, which figured prominently in our analysis of the PIN, APIN, and GPIN models above, are not available under the OWR model.

to private information arrival. Thus, unlike the PIN and APIN models, the GPIN and OWR models do not identify private-information arrival mechanically from turnover.

As neither the GPIN and OWR models suffer from the same problems as the PIN and APIN models, one obvious question is which model performs better in identifying private information arrival. Therefore, we use the GPIN and OWR $CPIE$s ($CPIE_{GPIN}$ and $CPIE_{OWR}$) to compare the two models' ability to identify private-information arrival in the context of insider trades and return 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, 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., liquidity shocks) should be subsequently reversed, while information related trades should not. Therefore, under this working hypothesis a model that properly identifies private information should have a $CPIE$s that is associated with smaller future price reversals.[7]

Our results suggest that measures of private information from the OWR or GPIN models are promising alternatives to $PIN$. This being said, the OWR model performs somewhat better in our tests than the GPIN model. The superior performance of the OWR model is perhaps not surprising given that it uses returns along with order flow data rather than simply order flow data as with GPIN. What is perhaps surprising given the theoretical arguments of Back, Crotty, and Li (2018) is that a model based on order flow alone (GPIN) seems to identify private-information arrival at all.

Even though the OWR model performs better than the GPIN model, the GPIN model is a promising alternative to the PIN and APIN models in applications that require measures of adverse selection that are not based on returns. For instance, there are a large variety of corporate finance applications that involve cross-sectional analysis of announcement day returns for various corporate events. If a corporate finance researcher interested in the

[7]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 3.3.

impact of information asymmetry on announcement day returns for a particular event, i.e. a merger announcement, was to run a cross-sectional regression of announcement day returns on $CPIE_{OWR}$, the coefficients and $R^2$ in the regressions would be biased since the dependent variable was used to compute $CPIE_{OWR}$. However, if the researcher chose to use $CPIE_{GPIN}$, this would not be a problem.[8]

Naturally, there are other alternatives to the PIN and APIN models besides the GPIN and OWR models. For instance, Cipriani and Guarino (2014) extend the predecessor of the PIN model, the Easley, Kiefer, and O'Hara (1997) model, to allow informed traders to receive imprecise signals. We do not consider their model because it requires meaningful periods during day without trade. Indeed, Easley, Kiefer, O'Hara, and Paperman (1996) note that the discrete time likelihood function of Easley, Kiefer, and O'Hara (1997), which is similar to that of Cipriani and Guarino (2014), cannot be computed for data sets with many trades per day.[9] Easley, Engle, O'Hara, and Wu (2008) develop an alternative version of the PIN model with a time-varying measure of private information arrival. We show in the Internet Appendix A that, even though the Easley, Engle, O'Hara, and Wu (2008) model performs better than the APIN and PIN models, the Easley, Engle, O'Hara, and Wu (2008) model also mechanically identify private information from turnover in the later part of our sample period. Like our paper, Back, Crotty, and Li (2018) estimate a series of models of private information arrival, including the PIN and APIN models. However, their empirical approach is very different from ours. We focus on comparing the implied moments of these models with the actual data, comparing the models with mechanical heuristics, and using opportunistic insider trades as well return reversals to examine the relative performance of the models, while they do not. On the other hand, their paper provides an extensive analyses of their hybrid-PIN model, so we do not consider it in this paper.

[8] Another application in which a measure of adverse selection based on order flow alone is needed is in Easley, Kiefer, and O'Hara (1997). They regress daily prices for Ashland, Inc. on the lagged price and a measure analogous to $CPIE_{PIN}$. Easley, Kiefer, and O'Hara (1997) note that if the PIN model used information about returns then the coefficients in their regression would be biased, because the independent variable in their regression would be mechanically related to the dependent variable.

[9] Both Easley, Kiefer, and O'Hara (1997) and Cipriani and Guarino (2014) estimate their models for one, single stock: Ashland, Inc. Cipriani and Guarino (2014) uses Ashland, Inc. data in 1995 when Ashland, Inc. had only about 90 trades per day. In contrast, for the average stock in our sample, the average number of trades is around 3,800 per day and the average trading activity intensified after 2000. Exxon-Mobil, for instance, has an average of about 62,000 trades per day in 2012.

This paper contributes to the extensive and growing literature in finance and accounting that employs measures of private information. We do so by showing that the two most commonly used adverse selection proxies in the literature, the $PIN$ and the $Adj.PIN$ are unreliable. In addition, this paper contributes to an emerging literature that uses daily measures of private information. In a contemporaneous paper, Brennan, Huh, and Subrahmanyam (2018) examine high-frequency measures of good and bad news in event study settings. In contrast, we use $CPIE$ to shed light on how the various models identify private information. A related literature shows that the PIN model does not fit the order flow data well. For instance, Gan, Wei, and Johnstone (2014) show that the PIN and APIN models poorly describe the empirical distribution of order flow. While these results are suggestive of problems with these models, the fact that they do not match some of the moments of the order flow distribution does not imply that $PIN$ and $Adj.PIN$ fail to capture the variable of economic interest, namely private-information arrival. We contribute to this literature because we show that these models' statistical limitations impact *how* these models identify private-information arrival by using $CPIE_{PIN}$ and $CPIE_{APIN}$. Furthermore, we also evaluate two alternatives to the PIN and APIN models – the GPIN and OWR models.

The remainder of the paper is as follows. Section 1 outlines our data. Section 2 shows that the PIN and APIN models do not match the variability of noise trading in the data and, as a result, produce inferences that mimic mechanical turnover heuristics. Section 3 analyzes the GPIN and OWR models. Section 4 concludes.

# 1 Data

To estimate the PIN, APIN, GPIN, and OWR models, we collect trade and quote data for all NYSE stocks between 1993 and 2012 from the NYSE TAQ database. We require that the firms in our sample have only one type of common stock (i.e., a single `PERMNO` and share code 10 or 11), are listed on the NYSE (exchange code 1), and have at least 200 days worth of non-missing observations for the year. Our sample contains 1,060 stocks per year on average, of which about 36% (25%) are in the top (bottom) three Fama-French size deciles. For each stock in the sample, we classify each trade as either a buy or a sell, following the Lee and Ready (1991) algorithm. We estimate the PIN, APIN and GPIN models for each stock $j$

using a sample consisting of the number of buys and sells for each day ($B_{j,t}$ and $S_{j,t}$). In our regression analysis, we also use the daily absolute order flow imbalance ($|B_{j,t} - S_{j,t}|$), and turnover ($turn_{j,t} = B_{j,t} + S_{j,t}$).

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

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

We also examine a sample of opportunistic insider trades, as defined in Cohen, Malloy, and Pomorski (2012), from the Thomson Reuters' database of insider trades. In order to classify a trader as opportunistic or routine, we require three years of consecutive insider trades. We classify a trader as routine if she places a trade in the same calendar month for at least three years. All non-routine insiders' trades are classified as opportunistic. Our event sample includes 32,944 opportunistic insider trades.

Table 1 contains summary statistics for all the variables used to estimate the models. Panel A gives summary statistics for our entire sample and for opportunistic insider trading days. Panel B displays the distributions of some moments of buys, sells, and turnover for each stock-year in the entire sample.

# 2 Do the PIN and APIN Models Mechanically Identify Private Information?

This section shows that the PIN and APIN models do not match the variability of noise trading in the data and, as a result, produce inferences that mimic mechanical heuristics that identify private-information arrival based on the level of turnover.

## 2.1 The PIN Model

The Easley, Kiefer, O'Hara, and Paperman (1996) PIN model posits the existence of a liquidity provider who receives buy and sell orders from both noise traders and informed traders. Fig. 1 shows a tree diagram of the model. At the beginning of each day, if there is no private signal (which occurs with probability $1-\alpha$), buy and sell orders arrive at the normal mean rate of noise trade ($\epsilon_B$ for buys, $\epsilon_S$ for sells and $\epsilon_B+\epsilon_S$ for turnover). If the informed receive a signal (positive with probability $\delta$ and negative with probability $1-\delta$), they join the noise traders and trade at the rate $\mu$. In this case, mean turnover is $\epsilon_B+\epsilon_S+\mu$. It is important to note at this point that, under the PIN model, private-information arrival is necessarily the only cause for increases in expected daily turnover.

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

$L_{I^+}(D_{PIN,j,t})$ and $L_{I^-}(D_{PIN,j,t})$ and their computation, see the Internet Appendix.

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

$$CPIE_{PIN,j,t} = \frac{L_{I^-}(D_{PIN,j,t}) + L_{I^+}(D_{PIN,j,t})}{L_{I^-}(D_{PIN,j,t}) + L_{I^+}(D_{PIN,j,t}) + L_{NI}(D_{PIN,j,t})} \tag{1}$$

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

We estimate the PIN model numerically via maximum likelihood for every firm-year in our sample. Specifically, we maximize $\prod_{t=1}^{T} L(D_{PIN,j,t})$. Maximization of this likelihood function is prone to numerical issues because of two features of the data. First, days with thousands of buys and sells are common. As a result, attempting to directly compute the exponentials and factorials in the Poisson distributions in $L_{NI}(D_{PIN,j,t})$, $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$ often generates values that are too large to be represented by a typical computer. To address this problem we follow Duarte and Young (2009) and compute $L_{NI}(D_{PIN,j,t})$, $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$ by first computing their logarithms. For instance, consider the computation of $L_{NI}(D_{PIN,j,t})$. The direct computation of $\ell_{NI} = ln[L_{NI}(D_{PIN,j,t})]$ does not result in numerical overflow problems even for very large numbers of trades because $B_{j,t}$ and $S_{j,t}$ enter $\ell_{NI}$ multiplicatively instead of as exponents in $L_{NI}(D_{PIN,j,t})$. Moreover, the negative terms in $\ell_{NI}$ net out with the positive terms, resulting in values of $\ell_{NI}$ that can be readily exponentiated to compute $L_{NI}(D_{PIN,j,t})$. As in the computation of $L_{NI}(D_{PIN,j,t})$, we compute $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$ as the exponential of $\ell_{I^+} = ln[L_{I^+}(D_{PIN,j,t})]$ and $\ell_{I^-} = ln[L_{I^-}(D_{PIN,j,t})]$. Second, the PIN model likelihood functions often take values very close to zero, which makes the estimation susceptible to local optima. To get around this problem, we follow Duarte and Young (2009) by using ten different sets of starting points and choosing the parameter estimates associated with the largest final likelihood value.[10]

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

These same two features of the data also plague direct computation of $CPIE_{PIN}$ in Equation 1 with numerical overflow and underflow problems. To address this problem we first define $\ell_{\max} = \max\{\ell_{NI}, \ell_{I^+}, \ell_{I^-}\}$. We then compute $CPIE_{PIN}$ as:

$$CPIE_{PIN,j,t} = \frac{e^{(\ell_{I^+} - \ell_{\max})} + e^{(\ell_{I^-} - \ell_{\max})}}{e^{(\ell_{NI} - \ell_{\max})} + e^{(\ell_{I^+} - \ell_{\max})} + e^{(\ell_{I^-} - \ell_{\max})}} \tag{2}$$

The equation above handles days with thousands of buys and sells because it replaces direct computation of the likelihoods ($L_{NI}(D_{PIN,j,t})$, $L_{I^+}(D_{PIN,j,t})$, and $L_{I^-}(D_{PIN,j,t})$) in Equation 1 with their logs ($\ell_{NI}, \ell_{I^+}, \ell_{I^-}$). It also handles days when the denominator of Equation 1 is such a small positive number that typical computer systems cannot distinguish it from zero. The computation of $CPIE_{PIN}$ using Equation 2 avoids this problem because the denominator of Equation 2 has a lower bound of one.

It is important to note that Equation 2 addresses a *computational* problem, not a mathematical problem. Equation 2 is not an approximation or an arbitrary normalization of Equation 1. In fact, a simple algebraic manipulation shows that these expressions are equivalent. Thus, Equation 2 is a mathematically-sound way to rewrite Equation 1 in order to avoid computational problems that would lead to a large number of missing $CPIE_{PIN}$ observations. Indeed, direct computation of Equation 1 would result in the complete loss of all $CPIE_{PIN}$ observations for the median stock by 2004.

Panel A of Table 2 contains summary statistics for the parameter estimates of the PIN model as well as the cross-sectional sample means and standard deviations of $CPIE_{PIN}$. These statistics show that, as expected, the mean $CPIE_{PIN}$ behaves like the parameter $\alpha$.[11]

Panels A and B of Fig. 2 plot the simulated and real order flow for Exxon-Mobil in 1993 and 2012 respectively, with buys on the horizontal axis and sells on the vertical axis. Simulated data are marked as transparent dots and real data are marked with '×.' The simulated data are generated using Exxon-Mobil's estimated PIN model parameters for 1993 and 2012. These data are useful in illustrating the intuition for how the PIN model works.

In particular, the real data in Panels A and B of Fig. 2 show that noise trade displays

[11] In unreported results, we observe that the PIN model $\alpha$ increases over time, rising from about 30% in 1993 to 50% in 2012. The increase in our PIN model $\alpha$ parameters is somewhat larger than that in Brennan, Huh, and Subrahmanyam (2018). This small difference arises because we have a smaller number of stocks since we apply sample filters similar to those in OWR. Without these filters, the increase in our PIN model $\alpha$ parameters from 1993 to 2012 is comparable to that in Brennan, Huh, and Subrahmanyam (2018).

a large amount of variation. To see this note that the real data lie mostly around the positively sloped dotted line. Glosten and Milgrom (1985) imply that informed trading causes order *imbalance*. However, variation along the positively sloped line necessarily involves simultaneous changes to *both* the number of buys *and* sells, not the imbalance between them. Therefore, this variation must be related to realizations of a common noise-trade factor in buys and sells. These noise trade shocks are related to the factors that impact turnover but are unrelated to private information arrival.

The simulated data in Panels A and B of Fig. 2 display far less variability in noise trade than the real data. Instead of falling along the positively sloped dotted line, the simulated data fall into three categories corresponding to the nodes of the tree in Fig. 1. The data in these three categories create the distinct dark clusters in Panels A and B. In each panel, two of the clusters are made up of days characterized by relatively large absolute order flow imbalance, with a large number of sells (buys) and relatively few buys (sells). These are private-information days. The third group of days, which creates the southwest clusters in Panels A and B, has relatively low numbers of buys and sells because there is no private-information arrival.

The extremely tight clustering of the simulated data in the southwest region in Panels A and B of Fig. 2 renders the PIN model unable to match the high level of variation in turnover due to noise trade that we see in the actual data. The PIN model's assumption that buys and sells are conditionally Poisson implies that, according to the model, all the observations should fall within these three tight clusters. Consider, for example, the no-information node in Panel B of Fig. 2. According to the PIN model, on such days, buys and sells have an expected arrival rate of 29,123 and 33,146, respectively. The no-information cluster is thus centered on this point. Adding the arrival rates of buys and sells shows that the model implies that no-information days have an average turnover (i.e., the sum of buys and sells) of 62,269 with a standard deviation of about 250 ($\sqrt{62,269}$).[12] Thus, the Poisson assumption causes the model to infer that 95% of days without private information have turnover between 61,779 and 62,759. In the real data plotted in Panel B, on the other

[12]The reader may recall that a Poisson random variable has a standard deviation equal to the square root of the mean and that Poisson random variables are approximately normal for large arrival rates.

hand, 95% of the days have turnover between 34,960 and 103,778. Moreover, the real data lie mostly around the positively sloped dotted line and hence variation in the real data is mostly due to noise trade. Thus, the PIN model is unable to match the large amount of variation in turnover due to noise trade that we see in the actual data.

The inability of the model to match the high levels of turnover variation stemming from noise trade is also apparent when we consider the model-implied versus actual moments. Consider the data in 2012 (Panel B of Fig. 2). While the model can match the means of buys, sells and turnover it cannot match their variances. Indeed, the implied mean of turnover under the PIN model is 104% of the actual mean, while the implied variance of turnover under the model is only 0.8% of the actual variance.[13] Moreover, under the PIN model, private information shocks are the only reason for increases in expected buys and sells and thus expected turnover. As a result, the PIN model implied covariance of buys and sells is necessarily non-positive (-342 in Panel B).[14] The data, on the other hand, strongly indicates the presence of noise trade shocks. That is, shocks to both buys and sells that increase turnover without increasing order imbalance. As a result, the covariance of buys and sells in the data is positive (76,840,307 for Exxon-Mobil in 2012).

The PIN model's inability to match the variation in noise trading has severe implications for the way the model identifies private-information arrival for Exxon-Mobil in 1993 and 2012. To see this, consider Panels C and D of Fig. 2, which plot $CPIE_{PIN}$ as function of turnover. These plots show that the PIN model is essentially 'sure' that any day with turnover even slightly above a particular threshold (near the mean) is a private-information day ($CPIE_{PIN} = 1$). On the other hand, any day with turnover below this threshold is classified as a day with no private information ($CPIE_{PIN} = 0$). Recall that, under the model, the arrival of private information is the only reason for increases in expected turnover. As a result, the model infers that any day with 'extreme' high turnover (i.e., turnover larger than the mean) is a private-information day and all other days are not. Thus, $CPIE_{PIN}$ mimics a dummy variable that is equal to one when turnover is above some threshold (near

[13] See the Internet Appendix for the formulas of the PIN model implied moments.

[14] Since noise trade arrives at a constant rate while informed trade increases the arrival rate of either buys or sells but not both, the PIN model imposes a negative covariance between buys and sells. As Duarte and Young (2009) show, the covariance between buys and sells under the PIN model is given by $cov_{B,S} = (\alpha\mu)^2(\delta - 1)\delta$, which is necessarily non-positive.

the mean) and zero otherwise. Thus, the model's inability to match the variation in noise trading causes it to mechanically identify private information from turnover.[15]

The PIN model's inability to match the variability of the noise trading is a problem, not only for Exxon-Mobil, but for nearly all of the stocks in our sample. To see this, consider the PIN model-implied mean and variance of turnover in Panel B of Table 2 compared to the empirical turnover mean and variance in Panel B of Table 1. The mean model-implied mean of turnover is about 91% of the actual mean (3,371/3,695), which indicates that the PIN model is able to capture the first moment of turnover. However, the mean model-implied variance is only 0.2% (84,948/46,848,275) of the mean empirical variance. More than just XOM, the PIN model's problem in matching the variation in noise trading has severe and widespread implications for the way the model identifies private-information arrival. To show this, we introduce the PIN-Mechanical Heuristic.

The PIN-Mechanical Heuristic, or PIN-Mechanical dummy, treats any day with above (below) average turnover as a private-information (no private-information) day:

$$CPIE_{Mech,PIN,j,t} = \begin{cases} 0, & \text{if } turn_{j,t} < \overline{turn}_j \\ 1, & \text{if } turn_{j,t} \geq \overline{turn}_j, \end{cases} \tag{3}$$

where $\overline{turn}_j$ is the average daily turnover computed over the same sample period as we used to compute the PIN model parameters. We then run the regression $CPIE_{PIN,j,t} = \beta_{0,j} + \beta_{1,j} \times CPIE_{Mech,PIN,j,t} + \varepsilon_{j,t}$ for each stock $j$ in the sample. For each stock $j$ and day $t$, we calculate $CPIE_{PIN,j,t}$ and $CPIE_{Mech,PIN,j,t}$ using data and estimates of the PIN model parameters for the entire calendar year containing day $t$.[16]

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

[15] Note that this mechanical identification of private information does not necessarily relate to the possibility that informed traders may sometimes choose to trade on days with high liquidity or turnover (see Collin-Dufresne and Fos, 2016). Naturally, it is possible that informed traders do trade on some days with high turnover. However, our point is that the PIN model mechanically identifies *all* days with above average turnover as definitely private-information days and all days with below average turnover as definitely not private-information days.

[16] Naturally, market makers and traders do not have all of this information on day $t$. Therefore $CPIE_{PIN,j,t}$ and $CPIE_{Mech,PIN,j,t}$ cannot be used to set prices or conduct trading strategies. However, they are useful to gauge the similarity between the PIN model and a mechanical heuristic of private-information arrival. Such an assessment is important to researchers who do observe order flow, PIN model parameters, and turnover over their entire sample period and thus can construct both measures for use in their work.

in the regression is the expected value of $CPIE_{PIN}$ when turnover is below the mean. Similarly, the sum of the coefficients $(\beta_{0,j}+\beta_{1,j})$ is the expected value of $CPIE_{PIN}$ when turnover is above the mean. The coefficient estimates in Specification 1 of Table 3 reveal that for days with turnover below the mean ($CPIE_{Mech,PIN}=0$), the median stock's $CPIE_{PIN}$ is close to zero, around 0.06. In contrast, for days with turnover above the mean ($CPIE_{Mech,PIN}=1$), $CPIE_{PIN}$ for the median stock is 0.79 (0.73 +0.06). Furthermore, the median $R^2$ is 59%.

A natural question is whether, despite the high $R^2$s in Specification 1, $CPIE_{Mech,PIN}$ oversimplifies the relation between $CPIE_{PIN}$ and turnover. To address the possibility of a more complicated, non-linear relation between $CPIE_{PIN}$ and *turn*, we regress $CPIE_{PIN}$ on *turn*, $turn^2$, and $CPIE_{Mech,PIN}$. Specification 2 of Table 3 displays the results of these regressions. The small difference of 5% in the median $R^2$s between Specifications 1 and 2 indicates that *turn* and $turn^2$ add little to the explanatory power of $CPIE_{Mech,PIN}$, a simple dummy variable based only on turnover.[17]

One potential explanation for the results in Specification 1 of Table 3 is that while turnover and the mechanical heuristic explain nearly 60% of the variation in $CPIE_{PIN}$, it is possible that it is the *unexplained* variation in $CPIE_{PIN}$ that captures the arrival of private information. Specification 3 addresses this possibility by including a series of control variables that are plausibly related to the arrival of private information. To come up with a list of such variables, we look to the OWR and PIN models for guidance. Specifically, the PIN model suggests that the daily absolute order flow imbalance ($|B-S|$) is related to private-information arrival.[18] Moreover, the OWR model suggests that the squared intraday and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$ and $r_o \times y_e$) vary with private-information arrival.[19] Thus, if the variation in $CPIE_{PIN}$ that is unexplained by turnover successfully captures the arrival

[17]Note that the interpretation of the coefficients ($\beta_0$ and $\beta_1$) from Specification 1 does not carry over to Specification 2 because $CPIE_{Mech,PIN}$ is, by construction, mechanically related to *turn* and $turn^2$. That is, $\beta_0$ is no longer the expected value of $CPIE_{PIN}$ when turnover is less than its mean and the sum of the coefficients $\beta_0+\beta_1$ is no longer the expected value of $CPIE_{PIN}$ when turnover is greater than its mean. As such, we focus on the difference in the $R^2$s across Specifications 1 and 2, which tells us the contribution of *turn* and $turn^2$ relative to $CPIE_{Mech,PIN}$ in explaining variation in $CPIE_{PIN}$.

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

[19]See Section 3.2 below.

of private information then we would expect that including these variables in the regression should substantially increase the $R^2$s from those in Specifications 1 and 2. The results in Specification 3 indicate that this is not the case. In fact, these controls increase the average $R^2$ for the median stock by only 2% over the 64% average $R^2$ in Specification 2. Moreover, the average $R^2$ for the stocks at the fifth and $95^{th}$ percentiles are similarly unaffected by the inclusion of the control variables.[20] In summary, our results strongly support the conclusion that the PIN model mechanically identifies the arrival of private information from turnover.

## 2.2 The APIN model

Duarte and Young (2009) extend the PIN model to address some of its shortcomings in matching the order flow data. The APIN model does so by allowing the intensity of noise trade arrivals to vary due to random disturbances (called symmetric order flow shocks) that simultaneously increase both the expected number of buyer- *and* seller-initiated noise trades. These shocks arrive at the beginning of the day with probability $\theta$. On days without (with) a symmetric order flow shock, buy and sell orders from uninformed traders arrive according to Poisson distributions with intensities $\epsilon_B$ ($\epsilon_B + \Delta_B$) and $\epsilon_S$ ($\epsilon_S + \Delta_S$). As with the PIN model, the APIN model posits that at the beginning of each day, informed investors receive a private signal with probability $\alpha$. If the private signal is positive, buy orders from the informed traders arrive according to a Poisson distribution with intensity $\mu_B$. If the private signal is negative, informed sell orders arrive according to a Poisson distribution with intensity $\mu_S$. If the informed traders receive no private signal, they do not trade.

Fig. 3 shows that the APIN model is best thought of as a mixture of two independent PIN models with different intensities of noise trading arrival and mixture weights $\theta$ and $1-\theta$. That is, on days with no symmetric order flow shock, the APIN model is similar to the PIN model with a noise trading intensity of $\epsilon_B + \epsilon_S$. These days correspond to the branches in the bottom of the tree in Fig. 3. On the other hand, on days with a symmetric order

[20]The $R^2$s in Table 3 also allow us to examine how pervasive the mechanical conflation of private-information arrival with turnover is in the cross section. Stocks with the lowest (highest) $R^2$s are those for which variation in $CPIE_{Mech,PIN}$ explains the least (most) variation in $CPIE_{PIN}$. To show graphically how this conflation varies in the cross section, we select two stocks whose $R^2$s of the regressions $CPIE_{PIN,j,t} = \beta_{0,j} + \beta_{1,j} \times CPIE_{Mech,PIN,j,t} + \varepsilon_{j,t}$ using data in 1993 (first year of our sample) and 2012 (last year of our sample) are at the $5^{th}$. The results are in the Internet Appendix.

flow shock, the APIN model is similar to a second PIN model with a higher noise trading intensity: $\epsilon_B + \epsilon_S + \Delta_B + \Delta_S$. These days correspond to the branches in the top of the tree in Fig. 3.

As with the PIN model, we define $CPIE_{APIN}$ as the probability of an information event conditional on both the model parameters and the data observed that day. An application of Bayes' rule results in a formula that expresses $CPIE_{APIN}$ as function of the likelihood of each branch in the tree in Fig. 3. The same numerical problems that plague the estimation of the PIN model also plague the estimation of the APIN model. As such, we adopt the same procedures that we use to estimate the PIN model and to calculate $CPIE_{PIN}$ to estimate the APIN model and to calculate $CPIE_{APIN}$. See the Internet Appendix for details about the likelihood function, and the $CPIE_{APIN}$ calculation.

Panel A of Table 4 contains summary statistics for the APIN model parameter estimates as well as the cross-sectional sample means and standard deviations of $CPIE_{APIN}$. These statistics show that, as expected, the mean $CPIE_{APIN}$ behaves like the parameter $\alpha$.

The graphs in Panels A and B of Fig. 4 are useful in illustrating the intuition for how the APIN model works. Panels A and B of Fig. 4 present the same real buy and sell data for XOM in 1993 and 2012 as those in Fig. 2. In contrast to Fig. 2, Fig. 4 presents simulated data from the APIN model rather than from the PIN model.

The simulated data from the APIN model falls into six discrete categories corresponding to the nodes of the tree in Fig. 3. The data in these categories create two groups of three distinct dark clusters in Panels A and B. The first group, the three black clusters to the southwest correspond to the bottom three nodes of the tree in Fig. 3. The second group, the three black clusters to the northeast, represent days with increased noise trade and correspond to the three nodes at the top of the tree in Fig. 3. Both the southwest and northeast groups have one black cluster that sits on the positively sloped dotted line. These are the no-information days. The cluster with negative (positive) private-information days sits north (east) of the non-information cluster. Thus, the APIN model mixes between the 'northeast' PIN model which has high levels of noise trade and the 'southwest' PIN model which has low levels of noise trade.

The simulated data in Panels A and B of Fig. 4 display far less variability in noise

trade than the real data. Like the PIN model, the APIN model is able to match the mean of turnover—the implied mean is 102% of the actual mean in 2012.[21] However, the APIN model-implied turnover variance for XOM is only 60% of the actual variance. Thus, while mixing between two PIN models improves the APIN model's ability to fit the data relative to the PIN model, the APIN model still dramatically underestimates the variation in noise trade. In Panels A and B of Fig. 4 this failure to match the noise trade variance is manifest in the model's inability to generate buy and sell data that vary continuously along the positive sloped dotted line. As a result, the model *perceives* any day with turnover slightly above (below) the mean of each group of three distinct dark clusters as extremely unlikely.

Panels C and D plot $CPIE_{APIN}$ as function of turnover. These panels show that the model's identification of private information is based solely on turnover. The lower (higher) turnover level indicated with a vertical line represents the expected turnover conditional on the absence (presence) of a symmetric order flow shock. The position of these lines along with the variation in $CPIE_{APIN}$ between zero and one across these lines indicates that the APIN model is performing the mechanical identification of private information from one of the two PIN models. To see this first consider the 'southwest' PIN model in Panels A and B. The APIN model considers any day in this part of the graph that doesn't overlap with these three dots as an extreme outlier. Thus, the 'southwest' PIN model is 100% certain that all days to the immediate northeast/(southwest) of its dashed line are information (non-information) days. Similarly, the 'northeast' PIN model is 100% certain that any day immediately northeast (southwest) of its dashed line is an information (non-information) day. This creates the distinctive light/dark/light/dark (cyan/magenta/cyan/magenta when printed in color) pattern in the shading of the data in Panels A and B of Fig. 4.

As we saw with the PIN model, the APIN model's inability to match the variability of the noise trade visible in the actual data is a problem, not only for Exxon-Mobil, but for nearly all of the stocks in our sample. To see this consider the APIN model-implied mean and variance of turnover in Panel B of Table 4 compared with the empirical turnover mean and variance in Panel B of Table 1. The mean model-implied mean of turnover is about 97% of the actual mean (3,575/3,695). However, the mean model-implied turnover

[21] Formulae for the implied moments of the APIN model are provided in the Internet Appendix.

variance from the APIN model is only around 42% (19,491,309/46,848,275) of the empirical variance. Thus, even though the APIN model improves on the PIN model's ability to match the empirical turnover variance, it still vastly underestimates the variability of turnover and thus the variability of noise trade. Panel B of Table 4 also displays the results of likelihood ratio (LR) tests between the APIN and the PIN model.[22] As the APIN model's improved performance in matching the moments of the data suggests, on average the LR test rejects the PIN model in favor of the APIN model at a $p$-value of 0.01. Indeed, 99% of the firm years in the cross section have LR tests that reject the PIN in favor of the APIN at the 6% level.

The APIN model's inability to capture the variation in noise trading has severe and widespread implications for the way the APIN model identifies private information arrival. To show this, let the indicator $SOS_{j,t}$ take the value of one if a symmetric order flow shock occurs for stock $j$ on day $t$ and zero otherwise. Let the APIN-Mechanical Heuristic be defined as:

$$CPIE_{Mech,APIN,j,t} = \begin{cases} 0, & \text{if } turn_{j,t} < E[turn|SOS_{j,t}=0] \\ 1, & \text{if } E[turn|SOS_{j,t}=0] \leq turn_{j,t} < \frac{E[turn|SOS_{j,t}=0]+E[turn|SOS_{j,t}=1]}{2} \\ 0, & \text{if } \frac{E[turn|SOS_{j,t}=0]+E[turn|SOS_{j,t}=1]}{2} \leq turn_{j,t} < E[turn|SOS_{j,t}=1] \\ 1, & \text{if } turn_{j,t} \geq E[turn|SOS_{j,t}=1]. \end{cases} \tag{4}$$

Analogous to our analysis of the PIN model, we compare time series variation in $CPIE_{APIN}$ with variation in $CPIE_{Mech,APIN}$ by running the following regression for each stock $j$ in our sample: $CPIE_{APIN,j,t} = \beta_{0,j} + \beta_{1,j} \times CPIE_{Mech,APIN,j,t} + \varepsilon_{j,t}$.[23]

The results in Table 5 show that, similar to our PIN model findings, $CPIE_{APIN}$ is very closely approximated by the APIN Mechanical Heuristic, not only for Exxon-Mobil, but throughout the cross section. For the median stock, the APIN Mechanical dummy explains nearly 55% of the variation in $CPIE_{APIN}$. Furthermore, the coefficient estimates

[22] The APIN model nests the PIN model. To see this, consider the APIN model and let $\theta = 0$ and $\mu_S = \mu_B$.

[23] The expected turnover conditional on no symmetric order flow shock is $E[turn|SOS_{j,t}=0] = \epsilon_B + \epsilon_S + \alpha(1-\delta)\mu_S + \alpha\delta\mu_B$ while the expected turnover conditional on symmetric order flow shock is $E[turn|SOS_{j,t}=1] = \epsilon_B + \epsilon_S + \Delta_B + \Delta_S + \alpha(1-\delta)\mu_S + \alpha\delta\mu_B$. Note that the mechanical heuristic above depends on the parameters of the APIN model. To address the possibility that our regression results are driven by this dependency, we also use an alternative mechanical heuristic based only on the turnover data, where the break points are determined using a k-means algorithm. The results are similar to those reported below and are in the Internet Appendix.

highlight the economically incongruous relation between turnover and the probability of private-information arrival implied by the APIN model. To see this note that the intercept for the median stock ($\beta_{0,j}$) is 0.135, while $\beta_{1,j}$ is 0.691. Thus for a typical stock, $CPIE_{APIN}$ jumps dramatically back and forth between 0.135 to 0.826 (0.135+0.691) based only on the level of daily turnover. It is difficult to see how this peripatetic relation between the probability of private-information arrival and turnover is economically sensible.

Specification 2 of Table 5 displays the results of regressions of $CPIE_{APIN}$ on $turn$, $turn^2$, and $CPIE_{Mech,APIN}$. The small difference of 1% in the median $R^2$s of Specifications 1 and 2 indicates that $turn$ and $turn^2$ add little to the explanatory power of $CPIE_{Mech,APIN}$. Specification 3 shows the results from regressions including the same series of control variables that we used to analyze the PIN model. These controls increase the $R^2$ for the median stock by only 5% over the 55% $R^2$ in Specification 1.[24] Therefore, the portion of the variation in $CPIE_{APIN}$ that is unexplained by turnover does not capture the arrival of private information either. In summary, these results strongly support the conclusion that, like the PIN model, the APIN model mechanically identifies the arrival of private information from turnover.

# 3 Two Alternatives to the PIN and APIN Models

Section 3.1 and Section 3.2 show that the GPIN and OWR models do not mechanically identify private-information arrival from turnover. Section 3.3 compares the GPIN and the OWR models.

## 3.1 The GPIN model

In this section, we present a generalization of the PIN model that addresses the limitations of both the PIN and APIN models described in Section 2. As in the APIN model, the GPIN model allows expected daily turnover due to noise trading to be random, while keeping the same information structure as the PIN model. However, in contrast to the APIN model, the GPIN model does not rely on mixing two discrete PIN models. Instead, it allows for

[24] As we do with the PIN model, we show graphically how the conflation of $CPIE_{APIN}$ and $CPIE_{Mech,APIN}$ varies in the cross section. The results are in the Internet Appendix.

a continuum of PIN models. That is, the GPIN model allows noise trade intensity to vary continuously rather than switch between high and low noise trade intensity regimes.

Fig. 5 presents the tree structure for the GPIN model. Under the GPIN model noise trade on any day $t$ is a Poisson random variable with intensity $\lambda_t$. Of these trades, $(1-\theta)\lambda_t$ are expected to be seller-initiated and $\theta\lambda_t$ buyer-initiated, where $\theta$ is a constant between zero and one. Identical to the PIN and APIN models, the informed traders receive a signal with probability $\alpha$. On days where the informed receive a signal (positive with probability $\delta$ and negative with probability $1-\delta$), they join the noise traders and initiate a number of trades given by a Poisson distribution with intensity $\eta\lambda_t$, where $\eta$ is a constant.

The parameter $\lambda_t$ is drawn from a *Gamma* distribution with shape parameter $r$ and scale parameter $p/(1-p)$. The fact that $\lambda_t$ is drawn from a *Gamma* distribution makes the model particularly tractable because this implies that the number of buys, sells and turnover are distributed as a mixtures of *Negative Binomial* distributions.[25] This dramatically simplifies the numerical estimation of the model. $CPIE_{GPIN}$ is calculated in the same way as in the PIN model. See the Internet Appendix for a detailed discussion of the model, the likelihood function, and the $CPIE_{GPIN}$ calculation. Panel A of Table 6 contains summary statistics for the parameter estimates of the GPIN model. Panel A also contains summary statistics of the cross-sectional sample means and standard deviations of $CPIE_{GPIN}$.[26]

Panels A and B of Fig. 6 present a stylized example to illustrate the central intuition for how the GPIN model works. Analogous to the plot in Figs. 2 and 4 for the PIN and APIN models, Panels A and B of Fig. 6 plot simulated and real order flow data for Exxon-Mobil during 1993 and 2012. The simulated data comprise three types of days, and thus three distinct clusters. In contrast to the PIN and APIN models, these clusters are not tightly clustered rounded regions. Instead, under the GPIN model the clusters form three positively sloped lines. The center line has a low proportion of imbalanced trades and thus represents days with no private information. The top and bottom lines represent private information

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

[26] 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 $CPIE_{GPIN}$ in Section 3.3. The summary statistics for the parameter estimates used in our event studies are qualitatively similar to those in Table 6.

days. That is, the top and bottom lines reflect a high proportion of imbalanced trades, with either a large number of sells and relatively few buys (the top line) or a large number of buys and relatively few sells (the bottom line).

In contrast to Figs. 2 and 4, the simulated data clusters in Panels A and B of Fig. 6 overlap substantially with the actual data. Panels C and D plot $CPIE_{GPIN}$ as function of turnover. As opposed to the analogous plot of the PIN and APIN models in Figs. 2 and 4, Panels C and D give no indication that the GPIN model mechanically identifies private-information arrival from turnover.

Consistent with the graphs in Fig. 6, the GPIN model improves on the PIN and APIN models' ability to match the empirical moments of the data throughout the cross section. Panel B of Table 6 displays the moments implied by the GPIN model. Note that the GPIN model-implied turnover mean is about 99.9% of the actual mean (3,690/3,695) and the mean GPIN model-implied variance is around 68% (31,792,976) of the empirical variance for the mean firm year in the sample.[27] Panel C of Table 6 displays the results of likelihood ratio (LR) tests between the GPIN and the PIN model.[28] As the GPIN model's improved performance in matching the moments of the data suggests, the LR test rejects the PIN model at the 1% level in favor of the GPIN model for 99% of the firm years in the cross section. This stands in contrast to the 99% of the firms in the cross section that reject the PIN model in favor of the APIN model at the 6% level.

Fig. 6 shows that, at least for XOM, the GPIN model does not mechanically conflate turnover with private information arrival. To show that XOM is not an isolated case, Table 7 presents results from time-series regressions of $CPIE_{GPIN}$ on the mechanical dummies. Specification 1 in Table 7 shows the coefficient estimates and $R^2$s of regressions of $CPIE_{GPIN}$ on $CPIE_{Mech,PIN}$ and $CPIE_{Mech,APIN}$. In contrast to Tables 3 and 5, the coefficient estimates are small and the $R^2$ is negligible. The results in Specifications 2 are similar despite the inclusion of $turn$, $turn^2$. This indicates that, unlike the $CPIE_{PIN}$ and $CPIE_{APIN}$, simple mechanical heuristics do not explain variation in $CPIE_{GPIN}$. Significantly, includ-

[27] See the Internet Appendix for the formulas of the GPIN model implied moments.

[28] The GPIN model nests the PIN model. To see this, consider the limiting case of the GPIN model in which $p \to 0$ and $r \to (\epsilon_B + \epsilon_S)/p$. Moreover, reparameterize the GPIN model as $\theta = \epsilon_B/(\epsilon_B + \epsilon_S)$ and $\eta = \mu/(\epsilon_B + \epsilon_S)$.

ing our control variables dramatically increases the $R^2$ from 1% in Specification 1 to 35% in Specification 3. Therefore, a substantial fraction of the variation in $CPIE_{GPIN}$ is both orthogonal to turnover and associated with variables that are plausibly related to private information arrival. In summary, our results suggest that the GPIN model, unlike the PIN and APIN models, does not suffer from the debilitating problem of mechanically associating turnover shocks with private information arrival.

## 3.2 The OWR model

Odders-White and Ready (2008) extend Kyle (1985) to allow for days with and without private-information arrival. Fig. 7 shows a time line for the events in the model. Under the OWR model, private information arrives before the opening of the trading day with probability $\alpha$. On days when private information arrives, the information is assumed to be publicly revealed after the close of trade. There are three key quantities of interest in the OWR model: daily net order flow ($y_e$), the intra day return ($r_d$), and the overnight return ($r_o$). In the model, the covariance matrix of these variables differs between days with and without private-information arrival. Econometricians can therefore use these variables to infer whether private information has arrived or not.[29]

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

Formally, let $\Theta_{OWR,j} = (\alpha_j, \sigma_{z,j}, \sigma_{u,j}, \sigma_{i,j}, \sigma_{p,d,j}, \sigma_{p,o,j})$ be the vector of OWR parameters

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

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

Let $I_{j,t}$ be an indicator function with value one when private information arrives on day $t$ for stock $j$. As is the case for the other examined models, $CPIE_{OWR,j,t} = P\left[I_{j,t} = 1 | D_{OWR,j,t}\right]$. Bayes' theorem implies that $CPIE_{OWR,j,t}$ is given by:

$$CPIE_{OWR,j,t} = \frac{L_I(D_{OWR,j,t})}{L_I(D_{OWR,j,t}) + L_{NI}(D_{OWR,j,t})} \tag{5}$$

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

As with the PIN and APIN models, we estimate the OWR model numerically via maximum likelihood. Specifically, we maximize $\prod_{t=1}^{T} L(D_{OWR,j,t})$, where $L(D_{OWR,j,t})$ is the sum of $L_{NI}(D_{OWR,j,t})$ and $L_I(D_{OWR,j,t})$. In contrast to the PIN and APIN models, we do not encounter any numerical issues in directly computing either $L(D_{OWR,j,t})$ or $CPIE_{OWR}$ with Equation 5. Table 8 contains summary statistics for the OWR parameter estimates and $CPIE_{OWR}$.[30]

[30] As expected, we see from Table 8 that the mean $CPIE_{OWR}$ behaves like $\alpha$ in the OWR model. Note that the estimated OWR $\alpha$ parameters are in general higher than those in OWR. This is due to the fact that our definition of $y_e$ is different from that in OWR (see discussion in Section 1 above). In fact, we get $\alpha$ estimates close to those reported in OWR if we define $y_e$ in the same way that they do. We also estimate the parameter vector $\Theta_{OWR,j}$ in the period $t \in [-312, -60]$ before opportunistic insider trades. These parameter estimates are used to compute the $CPIE$s used in our opportunistic insider trading event

Fig. 8 shows that, at least for XOM, the OWR model does not mechanically conflate turnover with private information arrival. It is important to note that the OWR model differs from the PIN, APIN, and GPIN models in that it does not attempt to model the number of buys and sells. Instead, the OWR model focuses on the net imbalance between buys and sells (i.e., $y_e$). Liquidity trade in the OWR model simply adds noise to the order imbalance and prevents the market maker from inverting the order flow to reveal the informed investors' private signal. This implies that moments such as the variance of turnover and the covariance between buys and sells, which figured prominently in our analysis of the PIN, APIN, and GPIN models above, are not available under the OWR model. Thus, the OWR model does not allow us to construct analogs to Panel B in Tables 2, 4, and 6 as well as to simulate the number of buys and sells data as we do in Figs. 2, 4 and 6.

While we cannot perform the analyses in Panel B of Tables 2, 4, and 6 for the OWR model, we can still use $CPIE_{OWR}$, $CPIE_{Mech,PIN}$, $CPIE_{Mech,APIN}$, $turn$, and $turn^2$ to determine whether the model mechanically conflates turnover with private information arrival. Table 9 presents results from time-series regressions of $CPIE_{OWR}$ on the mechanical heuristics. In contrast to Tables 3, and 5, the results in Table 9 show that $CPIE_{OWR}$ is poorly approximated by $CPIE_{Mech,PIN}$ and $CPIE_{Mech,APIN}$. Indeed, the median $R^2$ in Specification 1 is low, around 1.2%. Moreover adding $turn$ and $turn^2$ to the regression increases the $R^2$ for the median stock to only about 10%, considerably smaller than the 64% and 56% in Tables 3 and 5 for the PIN and APIN models. Hence, in contrast to the PIN and APIN models, turnover plays little role in identifying private-information arrival under the OWR model. Furthermore, including variables such as $|B-S|$, $|B-S|^2$, $r_d^2$, $r_o^2$, $y_e^2$,$r_d \times r_o$, $r_d \times y_e$, and $r_o \times y_e$ in the regression dramatically increases the $R^2$ from around 1% in Specification 1 to nearly 45% in Specification 3. This indicates that, unlike $CPIE_{PIN}$ and $CPIE_{APIN}$, a substantial fraction of the variation in $CPIE_{OWR}$ is both orthogonal to turnover and associated with variables that are plausibly related to private information arrival. Thus, the OWR model, unlike the PIN and APIN models, does not mechanically associate private-information arrival with turnover shocks.

study. The summary statistics of the parameter estimates for the event studies are similar to those in Table 8.

## 3.3 Comparing the GPIN and OWR models

To gain further insight into the GPIN and OWR models' performance, we consider two additional working hypotheses in the contexts of opportunistic insider trades and of return reversals.[31] Consider first the relation between $CPIE_{OWR}$, $CPIE_{GPIN}$ and opportunistic insider trades. Under the working hypothesis that opportunistic insiders trade up to the point that prices reveal their information, the $CPIE$s of both models should be higher before and coincident with an opportunistic trade, then decline immediately following the trade. Accordingly, we examine $CPIE_{OWR}$ and $CPIE_{GPIN}$ around opportunistic insider trades. Specifically, we estimate the parameter vectors $\Theta_{GPIN,j}$ and $\Theta_{OWR,j}$ in the period $t \in [-312, -60]$ before each opportunistic insider trade. We then use these parameter estimates to compute each model's $CPIE$s during the period ($t \in [-20, 20]$).

Panel A (B) of Fig. 9 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 $t \in [-40, 21]$. 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. In sum, these results suggest that both the OWR and GPIN models capture the arrival of private information around opportunistic insider trades. However, only the OWR model results are completely consistent with the idea that opportunistic insiders trade up to the point that prices fully reveal private information.

[31] Both of these working hypotheses are not as strongly established in the literature as the hypothesis that turnover varies for reasons unrelated to private information. Thus, these tests are only suggestive of the models' relative performance in identifying private arrival.

Next we examine the relation between $CPIE_{OWR}$, $CPIE_{GPIN}$ and future return reversals. A large number of papers have demonstrated that short horizon stock returns, on average, exhibit negative unconditional serial correlation (Jegadeesh and Titman (1995)), often called price reversals. On the other hand, the microstructure literature has long held that the arrival of private information causes permanent price changes.[32] Our working hypothesis here then is that the arrival of private information should be associated with smaller future return reversals. That is, the arrival of private information should be associated with less negative serial correlation in returns. Therefore, we estimate the following regression using $CPIE_{OWR}$ as well as $CPIE_{GPIN}$, including both firm and year fixed effects:

$$r_{j,t+1} = \alpha_{j,Year} + \beta_0 + \beta_1 \times r_{j,t} + \beta_2 \times CPIE_{j,t} + \beta_3 \times (CPIE_{j,t} \times r_{j,t}) + \epsilon_{j,t} \qquad (6)$$

Before continuing, however, there are two issues worth clarifying. First, note that the independent variable in this regression is the open-to-open, risk-adjusted return ($r_{j,t+1} = r_{d,j,t+1} + r_{o,j,t+1}$) on day $t+1$. Thus, there is no overlap between the intra-day and overnight returns that are used to compute $CPIE_{OWR,j,t}$ on day $t$ and the return on day $t+1$. This is important because if we were to regress $r_{j,t+1}$ on $CPIE_{OWR,j,t+1}$, the resulting relation would be mechanical due to overlapping data in the computation of both $r_{j,t+1}$ and $CPIE_{OWR,j,t+1}$. Second, while the OWR model relies in part on $r_{d,j,t} \times r_{o,j,t}$ to identify private-information arrival, it is a one-period model and has no predictions about the relation between $CPIE_{OWR,j,t}$ and the correlation between $r_{j,t}$ and $r_{j,t+1}$. Thus, for the regressions in this section we rely on our working hypothesis to yield implications for the effect of private-information arrival on the covariance between the daily returns $r_{j,t}$ and $r_{j,t+1}$, not on the OWR model *per se*.

Table 10 reports the coefficient estimates and t-statistics for these regressions. Most importantly, the results in Table 10 show that the estimates for $\beta_3$ in the OWR and GPIN models are positive and significant, indicating that $CPIE_{OWR}$ and $CPIE_{GPIN}$ are both associated with smaller future return reversals. Indeed, for the OWR model, the effect is particularly large. To see this note that a one standard deviation shock to $CPIE_{OWR}$ is associated with a 65% (8.161/12.555) decline in the subsequent reversal. A one standard deviation shock to $CPIE_{GPIN}$, on the other hand, is associated with a 6% (0.414/7.147) drop in

[32] See Hasbrouck (1988, 1991a,b).

the subsequent reversal. Finally, Table 10 presents the coefficient estimates from a regression including both $CPIE_{OWR,j,t}$ and $CPIE_{GPIN,j,t}$ and their interaction terms with $r_{j,t}$. After including both $CPIE$s in the regression, the coefficient estimate on $CPIE_{GPIN} \times r_t$ drops by a factor of four and is rendered insignificant. The coefficient estimate on $CPIE_{OWR} \times r_t$ remains almost unchanged from Specification 2. Thus, these results suggest that both the OWR and GPIN models capture the arrival of private information with persistent impact on prices. However, the OWR model appears to be more strongly associated with the arrival of such information.

# 4 Conclusion

This paper analyzes four structural microstructure models of private information arrival: the PIN model, the APIN model, the OWR model and a new variant of the PIN model, the GPIN model. We show that the PIN and APIN models cannot match the variability of noise trade in the data and, as a result, these models are no more useful in identifying private-information arrival than mechanical heuristics or placebos based on the level of turnover. In contrast, our examination reveals no evidence that either the OWR or GPIN suffer from these issues.

Further examination of the OWR and GPIN models reveals that the OWR model performs somewhat better than the GPIN model in actually identifying the arrival of private information. In sum, our results suggest that proxies for information asymmetry or private-information arrival based on the PIN and APIN models (e.g., $PIN$ and $Adj.PIN$) are unreliable. The GPIN model is a promising alternative to the PIN and APIN models 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 measures based on the APIN and PIN models.

# 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.

Back, Kerry, Kevin Crotty, and Tao Li, 2018, Identifying information asymmetry in securities markets, *The Review of Financial Studies* 31, 2277–2325.

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

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

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

Brennan, Michael J, Sahn-Wook Huh, and Avanidhar Subrahmanyam, 2018, High-frequency measures of informed trading and corporate announcements, *The Review of Financial Studies* 31, 2326–2376.

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

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

Cipriani, Marco, and Antonio Guarino, 2014, Estimating a structural model of herd behavior in financial markets, *The American Economic Review* 104, 224–251.

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

Collin-Dufresne, Pierre, and Vyacheslav Fos, 2016, Insider trading, stochastic liquidity, and equilibrium prices, *Econometrica* 84, 1441–1475.

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

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

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

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

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

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

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

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

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

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

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

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

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

Jegadeesh, N., and Sheridan Titman, 1995, Short-horizon return reversals and the bid-ask spread, *Journal of Financial Intermediation* 4, 116–132.

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

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

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

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.

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.

Table 1: **Summary Statistics.** Panel A presents the mean, standard deviation as well as different percentiles of the data used to estimate the PIN, APIN, GPIN, and OWR models for the full sample and opportunistic insider trade sample. The data include order imbalance ($y_e$), intra-day returns ($r_d$), overnight returns ($r_o$), as well as the number of buys ($B$) and sells ($S$). We compute intraday and overnight returns as well as the number of daily buys and sells for stocks between 1993 and 2012 using data from the NYSE TAQ database, CRSP and COMPUSTAT. The intra-day and overnight returns are risk adjusted using daily cross-sectional regressions. Our sample of opportunistic insider trades is constructed using the method detailed in Cohen, Malloy, and Pomorski (2011). Panel B contains summary statistics for each stock-year in the full sample. The data include the mean number of buys ($\overline{B}$), sells ($\overline{S}$) and turnover ($\overline{turn}$) as well their variances ($\sigma^2_B$, $\sigma^2_S$, $\sigma^2_{turn}$) and the covariance between buys and sells ($cov_{B,S}$).

A. Data for Model Estimation

| | N | Mean | Std | Q1 | Median | Q3 |
|---|---|---|---|---|---|---|
| | | | Full Sample | | | |
| $y_e$ | 5,286,191 | 2.766 | 31.259 | -10.433 | 3.282 | 18.996 |
| $r_d$ | 5,286,191 | -0.004 | 1.500 | -0.707 | -0.024 | 0.680 |
| $r_o$ | 5,286,191 | 0.003 | 1.297 | -0.566 | -0.024 | 0.525 |
| $B$ | 5,286,191 | 1,876 | 6,917 | 37 | 220 | 1,128 |
| $S$ | 5,286,191 | 1,842 | 6,894 | 36 | 194 | 1,033 |
| | | | Opportunistic Insider Trade Sample | | | |
| $y_e$ | 32,944 | 4.696 | 6.638 | -0.182 | 3.080 | 9.713 |
| $r_d$ | 32,944 | -0.006 | 0.179 | -0.099 | -0.003 | 0.092 |
| $r_o$ | 32,944 | 0.029 | 0.157 | -0.061 | 0.018 | 0.107 |
| $B$ | 32,944 | 3,177 | 7,488 | 267 | 916 | 2,742 |
| $S$ | 32,944 | 3,120 | 7,531 | 229 | 802 | 2,620 |

B. Stock-year Sample Moments

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\overline{B}$ | 21,206 | 1,864 | 5,946 | 3 | 41 | 236 | 1,199 | 24,848 |
| $\overline{S}$ | 21,206 | 1,831 | 5,952 | 3 | 40 | 208 | 1,095 | 24,734 |
| $\sigma^2_B$ | 21,206 | 12,074,193 | 235,656,812 | 10 | 569 | 11,943 | 284,581 | 141,637,980 |
| $\sigma^2_S$ | 21,206 | 11,693,254 | 224,595,479 | 8 | 409 | 8,425 | 236,495 | 139,723,237 |
| $cov_{B,S}$ | 21,206 | 11,586,477 | 228,505,016 | 2 | 261 | 7,080 | 225,970 | 133,813,028 |
| $\overline{turn}$ | 21,206 | 3,695 | 11,897 | 6 | 81 | 444 | 2,299 | 49,295 |
| $\sigma^2_{turn}$ | 21,206 | 46,848,275 | 915,328,225 | 23 | 1,530 | 34,824 | 974,185 | 542,363,794 |

Table 2: **PIN Model Parameter Estimates and Implied Moments.** This table presents the mean, standard deviation as well as different percentiles of the parameter estimates and implied moments for the PIN model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\delta$ represents the probability of good news, and $1-\delta$ represents the probability of bad news. The parameters $\epsilon_B$ and $\epsilon_S$ represent the expected number of daily buys and sells given no private information, and $\mu$ is the expected increase in the number of trades given the arrival of private information. $CPIE_{PIN}$ is the probability of private-information arrival on a particular day, conditional on the PIN model parameters and the observed buys and sells. $\overline{CPIE}$ and $Std(CPIE)$ are the mean and standard deviation of $CPIE_{PIN}$ computed for each firm-year. Panel A reports summary statistics for the parameter estimates and the $\overline{CPIE}$ and $Std(CPIE)$ across all firm-years. Panel B reports the model-implied mean, variance, covariance of buys and sells, as well as the model-implied mean and variance of turnover calculated with the estimated parameters.

A. Model Parameters

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.372 | 0.122 | 0.091 | 0.291 | 0.375 | 0.445 | 0.683 |
| $\delta$ | 21,206 | 0.607 | 0.209 | 0.043 | 0.484 | 0.625 | 0.762 | 0.977 |
| $\epsilon_B$ | 21,206 | 1,624.729 | 5,388.488 | 1.949 | 32.550 | 193.133 | 1,038.498 | 22,167.410 |
| $\epsilon_S$ | 21,206 | 1,596.070 | 5,368.939 | 2.690 | 35.476 | 185.862 | 956.037 | 21,964.720 |
| $\mu$ | 21,206 | 312.291 | 593.385 | 6.161 | 43.458 | 160.408 | 314.334 | 2,750.986 |
| $\overline{CPIE}$ | 21,206 | 0.382 | 0.135 | 0.093 | 0.293 | 0.379 | 0.449 | 0.756 |
| $Std(CPIE)$ | 21,206 | 0.451 | 0.052 | 0.274 | 0.427 | 0.470 | 0.490 | 0.500 |

B. Stock-year Implied Moments

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\overline{B}$ | 21,206 | 1,702 | 5,523 | 3 | 41 | 231 | 1,123 | 22,675 |
| $\overline{S}$ | 21,206 | 1,668 | 5,523 | 3 | 40 | 204 | 1,025 | 22,822 |
| $\sigma_B^2$ | 21,206 | 76,061 | 573,245 | 8 | 310 | 4,119 | 17,242 | 1,400,620 |
| $\sigma_S^2$ | 21,206 | 56,858 | 1,684,673 | 4 | 55 | 485 | 4,666 | 757,083 |
| $cov_{B,S}$ | 21,206 | -23,985 | 252,844 | -448,891 | -3,521 | -504 | -25 | -0 |
| $\overline{turn}$ | 21,206 | 3,371 | 11,043 | 6 | 81 | 434 | 2,159 | 45,644 |
| $\sigma_{turn}^2$ | 21,206 | 84,948 | 1,656,501 | 11 | 313 | 3,979 | 16,578 | 1,290,203 |

Table 3: **Regressions of $CPIE_{PIN}$ on the Mechanical Dummy.** This table reports results from the regression: $CPIE_{PIN,j,t} = \beta_0 + \beta_1 CPIE_{Mech,PIN,j,t} + \beta_2 X_{j,t} + \varepsilon_{j,t}$, where $CPIE_{Mech,PIN,j,t}$ is a dummy variable equal to one if stock $j$'s turnover on day $t$ is greater than the mean daily turnover of stock $j$ during the calendar year, and zero otherwise. $X$ represents a vector of covariates consisting of $turn$ and $turn^2$ and additional controls: $|B-S|$, $|B-S|^2$, squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$, and $r_o \times y_e$). We report median coefficient and $t$-statistic estimates (in parentheses), as well as the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles of $R^2$. We compute Newey-West standard errors with a lag length selected according to the Akaike Information Criterion (AIC) from a regression of $CPIE_{PIN}$ on a constant, trend, and quadratic trend.

| | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | 0.063 | 0.109 | 0.113 |
| | (8.68) | (11.38) | (12.10) |
| $CPIE_{Mech,PIN}$ | 0.730 | 0.661 | 0.645 |
| | (44.82) | (32.57) | (31.67) |
| $turn$ | - | 0.169 | 0.114 |
| | - | (8.64) | (5.20) |
| $turn^2$ | - | -0.086 | -0.060 |
| | - | (-5.05) | (-3.47) |
| Controls | No | No | Yes |
| $R^2, 5\%$ | 41.63% | 51.89% | 54.79% |
| $R^2, 50\%$ | 58.56% | 63.83% | 66.13% |
| $R^2, 95\%$ | 73.02% | 75.67% | 78.06% |

Table 4: **APIN Model Parameter Estimates and Implied Moments.** This table presents the mean, standard deviation as well as different percentiles of the parameter estimates and implied moments for the APIN model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\delta$ represents the probability of good news. The parameter $\epsilon_B$ ($\epsilon_S$) represents the expected number of daily buys (sells) given no private information, $\mu_B$ ($\mu_S$) represents the expected additional number of buys (sells) given good (bad) news, and $\Delta_B$ ($\Delta_S$) represents the expected additional number of buys (sells) given an order flow shock. $CPIE_{APIN}$ is the probability of private-information arrival on a particular day, conditional on the APIN model parameters and the observed buys and sells. $\overline{CPIE}$ and $Std(CPIE)$ are the mean and standard deviation of $CPIE_{APIN}$ computed for each firm-year. Panel A reports summary statistics for the parameter estimates and the $\overline{CPIE}$ and $Std(CPIE)$ across all firm-years. Panel B reports the APIN model implied mean, variance, covariance of buys and sells, as well as the implied mean and variance of turnover. Panel B also includes summary statistics for Likelihood Ratio Tests comparing the fit of the APIN model to the PIN model for each firm-year in the sample.

A. Model Parameters

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.456 | 0.092 | 0.199 | 0.409 | 0.464 | 0.509 | 0.670 |
| $\delta$ | 21,206 | 0.550 | 0.192 | 0.069 | 0.441 | 0.541 | 0.680 | 0.963 |
| $\theta$ | 21,206 | 0.249 | 0.137 | 0.004 | 0.149 | 0.253 | 0.344 | 0.566 |
| $\epsilon_B$ | 21,206 | 1,417.934 | 4,570.896 | 1.356 | 25.778 | 158.244 | 866.207 | 19,539.850 |
| $\epsilon_S$ | 21,206 | 1,396.894 | 4,569.861 | 1.954 | 27.610 | 147.615 | 807.330 | 19,617.390 |
| $\mu_B$ | 21,206 | 289.891 | 574.594 | 3.752 | 28.838 | 119.176 | 310.285 | 2,664.918 |
| $\mu_B$ | 21,206 | 283.912 | 573.656 | 3.689 | 26.924 | 106.996 | 301.787 | 2,647.224 |
| $\Delta_B$ | 21,206 | 2,147.940 | 10,058.220 | 4.018 | 41.065 | 189.856 | 988.834 | 30,725.600 |
| $\Delta_B$ | 21,206 | 2,096.510 | 9,934.216 | 3.208 | 33.544 | 159.952 | 907.448 | 29,830.650 |
| $\overline{CPIE}$ | 21,206 | 0.455 | 0.092 | 0.202 | 0.409 | 0.461 | 0.506 | 0.680 |
| $Std(CPIE)$ | 21,206 | 0.454 | 0.056 | 0.272 | 0.431 | 0.479 | 0.493 | 0.500 |

B. Stock-year Implied Moments

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\overline{B}$ | 21,206 | 1,804 | 5,635 | 3 | 42 | 235 | 1,184 | 23,742 |
| $\overline{S}$ | 21,206 | 1,771 | 5,648 | 3 | 40 | 206 | 1,079 | 23,668 |
| $\sigma_B^2$ | 21,206 | 5,008,291 | 102,352,705 | 9 | 432 | 8,550 | 155,932 | 56,223,088 |
| $\sigma_S^2$ | 21,206 | 4,821,736 | 105,048,958 | 3 | 164 | 3,465 | 95,653 | 52,046,962 |
| $cov_{B,S}$ | 21,206 | 4,830,641 | 103,177,649 | 0 | 148 | 3,819 | 108,080 | 52,496,427 |
| $\overline{turn}$ | 21,206 | 3,575 | 11,281 | 6 | 81 | 443 | 2,267 | 47,619 |
| $\sigma_{turn}^2$ | 21,206 | 19,491,309 | 413,403,973 | 14 | 933 | 20,166 | 475,325 | 212,128,415 |
| LRT | 21,206 | 16,245.950 | 22,986.340 | 9.203 | 1,062.791 | 5,084.253 | 22,398.850 | 90,281.360 |
| $p$-value | 21,206 | 0.010 | 0.098 | 0 | 0 | 0 | 0 | 0.056 |

Table 5: **Regressions of $CPIE_{APIN}$ on Mechanical Dummies.** This table reports results from the regression: $CPIE_{APIN,j,t} = \beta_0 + \beta_1 CPIE_{Mech,APIN,j,t} + \beta_2 X_{j,t} + \varepsilon_{j,t}$, where $CPIE_{Mech,APIN,j,t}$ is a dummy variable, analogous to $CPIE_{Mech,PIN}$. See the text for the definition of $CPIE_{Mech,APIN}$. $X$ represents a vector of covariates consisting of $turn$ and $turn^2$ and additional controls: $|B-S|$, $|B-S|^2$, squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$, and $r_o \times y_e$). We report median coefficient and $t$-statistic estimates (in parentheses), as well as the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles of $R^2$. We compute Newey-West standard errors with a lag length selected according to the Akaike Information Criterion (AIC) from a regression of $CPIE_{APIN}$ on a constant, trend, and quadratic trend.

| | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | 0.135 | 0.149 | 0.153 |
| | (13.45) | (14.14) | (15.32) |
| $CPIE_{Mech,APIN}$ | 0.691 | 0.663 | 0.655 |
| | (45.30) | (41.89) | (41.65) |
| $turn$ | - | 0.067 | 0.002 |
| | - | (4.76) | (0.10) |
| $turn^2$ | - | -0.029 | -0.002 |
| | - | (-2.92) | (-0.13) |
| Controls | No | No | Yes |
| $R^2, 5\%$ | 31.66% | 39.32% | 45.51% |
| $R^2, 50\%$ | 54.35% | 56.07% | 59.95% |
| $R^2, 95\%$ | 69.75% | 70.63% | 74.08% |

Table 6: **GPIN Model Parameter Estimates and Implied Moments.** This table presents the mean, standard deviation as well as different percentiles of the parameter estimates and implied moments for the GPIN model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\delta$ represents the probability of good news. The parameters $\theta$ and $\eta$ represent the relative fraction of expected buys when there is no information, and the relative increase in expected turnover when there is private information, respectively. The arrival rate of turnover on a given day $t$ ($\lambda_t$) is drawn from a *Gamma* distribution with shape and scale parameter $r$ and $p/(1-p)$. $CPIE_{GPIN}$ is the probability of private-information arrival on a particular day, conditional on the GPIN model parameters and the observed buys and sells. $\overline{CPIE}$ and $Std(CPIE)$ are the mean and standard deviation of $CPIE_{GPIN}$ computed for each firm-year. Panel A reports summary statistics for the parameter estimates and the $\overline{CPIE}$ and $Std(CPIE)$ across all firm-years. Panel B reports the GPIN model implied mean, variance, covariance of buys and sells, as well as the implied mean and variance of turnover. Panel B also includes summary statistics for Likelihood Ratio Tests and corresponding p-values comparing the fit of the GPIN model to the PIN model for each firm-year in the sample.

A. Model Parameters

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.314 | 0.209 | 0.00001 | 0.137 | 0.262 | 0.491 | 1.000 |
| $\delta$ | 21,206 | 0.562 | 0.186 | 0.175 | 0.430 | 0.546 | 0.683 | 1.000 |
| $\theta$ | 21,206 | 0.499 | 0.070 | 0.218 | 0.483 | 0.507 | 0.538 | 0.610 |
| $\eta$ | 21,206 | 0.446 | 0.368 | 0.00001 | 0.135 | 0.419 | 0.706 | 1 |
| $p$ | 21,206 | 0.931 | 0.861 | 0.460 | 0.899 | 0.981 | 0.997 | 1.000 |
| $r$ | 21,206 | 8.694 | 6.272 | 1.535 | 4.678 | 7.029 | 10.967 | 28.660 |
| $\overline{CPIE}$ | 21,206 | 0.339 | 0.195 | 0.00002 | 0.186 | 0.285 | 0.496 | 1.000 |
| $Std(CPIE)$ | 21,206 | 0.350 | 0.139 | 0.00000 | 0.303 | 0.376 | 0.466 | 0.499 |

B. Stock-year Implied Moments

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\overline{B}$ | 21,206 | 1,859 | 5,910 | 3 | 40 | 234 | 1,199 | 24,771 |
| $\overline{S}$ | 21,206 | 1,831 | 5,965 | 3 | 40 | 207 | 1,091 | 24,865 |
| $\sigma_B^2$ | 21,206 | 8,103,282 | 140,308,553 | 8 | 444 | 9,339 | 229,223 | 103,328,265 |
| $\sigma_S^2$ | 21,206 | 8,275,069 | 144,306,254 | 7 | 367 | 7,490 | 205,005 | 104,955,812 |
| $cov_{B,S}$ | 21,206 | 7,707,312 | 138,562,180 | 1 | 178 | 5,585 | 178,936 | 93,606,349 |
| $\overline{turn}$ | 21,206 | 3,690 | 11,870 | 6 | 81 | 442 | 2,293 | 49,221 |
| $\sigma_{turn}^2$ | 21,206 | 31,792,976 | 561,402,687 | 19 | 1,197 | 28,396 | 793,492 | 400,891,561 |
| LRT | 21,206 | 43,607.630 | 69,036.320 | 11.644 | 1,417.357 | 8,653.139 | 53,631.800 | 259,704.100 |
| $p$-value | 21,206 | 0.008 | 0.089 | 0 | 0 | 0 | 0 | 0.009 |

Table 7: **Regressions of $CPIE_{GPIN}$ on the Mechanical Dummy.** This table reports results from the regression: $CPIE_{GPIN,j,t} = \beta_0 + \beta_1 CPIE_{Mech,j,t} + \beta_2 X_{j,t} + \varepsilon_{j,t}$, where $CPIE_{Mech,j,t}$ is a vector of dummy variables consisting of $CPIE_{Mech,PIN}$ and $CPIE_{Mech,APIN}$. $X$ represents a vector of covariates consisting of $turn$ and $turn^2$ and additional controls: $|B-S|$, $|B-S|^2$, squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$, and $r_o \times y_e$). We report median coefficient and $t$-statistic estimates (in parentheses) as well as the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles of $R^2$. We compute Newey-West standard errors with a lag length selected according to the Akaike Information Criterion (AIC) from a regression of $CPIE_{GPIN}$ on a constant, trend, and quadratic trend.

| | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | 0.277 | 0.294 | 0.306 |
| | (17.42) | (17.81) | (21.42) |
| $CPIE_{Mech,PIN}$ | 0.069 | 0.047 | 0.031 |
| | (3.34) | (1.97) | (1.53) |
| $CPIE_{Mech,APIN}$ | 0.028 | 0.020 | 0.011 |
| | (1.50) | (1.04) | (0.69) |
| $turn$ | - | 0.090 | -0.162 |
| | - | (3.30) | (-5.92) |
| $turn^2$ | - | -0.055 | 0.040 |
| | - | (-2.36) | (1.76) |
| Controls | No | No | Yes |
| $R^2, 5\%$ | 0.04% | 0.43% | 15.59% |
| $R^2, 50\%$ | 1.16% | 4.41% | 34.99% |
| $R^2, 95\%$ | 17.47% | 25.21% | 66.40% |

Table 8: **OWR Parameter Estimates.** This table presents the mean, standard deviation as well as different percentiles of the parameter estimates for the OWR model. The sample consists of 21,206 firm-years from 1993 to 2012. The parameter $\alpha$ is the unconditional probability of private-information arrival on a particular day. The parameter $\sigma_u$ represents the standard deviation of order imbalance due to uninformed trades, which are observed with normally distributed noise with variance $\sigma_z^2$. The parameter $\sigma_i$ is the standard deviation of the informed trader's private signal, while $\sigma_{pd}$ and $\sigma_{po}$ are the standard deviations of the public news component of the idiosyncratic intraday and overnight returns, respectively. $CPIE_{OWR}$ is the probability of private-information arrival on a particular day, conditional on the OWR model parameters and the observed market data. $\overline{CPIE}$ and $Std(CPIE)$ represent the mean and standard deviation of $CPIE_{OWR}$ computed for each firm-year.

| | N | Mean | Std | 1% | Q1 | Median | Q3 | 99% |
|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 21,206 | 0.437 | 0.257 | 0.015 | 0.214 | 0.436 | 0.639 | 0.974 |
| $\sigma_u$ | 21,206 | 0.075 | 0.068 | 0.00001 | 0.022 | 0.062 | 0.109 | 0.309 |
| $\sigma_z$ | 21,206 | 0.239 | 0.143 | 0.00001 | 0.137 | 0.221 | 0.332 | 0.603 |
| $\sigma_i$ | 21,206 | 0.030 | 0.286 | 0.00001 | 0.013 | 0.021 | 0.027 | 0.047 |
| $\sigma_{pd}$ | 21,206 | 0.010 | 0.005 | 0.00001 | 0.006 | 0.009 | 0.012 | 0.026 |
| $\sigma_{po}$ | 21,206 | 0.006 | 0.004 | 0.00001 | 0.004 | 0.006 | 0.008 | 0.020 |
| $\overline{CPIE}$ | 21,206 | 0.451 | 0.258 | 0.018 | 0.227 | 0.455 | 0.656 | 0.974 |
| $Std(CPIE)$ | 21,206 | 0.137 | 0.047 | 0.00000 | 0.109 | 0.142 | 0.171 | 0.229 |

Table 9: **Regressions of $CPIE_{OWR}$ on the Mechanical Dummy.** This table reports results from the regression: $CPIE_{OWR,j,t} = \beta_0 + \beta_1 CPIE_{Mech,j,t} + \beta_2 X_{j,t} + \varepsilon_{j,t}$, where $CPIE_{Mech,j,t}$ is a vector of dummy variables consisting of $CPIE_{Mech,PIN}$ and $CPIE_{Mech,APIN}$. $X$ represents a vector of covariates consisting of $turn$ and $turn^2$ and additional controls: $|B-S|$, $|B-S|^2$, squared intra-day and overnight returns ($r_d^2$, $r_o^2$), squared order imbalance ($y_e^2$) and the three associated interaction terms ($r_d \times r_o$, $r_d \times y_e$, and $r_o \times y_e$). We report median coefficient and $t$-statistic estimates (in parentheses) as well as the $5^{th}$, $50^{th}$, and $95^{th}$ percentiles of $R^2$. We compute Newey-West standard errors with a lag length selected according to the Akaike Information Criterion (AIC) from a regression of $CPIE_{OWR}$ on a constant, trend, and quadratic trend.

| | (1) | (2) | (3) |
|---|---|---|---|
| Intercept | 0.437 | 0.420 | 0.442 |
| | (27.38) | (28.80) | (34.70) |
| $CPIE_{Mech,PIN}$ | 0.049 | 0.066 | 0.028 |
| | (3.02) | (4.11) | (2.09) |
| $CPIE_{Mech,APIN}$ | 0.011 | 0.019 | 0.009 |
| | (0.91) | (1.61) | (0.95) |
| $turn$ | - | -0.082 | -0.088 |
| | - | (-3.71) | (-4.24) |
| $turn^2$ | - | 0.055 | 0.040 |
| | - | (2.72) | (2.49) |
| Controls | No | No | Yes |
| $R^2, 5\%$ | 0.12% | 1.56% | 18.31% |
| $R^2, 50\%$ | 1.24% | 10.14% | 43.36% |
| $R^2, 95\%$ | 6.70% | 38.84% | 80.05% |

Table 10: **Return reversals.** This table reports panel predictive regressions of the open-to-open, risk-adjusted return of stock $j$ on day $t+1$ ($r_{j,t+1}$) on $r_{j,t}$, $CPIE$, and the interaction of $r_{j,t}$ and $CPIE$ for the GPIN and OWR models. The third specification includes both $CPIE_{GPIN}$ and $CPIE_{OWR}$. All specifications include Firm and Year fixed effects, and standard errors are clustered by Firm and Year.

| | (1) | (2) | (3) |
|---|---|---|---|
| $CPIE_{GPIN}$ | $0.052^{***}$ | | $0.047^{***}$ |
| | (6.019) | | (6.049) |
| $CPIE_{OWR}$ | | $0.046^{***}$ | $0.040^{***}$ |
| | | (4.356) | (3.856) |
| $r_t$ | $-7.147^{***}$ | $-12.555^{***}$ | $-12.597^{***}$ |
| | $(-7.461)$ | $(-6.057)$ | $(-6.239)$ |
| $CPIE_{GPIN} \times r_t$ | $0.414^{*}$ | | 0.140 |
| | (1.716) | | (0.537) |
| $CPIE_{OWR} \times r_t$ | | $8.161^{***}$ | $8.082^{***}$ |
| | | (4.158) | (4.110) |
| Firm-Year FE | Yes | Yes | Yes |
| $R^2$ | 0.5% | 0.6% | 0.6% |

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

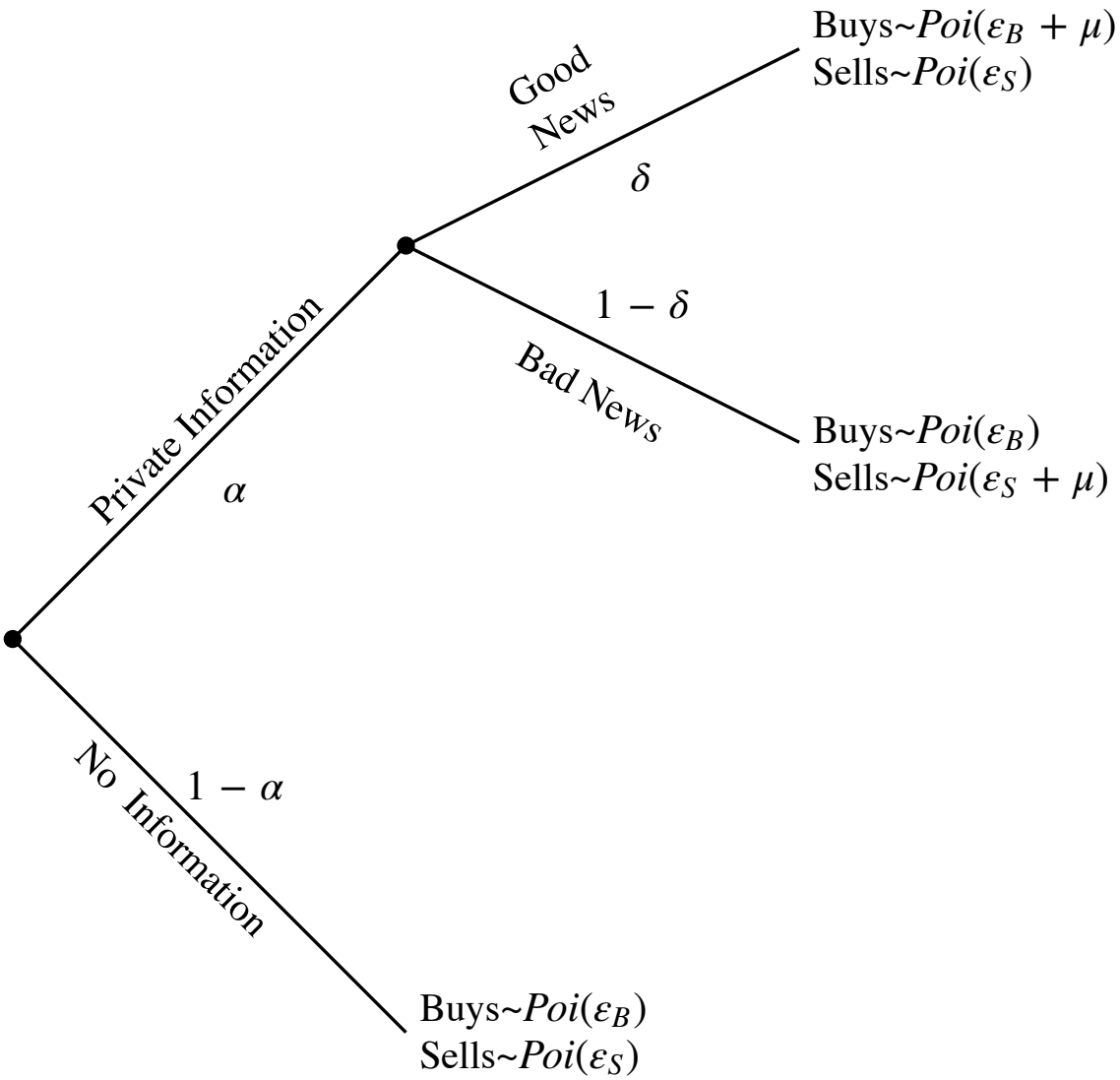

Figure 2: **PIN Model Example.** This figure compares real and simulated data for Exxon-Mobil (XOM) in 1993 and 2012 from the PIN model. In Panels A and B, the real data are marked as ×. The real data are shaded according to the $CPIE_{PIN}$, with darker markers (× magenta) representing high $CPIE$s and lighter markers (× cyan) low $CPIE$s. All the observations below (above) the downward-sloping dashed line have turnover below (above) the annual mean of daily turnover. The upward-sloping dotted line comes from a regression of sells on buys. High (low) probability states in the simulated data appear as a dark (light) "cloud" of points. The PIN model has three states: no news, good news, and bad news. Panels C and D plot the $CPIE$s for the real data as a function of turnover along with a dashed line indicating the mean turnover.

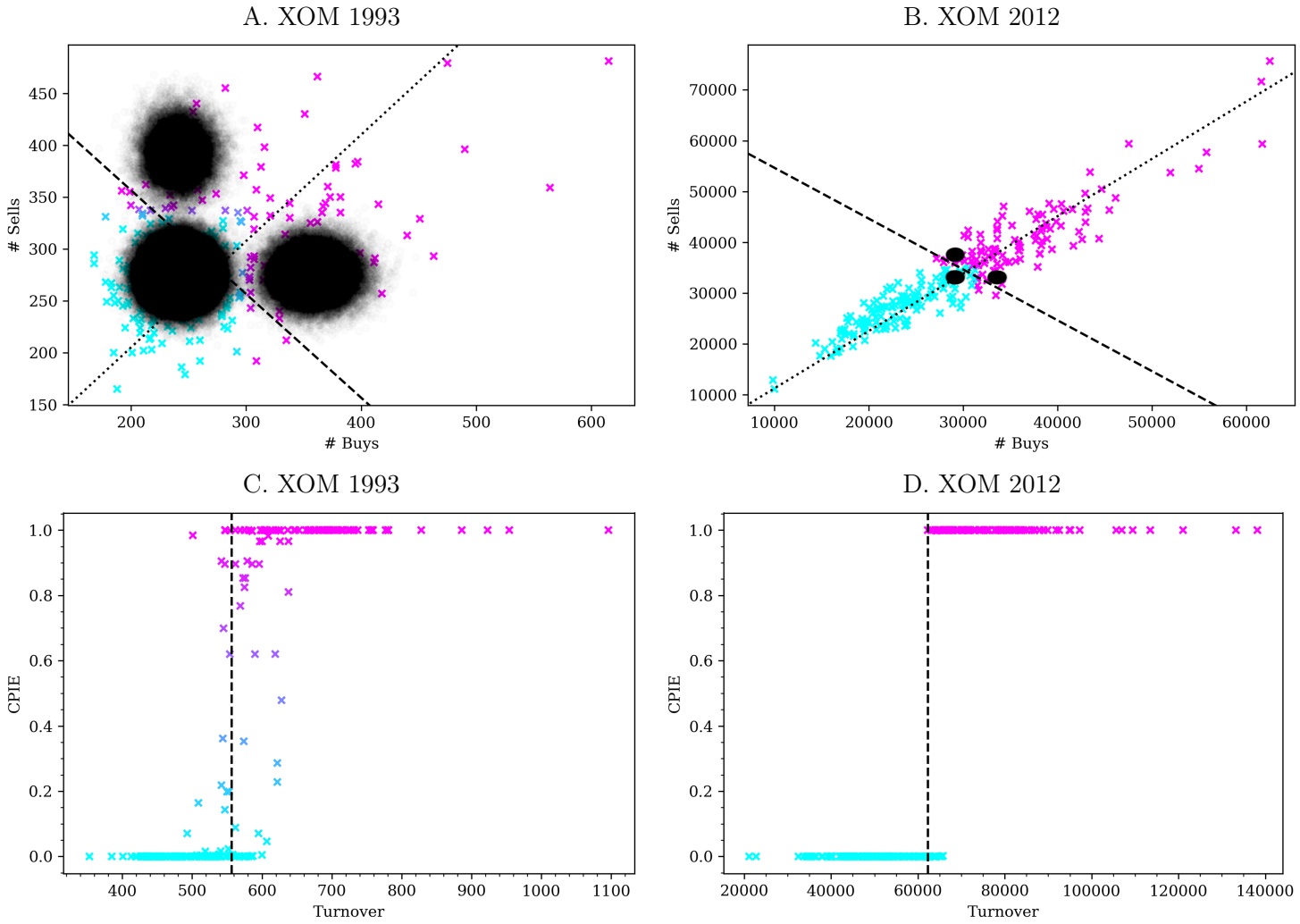

Figure 3: **APIN Model Tree.** The APIN model is a mixture of two independent PIN models. Shocks to the intensity of noise trading arrive with probability $\theta$. These shocks increase the expected amount of non-informed buys (sells) by $\Delta_B$ ($\Delta_S$). As with the PIN, private information arrives with probability $\alpha$. When there is no private information, and no symmetric order flow shock, buys and sells are distributed as Poisson random variables with intensity $\epsilon_B$ and $\epsilon_S$. When a symmetric order flow shock occurs without private information, buys and sells are distributed as Poisson random variables with intensity $\epsilon_B + \Delta_B$ and $\epsilon_S + \Delta_S$. Private information is good (bad) news with probability $\delta$ $(1-\delta)$. The expected number of buys (sells) increases by $\mu_B$ ($\mu_S$) in case of good (bad) news arrival.

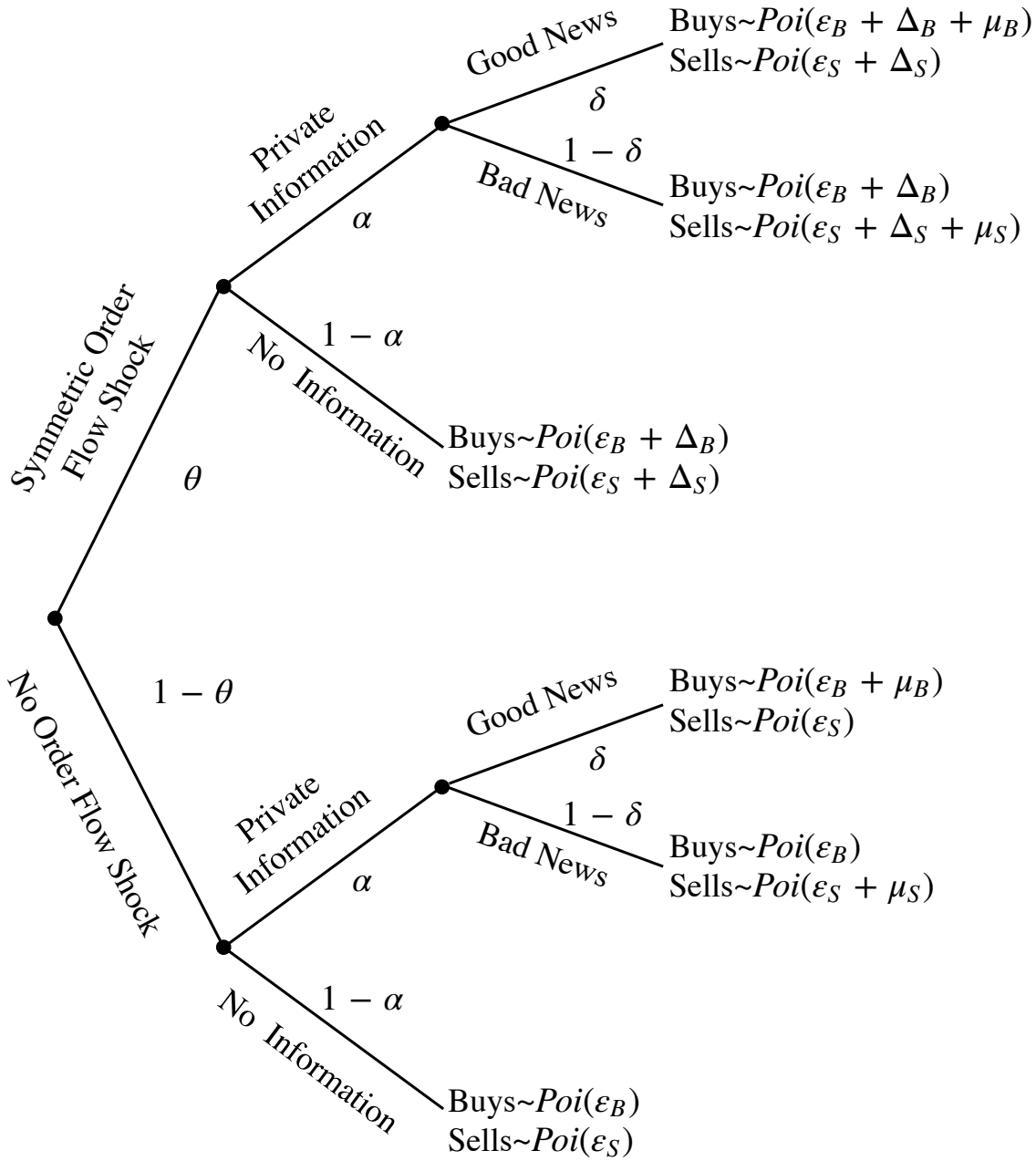

Figure 4: **APIN Model Example.** This figure compares real and simulated data for Exxon-Mobil (XOM) in 1993 and 2012 from the APIN model. In Panels A and B, the real data are marked as ×. The real data are shaded according to the $CPIE_{APIN}$, with darker markers (× magenta) representing high $CPIE$s and lighter markers (× cyan) low $CPIE$s. The upward-sloping dotted line comes from a regression of sells on buys. High (low) probability states in the simulated data appear as a dark (light) "cloud" of points. The APIN model has six states corresponding to the high and low order flow states, and good, bad, or no news arrival. Panels C and D plot the $CPIE$s for the real data as a function of turnover along with three dashed lines corresponding to mean turnover conditional on the presence (or absence) of a symmetric order flow shock, and the mean of the two conditional means.

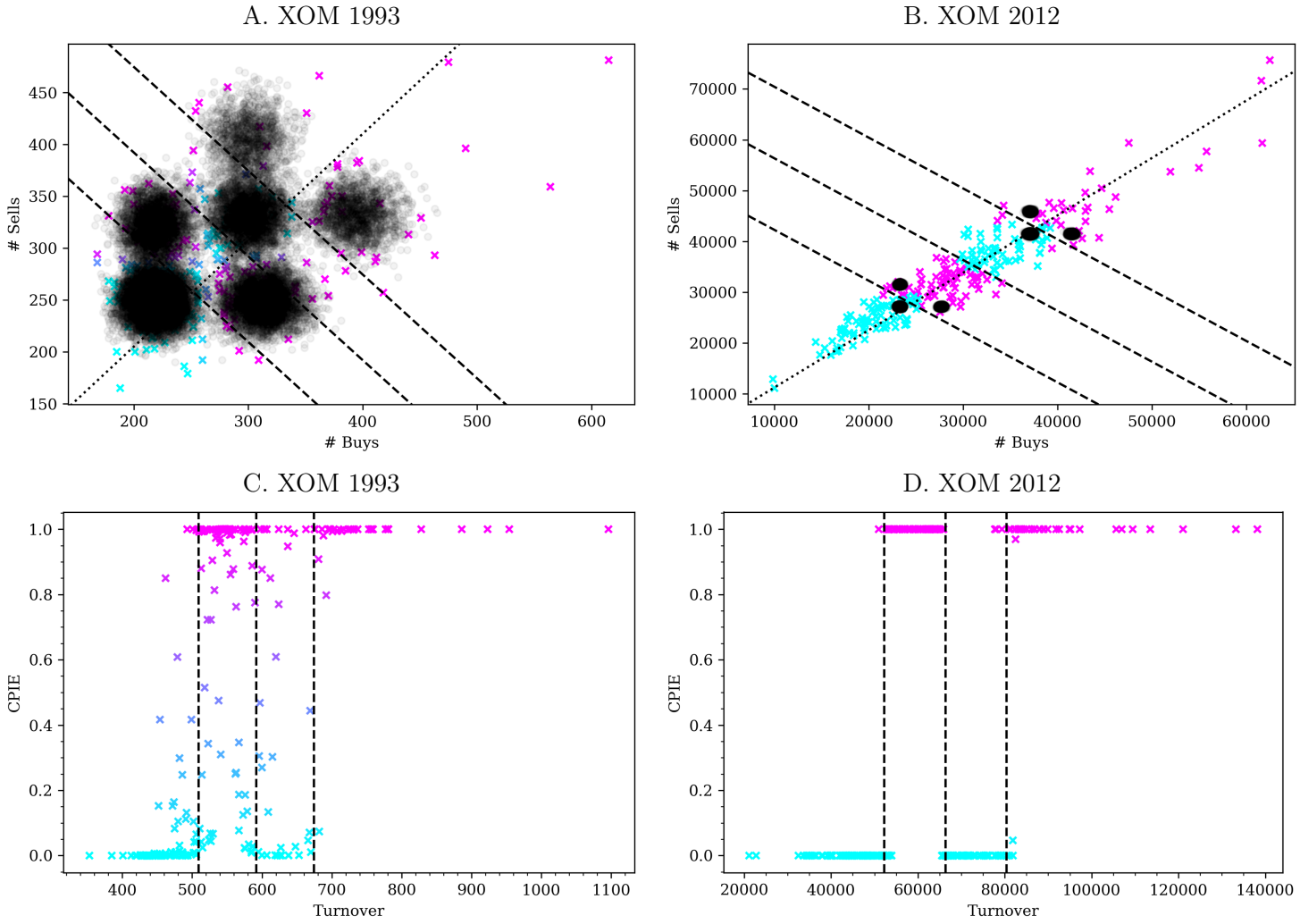

Figure 5: **GPIN Model Tree.** The GPIN model generalizes the PIN model by allowing the arrival rate of noise-trading order flow ($\lambda_t$) to be drawn from a *Gamma* distribution with shape and scale parameters $r$ and $p/(1-p)$ (e.g., $\lambda_t \sim \Gamma(r,p)$). As with the PIN model, private information arrives with probability $\alpha$. When there is no private information, buys and sells are distributed as Poisson random variables with intensity $\theta\lambda_t$ and $(1-\theta)\lambda_t$. Private information is good (bad) news with probability $\delta$ $(1-\delta)$. The expected number of buys (sells) increases proportionally by $\eta$ when there is news arrival.

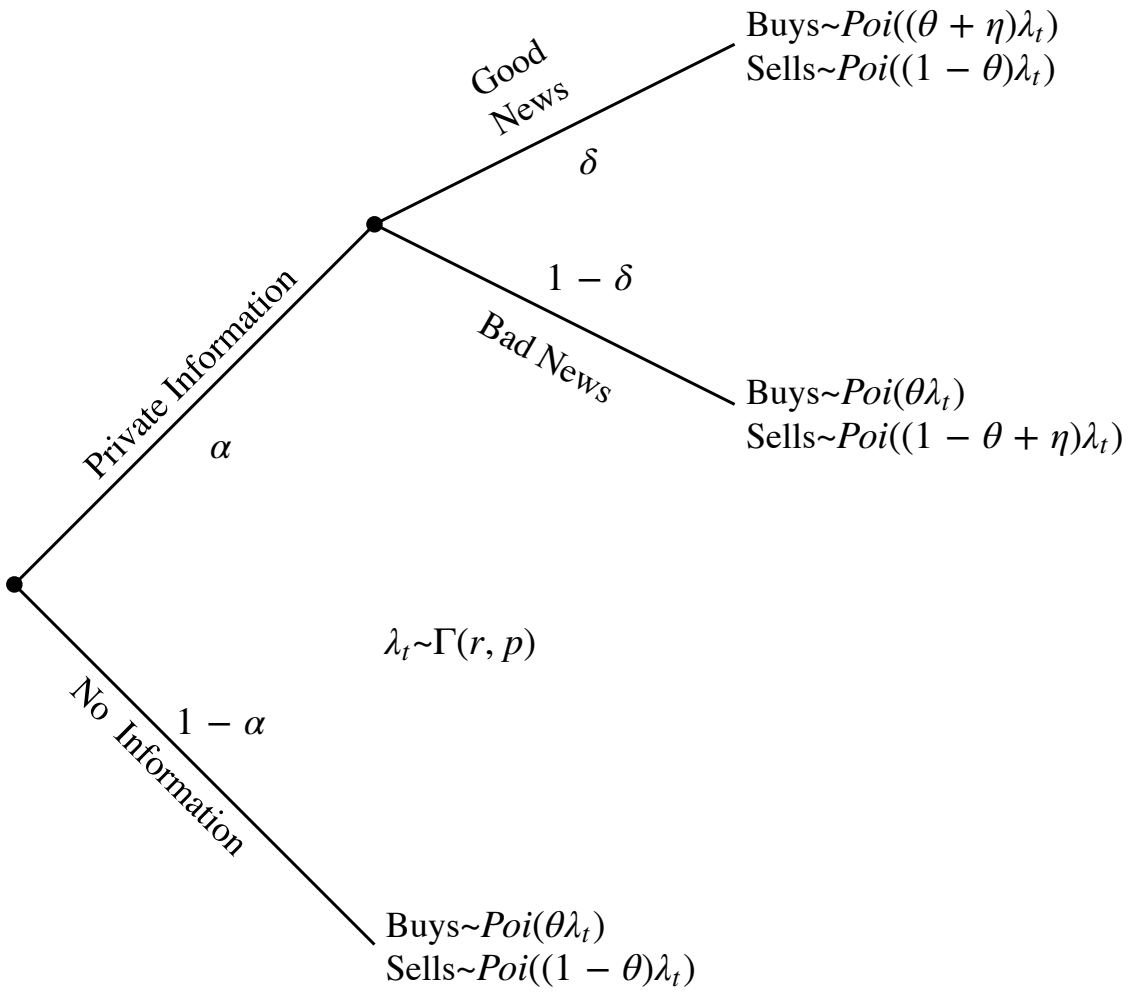

Figure 6: **GPIN Model Example.** This figure compares real and simulated data for Exxon-Mobil (XOM) in 1993 and 2012 from the GPIN model. In Panels A and B, the real data are marked as ×. The real data are shaded according to the $CPIE_{GPIN}$, with darker markers (× magenta) representing high $CPIE$s and lighter markers (× cyan) low $CPIE$s. The upward-sloping dotted line comes from a regression of sells on buys. High (low) probability states in the simulated data appear as a dark (light) "cloud" of points. The GPIN model has three states: no news, good news, and bad news. Panels C and D plot the $CPIE$s for the real data as a function of turnover along with a dashed line indicating the mean turnover.

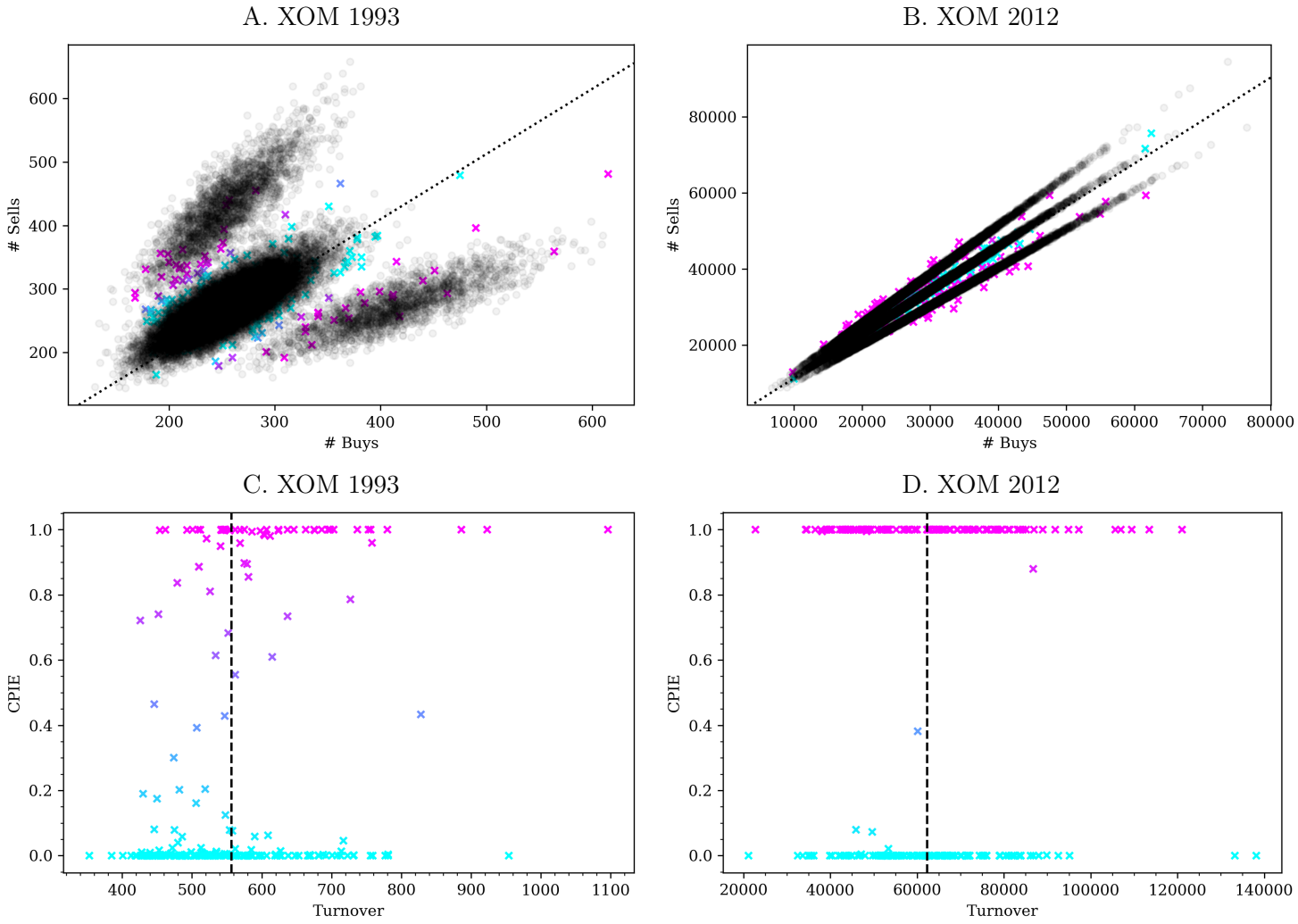

Figure 7: **OWR Model Tree.** In the OWR model, prior to markets opening, private information arrives with probability $\alpha$. Once markets open, investors submit their trades generating order imbalance ($y_e$), and the intraday return ($r_d$). After markets close, private information becomes public and is reflected in the overnight return ($r_o$). The variables ($y_e$, $r_d$, $r_o$) are normally distributed with mean zero. The covariance differs between days with private-information arrival, $\Sigma_I$, and days without the arrival of private information, $\Sigma_{NI}$. When there is no private-information arrival, there is a price reversal in the overnight return ($cov(r_d, r_o) < 0$) and when there is private-information arrival there is a continuation in the returns ($cov(r_d, r_o) > 0$).

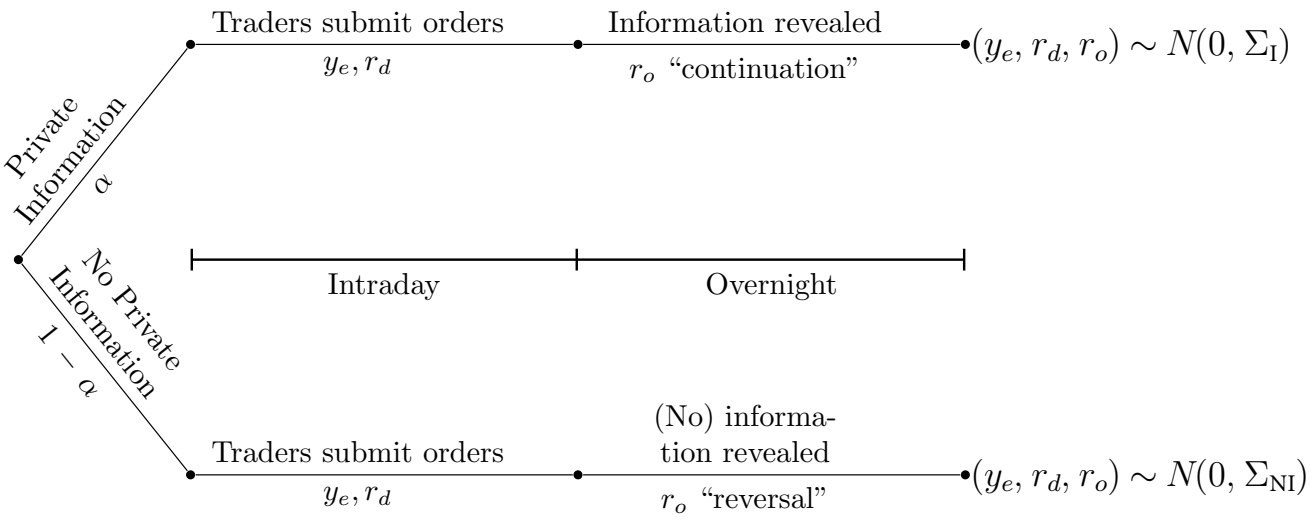

Figure 8: **OWR Model Example.** This figure plots data for Exxon-Mobil (XOM) in 1993 and 2012 from the OWR model. In Panels A and B, the data are marked as ×. The data are shaded according to the $CPIE_{OWR}$, with darker markers (× magenta) representing high and lighter markers (× cyan) low $CPIE$s. The upward-sloping dotted line comes from a regression of sells on buys. Panels C and D plot the $CPIE$s for the data as a function of turnover along with a dashed line indicating the mean turnover.

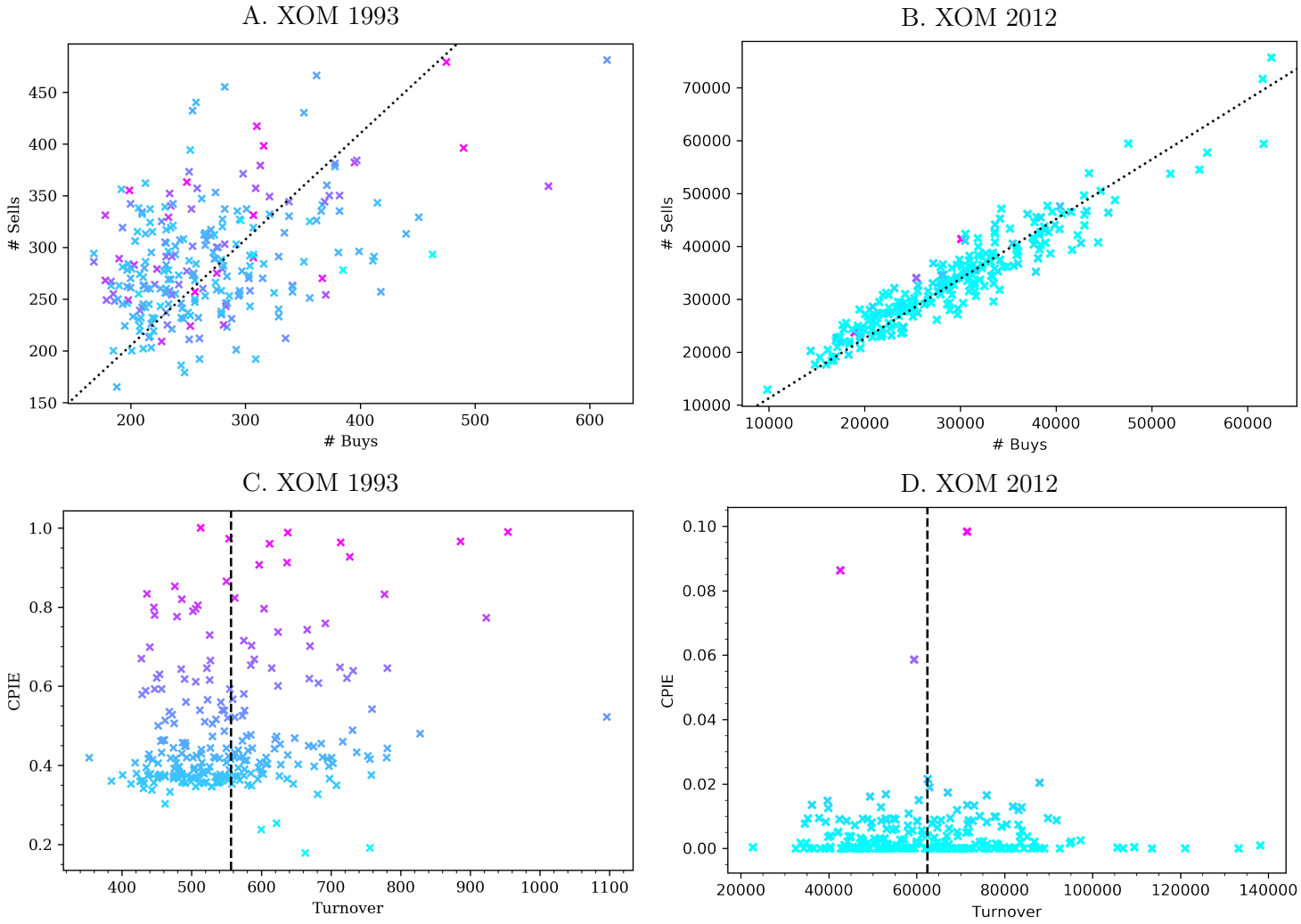

Figure 9: ***CPIE*$_{GPIN}$ and *CPIE*$_{OWR}$ around Insider Trades.** This figure plots the average $CPIE$s in event time surrounding opportunistic insider trades. The dashed lines are two standard errors from the mean estimated over the window [-40,-20].