Firm Characteristics and Informed Trading: Implications for Asset Pricing

Hadiye Aslan
University of Houston

David Easley
Cornell University

Soeren Hvidkjaer
Copenhagen Business School

Maureen O'Hara
Cornell University

September 2008

We would like to thank K.C. Chan, John Griffin, Inmoo Lee, Alexander Kempf, seminar participants at the University of Texas (Austin), National University of Singapore (NUS) and conference participants at the International Conference on Finance (University of Copenhagen), the Microstructure of Capital Markets Conference (City University London), and the European Finance Association Meeting in Ljubljana for helpful suggestions. An earlier version of this paper was entitled "Determinants of Informed Trade: Implications for Asset Pricing".

Firm Characteristics and Informed Trading: Implications for Asset Pricing

Abstract

This paper investigates the linkage of microstructure, accounting, and asset pricing. We determine the relationship between firm characteristics as captured by accounting and market data and a firm's probability of private information-based trade (PIN) as estimated from trade data. This allows us to determine what types of firms have high information risk. We then use these data to create an instrument for PIN, the PPIN, which we can estimate from firm-specific data. We show that PPINs have explanatory power for the cross-section of asset returns in long sample tests. We also investigate whether information risk vitiates the influence of other variables on asset returns. Our results provide strong support for information risk affecting asset returns, and suggest that PIN weakens, but does not remove, the role of size in asset returns.

Firm Characteristics and Informed Trading: Implications for Asset Pricing

1. Introduction

Microstructure research has increasingly focused on the important real effects of microstructure variables. While traditionally such research delineated the influence of microstructure on short term market variables, more recent research has shown that microstructure can explain a variety of long term effects due to its role in influencing the liquidity and price discovery of assets. Thus, microstructure research has evolved from determining trading costs and spreads to providing insights into areas such as asset pricing. Similarly, researchers in asset pricing are increasingly investigating a broader set of factors thought to affect asset market behavior. Asset pricing researchers have found success augmenting traditional measures of market risk with both accounting data and microstructure variables. Thus, asset pricing research has evolved from simple market risk models to complex factor models based on accounting and market data.

In this paper, we investigate the linkage of microstructure, accounting and asset pricing. While distinct in many ways, each of these areas tries to understand the behavior and pricing of financial assets. As we demonstrate in this paper, a unifying influence on these areas is the role played by information. Microstructure models (see Easley, Kiefer, O'Hara [1997a]) provide a measure of private information-based trading, the PIN measure, that has been shown to be useful in explaining asset returns. PIN is derived from market trade data, but it is fundamentally a proxy for a firm's information environment. Accounting and other market data also provide measures of a firm's information environment. Thus, accounting variables, Tobin's Q, and industry variables are widely used to measure a firm's economic performance and characteristics, and they have also been shown to matter for asset pricing. The questions we address in this research are how does PIN relate to these accounting variables, and what does this tell us about asset pricing?

There are three important reasons for pursuing this agenda. First, linking PIN and accounting measures allows us to develop operational measures for informed trading and information risk. Information risk arises when some investors have better information than others about a firm's prospects. How large this risk is depends upon a variety of

factors such as the nature and quality of a firm's accounting information, the availability of public information sources about the firm, the frequency of new information events, and the fraction of traders who have better information. The accounting and finance literatures have employed a number of proxies for private information (for example, analyst coverage and abnormal accruals), but these measures do not include the market information captured by PINs. Understanding how these measures relate to PINs provides insights into what kinds of firms have higher information risk. Moreover, linking PINs to variables widely believed to reflect private information provides an independent check on our interpretation of PIN as a measure of information risk.[1]

Second, determining how market and accounting characteristics relate to PIN allows us to determine a proxy for PIN. PIN measures are derived from trade data, and their estimation requires using tick-by-tick data to determine trade imbalances and activity. Recent structural changes in the market, however, have greatly reduced market depth and consequently increased the number of trades in active stocks. Such massive volume has severely compromised the maximum likelihood estimation technique needed to estimate PINs for active stocks. The technique we develop in this paper provides a mechanism to overcome this problem by developing a proxy for PIN, denoted PPIN. Such PPINs may be both easier to estimate and more widely applicable to a range of firms and markets.

Third, developing PPINs allows us to investigate the role of information risk in asset pricing using longer time periods. Easley and O'Hara (2004) demonstrate theoretically why information risk should affect asset returns, and Easley, Hvidkjaer and O'Hara (2002, 2008) provide strong empirical evidence in support of this effect. These empirical studies have limited sample periods, however, due to the unavailability of high frequency data in the U.S. before 1983. Using the PPIN technique developed here, we can estimate PINs over earlier time periods, allowing us to investigate the time spans typically employed in long horizon asset pricing studies. Our conjecture is that some of

[1] Some research (see Duarte and Young [2007]) argues that liquidity effects unrelated to information risk explain asset returns. Our work here refutes this contention by delineating how PINs relate to variables widely acknowledged in the literature as information proxies. See also Easley, Hvidkjaer, and O'Hara [2008] for evidence on the robustness of PINs in a factor pricing structure, and Li, Wang, Wu, and He [2008] who demonstrate distinct roles for information risk and liquidity in the Treasury bond market.

the variables found to matter for asset pricing are actually proxying for information risk, and our analysis here provides a way to investigate this hypothesis.

Our approach in this paper involves first describing the relationship between firm characteristics as captured by accounting and market data and a firm's PINs. We use a time-series of PINs estimated from market data from the period 1983-1999, and we regress those PINs on a wide variety of firm and industry specific variables suggested by the accounting and asset pricing literatures. Our analysis reveals a number of intriguing relationships between informed trading and industry structure, as well as with firm characteristics such as age, size, growth, profits, insider holding, institutional trading, and accruals.

We then use these data to create an instrument for PIN, the PPIN, which we can estimate from these firm-specific data. Our goal here is to create an instrument that has explanatory power for the cross-section of stock returns. We provide a number of diagnostic tests to evaluate the efficiency of the PPIN, and we show that PPINs are a successful proxy for PIN variables. Of particular importance, we find using in-sample asset pricing tests that PPINs perform as well as PIN variables.

This sets the stage for our analysis of the role of information risk in asset pricing. We construct a time-series of PPINs for each sample firm over the time period 1965-2004. We then use our PPINs as explanatory variables in standard asset pricing regressions. The asset pricing literature has suggested a wide variety of variables believed to explain asset returns. These include size and book to market (Fama and French [1993]), turnover (Chordia, Subrahmanyam, and Anshuman [2001]) and momentum (Jegadeesh and Titman [1993]; Carhart [1997]; Grundy and Martin [2001]). We test two hypotheses: first, is information risk priced in asset returns? and second, does information risk vitiate the influence of other variables on asset returns?

We find that information risk as captured by PPINs is both statistically and economically significant for asset prices. We also show that this PPIN effect is robust to the inclusion of explanatory factors such as Beta, Size, Book-to-Market, and Momentum, as well as to dollar volume and turnover. Our results with respect to these latter variables suggest that it is information and not these measures of illiquidity that matters for asset returns. Turning to our second hypothesis, we find that PPIN weakens, but does not

vitiate, the role of other variables in asset pricing. Of particular importance is that Size remains an important influence on asset returns. We interpret our results as suggesting that the Size effect is not created just by differences across firms in the extent to which information is private rather than public. It could, as Barry and Brown (1984) propose, proxy for the total amount of information available about a firm. But it may also arise for a variety of factors related to small firms such as financial constraints, bankruptcy concerns, or the like.

This paper is organized as follows. The next section discusses the nature of information risk, the role of public and private information, and the disparate approaches taken in the accounting and microstructure literatures to measure private information. This section also sets out the derivation and estimation of PIN variables. Section 3 presents the data and sample. Section 4 investigates the relationship between firm characteristics and PIN. Section 5 investigates asset pricing and information risk over both short and long sample periods. This section also discusses methodological issues connected with our analysis. Section 6 is a conclusion.

2. Information Risk and PIN

Standard asset pricing models accord no role to information in affecting a firm's cost of capital. This is because theoretical asset pricing analyses typically consider a homogeneous investor world in which the effect of differences in beliefs on asset prices cannot be analyzed. The standard asset pricing empirical literature builds from this idealized view of the economy, and it does not consider the possibility of information affecting beliefs, or by extension prices and returns on assets. In this world, the representative individual is compensated only for holding aggregate risk, idiosyncratic risk need not be held, and so in equilibrium there is no compensation for holding it.

If investors have differential information, however, they will have heterogeneous beliefs, and they may have differing views of aggregate and idiosyncratic risk. In equilibrium, even fully rational, but differentially informed, investors will forecast differing returns to assets, and they will hold different portfolios. Easley and O'Hara (2004) build on the classic rational expectations analysis of Grossman and Stiglitz (1980) to show how these differing portfolio choices resulting from differential information

change returns.[2] To see how prices are affected by private information, consider an uninformed investor faced with a choice between two assets which are identical except that one asset has less public information and more private information. The uninformed investor loses to the informed investors who know the private information, and so requires a greater expected return to hold the asset with more private, and less public, information. Easley and O'Hara (2004) show that in a rational expectations equilibrium, assets with more private, and less public, information should have greater expected returns.[3]

Both the market microstructure literature and the accounting literature offer empirical support for the claim that differential information matters for returns. In the microstructure literature the most direct evidence comes from Easley, Hvidkjaer and O'Hara (2002). They measure the probability of information-based trade using microstructure data and find that a 10 percentage point increase in the probability of information based trade leads to a 2.5% increase in annual expected returns. An alternative approach is taken by Amihud and Mendelson (1986, 1989) who show that firms with greater spreads have a greater cost of capital. Their analysis suggests that firms can reduce the spread by proving public information.

In the accounting literature, a number of variables are found to be related to the cost of capital. Botosan (1997) shows that for a sample of firms with low analyst following, greater disclosure of information reduces the cost of capital by an average of 28 basis points. Botosan, Plumlee and Xie (2004) find that proxies for information precision affect the cost of equity capital. Francis, Lafond, Olsson and Schipper (2004, 2005) show that firms with lower quality earnings, measured primarily by abnormal accruals, have a higher cost of capital.[4] Numerous accounting studies document the effect of information disclosure on returns; see Healy and Palepu (2001) for a survey of the empirical work and Verrecchia (2001) for a survey of the theoretical work on this topic.

---

[2] See also Admati (1985) and Wang (1993) who analyze the CAPM in an asymmetric information world.
[3] Diamond and Verrecchia (1991) consider an alternative mechanism by which information can affect returns by analyzing how disclosure affects the willingness of market makers to provide liquidity for a stock.
[4] An alternative view is found in Cohen (2005) who finds that firms providing higher quality financial information do not exhibit a lower cost of capital.

Our goal in this research is to relate our measure of private information based trading, PIN, to these, and other accounting measures, which might affect the cost of capital. PIN variables are derived from trade data, and essentially use the pattern and volume of buys and sells to infer the presence and frequency of information-based trades. An advantage of this approach is that it uses market data to infer information risk for a specific firm. A limitation of this approach, however, is that it does not use firm-specific variables such as accounting treatments, economic performance metrics, or outside analyst coverage that might naturally be expected to relate to the firm's information environment. Blending these approaches allows us to use both sources of information to measure a firm's information risk.[5] We show how to construct a proxy for PIN from accounting and market measures which we then use to explain expected returns.

A. The Derivation and Estimation of PIN Variables

We now describe a methodology for estimating the risk of private information-based trading. This approach uses a structural microstructure model to formalize the learning problem confronting a market maker in a world with informed and uninformed traders. In a series of papers, Easley et al demonstrate how such models can be estimated using trade data to determine the probability of information-based trading, or PIN, for specific stocks. The rest of this section sets out this approach, drawing heavily from Easley, Hvidkjaer, and O'Hara (2002). Readers conversant with the PIN methodology can proceed directly to the next section.

Microstructure models depict trading as a game between the market maker and traders that is repeated over trading days i=1,…,I. First, nature chooses whether there is new information at the beginning of the trading day, and these events occur with probability α. The new information is a signal regarding the underlying asset value, where good news is that the asset is worth $\overline{V}_i$, and bad news is that it is worth $\underline{V}_i$. Good news occurs with probability $(1-\delta)$ and bad news occurs with the remaining probability, δ. Trading for day i then begins with traders arriving according to Poisson processes throughout the day. The market maker sets prices to buy or sell at each time t in [0,T]

[5] See also Botosan and Plumlee (2003) who use PIN measures to proxy for information dispersion across traders in a study of information attributes and accounting based measures of the expected cost of equity capital

during the day, and then executes orders as they arrive. Orders from informed traders arrive at rate μ (on information event days), orders from uninformed buyers arrive at rate ε_b and orders from uninformed sellers arrive at rate ε_s. Informed traders buy if they have seen good news and sell if they have seen bad news. If an order arrives at time t, the market maker observes the trade (either a buy or a sale), and he uses this information to update his beliefs. New prices are set, trades evolve, and the price process moves in response to the market maker's changing beliefs. This process is captured in Figure 1.

The structural model described above allows us to relate observable market outcomes (i.e. buys or sells) to the unobservable information and order processes that underlie trading. The likelihood function for trade on a single trading day that is implied by this model is

$$
\begin{aligned}
L(\theta \mid B,S) &= (1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!} \\
&+ \alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^S}{S!} \\
&+ \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!}
\end{aligned}
\qquad (1)
$$

where B and S represent total buy trades and sell trades for the day respectively, and $\theta = (\alpha, \mu, \varepsilon_B, \varepsilon_S, \gamma)$ is the parameter vector. This likelihood is a mixture of distributions where the trade outcomes are weighted by the probability of it being a "good news day" $\alpha(1-\delta)$, a "bad news day" ($\alpha\delta$), and a "no-news day" ($1-\alpha$).

Imposing sufficient independence conditions across trading days gives the likelihood function across I days

$$
V = L(\theta \mid M) = \prod_{i=1}^{I} L(\theta \mid B_i, S_i) \qquad (2)
$$

where (B_i, S_i) is trade data for day i = 1,…,I and $M=((B_1,S_1), \ldots,(B_I,S_I))$ is the data set.[6] Maximizing (4) over θ given the data M thus provides a way to determine estimates for the underlying structural parameters of the model (i.e. α, μ, ε_B, ε_S, δ).

This model allows us to use observable data on the number of buys and sells per day to make inferences about unobservable information events and the division of trade between the informed and uninformed. In effect, the model interprets the normal level of buys and sells in a stock as uninformed trade, and it uses this data to identify the rates of uninformed order flow, ε_B and ε_S. Abnormal buy or sell volume is interpreted as information-based trade, and it is used to identify μ. The number of days in which there is abnormal buy or sell volume is used to identify α and δ. Of course, the maximum likelihood actually does all of this simultaneously.

The estimation of the model's structural parameters can be used to construct the probability that an order is from an informed trader, known as a PIN. In particular, given some history of trades, the market maker can estimate the probability that the next trade is from an informed trader. It is straightforward to show that the probability that the opening trade is information-based is given by

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_S + \varepsilon_B} \tag{3}$$

where $\alpha\mu + \varepsilon_S + \varepsilon_B$ is the arrival rate for all orders and $\alpha\mu$ is the arrival rate for information-based orders. PIN is thus a measure of the fraction of orders that arise from informed traders relative to the overall order flow.

Because PIN variables provide a direct measure of the risk of information-based trading, they are useful in addressing a wide range of microstructure issues. For example, in the standard microstructure model with competitive, risk neutral market makers, the

[6] The independence assumptions essentially require that information events are independent across days. Easley, Kiefer, and O'Hara (1997b) do extensive testing of this assumption and are unable to reject the independence of days.

opening spread is directly related to PIN.[7] Other uses of PIN have been to assess differential information of order flows across markets, to ascertain whether local or foreign investors trade more on private information, and to investigate the information content of foreign exchange trading, to name but a few applications. Of particular importance for our study here is that PIN as been shown to matter for asset pricing, an issue we return to in Section 5. In the next section we turn to the estimation of PINs and to the data we use in our analysis.

3. Data and Sample

We estimate our model for the sample of all ordinary common stocks listed on the New York Stock Exchange (NYSE) and the American Stock and Exchange (AMEX) for the years 1983 to 1999. In this estimation we exclude real estate investment trusts, stocks of companies incorporated outside of the United States, and closed-end funds. Also, we include only stocks which have at least 60 days with trade or quote data in a given year. This leaves a sample of between 1858 and 2371 stocks to be analyzed each year. In the next section we briefly describe the estimation of PIN.[8]

A. The PIN estimation

The likelihood function given in equation (2) depends on the number of buys and sells each day for each stock. To construct this data, we first extract transactions data from the Institute for the Study of Security Markets (ISSM) and Trade And Quote (TAQ) datasets. The ISSM and TAQ data provide a complete listing of quotes and trades for each traded security. For our analysis, we require the number of buys and sells for each

---

[7] In particular, in the case where the uninformed are equally likely to buy and sell ($\varepsilon_b = \varepsilon_s = \varepsilon$) and news is equally likely to be good or bad ($\delta = 0.5$), the percentage opening spread is

$$\frac{\Sigma}{V^*_i} = (PIN)\frac{(\overline{V_i} - \underline{V_i})}{V^*_i}$$

Where Σ is the opening spread, V^*_i is the unconditional expected value of the asset given by $V^*_i = \delta\underline{V_i} + (1-\delta)V_i$. Note that if PIN equals zero, either because of the absence of new information (α) or traders informed of it (μ), the spread is also zero. This reflects the fact that only asymmetric information affects spreads when market makers are risk neutral.

[8] The yearly, stock-by-stock results from this estimation can be found at www.smith.umd.edu/faculty/hvidkjaer/data.htm .

day, but the data record only transactions, not who initiated the trade. To sign trades as buys or sells, we use the standard Lee and Ready (1991) algorithm. Trades at prices above the midpoint of the bid-ask intervals are classified as buys, and trades below the midpoint are classified as sells. Trades occurring at the midpoint of the bid and ask are classified using the tick test, which compares the price with the previous transaction price to determine the trade direction. We apply this algorithm to each transaction in our sample to determine the daily numbers of buys and sells. Using a maximum likelihood procedure, we estimate the structural parameters of the model, $\Theta = (\alpha, \mu, \varepsilon_B, \varepsilon_S, \delta)$ simultaneously for each stock separately for each year in the period. The maximum likelihood estimation converges for 98.8 percent of the stocks in our sample. However, in stocks with very large number of trades, the optimization program encounters computational underflow, and we are unable to evaluate the likelihood function for those stocks.

B. The explanatory variables

We obtain data from several sources. Data on firm characteristics, returns and standard accounting variables come from the monthly CRSP and the annual COMPUSTAT files. *SIZE* is the market value of equity in firm *i* at the end of year *t*, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *TURNOVER* represents the number of shares traded divided by the number of shares outstanding, share turnover, for firm *i* in year *t*. *AGE* represents the number of years since the stock *i* was first covered by CRSP. We take the logarithm of *SIZE*, *STDEV*, *TURNOVER*, and 1+*AGE*, denoted *LSIZE*, *LSTDEV*, *LTURNOVER*, and *LAGE*, respectively. *GROWTH* is the percentage increase in sales (item 12) from year *t*-1 to year *t*. We measure the return on assets, *ROA,* as the ratio of operating income after depreciation income before extraordinary items to the book value of assets (item 178/ item 6). Tobin's Q is defined as the market value of the firm's assets divided by the replacement costs of the assets. We construct Tobin's Q, *TOBIN,* as the market value of assets divided by the book value of assets, where the market value of assets equals the sum of book value of assets and the market value of common equity (item 60) less the book value of common stock and balance sheet deferred taxes (item 74). To avoid

spurious inferences, *TOBIN*, *GROWTH* and *ROA* are winsorized at the top and bottom one percent of their respective distributions because these variables take extreme values for some firms.

ACCRUALS is the estimate of the discretionary component of total accruals based on Jones' (1991) model which maps current period working capital accruals into operating cash flow realizations. Recently, Francis, Lafond, Olson and Schipper (2004, 2005) have used this metric as a proxy for information risk. Specifically, for each of the 48 Fama-French (1997) industry definitions, we run the following regression for each industry and for each year *t*:

$$(4) \qquad \frac{TA_{i,t}}{Asset_{i,t-1}} = \phi_1 \frac{1}{Asset_{i,t-1}} + \phi_2 \frac{(\Delta \mathrm{Re}v_{i,t} - \Delta AR_{i,t})}{Asset_{i,t-1}} + \phi_1 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t}$$

The variable $TA_{i,t}$ is the total accruals which are defined as equal to ($\Delta CA_{i,t} - \Delta CL_{i,t} - \Delta Cash_{i,t} + \Delta STDEBT_{i,t} - DEPN_{i,t}$), where $\Delta CA_{i,t}$ is firm *i*'s change in current assets (item 4) between year *t-1* and year *t*, $\Delta CL_{i,t}$ is firm *i*'s change in current liabilities (item 5) between year *t-1* and year *t*, $\Delta Cash_{i,t}$ is firm *i*'s change in cash (item 1) between year *t-1* and year *t*, $\Delta STDEBT_{i,t}$ is firm *i*'s change in short term debt (item 34) between year *t-1* and year, and $DEPN_{i,t}$ is firm *i*'s depreciation and amortization expense (item 14). $\Delta \mathrm{Re}v_{i,t}$ is firm *i*'s change in revenues (item 12) between year *t-1* and year *t*, $PPE_{i,t}$ is firm *i*'s gross value of property plant and equipment (item 7) in year *t* and $Assets_{i,t-1}$ is firm *i*'s total assets at the beginning of year *t*. The industry- and year-specific parameter estimates from equation (4) are used to obtain discretionary accruals:

$$(5) \qquad ACCRUALS_{i,t} = \frac{TA_{i,t}}{Assets_{i,t-1}} - NA_{i,t}$$

where $NA_{i,t} = \hat{\phi}_1 \frac{1}{Asset_{i,t-1}} + \hat{\phi}_2 \frac{(\Delta \mathrm{Re}v_{i,t} - \Delta AR_{i,t})}{Asset_{i,t-1}} + \hat{\phi}_1 \frac{PPE_{i,t}}{Asset_{i,t-1}} + \varepsilon_{i,t}$ is normal accruals for firm *i* in year *t* and $\Delta AR_{i,t}$ is firm *i's* change in accounts receivable (item 2) between

year *t-1* and year *t.* Following Francis *et al* (2004, 2005), we use the absolute value of this measure, |*ACCRUALS*|, in our empirical analyses.

PERINST represents the institutional ownership as a fraction of shares outstanding. Our institutional holdings data consists of end-of-year total institutional stock holdings for every publicly traded U.S. firm between 1980 and 1999. We obtain the data from Thompson Financial, which gathers the information from institutional 13F SEC filings. All institutions with holdings of $100 million or more under management are required to file. The filings are submitted quarterly and include institutional holdings in every U.S. firm, as long as the holdings are more than $200,000 or 10,000 shares. We combine the institutional holding data with the CRSP data, and if a firm in CRSP cannot be matched with the 13F data, we assume that the institutional holdings are zero.

We obtain data on analyst coverage, *ANALYST*, from the I/B/E/S Historical Summary files. For each stock *i* on CRSP/COMPUSTAT merged data, we measure the analyst coverage for firm *i* in any given year *t* as the number of analysts who provide fiscal-year-1-ahead earnings estimation for this particular firm. If no I/B/E/S value is available (i.e., CRSP is not matched with I/B/E/S data), we set coverage equal to zero. We use the logarithm of 1+*ANALYST*, *LANALYST*, in the analysis below.

PERINSIDER represents the fraction of shares outstanding held by company insiders. The Securities Exchange Act of 1934 requires all officers, directors, and holders of ten percent or more of the company's stock to report their trades within 10 days after the end of month in which the security trade took place.[9] For the purpose of our analysis, insiders are defined as officers and directors. While large shareholders are required to report their trades to the SEC, they are not directly involved in the day-to-day operations of the firm and less likely to have access to superior information. The insider data used in this analysis come from two different sources. For the period 1983-1986, the data are obtained from the U.S. Securities and Exchange Commission's (SEC) Ownership Reporting System (ORS) tapes. The 1987-1999 data are obtained from the Insider Filing Data provided by Thompson Financial. We compute yearly holdings data as the average

[9] Since the enactment of Sarbanes-Oxley in July 2002, insiders have been required to report their trades by the second business day following the trade.

of the holdings at the beginning and at the end of each year. If we cannot find a match with the CRSP data, we assume the insider holdings are zero.[10]

Finally, we consider another likely influence on the information risk of firms by looking at the role of industry effects. Industry effects can arise for a variety of reasons, such as correlated information events, industry norms or standards in accounting and disclosure practices, or the like. It seems likely that the level of information risk should differ across industries, just as other variables related to risk typically do. Using SIC codes, we sorted firms into industries annually from 1983 to 1999 based on the Fama-French 17-industry classification, and we use indicator variables for 16 industries in the analysis below. Our final sample has 35,722 firm-years of data.

C. *Summary statistics and correlations*

Table 1 provides summary statistics for the variables listed above. In Panel A, the accounting and market variables all have sensible means, with some of the variables exhibiting vary large divergences across the firms in our sample. The estimated PIN variable has a mean of 0.211 and a standard deviation of 0.076. The PINs range from 0.039 to 0.705, suggesting a wide diversity in the probability of information-based trading across the stocks in our sample. In Panel B of Table 1, we report the average number of stocks assigned to each industry every year, along with the average percentage of total market capitalization. As is apparent, industries differ in size, with fabricated products being the smallest industry (both in terms of the number of firms and in terms of market capitalization) and banks and other financials being the largest. Approximately 18% of our sample by market value is classified into the residual "Other" category.

The simple correlations in Table 2 provide some first insights into the relationship between PIN and these accounting and market variables. PIN is positively correlated with the percentage of Insider Holdings, with Accruals, and with Standard Deviation. These findings seem sensible, as firms with more insider holdings are more likely to have more asymmetric information. Similarly, the accounting literature has used accruals as a measure of asymmetric information, an interpretation consistent with our finding here.

[10] The analysis was also performed removing observations without matches for analysts, institutional holdings or insider holdings. The results are qualitatively identical to those reported below.

PIN and Size are strongly negatively correlated, a result also found in earlier work (see Easley et al [2002]). This is consistent with private information being a more important property of small firms. PIN is also negatively correlated with firm Age, Analyst Coverage, and Turnover. Each of these variables is positively correlated with SIZE, suggesting that older, larger firms have more analyst coverage and greater trading volume. Institutional holdings and PIN are also negatively correlated, while such holdings are positively correlated with Size and Turnover. Finally, PIN has a small negative correlation with Tobin's Q, with Growth, and with ROA.

4. Relation between PIN and Accounting and Market Variables

The analysis above suggests that the probability of information-based trading as measured by PIN exhibits sensible, albeit potentially complex relationships with accounting and market data typically used to measure firm characteristics and performance. In this section, we investigate these relationships more fully by finding the set of accounting and market variables that provide the best model for PINs. We use these variables to create a proxy for PIN based on accounting and market data and we then use this proxy in asset pricing. This research also allows us to describe which types of firms have high information risk. We do not have a theoretical model of how these accounting and market variables create information risk or how they may be jointly determined with PIN, so we do not claim causation. Instead, our goal is to discern the characteristics of firms which have greater private information risk.

A. Cross-Sectional Regressions

We first investigate the relationship between these firm characteristics and PIN using cross-sectional regressions over the sample period 1983-1999. In each year we have a PIN estimate for each firm in our sample, yielding 35,722 firm-years of data. We analyze this data using both pooled regressions and Fama-MacBeth (1973) regressions. The pooled regressions use OLS estimation over the entire firm-year sample, while the Fama-MacBeth approach uses yearly regressions, and then averages the estimated coefficients over the sample period. The estimating equation is given by:

$$PIN_t = b_0 + b_1 LSIZE_t + b_2 GROWTH_t + b_3 LAGE_t + b_4 LANALYST_t + b_5 LTURNOVER_t + b_6 PERINSIDER_t + b_7 PERINST_t + b_8 ACCRUALS_t + b_9 ROA_t + b_{10} STDEV_t + b_{11} TOBIN_t + \sum_{i=1}^{16} I_{(i=d)} c_d + \eta_t$$

$I_{(i=d)}$ is an indicator variable equal to one if $i = d$, where $d = 1,\ldots,16$, corresponding to each of the 16 industries, and zero otherwise. Thus, the estimates measure the incremental effect of an industry relative to the "Other" industry.

Table 3 presents the results of these cross-sectional regressions. All explanatory variables in the regressions are normalized by their cross-sectional means and standard deviations each year. This allows one to interpret the coefficients as the marginal effect on PIN of a one standard deviation move in the explanatory variable. As the results from the two approaches are virtually identical, we focus our discussion on the Fama-MacBeth estimates. The data show that most, but not all, of our explanatory variables have a statistically significant effect on PIN. Size and AGE are both strongly negatively related to PIN, consistent with larger, older firms having less private information. Analysts also has a strong, negative relation, suggesting that firms with greater analyst following have lower information risk. Earlier work (see Easley et al [1998]) suggests that analysts may serve to turn private information into public information, a result consistent with our finding here. Additionally, analysts may attract additional uninformed order flow to a stock, an effect that would also reduce the information risk of a stock. Turnover is also negatively related to PIN, again consistent with the notion that stocks with greater trading activity tend to have more uninformed order flow.

The percentage of Insider Holding continues to be positively related to PIN, and now, too, so is the percentage of Institutional Holdings. The issue of who are informed traders is a question of perennial interest in microstructure, and our results here accord well with the general view that it is insiders and institutional traders who are more likely to be acting on private information. Profitable firms (as captured by ROA) and high growth firms also have higher PINs, consistent with informed traders seeking out such firms due to their high potential for returns to information-based trading. A similar

explanation may attach to the positive relation between Standard Deviation and PIN, as firms with greater volatility present greater profit opportunities for informed traders. As was found in the simple correlations, Tobin's Q is negatively related to PINs. Accruals, however, exhibit no statistically significant effect, a somewhat surprising result given the extensive use of this variable in accounting as a proxy for asymmetric information.

Because PIN variables measure the risk of information-based trading, we would expect stable, less dynamic industries to exhibit lower PINs, and more volatile industries to have the opposite result. This is exactly what we find. Utilities exhibit a very strong, negative relation with PIN, consistent with there being little asymmetric information in this industry. Conversely, strong positive relations with PIN are connected with firms in Oil and Petroleum Products, Construction, Textiles, and Retail. A number of basic industries, Autos, Food, Chemicals, and Steel, have PINs which are not significantly different from those of the baseline "Other" category.

The results in Table 3 are consistent with high information risk firms being younger, smaller, less covered by analysts, higher profit, growth, volatility firms, and firms belonging to volatile industries. In addition, such high information risk firms are more widely held by insiders and institutions. While the data indicate a role for all of these factors, the most significant influence is that of Size. Smaller firms have higher PINs, suggesting that the information structure of smaller firms may differ in important ways from that of larger firms.

B. The Role of Size

To investigate these size effects more fully, we ran our estimating equation separately for the smallest 50% of firms and for the largest 50% of firms. The results are given in Table 4. Looking first at the smallest firms, Panel A shows that Size remains very significant and negative, but less so than before. ROA, or profitability, also continues to have a significant, positive effect on PINs. Age, however, now has an insignificant role, as does Standard Deviation and Tobin's Q. Accruals, also, are not an important component of information risk for smaller firms.

Both institutional and insider holdings have significant positive influences on small firm PINs, but this effect is not found with large firms. As Panel B indicates, for

large firms insider holdings remain significantly positive, but this effect is greatly attenuated for institutional holdings, where it is now only marginally significant. These findings suggest that institutions play very different roles with respect to small and large firms. Conversely, Standard Deviation has a strong, positive effect on PIN for large firms, in contrast with an insignificant role for small firms, while ROA matters for small firms, but not for large ones. The negative effect of Size is even greater for large firms, further amplifying our result that information-based trading is a greater risk for small firms. Again, we find that accruals are not significantly related to PINs for either large or small firms.

These results demonstrate that the risk of information-based trading as captured by PIN has a well-defined relation with firm-specific accounting and market data. That this relationship differs between large and small firms is consistent with a greater fraction of available information being public for large firms, and greater fraction of information being private for small firms. This difference in information composition is predicted to matter for asset returns, an issue we address later in the paper.

C. Time-period Effects

The dramatic changes in both the market and the economy over our sample period lead to a natural concern that the information environment surrounding firms might be affected, resulting in temporal instability of our estimates. To address these concerns, we ran our estimating equation over the sub-periods 1983–1990 and 1991–1999. The results are given in Table 5.

Most of the estimated relationships are stable across both sub-periods, although there are a number of exceptions. Age, for example, is significant only in the later period, a pattern also exhibited by the Growth variable. The coefficients suggest that younger, faster growing firms had higher information risk in our later period, results that are consistent with the aberrant behavior typifying the tech boom. Interestingly, while both insider and institutional holdings retain consistent signs, insider effects are stronger in the latter period while institutions are more significant in the earlier time period. Tobin's q also only is significant in the early period, where it has a negative sign. Size,

turnover, analysts, ROA and standard deviation exhibit consistent behavior over both time intervals.

The behavior of the Accruals variable is particularly intriguing. In our overall sample results, Accruals was not significant for either large or small firms. Segmenting by time, however, reveals a very different story. In the early period, Accruals has a weakly significant negative relation with the PIN measure of information risk. This changes over time, however, so that in the later period Accruals has a stronger, now positive relationship with PIN. Accruals have been used in the accounting literature as a proxy for asymmetric information, a role consistent with our estimates in the later period. As accounting practices do change over time, these findings suggest that the behavior and information content of accruals in the 1990's was very different than it was in earlier times.

D. Summary – What types of firms have high information risk?

Overall, our analysis provides a clear profile of the types of firms with greater information risk. Smaller, younger firms are more likely to have higher information risk, as are faster growing and more profitable firms. Firms with more insider holdings and/ or greater institutional holdings are also subject to greater information-based trading, consistent with the general view that insiders and institutions are more likely to be informed traders. Firms followed by few analysts also generally have high information risk, as do firms with lower Tobin's q. Information risk is also higher in firms with low turnover and in firms with high abnormal accruals in the 1991-1999 period. Finally, Oil and Petroleum Products, Construction, Textiles, and Retail firms are also more likely to have higher information risk as captured by the PIN variable.

As noted earlier, one goal of this research is to relate the trade-based measures of information-based trading from microstructure analyses to the more accounting-based measures often used to describe a firm's information and economic environment. Our analysis above shows that we can successfully characterize what types of firms have higher information risk. We now turn to a second goal of our analysis which is investigating how this information risk affects asset pricing.

5. Asset Pricing and Information Risk

Asset pricing analyses generally rely on long sample periods to investigate the influence of various factors on cross-sectional returns. As noted earlier, a limitation of the PIN measure is that the trade-based data needed to construct PIN is not available prior to 1983. Thus, microstructure analyses of the role of information risk in asset pricing have been limited to only relatively short sample periods starting in 1983, as opposed to studies such as Fama and French which use data onwards from 1965. In this section, we address this problem by forming an accounting and market statistics based proxy for a firm's information risk. Our particular analysis focuses on three questions: First, can we construct an instrument for PIN that has explanatory power for the cross-section of returns? Second, can we use that instrument to extend our sample period and so investigate the effects of information risk over the time periods typically used in asset pricing studies? And, third, can we determine if information risk vitiates the influence of other variables on asset returns?

A. Creating PPINs

In principle, the regression analysis of the previous section provides a template for creating such an economic proxy for the PIN variable. Our analysis requires both going forward in time to create a PIN proxy for the years 2000-2004, and backwards to create PPINs for the years 1965- 1982. Creating PPINs going forward is straightforward, but a practical problem arises in going backwards because some of the data series in our regressions (most notably, analysts following and percentage insider and institutional holdings) are not available in earlier time periods.[11] To address this concern, we calculate our PIN proxy (PPIN) using a subset of variables with continuous data series. These continuous time series, and their associated coefficient in the PIN-regression, are given in Table 6 for the sample period 1983 - 1999. As is apparent, the coefficients on this restricted set of variables are consistent with our earlier results, and the relatively high R^2 suggests that PIN variables can be accurately described by a combination of accounting and market variables.

[11] Note, however, that this is not an issue if we are interested in estimated PPINs going forward in time. Thus, for studies implementing PPINs on more recent data we recommend using the full data set of variables.

We now use these coefficients, and the values of the independent variables, to create a proxy for PIN, denoted PPIN, over the time period 1965-2004. Figure 2 provides a comparison of the distributions of PPIN and PIN over the sample period 1983 - 1999. As is apparent, the PPIN distribution is less skewed than the PIN distribution, and it has a smaller standard deviation ($\sigma_{PIN} = 0.076$, $\sigma_{PPIN} = 0.045$). This is largely due to there being fewer outliers in the estimated PPINs, a not unexpected result given that PPIN is the prediction of PIN from our regression. The means of the two distributions are not significantly different, however, with $\mu_{PIN} = 0.211$ and $\mu_{PPIN} = 0.208$.

Figure 3 shows the time-series distribution of the PPINs over the entire sample period 1965-2004. The PPINs are remarkably stable, with mild variability in the distributions occurring only at the very beginning and the very end of the sample period. The mean PPIN, captured by the *p50* line, is virtually constant over the sample period. Overall, the estimated PPINs appear to be well-defined and stable, but what is not yet clear is whether the PPINs actually capture the economic properties of the PIN variables.

B. In-sample Asset pricing

We assess the economic efficiency of the PPINs by investigating how well the PPINs perform in explaining cross-sectional asset returns. We use a standard Fama-Macbeth methodology and include as explanatory variables Beta, LSIZE, and Book-to-Market (BM) similar to Fama and French (1992).[12] We also augment these variables with a momentum measure, RET12, based on the previous 12-month return in the stock. These results are reported in Table 7.

To form a baseline, we report in Panel A on Table 7 cross-sectional asset pricing results using our PIN variable. These results are similar to those first reported in Easley, Hvidkjaer and O'Hara (2002), and they show that the PIN variable has a positive and statistically significant effect on asset returns. This positive role is consistent with investors demanding a higher return to hold stocks subject to greater information risk. Note that over this time period, Book-to-Market is significant and positive, Size is

[12] The construction of the explanatory variables in the asset pricing regressions is outlined in the Appendix.

positive but only marginally significant, and Beta is not significant.[13] The positive sign on Size is the opposite of that predicted by Fama and French (1993), who argued that small firms, not large firms, should command higher returns. The insignificant coefficient on Beta, while inconsistent with standard asset-pricing theory, is consistent with the findings of Fama and French (1992), Chalmers and Kadlec (1998) and Datar, Naik and Radcliffe (1998) who investigate similar sample periods. These authors find, as we do, that market risk is not statistically significant over this period. The Book-to-Market result is similar to that found by other researchers, most notably Fama and French.

Panel A also reports the results when a momentum measure is included in the asset pricing regressions. Momentum has posed a challenge to many asset pricing models, and a natural concern is that our results may somehow be due to momentum instead of information risk. The results show that this is not the case. PIN remains positive and economically significant, albeit with a slightly smaller coefficient. A similar effect is found on Book-to-Market. However, including momentum does vitiate the statistical significance of Size, while Beta continues to be not significant.

Having reviewed the asset pricing influence of PIN, we now consider these cross-sectional asset pricing results using PPINs. For comparison purpose, we present our results based on PPINs estimated using the full set of variables (Panel B) and the continuous set of variables (Panel C). Using the full set of variables, Panel B shows that the PPINs appear to be remarkably robust, exhibiting significant positive coefficients just as we found with the PIN variables. Including momentum has little effect on the PPINs, suggesting robustness both to our PIN proxy and to the influence of information risk on asset pricing. Panel C presents the cross-sectional results when PPINs are calculated based on the continuous variable set. The results here are stronger still, with the coefficient on the PPINs again positive and strongly significant in either specification.

The coefficient on PPIN in Panels B and C of Table 7 is significantly larger that the coefficient on PIN in Panel A. This may be due to the reduced volatility of PPIN as well

[13] Easley, Hvidkjaer and O'Hara (2002) found no significant effect of book-to-market. In the current sample, book-to-market also becomes insignificant once we exclude Amex firms, which were not included in the Easley, Hvidkjaer and O'Hara (2002) sample. This is consistent with Loughran (1997), who finds that the book-to-market effect is only present in small firms.

as to the high correlation of PPIN with Size. Note also that using PPINs instead of PINs actually strengthens the statistical significance of the Size variable in the results reported in both Panels B and C. This effect also may be due to the reduced volatility of PPIN relative to PIN. Because the composite variable PPIN is estimated from a variety of information-linked variables including Size, there is also the possible influence of multi-collinearity. We address this concern more fully later in this section.

We interpret these in-sample results as strong evidence that the PINs and PPINs capture the same fundamental economic influences on asset prices. The success of our PPIN proxy now allows us to investigate these information risk effects on asset pricing over a longer sample period. Such long-sample period tests provide both a more rigorous test of our hypothesis that information risk matters in asset pricing, and allows comparison of our results to those of more standard asset pricing models.

C. Information risk and asset pricing: Extended sample period results

Our hypothesis is that information risk affects asset returns because given the total amount of information available about an asset, uninformed investors require compensation to hold assets in which there is more trade based on private information. We test our hypothesis by seeing whether PPIN, our estimated proxy for information risk, has a significant and positive effect on asset returns. To capture the variety of influences that are argued to affect asset returns in the literature, we present results from three pairs of cross-sectional asset pricing specifications.

As a useful preliminary, Figure 4 traces the coefficients over time through plots of the annual cross-sectional regression coefficients for PPIN, LSIZE, Beta Ret12 and Book-to-Market. All series show some variability, but in general are consistent across the sample period. The coefficient for the PPIN variable tends to vary across years, with strong positive values attaching to our information risk variable for most of the period.

Table 8 presents results for cross-sectional asset pricing regressions over the sample period 1965-2004. We first note that across all specifications tested in Table 8 the coefficient on PPIN is both positive and statistically significant. Thus, whether in combination with β, LSIZE, and Book-to-market, or adding Momentum, or adding

Dollar-volume, PPINs exhibit exactly the behavior predicted by our information risk theory.

Interestingly, over this long sample period, β does not exhibit the expected positive sign, nor is it statistically significant in any specification. This puzzling result is not unique to our study, but it does raise concerns about traditional asset pricing models in which only systematic risk affects asset returns. A popular alternative framework is suggested by Fama-French (1992) who find that asset returns are influenced by β, Size, and Book-to-Market. We provide results on these variables in Specification I.1, with Specification II.1 augmenting the Fama-French variables with Momentum. The results show the while β is not significant, we do find the predicted negative sign on LSIZE and positive sign on BM, with both results statistically significant. Momentum appears to have little effect on these results.

Including PPIN into the FF-three factor specification changes these results, particularly as they relate to the LSIZE variable. PPIN is positive and statistically significant in combination with these variables. Comparing Specification I.1 (β, LSIZE, BM) to Specification I.2 (β, LSIZE, BM, PPIN), we find that including PPIN reduces the significance of all the three FF variables, and changes the sign of the LSIZE and β coefficients. Including PPIN into the four-factor specification (the FF-three factor model plus momentum as measured by the previous year's return) yields similar results. Most important for us is that PPIN remains positive and significant. These results are presented in specifications II.1 and II.2 of Table 8.

In previous research using a shorter time period we found similar effects of including PIN in various asset pricing specifications.[14] Such a sign reversal of the Size variable has been found by other authors, most notably Brennan, Chordia, and Subrahmanyam (BCS) (1998), who show that the inclusion of dollar volume (DVOL) changes the sign of the Size variable in asset pricing regressions. BCS also find that DVOL has significant explanatory power for asset returns, with high dollar volume stock having low future returns. What accounts for this relationship is unclear, but BCS argue that one explanation is that DVOL is a proxy for liquidity. A natural concern is whether our PPIN is capturing a similar effect, and hence not related to information risk as we conjecture.

[14] Several of these results are reviewed in Panel A of Table 7.

To test for these effects, we included DVOL in our asset pricing regressions. We report in specifications III.1 and III.2 of Table 8 the effect of including DVOL in the four-factor model and the four-factor-plus-PPIN model. Several results are noteworthy. First, we find that PPIN is robust with respect to the inclusion of DVOL, suggesting that PPIN is not acting as a proxy for liquidity effects. Second, we find that without controlling for PPIN, LSIZE and DVOL are not significantly related to returns.[15] Finally, we find that when PPIN is included LSIZE and DVOL have significantly positive effects on returns.

These findings show that information risk, as captured by PPIN clearly matters for asset returns, but that other variables, most notably Size, continue to exert an influence as well. These other variables also appear to have complex interactions both with PPIN and each other, an issue we now consider in more detail.

D. Multi-collinearity analysis

We first address the possible impact of multicollinearity between PPIN and LSIZE. Although pair-wise collinearity can be determined from viewing a correlation matrix of the independent variables, correlation matrices will not reveal higher-order multicollinearity. Several classical tests exist for diagnosing multicollinearity problems, but we focus on the most common test, namely the variance-inflation factor. The variance-inflation factor (*VIF*) is defined as:

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is obtained from regressing explanatory variable x_j on the other explanatory variables in the regression model. The *VIF* measures the factor by which the parameter's variance is multiplied relative to a regression without multicollinearity. If x_j is highly correlated with the other variables, then R_j^2 and in turn the *VIF* will be large. This inflates the variance of the parameter in the original regression, making it difficult to obtain a significant *t*-ratio. As a rule of thumb, *VIF* values greater than 10 may merit further

[15] If we remove SIZE from III.1 we find that DVOL is significantly negatively related to returns. Removing SIZE from III.2 has little effect on the coefficients of PPIN or DVOL.

investigation. In our tests *VIF* values turned out to be less than 10. Specifically, the *VIF* statistics for Beta, BM, LSIZE and PPIN variables are 1.09, 1.19, 6.29 and 6.82, respectively.

E. PPIN portfolio returns

An alternative approach to determining whether PPINs are related to returns is to sort stocks into portfolios based on PPINs and ask how returns differ across these portfolios. This approach has the advantage that it does not require the specification of a regression equation and that it reduces the idiosyncratic noise from the firm level. In this section we form portfolios by sorting stocks into deciles each year based on their PPINs in the previous year. Within each decile, stocks are value weighted to create portfolio returns. The raw returns on these PPIN portfolios are reported in row 1 of Table 9. Unadjusted raw returns to PPIN portfolios increase monotonically, from 0.92% for the lowest PPIN portfolio to 2.32% for the highest PPIN portfolio, leading an economically and statistically significant monthly excess return of 1.41 percent to the zero-investment portfolio (i.e., portfolio 10-1).

Of course, returns to the PPIN portfolios may be correlated with other risk factors. Therefore, to control for the possibility that this return is due to holding some other risk, we adjust returns following the method of Daniel, Grinblatt, Titman and Wermers (1997), henceforth DGTW. This method simultaneously controls for the effects of size, book-to-market, and return momentum. To construct DGTW benchmark portfolios, we first rank all stocks based on their market capitalization and assign them to size quintiles using NYSE size quintile breakpoints. Within each size quintile, we rank stocks based on their book-to-market ratios, and assign them to book-to-market quintiles, yielding a total of 25 size- and book-to-market sorted portfolios. We further sort stocks in each of the 25 portfolios into quintiles, based on the prior 12-month return of each stock. Each stock is now uniquely identified with one of the resulting 125 portfolios, and the size-, book-to-market-, and momentum-adjusted return for stock *i* in month *t* is then its return minus the value-weighted mean return of the other stocks in the portfolio. The characteristic-adjusted return to our zero-investment portfolio is still large at 0.72% per month.

To further understand the complex joint effects of firm size and information on returns, we also conduct dual sorts by size and PPIN. In addition to controlling for the effect of size, this approach reveals any interaction effects between the two variables. We first sort stocks into five quintiles based on their market capitalization. Then within each quintile, we sort stocks into five quintiles based on their PPIN measure. These portfolios are rebalanced every month. Table 10 reports average percentage monthly returns for each Size and PPIN group. The information effect is clearly mostly pronounced for small firms; within every size quintile except the largest the portfolio with the highest PPIN has the highest returns. The results also indicate a large spread in average returns between the highest and lowest PPIN portfolios. The monthly return to the 5-1 portfolio varies from 1.82% for small stocks to 0.25% for large stocks. All of these returns are significantly different from zero. These results reinforce our previous results about the effect of PPIN on asset pricing. PPIN is priced and the effect is both statistically and economically significant.

6. Conclusion

The Probability of Information-based Trade variable (PIN) is created from trade data using our structural model relating observable buys and sells to the unobservable variables that make up PIN. One contribution of this paper is to provide an independent verification of our interpretation of PIN. In particular, we show that the correlations between PIN and other variables that have been used as proxies for a firm's information environment are consistent with our interpretation. Smaller, younger firms, firms with more insider holdings, firms with greater institutional holdings, firms followed by few analysts and firms in Oil and Petroleum Products, Construction, Textiles, and Retail are all more likely to have higher information risk.

Because estimating PIN is difficult with modern data, and impossible before trade data was first collected in 1983, having a good proxy for PIN for the period before 1983 and for most recent years would be valuable. Our collection of accounting, firm characteristic and industry characteristic together explain nearly one-half of the variation in PIN across firms. So we create a proxy for PIN by regressing PIN on these characteristics. Using the coefficients from this regression, and firm and industry

characteristics, we create a long time series of PPINs. The real test of whether PPIN is actually a good proxy for PIN is how effective it is in explaining asset prices. We show that in fact PPIN is at least as successful as PIN in our withinsample asset pricing regressions.

Since we have a good proxy for PIN, we ask how it performs in long time series asset pricing regressions in which we include periods both before and after the period used to construct our proxy. Our main conclusion is that our proxy for information-based-trade, PPIN, performs as expected in asset pricing. Firms with higher PPINs have higher returns, and this conclusion is robust to every asset pricing structure that we have explored. The only other variable that performs consistently well in our asset pricing investigation is Book-to-Market, all of the other usual variables (beta, size, momentum, and dollar volume) either are insignificant or have an unexpected sign in the presence of PPIN. We take this to be strong evidence that PPIN provides a valuable input into asset pricing.

We believe that our PPIN results provide part of the explanation for why some of the firm and industry characteristics that other researchers have found useful seem to work in asset pricing regressions. Our view is that those variables helped to explain asset prices in part because they were standing in for the information effects that are captured more directly by PIN. Some of these variables may also play a role that is separate from PIN; in particular they could proxy for the total amount of information produced about a firm, so it is not necessarily the case that their effectiveness will vanish once PIN is considered. Such an explanation seems particularly relevant for the Size variable. Starting with Barry and Brown (1984), researchers have conjectured that Size is an information effect, but our results here suggest that Size is not just a differential information effect.

That a range of variables can influence asset prices is consistent with our view that asset prices are influenced by a variety of imperfections not well captured in the standard asset pricing approach. Asymmetric information is one such imperfection, but other frictions such as financing constraints, bankruptcy concerns, and even differences arising from corporate governance structures may differ in systematic ways across firms.

Incorporating such frictions into asset pricing analyses seems a particularly fruitful area for future research.

REFERENCES

Admati, Anat, 1985, "A noisy rational expectations equilibrium for multi-asset securities markets", *Econometrica*, 53, 629–658.

Amihud, Y., and H. Mendelson, 1986, "Asset pricing and the bid-ask spread", *Journal of Financial Economics*, 17, 223-249.

Amihud, Y., and H. Mendelson, 1989, "The effects of beta, bid-ask spread, residual Risk, and size on stock returns", *Journal of Finance*, 44, 479-486.

Barry, C. and S. J. Brown, 1984, "Differential information and the small firm effect", *Journal of Financial Economics*, 1, 283-294.

Botosan, C. A., 1997, "Disclosure level and the cost of equity capital", *The Accounting Review*, 72, 323-349.

Botosan, C.A. and M.A. Plumlee, 2003, "Are information attributes prices?", Working Paper, Eccles School of Business.

Botosan, C.A., Plumlee, M.A. and Y. Xie, 2004, "The role of information precision in determining the cost of equity capital", *Review of Accounting Studies*, June/September.

Brennan, M., Chordia, T. and A. Subrahmanyam, 1998, "Alternative factor specifications, security characteristics, and the cross-section of expected returns", *Journal of Financial Economics*, 49, 345-373.

Carhart, M., 1997, "On persistence in mutual fund performance", *Journal of Finance*, 52, 57-88.

Chalmers, J. M., and G. B. Kadlec, 1998, "An empirical examination of the amortized spread", *Journal of Financial Economics*, 48, 159-188.

Cohen, D. A., 2005, "Quality of financial reporting choices: Determinants and economic consequences", NYU Stern School Working Paper.

Chordia, T., A. Subrahmanyam and V. R. Anshuman, 2001, "Trading activity and expected stock returns", *Journal of Financial Economics*, 59, 3-32.

Daniel, K., Grinblatt, M., Titman, S., and R. Wermers, 1997, "Measuring mutual fund performance with characteristic-based benchmarks", *Journal of Finance*, 52, 1035-58.

Datar, V., N. Naik, and R. Radcliffe, 1998, "Liquidity and stock returns: An alternative test", *Journal of Financial Markets*, 1, 203-219.

Diamond, D. W. and R. Verrecchia, 1991, “Disclosure, liquidity, and the cost of capital”, *Journal of Finance*, 46, 1325-1359.

Dimson, E., 1979, “Risk measurement when shares are subject to infrequent trading”, *Journal of Financial Economics*, 7, 197-226.

Duarte, J. and L. Young, 2007, “Why is PIN Priced?” *Journal of Financial Economics, forthcoming.*

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

Easley, D., S. Hvidkjaer and M. O'Hara, 2008, "Factoring information into returns", *Journal of Financial and Quantitative Analysis*, forthcoming.

Easley, D., Kiefer, N. and M. O’Hara, 1997a, “One day in the life of a very common stock”, *Review of Financial Studies,* 10, 805-835.

Easley, D., Kiefer, N. and M. O’Hara, 1997b, “The information content of the trading process”, *Journal of Empirical Finance*, 4, 159-186.

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

Easley, D., M. O'Hara, and J. Paperman, 1998, “Financial analysts and information-based trade”, *Journal of Financial Markets*, 1, 175-201

Fama, E. F., and K. R. French, 1992, “The cross-section of expected stock returns”, *Journal of Finance*, 47, 427-465.

Fama, E. F., and K. R. French, 1993, “Common risk factors in the return on stocks and bonds”, *Journal of Financial Economics*, 33, 3-56.

Fama, E.F. and K.R. French, 1997, “Industry costs of equity”, *Journal of Financial Economics*, 93, 153-194.

Fama, Eugene F., and James D. MacBeth, 1973, “Risk, return, and equilibrium: Empirical tests”, *Journal of Political Economy*, 81, 607–636.

Francis, J., R. LaFond, P. Olsson and K. Schipper, 2005, “The market pricing of accruals quality”, *Journal of Accounting and Economics*, 39, 295-327.

Francis, J., R. LaFond, P. Olsson and K. Schipper, 2004, "Costs of equity and earnings attributes", *Accounting Review*, 79, 967-1010.

Grossman, S. and J. Stiglitz, 1980, “On the impossibility of informationally efficient markets”, *American Economic Review*, 70, 393-408.

Grundy, B. and S. Martin, 2001, "Understanding the nature of the risk and the source of the rewards to momentum investing", *Review of Financial Studies*, 14, 29-78.

Jegadeesh, N. and S. Titman, 1993, "Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency," *Journal of Finance, 48*, 65-91.

Healy, P. M., and Palepu, K. G., 2001, "Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature," *Journal of Accounting and Economics*, 31, 405-440.

Jones, J., 1991, "Earnings management during import relief investigations", *Journal of Accounting Research*, 29, 193–228.

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

Li, H., Wang, J., Wu, C., and Y. He, 2008, "Are Liquidity and Information Risks Priced in the Treasury Bond Market?" *Journal of Finance, forthcoming.*

Loughran, T., 1997, "Book-to-market across firm size, exchange, and seasonality: Is there an effect?", *Journal of Financial and Quantitative Analysis*, 32(3), 249-268.

Verrecchia, R. E., 2001, "Essays on disclosure," *Journal of Accounting and Economics*, 32, 97-180.

Wang, Jiang, 1993, "A model of intertemporal asset prices under asymmetric information", *Review of Economic Studies*, 60, 249-287.

Appendix: Construction of variables in the asset pricing regressions

Because PIN is estimated over the calendar year, we update all yearly variables in January for the asset pricing tests. PIN estimated in year *t*-1 is used in the asset pricing regressions for year *t*. SIZE is the market value of equity at the end of year *t*-1. Book-to-Market ratios are computed in a manner similar to Fama and French (1992), except that we use book values and market values as of the end of June in year *t*-1.

We calculate betas using the following approach. Pre-ranking betas are estimated for individual stocks using monthly returns from at least two years to, when possible, five years, before the test year. Thus, for each stock we use at least 24 monthly return observations in the estimation. We regress these stock returns on the contemporaneous and lagged value-weighted CRSP NYSE/Amex index. Pre-ranking betas are then given as the sum of the two coefficients (this approach, suggested by Dimson (1979), is intended to correct for biases arising from non-synchronous trading). Next, 40 portfolios are sorted every January on the basis of the estimated betas, and monthly portfolio returns are calculated as value-weighted averages of individual stock returns. Post-ranking portfolio betas are estimated from the full sample period, such that one beta estimate is obtained for each of the 40 portfolios. Portfolio returns are regressed on contemporaneous and lagged values of CRSP index returns. The portfolio beta is then the sum of the two coefficients. We use individual stocks in the cross-sectional regressions, so individual stock betas are taken as the beta of the portfolio to which they belong. Because the portfolio compositions change each year, individual stock betas vary over time.

Table 1
Summary Statistics

The table contains time series means of cross sectional statistics computed each year in 1983-1999. In Panel A, *PIN* is the probability of information-based trading in stock *i* of year *t*. *SIZE* is the market value of equity in firm *i* at the end of year *t* in billions, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *AGE* represents the number of years since the stock *i* was first covered by CRSP. *TURNOVER* represents the average monthly turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. *PERINST* represents the institutional ownership as a percentage of shares outstanding, *PERINSIDER* represents the percentage of shares held by company insiders, and *ANALYST* is the number of analysts following stock *i* in year *t*. In Panel B, industries are formed annually from 1983 through 1999 based on the Fama-French 17-industry classification. The average number of stocks assigned to each industry every year is reported, along with the average percentage of total market capitalization.

Panel A: Continuous variables

	Mean	Std Dev	Min.	Q1	Median	Q3	Max.
PIN	0.211	0.076	0.039	0.158	0.199	0.249	0.705
SIZE	1.612	5.081	0.001	0.063	0.273	1.127	94.245
GROWTH	0.112	0.313	-0.681	-0.016	0.072	0.181	1.771
AGE	20.841	16.933	0.000	8.176	17.382	26.765	66.000
ANALYST	7.237	8.347	0.000	1.000	3.824	11.471	41.588
TURNOVER	0.681	0.623	0.011	0.304	0.534	0.865	10.125
PERINSIDER	0.066	0.117	0.000	0.002	0.014	0.076	0.897
PERINST	0.375	0.240	0.000	0.166	0.363	0.564	0.998
ACCRUALS	0.056	0.048	0.000	0.017	0.041	0.085	0.160
ROA	0.069	0.106	-0.445	0.031	0.078	0.121	0.319
STDEV	0.402	0.268	0.034	0.248	0.340	0.480	4.316
TOBIN	1.324	0.437	0.671	1.002	1.173	1.554	2.227

Panel B: Industries

	Industry	*Avg. No of Stocks*	*Avg. % of Market Cap.*
1	Food	70	5.24
2	Mining and Minerals	37	0.94
3	Oil and Petroleum Products	114	8.19
4	Textiles, Apparel & Footwear	76	0.80
5	Consumer Durables	81	3.86
6	Chemicals	48	3.40
7	Drugs, Soap, Perfumes, Tobacco	68	9.57
8	Construction and Constr. Materials	112	2.55
9	Steel Works Etc	50	1.06
10	Fabricated Products	36	0.58
11	Machinery and Business Equipment	248	8.56
12	Automobiles	39	2.95
13	Transportation	72	2.97
14	Utilities	147	7.81
15	Retail Stores	126	5.74
16	Banks, Insurance Cos, and Other Financials	324	17.84
17	Other	453	17.97

Table 2
Spearman Correlations

The table shows the time-series means of Spearman rank correlations computed each year 1983-1999. *PIN* is the probability of information-based trading in stock *i* of year *t*. *SIZE* is the market value of equity in firm *i* at the end of year *t* in millions, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *AGE* represents the number of years since the stock *i* was first covered by CRSP. *TURNOVER* represents the average monthly turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. *PERINST* represents the institutional ownership as a percentage of shares outstanding, *PERINSIDER* represents the percentage of shares held by company insiders, and *ANALYST* is the number of analysts following stock *i* in year *t*.

	SIZE	*GROWTH*	*AGE*	*ANALYST*	*TURN-OVER*	*PER-INSIDER*	*PERINST*	*ACCRUALS*	*ROA*	*STDEV*	*TOBIN*
PIN	-0.642	-0.020	-0.266	-0.620	-0.300	0.119	-0.408	0.192	-0.179	0.306	-0.127
SIZE		0.107	0.331	0.874	0.333	-0.120	0.649	-0.276	0.331	-0.464	0.214
GROWTH			-0.149	0.082	0.124	0.141	0.112	0.065	0.286	0.006	0.160
AGE				0.284	0.045	-0.114	0.244	-0.199	0.072	-0.317	-0.076
ANALYST					0.425	-0.116	0.658	-0.248	0.262	-0.374	0.137
TURNOVER						-0.020	0.485	0.044	0.101	0.222	0.146
PERINSIDER							-0.032	0.076	0.092	0.079	0.117
PERINST								-0.151	0.294	-0.282	0.098
ACCRUALS									-0.119	0.303	0.107
ROA										-0.234	0.414
STDEV											0.099

Table 3
Cross-Sectional Regressions

The dependent variable is the probability of information-based trading, *PIN*, in stock *i* of year *t* from 1983 through 1999. *LSIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *LAGE* represents the logarithm of 1+the number of years since the stock *i* was first covered by CRSP. *LTURNOVER* represents the logarithm of share turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. *PERINST* represents the institutional ownership as a percentage of shares outstanding, *PERINSIDER* represents the percentage of shares held by company insiders, and *LANALYST* is the logarithm of 1+number of analysts following stock *i* in year *t*. Industry indicator variables are formed annually based on the Fama-French 17-industry classification. All continuous independent variables are normalized by their yearly cross-sectional means and standard deviations.

	Fama-MacBeth Estimates		***Pooled Reg. Estimates***	
	Parameter Estimates	***t-value***	***Parameter Estimates***	***t-value***
INTERCEPT	20.88	51.38	18.05	125.07
LSIZE	-3.51	-16.23	-3.46	-45.19
GROWTH	0.17	2.88	0.20	5.68
LAGE	-0.13	-2.34	-0.07	-1.79
LANALYST	-0.85	-12.70	-0.90	-12.41
LTURNOVER	-1.57	-14.22	-1.54	-34.48
PERINSIDER	0.15	3.54	0.19	5.54
PERINST	0.52	5.29	0.51	10.1
ACCRUALS	0.03	0.73	0.05	1.25
ROA	0.29	5.45	0.32	8.5
STDEV	0.37	3.52	0.38	7.98
TOBIN	-0.19	-4.99	-0.22	-5.79
Industry				
Food	0.06	0.35	0.01	0.05
Mining and Minerals	-0.08	-0.29	0.00	-0.02
Oil and Petroleum Products	0.80	3.02	0.68	4.50
Textiles, Apparel & Footwear	0.63	3.06	0.52	2.97
Consumer Durables	0.38	1.79	0.27	1.56
Chemicals	0.05	0.29	0.10	0.48
Drugs, Soap, Perfumes, Tobacco	0.05	0.22	0.07	0.39
Construction and Constr. Materials	0.45	3.22	0.33	2.08
Steel Works Etc	-0.09	-0.37	-0.13	-0.64
Fabricated Products	0.04	0.13	-0.17	-0.71
Machinery and Business Equipment	0.07	0.45	0.01	0.08
Automobiles	0.36	1.56	0.17	0.72
Transportation	0.20	1.35	0.25	1.38
Utilities	-1.65	-7.50	-1.69	-10.85
Retail Stores	0.57	4.39	0.48	3.42
Banks, Insurance Companies, and Other Financials	-0.22	-1.45	-0.19	-1.13
Adjusted R^2	0.46		0.47	

Table 4
Cross-Sectional Regressions by Firm Size

The dependent variable is the probability of information-based trading, *PIN*, in stock *i* of year *t* from 1983 through 1999. *LSIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *LAGE* represents the logarithm of 1+the number of years since the stock *i* was first covered by CRSP. *LTURNOVER* represents the logarithm of share turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. *PERINST* represents the institutional ownership as a percentage of shares outstanding, *PERINSIDER* represents the percentage of shares held by company insiders, and *LANALYST* is the logarithm of 1+number of analysts following stock *i* in year *t*. All independent variables are normalized by their yearly cross-sectional means and standard deviations. Results for industry indicator variables are suppressed for shortness of exposition.

Panel A: 50% smallest firms by market value

	Fama-MacBeth estimates		*Pooled Regression estimates*	
	Parameter Estimate	*t-value*	*Parameter Estimate*	*t-value*
INTERCEPT	24.51	62.69	23.26	95.86
LSIZE	-2.41	-13.49	-2.38	-27.73
GROWTH	0.21	2.63	0.26	4.33
LAGE	-0.09	-1.04	-0.02	-0.28
LANALYST	-0.65	-10.29	-0.68	-8.6
LTURNOVER	-2.00	-12.56	-1.95	-27.66
PERINSIDER	0.20	3.03	0.23	4.04
PERINST	0.68	8.91	0.65	8.66
ACCRUALS	0.01	0.11	0.02	0.3
ROA	0.45	5.84	0.51	8.24
STDEV	0.10	0.73	0.10	1.36
TOBIN	-0.07	-0.84	-0.05	-0.9
Adjusted R^2	0.27		0.26	

Panel B: 50% largest firms by market value

	Fama-MacBeth estimates		*Pooled Regression estimates*	
	Parameter Estimate	*t-value*	*Parameter Estimate*	*t-value*
INTERCEPT	17.26	32.00	12.78	88.27
LSIZE	-1.73	-20.44	-1.74	-35.06
GROWTH	0.11	1.75	0.15	4.38
LAGE	-0.26	-3.28	-0.26	-6.66
LANALYST	-0.69	-10.65	-0.68	-12.95
LTURNOVER	-0.79	-7.92	-0.79	-16.88
PERINSIDER	0.12	3.52	0.11	3.11
PERINST	0.17	1.71	0.19	4.20
ACCRUALS	0.01	0.26	0.02	0.64
ROA	-0.1	-1.56	-0.14	-3.20
STDEV	0.6	8.42	0.57	12.62
TOBIN	0.01	0.20	0.04	0.94
Adjusted R^2	0.39		0.44	

Table 5
Cross-Sectional Regressions by Sub-period

The dependent variable is the probability of information-based trading, *PIN*, in stock *i* of year *t* from 1983 through 1999. *LSIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *LAGE* represents the logarithm of 1+the number of years since the stock *i* was first covered by CRSP. *LTURNOVER* represents the logarithm of share turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. *PERINST* represents the institutional ownership as a percentage of shares outstanding, *PERINSIDER* represents the percentage of shares held by company insiders, and *LANALYST* is the logarithm of 1+number of analysts following stock *i* in year *t*. All independent variables are normalized by their yearly cross-sectional means and standard deviations. Results for industry indicator variables are suppressed for shortness of exposition.

Panel A: 1983-1990

	Fama-MacBeth estimates		*Pooled Regression estimates*	
	Parameter Estimate	*t-value*	*Parameter Estimate*	*t-value*
INTERCEPT	21.99	82.66	22.70	127.65
LSIZE	-3.00	-12.54	-2.99	-25.88
GROWTH	0.06	0.69	0.02	0.38
LAGE	0.02	0.37	0.05	0.78
LANALYST	-0.95	-10.82	-0.98	-8.73
LTURNOVER	-1.41	-7.33	-1.38	-20.60
PERINSIDER	0.06	0.84	0.07	1.30
PERINST	0.75	6.65	0.75	9.69
ACCRUALS	-0.05	-1.14	-0.07	-1.22
ROA	0.37	5.03	0.43	7.03
STDEV	0.42	3.16	0.41	5.82
TOBIN	-0.19	-2.66	-0.22	-3.76
Adjusted R^2	0.38		0.37	

Panel B: 1991-1999

	Fama-MacBeth estimates		*Pooled Regression estimates*	
	Parameter Estimate	*t-value*	*Parameter Estimate*	*t-value*
INTERCEPT	19.9	35.7	17.96	122.69
LSIZE	-3.96	-14.19	-3.89	-38.06
GROWTH	0.27	3.75	0.29	6.44
LAGE	-0.26	-4.44	-0.28	-5.77
LANALYST	-0.76	-8.15	-0.84	-8.80
LTURNOVER	-1.72	-15.87	-1.69	-28.75
PERINSIDER	0.23	6.78	0.23	5.04
PERINST	0.33	2.51	0.31	4.71
ACCRUALS	0.10	1.69	0.12	2.43
ROA	0.23	3.04	0.24	5.04
STDEV	0.33	1.98	0.29	4.66
TOBIN	-0.18	-5.11	-0.18	-3.67
Adjusted R^2	0.53		0.54	

Table 6
Cross-Sectional Regressions – Continuous Data

The dependent variable is the probability of information-based trading, *PIN*, in stock *i* of year *t* from 1983 through 1999. *LSIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*, *STDEV* is the annualized standard deviation of daily returns for firm *i* in year *t*. *LAGE* represents the logarithm of 1+the number of years since the stock *i* was first covered by CRSP. *LTURNOVER* represents the logarithm of share turnover for firm *i* in year *t*. *GROWTH* is the percentage increase in sales, *ROA* is the return on assets, *TOBIN* is the Tobin's Q, *ACCRUALS* is the absolute value of the abnormal accruals based on the Jones (1991) model. Industry indicator variables are formed annually based on the Fama-French 17-industry classification. All independent variables are normalized by their yearly cross-sectional means and standard deviations.

	Fama-MacBeth estimates		***Pooled Regression estimates***	
	Parameter Estimate	***t-value***	***Parameter Estimate***	***t-value***
INTERCEPT	20.92	51.38	18.10	124.97
LSIZE	-3.97	-18.18	-3.99	-82.19
GROWTH	0.20	3.44	0.23	6.60
LAGE	-0.11	-1.95	-0.06	-1.57
LTURNOVER	-1.59	-13.55	-1.56	-39.11
ACCRUALS	0.03	0.81	0.05	1.30
ROA	0.36	5.82	0.39	10.39
STDEV	0.31	3.08	0.32	6.82
TOBIN	-0.20	-4.87	-0.22	-5.96
Industry				
Food	0.03	0.19	-0.04	-0.21
Mining and Minerals	-0.19	-0.66	-0.13	-0.54
Oil and Petroleum Products	0.62	2.23	0.49	3.26
Textiles, Apparel & Footwear	0.63	3.02	0.55	3.12
Consumer Durables	0.40	1.85	0.31	1.80
Chemicals	0.05	0.28	0.09	0.43
Drugs, Soap, Perfumes, Tobacco	0.01	0.04	0.04	0.20
Construction and Constr. Materials	0.50	3.53	0.37	2.34
Steel Works Etc	-0.02	-0.09	-0.08	-0.36
Fabricated Products	0.14	0.51	-0.07	-0.28
Machinery and Business Equipment	0.10	0.65	0.02	0.20
Automobiles	0.36	1.54	0.17	0.71
Transportation	0.25	1.71	0.29	1.59
Utilities	-2.28	-10.06	-2.31	-15.60
Retail Stores	0.64	4.48	0.54	3.83
Banks, Insurance Companies, and Other Financials	-0.23	-1.44	-0.20	-1.19
Adjusted R^2	0.45		0.46	

Table 7
In-sample Asset Pricing Tests of PIN and PPIN

The table contains time series averages of the coefficients in cross-sectional asset pricing tests using the standard Fama-MacBeth (1973) methodology for the sample period 1983-1999. The dependent variable is the percentage monthly return in excess of the one-month T-bill rate. Betas are portfolio betas calculated from the full period using 40 portfolios. *PIN* is the probability of information-based trading in stock *i* of year *t*-1 estimated from trade data. PPIN is the probability of information-based trading in stock *i* of year *t*-1 estimated from trade data. L*SIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*-1. *BM* is the logarithm of the ratio of book value of common equity to market value of equity for firm *i* in year *t*-1. *RET12* is the return in month *t*-12 to *t*-1. T-values are given in parentheses. Panel A gives asset pricing results using PIN estimates. Panel B gives asset pricing results using the PPIN proxy estimated with the full set of economic variables given in Table 3. Panel C gives asset pricing results using the PPIN proxy estimated with the restricted set of economic variables given in Table 6.

Panel A: PIN

Beta	LSize	BM	PIN	Ret12
-0.190	0.104	0.248	0.017	
(-0.80)	(1.64)	(2.94)	(2.49)	
-0.190	0.078	0.221	0.013	0.007
(-0.84)	(1.25)	(2.69)	(2.03)	(3.82)

Panel B: PPIN based on full set of explanatory variables

Beta	LSize	BM	PPIN	Ret12
-0.251	0.246	0.255	0.086	
(-1.01)	(2.27)	(3.14)	(2.67)	
-0.266	0.196	0.225	0.067	0.005
(-1.12)	(1.85)	(2.81)	(2.21)	(2.76)

Panel C: PPIN based on restricted set of explanatory variables

Beta	LSize	BM	PPIN	Ret12
-0.228	0.296	0.250	0.121	
(-0.92)	(2.84)	(3.09)	(3.42)	
-0.244	0.244	0.219	0.101	0.005
(-1.04)	(2.37)	(2.75)	(2.91)	(2.70)

Table 8
PPINs and Asset Pricing

The table contains time series averages of the coefficients in cross-sectional asset pricing tests using the standard Fama-MacBeth (1973) methodology for the extended sample period 1965-2004. The dependent variable is the percentage monthly return. Betas are portfolio betas calculated from the full period using 40 portfolios. PPIN is the probability of information-based trading in stock *i* of year *t*-1 estimated from accounting and market data. L*SIZE* is the logarithm of market value of equity in firm *i* at the end of year *t*-1. *BM* is the logarithm of the ratio of book value of common equity to market value of equity for firm *i* in year *t*-1. *RET12* is the return in month *t*-12 to *t*-1. *Dvol* is the dollar volume in month *t*-2 relative to the test month. T-values are given in parentheses.

Specification	Beta	LSize	BM	PPIN	Ret12	Dvol
I.1	-0.128	-0.121	0.234			
	(-0.63)	(-2.89)	(4.21)			
I.2	0.047	0.244	0.227	0.159		
	(0.28)	(2.61)	(4.11)	(4.55)		
II.1	-0.151	-0.120	0.229		0.004	
	(-0.82)	(-2.97)	(4.24)		(2.60)	
II.2	0.023	0.221	0.227	0.147	0.003	
	(0.14)	(2.52)	(4.22)	(4.43)	(1.96)	
III.1	-0.135	-0.077	0.216		0.004	-0.050
	(-0.82)	(-1.04)	(4.02)		(2.89)	(-0.93)
III.2	-0.049	0.171	0.228	0.217	0.001	0.167
	(-0.31)	(1.90)	(4.24)	(6.21)	(0.82)	(3.05)

Table 9
Characteristic-adjusted returns to PPIN sorted portfolios

This table contains mean monthly percentage returns to PPIN-sorted deciles for the extended sample period 1965-2004. PPIN is the probability of information-based trading in stock *i* of year *t*-1 estimated from accounting and market data. DGTW benchmark portfolios are composed by first assigning them to size quintiles using NYSE size quintile breakpoints. Within each size quintile, stocks are assigned to quintiles based in their book-to-market ratio, yielding 25 size- and book-to-market sorted portfolios. We further sort stocks in each of the 25 portfolios into quintiles based on the prior 12-month return of each stock. Each stock is now uniquely identified with one of the resulting 125 portfolios, and the DGTW-adjusted return for stock *i* in month *t* is then its return minus the value-weighted mean return of the other stocks in the benchmark portfolio.

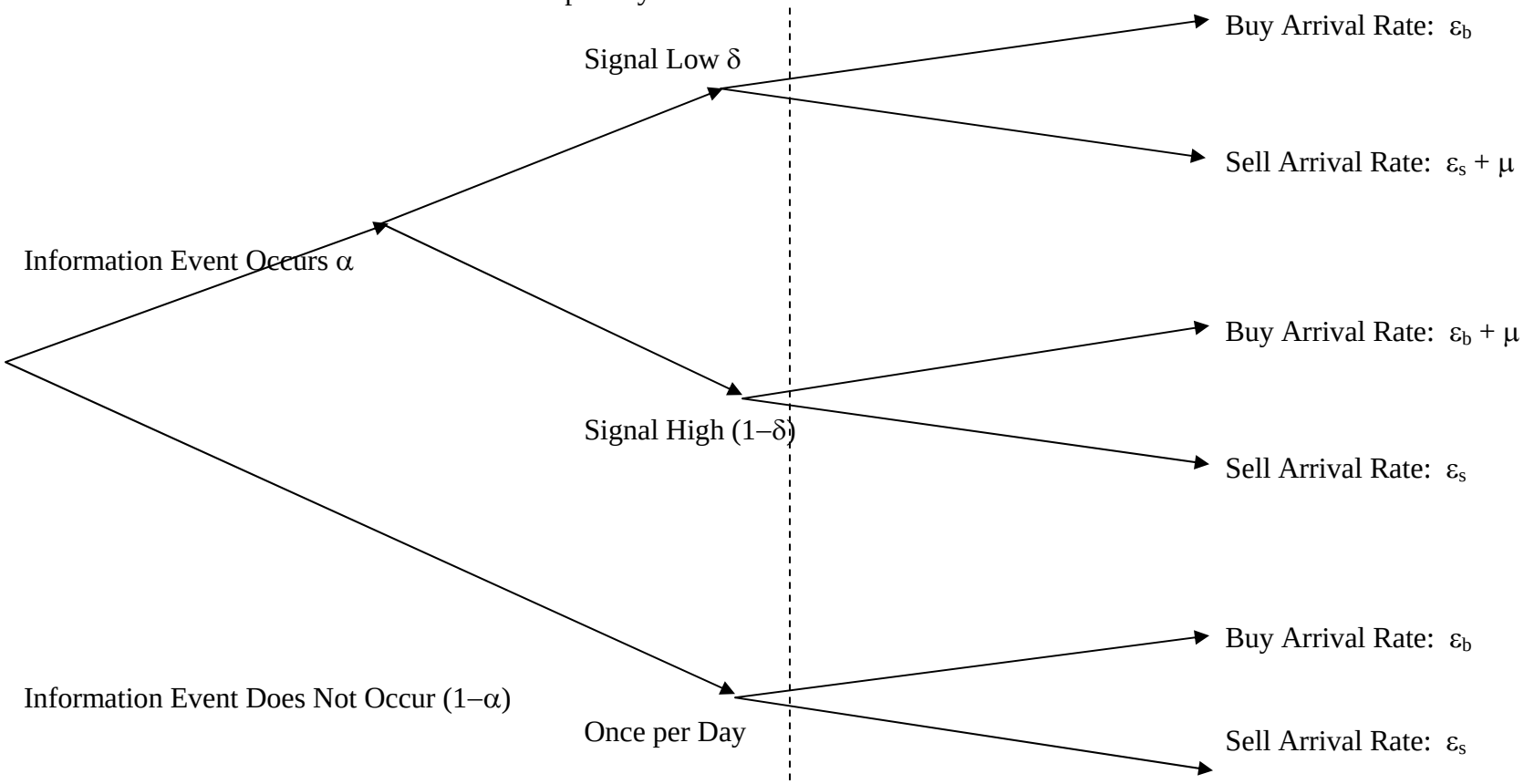
	1	2	3	4	5	6	7	8	9	10	10-1	*t*(10-1)
Raw	0.91	1.10	1.15	1.19	1.20	1.23	1.35	1.48	1.60	2.32	1.41	4.84
DGTW-	-0.07	0.06	-0.02	-0.00	-0.01	-0.06	0.01	0.08	0.09	0.65	0.72	4.95

Table 10
Average Returns to PPIN and Size sorted Portfolios

This table shows average percentage monthly returns on equally weighted portfolios balanced monthly by size and information effects. Information effects are captured by PPIN for the extended sample period 1965-2004. Each month we first sort stocks into five quintiles on the basis of size, where 5 is the largest firms. Then, within each size quintile, we sort stocks into five portfolios sorted by PPINs. The column '5-1' refers to the difference in monthly returns between PPIN portfolio 5 and 1.

	PPIN quintiles						
Size quintiles	1	2	3	4	5	5-1	*t*(5-1)
1	1.05	1.52	1.60	2.16	2.87	1.82	5.42
2	0.83	1.20	1.55	1.59	1.63	0.80	4.63
3	1.03	1.18	1.37	1.45	1.57	0.54	4.87
4	1.04	1.21	1.19	1.30	1.31	0.27	4.93
5	0.90	1.00	1.01	1.16	1.15	0.25	4.70

Figure 1. Tree Diagram of the Trading Process. α is the probability of an information event, δ is the probability of a low signal, μ is the rate of informed trade arrival, ε_b is the arrival rate of uninformed buy orders and ε_s is the arrival rate of uninformed sell orders. Nodes to the left of the dotted line occur once per day.



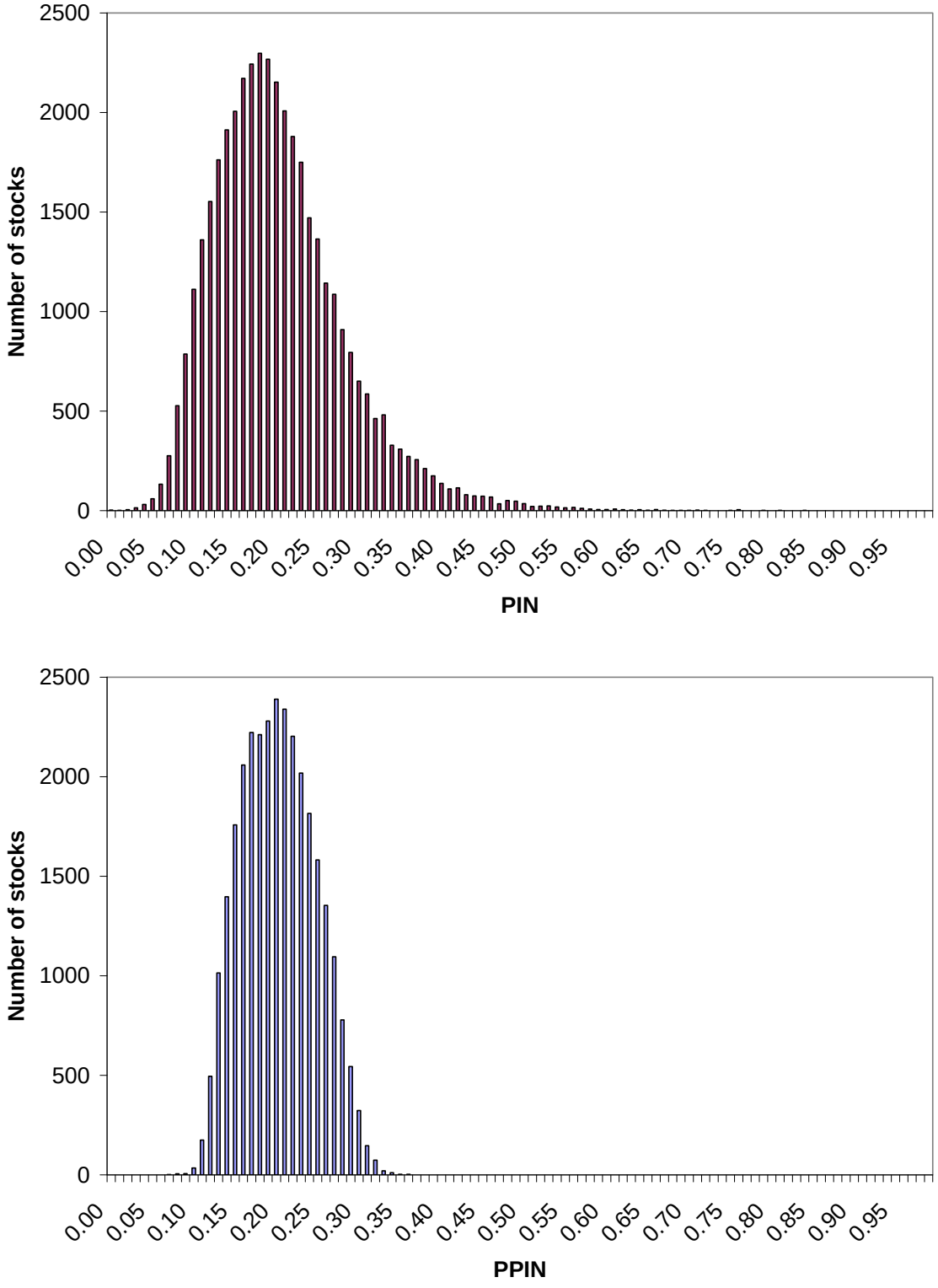


Figure 2: Distribution of PIN and PPIN

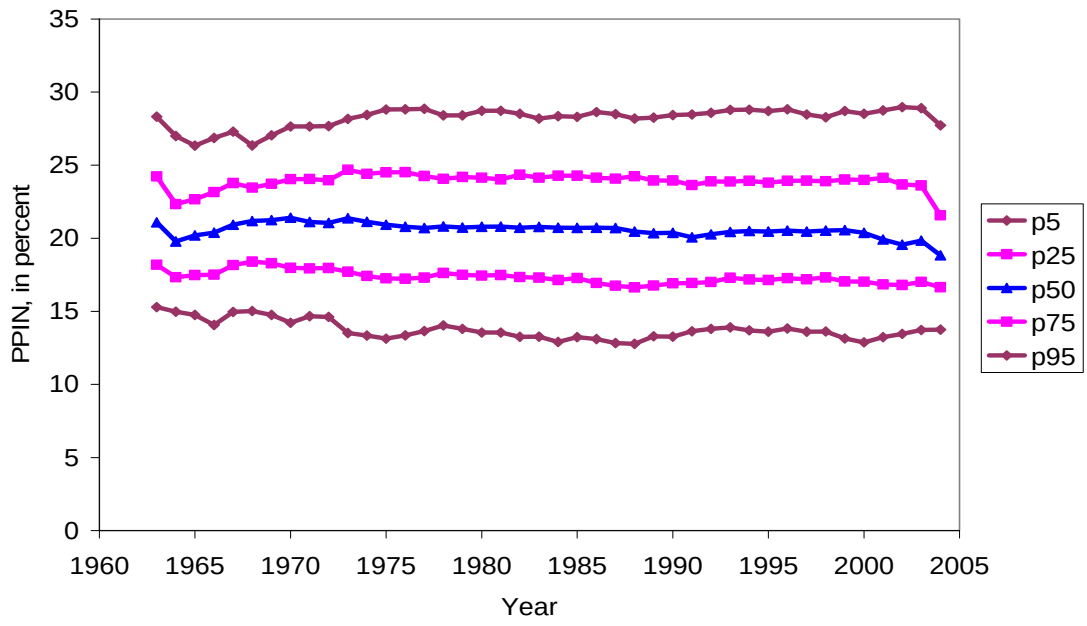


Figure 3: Distribution of PPIN across extended sample period
The figure shows the cross-sectional distribution of the proxy estimating probability of information based trading, PPIN. The figure shows the 5th, 25th, 50th, 75th and 95th percentiles each year in the sample period for the cross-sectional distribution of PPIN.

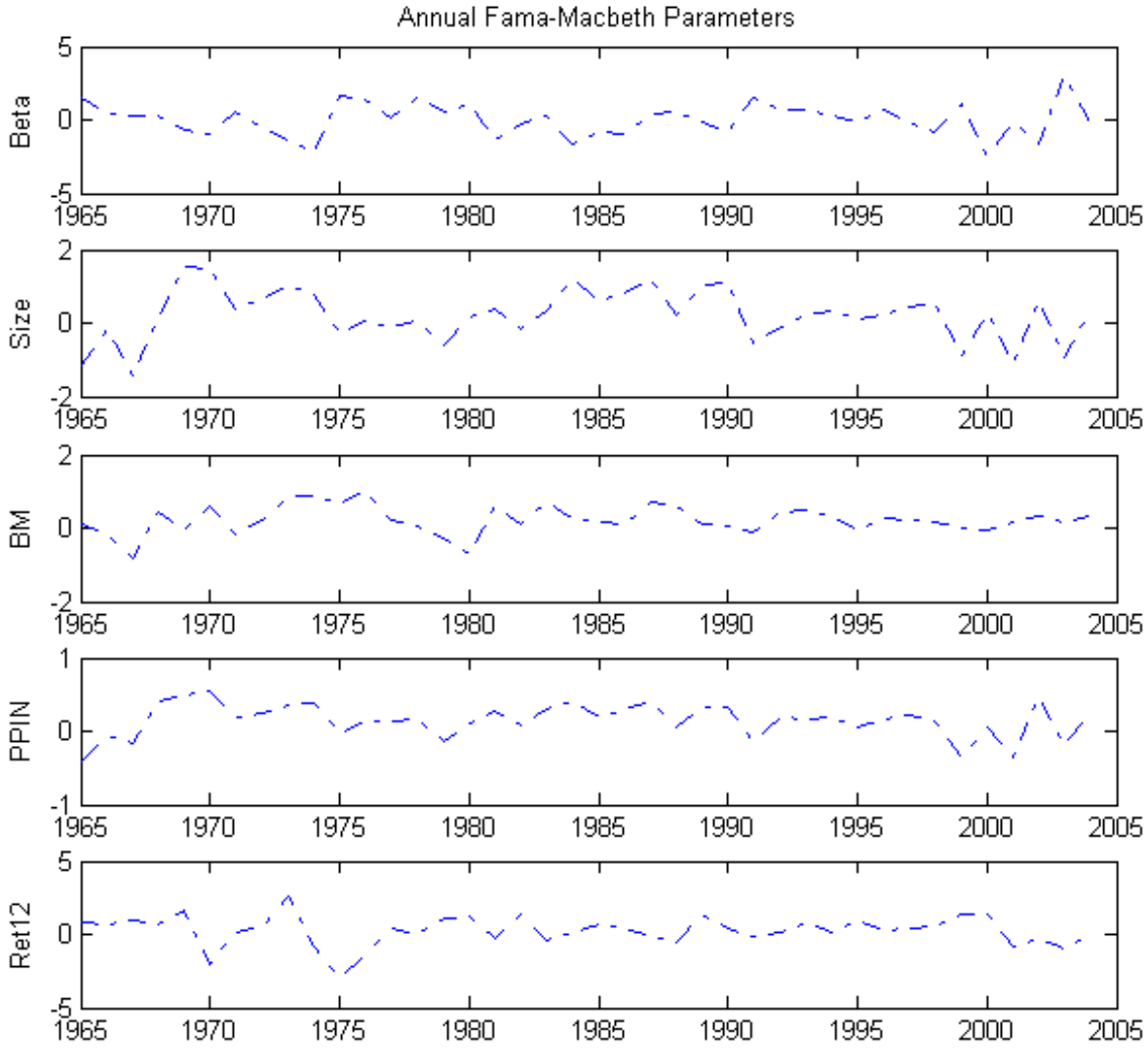


Figure 4: Annual Fama-Macbeth (1973) Coefficients