An Improved Estimation Method and Empirical Properties of PIN*

Yuxing Yan† and Shaojun Zhang‡

Abstract

The probability of information-based trading (PIN) is estimated via numerical maximization of a likelihood function. The maximization solutions frequently fall on the boundary of the parameter space and such boundary solutions cause a systematic bias in PIN estimates. We develop an algorithm to obtain PIN estimates that are free of this bias. We analyze the estimates for over 80,000 stock-quarter pairs and document evidence for temporal and seasonal variations in PIN. These variations are related to the January 2001 decimalization and the year-end tax-loss selling. We also find systematic differences between our estimates and others.

JEL Classification: C13, C61, G12, G14

Keywords: decimalization, market microstructure, numerical maximization, PIN, information-based trading, tax-loss selling

---

* We thank Chuan Yang Hwang for constructive and helpful comments and Stephen Brown for providing his PIN data.

† Wharton Research Data Services, 216 Vance Hall, 3733 Spruce Street, Philadelphia, PA, 19104. Phone: (215) 8986359, Fax: (215) 5736073, Email: yxyan@wharton.upenn.edu.

‡ Division of Banking and Finance, Nanyang Business School, Nanyang Technological University, S3-B1A-07 Nanyang Avenue, Singapore 639798. Phone: (65) 6790-4240, Fax: (65) 6792-4217, Email: asjzhang@ntu.edu.sg.

An Improved Estimation Method and Empirical Properties of PIN

Abstract

The probability of information-based trading (PIN) is estimated via numerical maximization of a likelihood function. The maximization solutions frequently fall on the boundary of the parameter space and such boundary solutions cause a systematic bias in PIN estimates. We develop an algorithm to obtain PIN estimates that are free of this bias. We analyze the estimates for over 80,000 stock-quarter pairs and document evidence for temporal and seasonal variations in PIN. These variations are related to the January 2001 decimalization and the year-end tax-loss selling. We also find systematic differences between our estimates and others.

The probability of information-based trading (PIN) is a measure of the information asymmetry between informed and uninformed traders in individual stocks. It has been used to study various topics such as how spreads differ between frequently and infrequently traded stocks, how market venues influence informed trading, whether financial analysts have private information, how stock splits affect trading, and whether information risk is a determinant of stock return.[1] An increasing number of recent studies, for example, Brown et al. (2004) and Vega (2005), use PIN to study a broader range of topics in corporate finance, empirical asset pricing, and so on.[2]

The value of PIN is derived from the market microstructure model in Easley and O'Hara (1992). Since PIN cannot be measured directly, it must be estimated by numerical maximization of the likelihood function specified by the underlying microstructure model.[3] In this paper, we identify a bias that may arise from numerical maximization of the likelihood function. We find that the maximization solutions are sensitive to initial values of the numeric procedure and frequently fall on the boundary of the parameter space. Such boundary solutions can cause substantial bias in the estimate of PIN. Logistic regressions show that the occurrence of boundary solutions is influenced by market capitalization and trading volume. Since stocks have smaller market capitalization and lower trading volume in early years than in recent years, we

---

[1] See Easley et al., 1996a, 1996b, 1997a, 1997b, 1998, 2001, 2002, 2004, and 2005.

[2] Brown et al. (2004) show that PIN is negatively associated with conference call activity and conclude that conference calls reduce information asymmetry and thus lower the cost of capital. Vega (2005) studies the effect of public surprises, media coverage, and private information on the post-earnings-announcement drift and uses PIN to proxy for private information prior to earnings announcements.

[3] A few studies point out that using the estimate of PIN in empirical analysis may introduce the errors-in-variable bias in regression results. For example, Easley et al. (2002) include an instrumental variable in their regression to correct for the potential errors-in-PIN bias. Vega (2005) uses bootstrapping to control for this bias. Brown et al. (2004) check the robustness of their empirical findings by filtering out questionable PIN estimates.

find that boundary solutions appear more frequently and show stronger effect in early years.

We develop an algorithm that can help avoid such boundary solutions. At the core of the algorithm are the moment equations derived from the marginal distributions of the daily number of trades according to the underlying market microstructure model. We use the moment equations to identify proper initial values in the parameter space. We apply the new algorithm to obtain estimates of the model parameters and PIN for a large number of stocks listed on the New York Stock Exchange (NYSE) and the American Stock Exchange (AMEX) in every calendar quarter between 1993 and 2004. The number of stocks in a quarter ranges from 1,481 to 1,923. We study these estimates and find several empirical properties of PIN.

First, we observe substantial changes in PIN and other trading-related parameters that coincide with the January 2001 decimalization on the NYSE and AMEX.[4] Bessembinder (2003) reports that, shortly after the decimalization, trade execution costs declined substantially and market quality improved on the NYSE. Harris (1997) conjectures that trading volumes should respond to the reduction in trading costs, but it may take a long while before traders adjust to a change in trading costs. Consistent with his conjecture, we find that uninformed and informed trades increased after the decimalization and exhibited a long upward drift. Because uninformed trades grew more than informed trades, the ratio of the daily number of uninformed trades to the daily number of informed trades (called the uninformed-to-informed ratio) also increased. The uninformed-to-informed ratio of an average stock jumped by an average of 5.2% in the

[4] Before the decimalization, both NYSE and AMEX were trading and quoting listed stocks in fractions of a dollar. On January 29 2001, both exchanges began trading and quoting listed stocks in increments of 1 cent.

first quarter of 2001 (denoted by 2001Q1) and increased dramatically by an additional 49.3% from 2001Q2 to 2004Q4. Because PIN is negatively related to the uninformed-to-informed ratio, PIN of an average stock declined by 19.8% from 2001Q1 to 2004Q4. This indicates significant reduction in information asymmetry on the NYSE/AMEX after the decimalization. We also find that the reduction in information asymmetry is mainly due to the significant increase in the uninformed-to-informed ratio.

Second, we find a strong seasonal pattern in PIN. Specifically, PIN tends to decline in the first quarter of a year relative to the previous fourth quarter.[5] Empirical evidence shows that the decline of PIN in the first quarter is also primarily due to contemporaneous increase in the uninformed-to-informed ratio.

We conduct an additional analysis to examine the relation between this seasonal pattern and the year-end tax-loss selling. We calculate the cumulative 10-month return from February to November in each year and classify a stock as winner if its cumulative return is greater than 10% or loser if the return is less than -10%. We find that, in the fourth quarter and the subsequent first quarter, winners and losers exhibit significant difference in market trading. Specifically, in the fourth quarter, losers on average experience greater increase in informed trading and uninformed selling than winners, and both groups of stocks experience about the same amount of increase in uninformed buying. In the subsequent first quarter, the majority of winners experience substantial increase in both informed and uninformed trading, while at least half of losers experience no increase in trading at all. This evidence is consistent with the year-end tax-loss selling hypothesis that investors sell losers at year end to save on tax. It also suggests that

[5] In the twelve-year period between 1993 and 2004, the only exception is the first quarter of 2001 when the decimalization took place.

investors seem to ignore losers in the following first quarter and chase winners instead.[6] Because PIN is negatively related to the uninformed-to-informed ratio and the ratio increases substantially more for winners than losers in the first quarter, PIN declines more for winners than losers in the first quarter.

In the end, we compare our quarterly PIN estimates with Stephen Brown's. Brown provides the only publicly available source of quarterly PIN estimates.[7] We find his PIN estimates are on average larger than ours in the years before 1999. Since boundary solutions are abundant in these years, the evidence suggests that his PIN estimates are subject to the upward bias associated with boundary solutions. To further examine the effect of boundary solutions on Brown's PIN estimates, we compare his PIN estimates with ours under two scenarios, where his estimates are more likely to be affected by boundary solutions in one scenario than in the other. We find that the differences between his PIN estimates and ours are larger in the first scenario than in the second.

In addition, there is an unusual difference between our PIN estimates and Brown's. In the three-year period between 1999 and 2001, we find that Brown's PIN estimates are unusually larger than ours, and that the quarterly standard deviations of his PIN estimates between 1999 and 2001 are twice as high as those of our PIN estimates. Because a regression coefficient estimate is positively related to the standard deviation of the dependent variable according to statistics theory, the coefficient estimates in a

[6] It is interesting to study whether such difference in trading is related to the abnormal return at the turn of the year is in order. Since it is out of the scope of this paper, we leave it for future research.
[7] Brown's PIN estimates are available at http://userwww.service.emory.edu/~sbrow22/index.html. Soeren Hvidkjaer provides another set of PIN estimates at his website, but his PIN estimate is on a yearly basis.

regression study may be different depending on whether our PIN estimates or Brown's are used as the dependent variable.

The remainder of this paper is organized as follows. Section I briefly describes the underlying market microstructure model and how PIN is estimated. Section II provides the details of our new algorithm. Section III describes the data. Sections IV, V, and VI present the empirical findings on the boundary solutions, the empirical properties of PIN, and the comparison between our PIN estimates and Stephen Brown's, respectively. Section VII concludes the paper.

I. The Estimation of PIN

In this section, we specify the underlying market microstructure model and describe how PIN is derived and estimated with actual trade data.[8] The underlying model views trading of a financial asset as a game between a market maker and traders that repeats over trading days. On any day, before trading starts, nature decides whether an information event occurs and reveals new information about the underlying asset value. The probability that an information event occurs is denoted α. If the information event occurs, it can be bad news with the probability δ or good news with the probability $1-\delta$.

Trading begins after the event occurs. Trades arrive according to Poisson processes throughout the day. Uninformed traders who are not aware of the new information submit buy orders at the daily arrival rate ε_b and sell orders at the daily arrival rate ε_s. On the day when an information event occurs, informed traders who know the new information submit orders at the daily arrival rate μ. The informed

[8] Please refer to Easley and O'Hara (1992) and Easley et al. (1997b, 2002) for detailed discussions about the motivation, structure, and interpretation of the underlying model and PIN.

traders buy at good news and sell at bad news. The market maker sets prices to buy or sell at any time during the day and executes orders as they arrive. The market maker updates his beliefs about the underlying asset value by observing the order flow and revises quote prices accordingly. The process of trading and learning continues throughout the day.

Mathematically, the model specifies that, on any day i, the likelihood of observing the number of buy trades B_i and the number of sell trades S_i is given by

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

where $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$ represents the five structural parameters in the model. Assuming days are independent, the joint likelihood of observing a series of daily buys and sells over trading days $i = 1,\ldots, I$ is the product of the daily likelihoods,

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

where $M = ((B_1, S_1),\ldots,(B_I, S_I))$ represents the data set. The probability of information-based trading (PIN) is defined as

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

Intuitively, PIN measures the fraction of trades in a day that arise from informed traders.

Maximizing the joint likelihood in equation (2) over the parameters in θ provides estimates of these structural parameters. However, there is no closed form solution to the maximization problem. A numerical maximization technique must be used to obtain a solution. Easley et al. (2003, 2005) recommend the following factorization of the joint likelihood function to facilitate numerical maximization

$$
\begin{aligned}
L((B_i,S_i)_{i=1}^{I}\big|\theta) &= \sum_{i=1}^{I}\left[-\varepsilon_b-\varepsilon_s+M_i\left(\ln x_b+\ln x_s\right)+B_i\ln\left(\mu+\varepsilon_b\right)+S_i\ln\left(\mu+\varepsilon_s\right)\right] \\
&+\sum_{i=1}^{I}\ln\left[\alpha\left(1-\delta\right)e^{-\mu}x_x^{S_i-M_i}x_b^{-M_i}+\alpha\delta e^{-\mu}x_b^{B_i-M_i}x_s^{-M_i}+\left(1-\alpha\right)x_s^{S_i-M_i}x_b^{B_i-M_i}\right]
\end{aligned}
\quad , \quad (4)
$$

where $M_i=\left(\min\left(B_i,S_i\right)+\max\left(B_i,S_i\right)\right)/2$, $x_s=\dfrac{\varepsilon_s}{\mu+\varepsilon_s}$, and $x_b=\dfrac{\varepsilon_b}{\mu+\varepsilon_b}$. The advantage of using this factorization is to increase computing efficiency and reduce truncation error. It is particularly important for stocks that have a large number of buys and sells because, without factorization, we would need to take the trading frequency parameters (i.e., μ, ε_b, or ε_s) to large powers that equal the actual number of trades (i.e., B_i or S_i). This frequently causes underflow or overflow problems in most computing environments.

Several numerical methods have been used for solving the maximization problem. Easley et al. (1996b, 1998, 2001) apply a logit transformation to restrict the probability parameters α and δ to (0, 1) and a logarithmic transformation to restrict the arrival rate parameters ε and μ to (0, ∞). They use the quadratic hill-climbing algorithm GRADX from the GQOPT package to maximize the likelihood function. Easley et al. (2003) factorize the likelihood function and maximize it by using the simplex search method implemented in the MATLAB *fminsearch* function. Brown et al. (2004) use the modified Newton-Raphson method implemented in the STATA *ml* procedure to maximize the likelihood function. In our study, we use the SAS *NLP* procedure to maximize the factorized likelihood function in equation (4). The SAS *NLP* procedure allows seven optimization methods. We chose the modified Newton-Raphson method because it is suitable for our problem and is widely used. It also makes our PIN estimate comparable with the PIN estimate used by Brown et al. (2004).

The Newton-Raphson method is an iterative search process that starts from certain initial values of the parameters and moves along a mathematically specified path in the parameter space in search of the largest value of the objective function. The process stops until certain convergence criteria are satisfied. Depending on the initial values, the process may or may not stop at the global optimum of the objective function.

In the following, we use one numerical example to illustrate the impact of initial values. We examine the numerical maximization results for the stock with the symbol MNR in the Trade and Quotes (TAQ) database in the first quarter of 1995. We use the SAS *NLP* procedure to maximize the factorized likelihood function conditional on the number of buy trades and sell trades in each day.[9] Since there are 63 trading days in this quarter, the data set consists of the number of buy trades B_i and the number of sell trades S_i for i = 1, …., 63. The average of the daily number of buy trades is 29.7 and the average of the daily number of sell trades is 25.6. We choose 80 different sets of the initial values for the five parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$ and run the SAS *NLP* procedure 80 times, each time with one different set of initial values. We will discuss the procedure of choosing these initial values in the next section. The maximization results from the 80 runs are given in Appendix A. We find that in 34 of the 80 runs, the maximization procedure produces a boundary solution for α, where α is equal to 0 or 1. More specifically, the solution of α is 0 for 22 of these runs and is 1 for the other 12 runs.

[9] We use the Lee and Ready (1991) algorithm to classify each trade into a buy or sell transaction.

In addition, we find that these boundary solutions have a large influence on the estimate of PIN. We calculate PIN with the estimated structural parameters according to equation (1). Table A.I in Appendix A reports the PIN estimates. When the solution of α is neither 0 nor 1, the PIN estimate is equal to either 0.131 or 0.157. To decide which one of these two values is the best PIN estimate, we examine the associated value of the log likelihood function. Because the maximum value of the log likelihood function is 8128.0, the best estimate of PIN ought to be 0.131. However, when the solution of α equals 0, the PIN estimate becomes 0, much smaller than 0.131. On the other hand, when the solution of α equals 1, the PIN estimate falls between 0.429 and 0.801, several times larger than 0.131.

The example demonstrates that the boundary solutions of α can influence the estimate of PIN substantially. The worst part is that it is unrealistic for α to have a boundary solution. Recall that α is the probability of an information event. When α equals 0, it means that information events never occur during the quarter. This is unlikely because many events can influence a firm's value, such as the publication of a quarterly report and announcements of new products or personnel changes by the firm and its competitors, suppliers, customers, and regulators. When α equals 1, it means that information events happen every day. Although we cannot rule out the possibility that this may happen in certain quarters for individual firms, it is rare.

Now the question is how can we avoid these problematic boundary solutions? The widely used approach is to run the maximization routine starting from different initial values and choose the solution that is realistic and at the same time maximizes the likelihood function. Easley et al. (2001) explicitly states, "To insure that we, in fact, find a global maximum for each stock, we run the optimization routine starting from many different points in the parameter space." However, the success of this approach depends on how representative the chosen initial values are of the parameter space. The initial values should be properly chosen so that the parameter space is explored thoroughly and efficiently. In the following section, we develop an algorithm that can set proper initial values.

II. A New Algorithm

The parameter space of the likelihood function in equation (4) is not simple, and it is a challenge to set initial values that are representative of this space. The probability parameters α and δ lie in the closed interval between 0 and 1, while the arrival rates μ, ε_b, and ε_s are unbounded, ranging from 0 to the positive infinity. For the bounded parameters, we can divide the closed interval equally and choose equal-distanced values within the interval. However, picking initial values for the unbounded parameters is not straightforward. What makes things more difficult is how to make sure that the chosen values of the five parameters together form a coherent set of reasonable initial values. The following algorithm accomplishes this purpose.

We use the moment conditions for the probability distributions of daily buys and sells to solve for the initial parameter estimates. The joint probability distribution of B and S is actually the same as the likelihood function in equation (1), from which we derive the marginal expected values of B and S as follows:

$$E(B) = \alpha(1-\delta)\mu + \varepsilon_b, \tag{5}$$

$$E(S) = \alpha\delta\mu + \varepsilon_s. \tag{6}$$

Appendix B provides the details of the mathematical derivation. We now use these two equations to set the initial values of the five parameters. We first divide the interval [0, 1] equally and choose equal-distanced values in the interval for α and δ. We then use the sample average $\overline{B}$ and $\overline{S}$ to replace *E(B)* and *E(S)* in the above two equations. Since the first part on the right-hand side of equation (5) is always positive (i.e., $\alpha(1-\delta)>0$), ε_b must be less than $\overline{B}$. This suggests that we choose the initial values of ε_b to be fractions of $\overline{B}$. At last, we solve equations (5) and (6) simultaneously for the initial values of ε_s and μ. Therefore, a whole set of initial values can be specified as follows

$$\alpha^0 = \alpha_i,\ \delta^0 = \delta_j,\ \varepsilon_b^0 = \gamma_k \cdot \overline{B},\ \mu^0 = \frac{\overline{B}-\varepsilon_b^0}{\alpha^0 \cdot \left(1-\delta^0\right)},\ \text{and}\ \varepsilon_s^0 = \overline{S} - \alpha^0 \cdot \delta^0 \cdot \mu^0, \qquad (7)$$

where the three variables α_i, δ_j, and γ_k represent fractions of one. More specifically, each variable equals one of the five fractions (0.1, 0.3, 0.5, 0.7, 0.9). The combinations of α_i, δ_j, and γ_k yield 125 sets of initial values.[10] Some of them are unacceptable because ε_s^0 has a negative value. In the above example for the stock MNR, 45 of the 125 sets are unacceptable, the remaining 80 sets are used to start the numerical maximization process, but 34 of them lead to boundary solutions.

We implement the following algorithm to obtain the estimate of PIN for any given stock. In the first step, we first construct the 125 sets of initial values with the data of daily buys and sells of the stock. In the second step, we run the maximization procedure for all acceptable sets and record the corresponding solutions. In the third step,

---

[10] We have tried to let α_i, δ_j, and γ_k each take one of the nine fractions (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9). The combination produces 729 sets of initial values, which substantially increases computation time. We used the 729 sets of initial vales for the first quarter of 2001 and did not find any systematic difference in the results between the 125 sets and the 729 sets. All results reported in this paper are obtained by using 125 sets of initial values.

we exclude all boundary solutions and choose the set of parameters that produces the highest value of the objective function among the nonboundary solutions. If all solutions are on the boundary, we interpret it as evidence that the optimum must occur on the boundary and thus choose the one that has the highest value of the objective function among all solutions. At last, we calculate the PIN estimate with the estimated parameters.

III. The Data and Preliminary Analysis

The estimation of PIN requires the number of buy trades and the number of sell trades in each day. The TAQ database includes bid quotes, ask quotes and trade prices, but does not tell whether a trade is buyer initiated or seller initiated. We classify each trade as buyer- or seller-initiated using the standard Lee and Ready (1991) algorithm. The algorithm classifies any trade that takes place above (below) the midpoint of the current bid-ask quotes as a buy (sell) because trades originating from buyers (sellers) are most likely to be executed at or near the ask (bid). For trades taking place at the midpoint, the most recent trade price is used to classify the trade. We follow Lee and Ready's (1991) suggestion to use the five-second lag of reported quote times to adjust for differences in reporting times between quotes and trades.

We estimate PIN for NYSE/AMEX listed stocks that have data in the TAQ database between January 1, 1993 and December 31, 2004. We focus on the NYSE/AMEX stocks because the underlying structural model conforms the most to the trading environment of the NYSE/AMEX. Stocks in the TAQ database are matched with those in the CRSP database by the historical eight-digit CUSIP. Since the TAQ database may reuse one symbol for different stocks whereas the CRSP database assigns a unique permanent number to every stock, this matching prevents us from mixing the trade data of two firms that are different but share the same TAQ symbol. It also lets us screen stocks based on the stock-specific information in the CRSP database. We keep only stocks with the CRSP share code 10 or 11, which means that the closed-end funds, real estate investment trusts, American depository receipts and foreign stocks are excluded. We also collect the closing price and the number of shares outstanding at the end of each calendar quarter from the CRSP database.

We calculate PIN for each stock on a quarterly basis and require that the stock have the trades and quotes data for at least 50 trading days in one quarter. A large number of daily observations help to reduce the estimation errors.[11] In total, 80,306 firm-quarters meet our requirements. Table I reports the number of stocks that obtain the PIN estimates and the number of stocks that do not obtain PIN estimates. For example, in the first quarter of 1993 (i.e. 1993Q1), we obtain PIN for 1,486 stocks and cannot obtain PIN for one stock, whereas in 2004Q4 the respective numbers are 1,625 and 65. Overall, we have PIN for 79,512 stock-quarters, accounting for 99% of the total number. By comparison, Easley et al. (2005) have a sample of nearly 40,000 stock-years between 1983 and 2001 and are able to obtain PIN estimates for all but 475. Our success rate is about the same as theirs.

[Insert Table I about here]

Consistent with Easley et al. (2005), we observe in Table I that the stocks that do not obtain PIN estimates account for a significant portion of the total market capitalization. A stock's market capitalization at the end of each calendar quarter is the product of the quarter-end closing price and the number of shares outstanding from the

[11] By comparison, Easley et al. (2005) requires a minimum of 60 trading days in one year.

CRSP database. In early years, the stocks that do not have PIN estimates account for a small percentage, mostly below 2%, of the total market capitalization of our sample stocks, yet their shares increased sharply in recent years. In 2004Q4, the number of stocks without PIN estimates is 65 out of a total of 1,690 stocks, but they account for 41.8% of the total market capitalization. Easley et al. (2005) point out that the main cause responsible for the failure to obtain PIN estimates is the large number of trades per day. Large numbers of buy or sell trades can produce a numerical value that exceeds the largest number a computer software program allows (i.e., overflow) or a numerical value that is smaller than the smallest number a computer software program can handle (i.e., underflow). Both overflow and underflow can cause the computer software to stop running.

IV. The Boundary Solutions

As described in Section II, the new algorithm we propose identifies 125 sets of initial values and run the numerical maximization procedure 125 times for each stock, each time with a different set of initial values. In the process of running our SAS program for the maximization, we kept a record of how many times the maximization produces a boundary solution. In this section, we study how frequently and when boundary solutions occur.

In Table II, we report the quarterly number of stocks that obtain PIN estimates and the mean and median of the number of times that each stock has a boundary solution out of the 125 runs. The mean and median are calculated over all stocks that obtain PIN estimates in each quarter. The means are larger than the medians, indicating that the number of boundary solutions has a right-skewed distribution. The mean and median remained almost the same in the quarters before 1999, but started declining from 1999Q1. The median was eight in 1999Q1 but became zero after 2002Q2, and the mean reduced from 15.8 in 1999Q1 to 3.5 in 2004Q4.

[Insert Table II about here]

The evidence shows that the occurrence frequency of boundary solutions varies greatly among stocks. It is interesting to know which stocks are likely to have boundary solutions. We study this question by comparing the characteristics of stocks that fall into two extreme categories: Stocks in the first category have no boundary solutions, whereas those in the second category have more than 15 boundary solutions. In Table II, we report, for each quarter, the number of stocks in each category, the proportion of each category accounting for the total number of stocks in the quarter, the average market capitalization and the average number of daily trades over all stocks in each category. The market capitalization in billions of dollar is calculated at the end of each quarter as the product of the quarter-end closing price and the number of shares outstanding. The number of daily trades is the sum of buyer-initiated and seller-initiated trades. We find that the number of stocks having no boundary solutions increases substantially from 243 (or 16.4% of stocks in the same quarter) in 1993Q1 to 1,001 (or 61.60%) in 2004Q4, whereas the number of stocks having more than 15 boundary solutions decreased from 431 (or 29.0%) to 70 (or 4.3%) over the same period.

We also observe that stocks with zero boundary solutions have larger market capitalization and more trades per day than those with more than 15 boundary solutions. It seems to suggest that the chance of having boundary solutions is negatively related to these two variables. However, it is widely known that stocks of large market capitalization are more actively traded. The influence of the two variables may be confounded. We thus conduct a logistic regression analysis to study the effect of the two variables jointly. Three logistic regression models are estimated. The dependent variable of all three models equals 1 for stocks having more than 15 boundary solutions, and 0 for stocks having zero boundary solutions. The three models have different explanatory variables. Model 1 has the logarithm of the market capitalization as the only explanatory variable. Similarly, the only explanatory variable in Model 2 is the logarithm of the number of trades per day. Model 3 includes both variables. Table III reports the estimation results of the three models. We omit the intercepts of the three models from Table III because they have no meaningful implications for our current analysis. The results show that both the market capitalization and the number of trades per day have negative coefficients in the single variable logistic regressions, i.e., Models 1 and 2. But in the joint model, market capitalization has a positive coefficient while the number of trades per day has a negative one. It means that, of any two stocks that have the same number of daily trades but differ in market capitalization, the one with larger market capitalization is more likely to have boundary solutions than the other.

The coefficient estimates in Table III show that the number of trades per day has a stronger effect than market capitalization. In addition, Table III reports the residual deviance of these logistic regressions and the associated chi-square tests. An insignificant residual deviance indicates a good fit of the logistic regression to the data. We find the residual deviance of the model with both variables is insignificant in most cases.

[Insert Table III about here]

The above evidence from the logistic regression analysis shows that the appearance of boundary solutions is related to the market capitalization of the stock and the level of trading. It has implications on how boundary solutions should be handled in empirical analysis. Brown et al. (2004) observed boundary solutions in the parameter estimates they obtained and, being concerned of the potential effect the boundary solutions might have, chose to filter out stocks with boundary solutions. They report, "We eliminate observations with extreme EKO parameter estimates using the following filters: (1) if $50\varepsilon > \mu$ or $50\mu > \varepsilon$; where $\varepsilon = \varepsilon_b + \varepsilon_s$; (2) if $\alpha < 0.02$ or $\alpha > 0.98$; (3) if $\delta < 0.02$ or $\delta > 0.98$; and (4) if $\min(\varepsilon, \mu) < 1$. These filters result in the elimination of 5,393 firm-quarter observations in the pooled, cross-sectional sample (representing 14% of the initial sample) and 1,981 observations in the time-series sample (19% of the initial sample)" (Brown et al., 2004, p. 350). The practice of filtering the observed boundary solutions is likely to impose a sample selection bias because stocks with low trading volume and low market capitalization are more likely to have boundary solutions.

V. Empirical Properties of PIN

Our estimates of PIN and other trading-related parameters cover a large number of stocks in the 48 calendar quarters between 1993 and 2004. The number of stocks in each quarter ranges from 1,481 to 1,923 (Table I). In the following, we study these estimates to explore temporal and cross-sectional pattern in PIN and other parameters.

A. PIN and the January 2001 Decimalization

Figure 1 displays the time series plots of the cross-sectional percentiles of PIN, α and δ in each quarter between 1993 and 2004. We examine five percentiles: the 5th, 25th, 50th (i.e., median), 75th, and 95th. The percentiles of the PIN estimate except the 95th follow a clear downward trend over time. The median PIN decreased from 0.169 in 1993Q1 to 0.115 in 2004Q4. The 95th PIN percentile has been relatively stable, suggesting that some stocks in the market consistently have high information asymmetry.

[Insert Figure 1 about here]

We observe a sharp difference between before and after 2001Q1 in the time series plots of the percentiles of α in Figure 1. The median of α jumped by 12.6% in 2001Q1 (relative to 2000Q4), and continued to rise by an additional 10.1% from 2001Q2 to 2004Q4. The increase in α is likely the consequence of the decimalization that took place on the NYSE and AMEX in January 2001. On January 29, 2001, the NYSE and AMEX began trading and quoting all listed stocks in increments of 1 cent. Before that, both exchanges were trading and quoting its listed stocks in fractions of a dollar. Bessembinder (2003) reports that, shortly after the decimalization, trade execution costs declined substantially. Lower trading costs enable investors to trade on private information that has a small value effect. For example, suppose an investor receives a piece of private information before trading begins and knows that the price will jump by 50 basis points over the opening price when the information becomes public during the trading session. The investor will buy at the opening price and sell after the price jumps only if his trading cost, as a percentage of the opening price, is below 50 basis points. Bessembinder (2003) reports in his Table 3 that the simple average effective spread as a percentage of price on the NYSE was 65 basis points before decimalization and 39 basis points after the decimalization. In this case, the informed investor would not trade before the decimalization but would trade after that. As a result, we observe more information-based trading after the decimalization, and this explains why the probability of an information event increased after 2001Q1.

In addition, we observe structural changes around 2001Q1 for the daily arrival rate of informed trades μ and the daily arrival rate of uninformed buy and sell trades ε_b and ε_s, as well. Figure 2 displays the time series plots of the quarterly cross-sectional percentiles of the three trading frequency parameters between 1993 and 2004. All three arrival rates increased slowly before 2001Q1 and exploded after 2001Q1. The annual compounded growth rates of the medians of μ, ε_b and ε_s were 7.8%, 11.2% and 12.1% for the period between 1993Q1 and 2000Q4. In contrast, the growth rates were 29.7%, 42.6% and 38.0% between 2001Q1 and 2004Q4. The long drift after the decimalization in the trading frequency parameters, μ, ε_b and ε_s, is consistent with Harris's (1997) conjecture that it takes a long while before investors recognize the effect of tick size changes on trading costs and adjust to it.

Figure 2 shows a few other patterns that have some implications on market trading activities. First, the cross-sectional distributions of μ, ε_b and ε_s are highly skewed to the right. For all three trade arrival rates, the difference between the 95th percentile and the median is much larger than the difference between the median and the 5th percentile. This implies that market trading concentrates in a relatively small number of active stocks. Second, the 95th percentile behaves very differently from the 5th percentile. While the 95th percentiles of all three arrival rates have gone up significantly between 1993 and 2004, the 5th percentiles have had little change. This suggests that stocks that were actively traded attracted more trading during this period, while those with low trading volume continue to be neglected. Third and last, there is a difference in the growth pattern between the informed and uninformed trades. The 75th and 95th percentiles of μ have maintained at a stable level after 2002Q3, but these two percentiles of ε_b and ε_s continue to grow at a fast pace. This seems to suggest that uninformed trading has been driving the growth in market trading activities and may continue to do so in near future.

[Insert Figure 2 about here]

B. Seasonal Variation in PIN

One subtle pattern in Figure 1 reveals that PIN almost always decreases in the first quarter of a year. We now study the quarterly pattern in the variation of PIN and other related parameters more closely. The mathematical formula of PIN can be rewritten as follows

$$PIN = \frac{\alpha\mu}{\alpha\mu + \varepsilon_b + \varepsilon_s} = \frac{1}{1 + \frac{1}{\alpha} \cdot \frac{\varepsilon_b + \varepsilon_s}{\mu}}. \tag{8}$$

The formula shows that PIN is related to the probability of an information event α and the ratio of the daily arrival rate of uninformed trades to the daily arrival rate of informed trades (i.e., $(\varepsilon_b + \varepsilon_s)/\mu$). We study the quarterly percentage change of PIN, α and $(\varepsilon_b + \varepsilon_s)/\mu$. The quarterly percentage change of a variable V is defined as $\Delta V_t = (V_t - V_{t-1})/V_{t-1}$, where t indexes calendar quarters. Figure 3 presents the bar plots for the median percentage changes of PIN, α, and $(\varepsilon_b + \varepsilon_s)/\mu$ in every quarter between 1993Q2 and

2004Q4. The height of each bar equals the median of the quarterly percentage changes across all sample stocks in the quarter. It is clear in Figure 3 that the percentage change of PIN is negative in all first quarters except 2001Q1. The percentage change of α is either negative or positive in the first quarter, yet the percentage change of $\left(\varepsilon_b+\varepsilon_s\right)/\mu$ is always positive in all first quarters. The evidence suggests that the decrease in the probability of informed trading in the first quarter is mainly due to the increase in the uninformed-to-informed ratio.

[Insert Figure 3 about here]

Table IV reports, for each of the four calendar quarters, the mean and median of the quarterly percentage changes of the six variables: PIN, α, μ, ε_b, ε_s, and $\left(\varepsilon_b+\varepsilon_s\right)/\mu$. The mean and median are calculated after removing the extreme 1% observations at both tails. The means are much larger than the medians, indicating that the sample distributions of these variables are right skewed. This is consistent with the time series plots of the percentiles in Figure 1. Table IV confirms the seasonal pattern that, in the first quarter, PIN decreases and the uninformed-to-informed ratio increases. In addition, Table IV shows that all three arrival rates of trades, μ, ε_b and ε_s, increase in the first quarter, but the arrival rates of uninformed trades ε_b and ε_s increase more than the arrival rate of informed trades μ. This explains the increase of the uninformed-to-informed ratio.

Moreover, Table IV shows that the change in the arrival rates of both informed trades and uninformed trades exhibits a U-shape seasonal pattern across the four quarters in a year. That is, the arrival rates increase significantly more in the first and fourth quarter than in the second and third quarter. In addition, the increase in informed trades is almost the same in the first and the fourth quarters, whereas the increase in uninformed trades is much larger in the first quarter than in the fourth. As a result, the uninformed-to-informed ratio increases more in the first quarter than in the fourth.

[Insert Table IV about here]

The evidence in both Figure 3 and Table IV supports a strong seasonal pattern that PIN always declines in the first quarter of a year. In an attempt to explain this phenomenon, we hypothesize that the seasonal pattern may be related to tax-loss selling at year end.[12] Table V reports relevant empirical evidence. We calculate the cumulative 10-month return from February to November in each year and classify a stock as winner if its cumulative return is greater than 10% or loser if the return is less than -10%. We study the quarterly changes of PIN and other parameters in the fourth quarter and the first quarter of the following year. Table V reports the mean and median of these quarterly changes by quarter and type of stock. The mean and median are calculated after removing the most extreme 1% observations at both tails.

We find that, in the fourth quarter and the subsequent first quarter, winners and losers exhibit significant difference in market trading. Specifically, in the fourth quarter, losers on average experience greater increase in informed trading and uninformed selling than winners, while both groups of stocks experience about the same amount of increase in uninformed buying. In the subsequent first quarter, the majority of winners experience substantial increase in both informed and uninformed trading, while at least half of losers experience no increase in trading at all. This evidence is consistent with the year-end tax-

[12] The year-end tax-loss selling is extensively studied in literature and is found to contribute to the January effect. Please refer to recent related studies including Grinblatt and Moskowitz (2004), Ginblatt and Keloharju (2003), Poterba and Weisbenner (2001), and references therein.

loss selling hypothesis that investors sell loser stocks at year end to save on tax. It also indicates that investors seem to ignore loser stocks in the following first quarter and chase winners instead. Furthermore, because PIN is negatively related to the ratio of uninformed trades to informed trades and the ratio increases substantially more for winners than losers in the first quarter, PIN declines more for winners than losers in the first quarter.

[Insert Table V about here]

VI. Comparison with Stephen Brown's PIN Estimates

The numerical example in Section I demonstrates that if boundary solutions are mistakenly accepted as the maximum likelihood estimates, the resulting PIN estimate can be seriously biased. In other words, if the maximization solution of α is equal to 0, PIN will be 0 and thus biased downward; if the maximization solution of α is equal to 1, PIN will be biased upward. The evidence in Section IV shows that boundary solutions occurred with high frequency in the years before 1999. More specifically, the average occurrence frequency of boundary solutions is about 15 out of 125 times before 1999. Therefore, if boundary solutions were not handled properly, PIN estimates would show an upward bias before 1999. Our PIN estimates are largely free of this bias because we use the new algorithm described in Section II to avoid boundary solutions.

We now compare our PIN estimates with another set of PIN estimates provided by Stephen Brown. Brown made his quarterly PIN estimates publicly available on his website. If Brown's PIN estimates contain the bias associated with boundary solutions, there should be systematic differences between his estimates and ours.

Table VI reports summary statistics of the two sets of PIN estimates in every quarter between 1993Q1 and 2003Q4.[13] It also reports summary statistics for the same-stock difference between our and Brown's PIN estimates. We match stocks in our sample with those in Brown's by their eight-digit CUSIP. We find the mean and median of the same-stock differences between our and Brown's PIN estimates are negative and declining in magnitude before 1998Q4. However, between 1999Q1 and 2002Q4, the quarterly mean differences are negative but have unusually large magnitude. For example, the mean difference is -0.004 in 1998Q4, but changes dramatically to -0.029 in 1999Q1. It is even more puzzling that, between 1999Q1 and 2001Q4, the quarterly standard deviations of the PIN differences are almost twice as large as the standard deviations in the years both before and after the three-year period. This relates to the unusually large standard deviations of Brown's PIN estimates in this three-year period. Our PIN estimates have almost the same standard deviations in these three years as in other years. Brown et al. (2004) analyze the effect of conference calls on information asymmetry using a regression that includes PIN as the dependent variable and the number of conference calls in the prior quarter as an independent variable. The estimated coefficient of the conference call variable indicates that PIN is 0.59 percentage point lower for each conference call held during the prior quarter. If the same regression were estimated using our PIN estimates, there might be a smaller impact of conference calls on PIN because statistics theory suggests that a regression coefficient estimate is positively related to the standard deviation of the dependent variable.

[Insert Table VI about here]

[13] Since Brown's data are available only between 1993 and 2003, we cannot compare our PIN estimates in 2004.

The observation that Brown's PIN estimates are larger than ours before 1999 is consistent with our conjecture that Brown's PIN estimates may contain an upward bias due to boundary solutions. We conduct another analysis to further examine this issue. We compare our and Brown's PIN estimates for two groups of stocks. Stocks in the first group have no boundary solutions out of 125 maximizations, whereas stocks in the second group have more than 15 boundary solutions. Since stocks in the second group are more likely to have boundary solutions, Brown's PIN estimates for these stocks are more likely to be biased. Thus we expect a larger difference between our and Brown's PIN estimates in the second group than in the first. The evidence reported in Table VII confirms our expectation well, especially in the years before 1999Q1 when boundary solutions are abundant.

[Insert Table VII about here]

VII. Conclusion

In this paper, we propose an improved method of estimating PIN and study empirical properties of PIN. Because PIN cannot be measured directly, it must be estimated by numerical maximization of the likelihood function derived from an underlying market microstructure model. We find that the maximization solutions frequently fall on the boundary of the parameter space and such boundary solutions cause a systematic bias in the estimate of PIN. Our estimation method overcomes this bias by exploring the parameter space thoroughly and effectively so as to avoid boundary solutions. We compare our PIN estimates with Stephen Brown's, and find systematic differences that are associated with boundary solutions.

We obtain the estimates of PIN and other parameters for about 80,000 stock-quarters of the NYSE/AMEX listed firms between 1993 and 2004. Analysis of this large set of estimates reveals two interesting empirical regularities about PIN. First, PIN has been consistently declining since the decimalization that took place in January 2001 on the NYSE and AMEX. In fact, PIN of an average stock decreased by about 20% between 2001 and 2004. During the same period, the probability of having new information event in a trading day, one of the two components of PIN, increased by about 24% for an average stock. A possible explanation is that the decimalization lowers transaction costs and makes it more convenient for informed traders to trade on their private information. On the other hand, lower transaction costs attract a lot more uninformed trades than informed trades. In consequence, the ratio of the arrival rate of uninformed trades to the arrival rate of informed trades, which is the other component of PIN, increased dramatically by about 57% for an average stock during the same period. Since PIN is negatively related to the uninformed-to-informed ratio and positively related to the probability of new information event, the post-decimalization decline in the probability of informed trading is thus due to the significant increase in the ratio.

Second, we find a strong seasonal pattern in PIN, that is, PIN tends to decline in the first quarter of a year. The decline is more prominent for winning stocks than for losing stocks. We investigate the relation between this seasonal pattern and year-end tax-loss selling, and find that investors sell losing stocks at year end for tax benefits but chase winning stocks at the beginning of next year. This behavior drives the difference in the buy and sell transactions between winning and losing stocks, and provides a sound explanation of the seasonal pattern in PIN.

Our empirical findings have important implications on the use of PIN in future research. First, the temporal and seasonal patterns in PIN may have confounding effects on the intended research hypotheses. Careful thoughts and proper research design may be required to account for the unwanted effects. Second, PIN is a composite variable consisting of a few components. The variation in PIN must be related to the variation in individual components to certain extent. Therefore studies that investigate or make use of the variation in PIN may gain additional insights by examining the variation of individual components. Last but not least, the systematic differences between our PIN estimates and Stephen Brown's highlight not only the impact of boundary solutions but also the importance of studying the quality of PIN estimates.

Appendix A
Effect of initial values on numerical maximization solutions

This appendix reports a numerical example that illustrates the effect of initial values on numerical maximization solutions. This example also illustrates that numerical maximization solutions can fall on the boundary of the parameter space frequently. The example uses data for the stock with the TAQ symbol MNR. The data is for the first quarter of 1995. The number of trading days is 63. The average numbers of daily buyer-initiated and seller-initiated trades are 29.7 and 25.6, respectively. Classification of buyer-initiated and seller-initiated trades is based on Lee and Ready (1991) methodology. We choose 80 different sets of the initial values for the five parameters $\theta = (\alpha, \delta, \mu, \varepsilon_b, \varepsilon_s)$ according to the algorithm described in Section II. Table A.I reports the maximization results for each of the 80 sets, where PIN is the probability of informed trading, α is the probability of an information event, δ is the probability of the information event being bad news, μ is the arrival rate of informed trades, ε_b and ε_s are the arrival rates of uninformed buy and sell trades, and Loglik is the logarithm of the maximum value of the likelihood function.

Table A.I
Numerical maximization solutions for the stock MNR

	Initial Values					Maximization solutions						
No.	α	δ	μ	ε_b	ε_s	Loglik	α	δ	μ	ε_b	ε_s	PIN
1	0.1	0.1	297.14	2.97	22.63	8094.6	0.000	0.100	297.14	29.71	25.60	0.000
2	0.1	0.1	231.11	8.91	23.29	8094.6	0.000	0.100	231.11	29.71	25.60	0.000
3	0.1	0.1	165.08	14.86	23.95	8094.6	0.000	0.100	165.08	29.71	25.60	0.000
4	0.1	0.1	99.05	20.80	24.61	8094.6	0.000	1.000	100.22	29.71	25.60	0.000
5	0.1	0.1	33.02	26.74	25.27	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
6	0.3	0.1	99.05	2.97	22.63	8094.6	0.000	0.000	98.62	29.71	25.60	0.000
7	0.3	0.1	77.04	8.91	23.29	8094.6	0.000	0.000	66.20	29.71	25.60	0.000
8	0.3	0.1	55.03	14.86	23.95	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
9	0.3	0.1	33.02	20.80	24.61	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
10	0.3	0.1	11.01	26.74	25.27	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
11	0.5	0.1	59.43	2.97	22.63	7777.4	1.000	0.000	50.12	0.01	22.83	0.687
12	0.5	0.1	46.22	8.91	23.29	8094.6	1.000	0.000	23.75	5.96	25.60	0.429
13	0.5	0.1	33.02	14.86	23.95	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
14	0.5	0.1	19.81	20.80	24.61	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
15	0.5	0.1	6.60	26.74	25.27	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
16	0.7	0.1	42.45	2.97	22.63	8010.5	1.000	0.000	39.06	0.00	23.26	0.627
17	0.7	0.1	33.02	8.91	23.29	8094.6	1.000	0.000	25.41	4.31	25.60	0.459
18	0.7	0.1	23.58	14.86	23.95	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
19	0.7	0.1	14.15	20.80	24.61	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
20	0.7	0.1	4.72	26.74	25.27	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
21	0.9	0.1	33.02	2.97	22.63	8092.8	1.000	0.000	30.77	0.00	24.86	0.553
22	0.9	0.1	25.68	8.91	23.29	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
23	0.9	0.1	18.34	14.86	23.95	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
24	0.9	0.1	11.01	20.80	24.61	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
25	0.9	0.1	3.67	26.74	25.27	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
26	0.1	0.3	382.04	2.97	14.14	8094.6	0.000	0.300	382.04	29.71	25.60	0.000
27	0.1	0.3	297.14	8.91	16.69	8094.6	0.000	0.300	297.14	29.71	25.60	0.000
28	0.1	0.3	212.24	14.86	19.24	8094.6	0.000	0.300	212.25	29.71	25.60	0.000
29	0.1	0.3	127.35	20.80	21.78	8094.6	0.000	0.301	127.35	29.71	25.60	0.000
30	0.1	0.3	42.45	26.74	24.33	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
31	0.3	0.3	127.35	2.97	14.14	7328.8	1.000	0.000	58.72	0.01	14.59	0.801
32	0.3	0.3	99.05	8.91	16.69	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
33	0.3	0.3	70.75	14.86	19.24	8094.6	0.000	0.000	90.48	29.71	25.60	0.000
34	0.3	0.3	42.45	20.80	21.78	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
35	0.3	0.3	14.15	26.74	24.33	8128.0	0.487	0.090	14.83	23.12	24.97	0.131

Table A.I (continued)
Numerical maximization solutions for the stock MNR

	Initial Values					Maximization outcomes						
No.	α	δ	μ	ε_b	ε_s	Loglik	α	δ	μ	ε_b	ε_s	PIN
36	0.5	0.3	76.41	2.97	14.14	7322.9	1.000	0.000	60.05	0.01	15.39	0.796
37	0.5	0.3	59.43	8.91	16.69	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
38	0.5	0.3	42.45	14.86	19.24	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
39	0.5	0.3	25.47	20.80	21.78	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
40	0.5	0.3	8.49	26.74	24.33	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
41	0.7	0.3	54.58	2.97	14.14	7476.1	1.000	0.099	53.22	0.01	14.41	0.787
42	0.7	0.3	42.45	8.91	16.69	8003.6	1.000	0.000	35.89	0.00	19.46	0.648
43	0.7	0.3	30.32	14.86	19.24	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
44	0.7	0.3	18.19	20.80	21.78	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
45	0.7	0.3	6.06	26.74	24.33	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
46	0.9	0.3	42.45	2.97	14.14	7816.8	1.000	0.000	41.53	0.01	15.83	0.724
47	0.9	0.3	33.02	8.91	16.69	8094.6	1.000	0.000	29.41	0.31	25.60	0.532
48	0.9	0.3	23.58	14.86	19.24	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
49	0.9	0.3	14.15	20.80	21.78	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
50	0.9	0.3	4.72	26.74	24.33	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
51	0.1	0.5	416.00	8.91	4.80	8094.6	0.000	0.500	416.00	29.71	25.60	0.000
52	0.1	0.5	297.14	14.86	10.75	8094.6	0.000	0.500	297.14	29.71	25.60	0.000
53	0.1	0.5	178.29	20.80	16.69	8094.6	0.000	0.500	178.29	29.71	25.60	0.000
54	0.1	0.5	59.43	26.74	22.63	8094.6	0.000	0.000	66.60	29.71	25.60	0.000
55	0.3	0.5	138.67	8.91	4.80	8094.6	0.000	0.003	95.54	29.71	25.60	0.000
56	0.3	0.5	99.05	14.86	10.75	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
57	0.3	0.5	59.43	20.80	16.69	8094.6	0.000	0.997	66.06	29.71	25.60	0.000
58	0.3	0.5	19.81	26.74	22.63	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
59	0.5	0.5	83.20	8.91	4.80	8094.6	0.000	0.000	87.65	29.71	25.60	0.000
60	0.5	0.5	59.43	14.86	10.75	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
61	0.5	0.5	35.66	20.80	16.69	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
62	0.5	0.5	11.89	26.74	22.63	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
63	0.7	0.5	59.43	8.91	4.80	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
64	0.7	0.5	42.45	14.86	10.75	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
65	0.7	0.5	25.47	20.80	16.69	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
66	0.7	0.5	8.49	26.74	22.63	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
67	0.9	0.5	46.22	8.91	4.80	8122.4	0.710	0.774	12.24	27.75	18.88	0.157
68	0.9	0.5	33.02	14.86	10.75	8122.4	0.710	0.774	12.24	27.75	18.87	0.157
69	0.9	0.5	19.81	20.80	16.69	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
70	0.9	0.5	6.60	26.74	22.63	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
71	0.1	0.7	297.14	20.80	4.80	8094.6	0.000	0.700	297.14	29.71	25.60	0.000
72	0.1	0.7	99.05	26.74	18.67	8094.6	0.000	0.738	99.06	29.71	25.60	0.000
73	0.3	0.7	99.05	20.80	4.80	8094.6	0.000	0.098	101.56	29.71	25.60	0.000
74	0.3	0.7	33.02	26.74	18.67	8128.0	0.487	0.090	14.83	23.12	24.97	0.131
75	0.5	0.7	59.43	20.80	4.80	8094.6	0.000	1.000	192.81	29.71	25.60	0.000
76	0.5	0.7	19.81	26.74	18.67	8122.4	0.710	0.774	12.24	27.75	18.87	0.157
77	0.7	0.7	42.45	20.80	4.80	8026.4	1.000	1.000	30.30	23.73	0.00	0.561
78	0.7	0.7	14.15	26.74	18.67	8122.4	0.710	0.774	12.24	27.75	18.87	0.157
79	0.9	0.7	33.02	20.80	4.80	8122.4	0.710	0.774	12.24	27.75	18.88	0.157
80	0.9	0.7	11.01	26.74	18.67	8122.4	0.710	0.774	12.24	27.75	18.88	0.157

Appendix B
Derivation of the marginal expected values of B and S in Equations (5) and (6)

The structural market microstructure model specifies the joint probability distribution of the daily number of buy and sell trades, i.e. B and S, as follows

$$p(B,S) = \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_{ss}}\frac{\varepsilon_S^S}{S!}$$
$$+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_{ss})}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_S^S}{S!}$$

The marginal distribution of B is obtained by summing the above probabilities over all possible values of S, i.e.,

$$p(B) = \sum_{s=0}^{\infty} p(B,S)$$
$$= \sum_{s=0}^{\infty}\Bigg[\alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_{ss}}\frac{\varepsilon_S^S}{S!}$$
$$+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_{ss})}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_S^S}{S!}\Bigg]$$
$$= \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_{ss}}\sum_{s=0}^{\infty}\frac{\varepsilon_S^S}{S!}$$
$$+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_{ss})}\sum_{s=0}^{\infty}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\sum_{s=0}^{\infty}\frac{\varepsilon_S^S}{S!}$$
$$= \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}$$

The last step in the above derivation uses the Taylor series expansion $e^x = \sum_{n=0}^{\infty} x^n / n!$.

Consequently, the marginal expected value of B is given by

$$E(B) = \sum_{s=0}^{\infty} p(B)\times B$$
$$= \sum_{B=0}^{\infty}\left[\alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}\right]\times B$$
$$= \sum_{B=1}^{\infty}\left[\alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{(B-1)!}+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{(B-1)!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{(B-1)!}\right]$$
$$= \alpha(1-\delta)e^{-(\mu+\varepsilon_b)}(\mu+\varepsilon_b)\sum_{B=1}^{\infty}\frac{(\mu+\varepsilon_b)^{B-1}}{(B-1)!}+\alpha\delta e^{-\varepsilon_b}\varepsilon_b\sum_{B=1}^{\infty}\frac{\varepsilon_b^{B-1}}{(B-1)!}+(1-\alpha)e^{-\varepsilon_b}\varepsilon_b\sum_{B=1}^{\infty}\frac{\varepsilon_b^{B-1}}{(B-1)!}$$
$$= \alpha(1-\delta)(\mu+\varepsilon_b)+\alpha\delta\varepsilon_b+(1-\alpha)\varepsilon_b$$
$$= \alpha(1-\delta)\mu+\varepsilon_b$$

Similarly, we can derive the marginal distribution and the expected value of the daily number of sell trades, S, as follows.

$$
\begin{aligned}
p(S) &= \sum_{B=0}^{\infty} p(B,S) \\
&= \sum_{B=0}^{\infty}\Bigg[\alpha(1-\delta)e^{-(\mu+\varepsilon_b)}\frac{(\mu+\varepsilon_b)^B}{B!}e^{-\varepsilon_{ss}}\frac{\varepsilon_S^S}{S!} \\
&+\alpha\delta e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-(\mu+\varepsilon_{ss})}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_b}\frac{\varepsilon_b^B}{B!}e^{-\varepsilon_s}\frac{\varepsilon_S^S}{S!}\Bigg] \\
&= \alpha(1-\delta)e^{-\varepsilon_{ss}}\frac{\varepsilon_S^S}{S!}e^{-(\mu+\varepsilon_b)}\sum_{s=0}^{\infty}\frac{(\mu+\varepsilon_b)^B}{B!} \\
&+\alpha\delta e^{-(\mu+\varepsilon_{ss})}\frac{(\mu+\varepsilon_s)^S}{S!}e^{-\varepsilon_b}\sum_{s=0}^{\infty}\frac{\varepsilon_b^B}{B!}+(1-\alpha)e^{-\varepsilon_s}\frac{\varepsilon_S^S}{S!}e^{-\varepsilon_b}\sum_{s=0}^{\infty}\frac{\varepsilon_b^B}{B!} \\
&= \alpha(1-\delta)e^{-\varepsilon_{ss}}\frac{\varepsilon_S^S}{S!}+\alpha\delta e^{-(\mu+\varepsilon_{ss})}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_s}\frac{\varepsilon_S^S}{S!}
\end{aligned}
$$

$$
\begin{aligned}
E(S) &= \sum_{s=0}^{\infty} p(S)\times S \\
&= \sum_{S=0}^{\infty}\left[\alpha(1-\delta)e^{-\varepsilon_S}\frac{\varepsilon_s^S}{S!}+\alpha\delta e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^S}{S!}+(1-\alpha)e^{-\varepsilon_s}\frac{\varepsilon_s^S}{S!}\right]\times S \\
&= \sum_{S=1}^{\infty}\left[\alpha(1-\delta)e^{-\varepsilon_S}\frac{\varepsilon_s^S}{(S-1)!}+\alpha\delta e^{-(\mu+\varepsilon_s)}\frac{(\mu+\varepsilon_s)^S}{(S-1)!}+(1-\alpha)e^{-\varepsilon_s}\frac{\varepsilon_s^S}{(S-1)!}\right] \\
&= \alpha(1-\delta)e^{-\varepsilon_S}\varepsilon_s\sum_{S=1}^{\infty}\frac{\varepsilon_s^{S-1}}{(S-1)!}+\alpha\delta e^{-(\mu+\varepsilon_s)}(\mu+\varepsilon_s)\sum_{S=1}^{\infty}\frac{(\mu+\varepsilon_s)^{S-1}}{(S-1)!}+(1-\alpha)e^{-\varepsilon_S}\varepsilon_s\sum_{S=1}^{\infty}\frac{\varepsilon_s^{S-1}}{(S-1)!} \\
&= \alpha(1-\delta)\varepsilon_s+\alpha\delta(\mu+\varepsilon_s)+(1-\alpha)\varepsilon_s \\
&= \alpha\delta\mu+\varepsilon_s
\end{aligned}
$$

References

Bessembinder, Hendrik, 2003, Trade execution costs and market quality after decimalization, *Journal of Financial and Quantitative Analysis* 38, 747-777.

Brown, Stephen, Stephen A. Hillegeist, and Kin Lo, 2004, Conference calls and information asymmetry, *Journal of Accounting and Economics* 37, 343-366.

Easley, David, Robert F. Engle, Maureen O'Hara, and Liuren Wu, 2001, Time-varying arrival rates of informed and uninformed trades, Working paper, Cornell University.

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

Easley, David, Soeren Hvidkjaer, and Maureen O'Hara, 2005, Factoring information into returns, Working paper, Cornell University.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1996a, Cream-skimming or profit-sharing? The curious role of purchased order flow, *Journal of Finance* 51, 811–33.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, 1997a, The information content of the trading process, *Journal of Empirical Finance* 4, 159–186.

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

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara, and Joseph B. Paperman, 1996b, Liquidity, information, and infrequently traded stocks, *Journal of Finance* 51, 1405–1436.

Easley, David, and Maureen O'Hara, 1992, Time and the process of security price adjustment, *Journal of Finance* 47, 577-604.

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

Easley, David, Maureen O'Hara, and Joseph Paperman, 1998, Financial analysts and information-based trade, *Journal of Financial Markets* 1, 175–201.

Grinblatt, Mark, and Moskowitz J. Tobias, 2004, Predicting stock price movements from past returns: the role of consistency and tax-loss selling, *Journal of Financial Economics* 71, 541-579.

Grinblatt, Mark, and Keloharju Matti, 2004, Tax-loss trading and wash sales, *Journal of Financial Economics* 71, 51-76.

Harris, Lawrence, 1997, Decimalization: A review of the arguments and evidence, Working paper, University of Southern California.

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

Poterba, James M., and Scott J. Weisbenner, 2001, Capital gains tax rules, tax-loss trading, and turn-of-the-year returns, *Journal of Finance* 56, 353-368.

Vega, Clara, 2005, Stock price reaction to public and private information, *Journal of Financial Economics* (forthcoming).

Table I
Number and Market Share of Stocks with/without PIN Estimates

We estimate PIN for NYSE/AMEX listed stocks that have data in the TAQ database between January 1, 1993 and December 31, 2004. Stocks in the TAQ database are matched with those in the CRSP database by the historical eight-digit CUSIP. We keep only stocks with the CRSP share code 10 or 11, which means that the closed-end funds, real estate investment trusts, American depository receipts and foreign stocks are excluded. Stock are required to have the trades and quotes data for at least 50 trading days in one quarter. This table reports the number and market share of stocks that obtain or do not obtain the PIN estimates in every calendar quarter between 1993 and 2004. YYYYQ represents year and quarter. N represents the number of stocks and SHARE represent the market share. Market share is calculated as the percentage of the total market capitalization of our sample stocks at quarter end.

	Stocks with PIN obtained		Stocks with PIN missing			Stocks with PIN obtained		Stocks with PIN missing	
YYYYQ	N	SHARE	N	SHARE	YYYYQ	N	SHARE	N	SHARE
19931	1486	98.0%	1	2.0%	19991	1826	90.2%	11	9.8%
19932	1481	97.1%	2	2.9%	19992	1824	86.5%	11	13.5%
19933	1516	99.1%	1	0.9%	19993	1778	89.9%	9	10.1%
19934	1592	100.0%	0	-	19994	1823	77.5%	16	22.5%
19941	1610	100.0%	0	-	20001	1824	71.2%	22	28.8%
19942	1490	100.0%	0	-	20002	1706	76.5%	19	23.5%
19943	1483	100.0%	0	-	20003	1615	75.5%	19	24.5%
19944	1542	100.0%	0	-	20004	1644	80.2%	20	19.8%
19951	1557	99.2%	1	0.8%	20011	1571	81.0%	14	19.0%
19952	1596	100.0%	0	-	20012	1565	82.3%	15	17.7%
19953	1635	100.0%	0	-	20013	1525	78.4%	17	21.6%
19954	1637	99.1%	3	0.9%	20014	1535	76.1%	22	23.9%
19961	1694	99.3%	2	0.7%	20021	1525	80.2%	21	19.8%
19962	1733	97.8%	1	2.2%	20022	1568	75.7%	22	24.3%
19963	1689	97.7%	4	2.3%	20023	1526	69.5%	34	30.5%
19964	1777	99.4%	1	0.6%	20024	1534	69.6%	39	30.4%
19971	1800	99.2%	1	0.8%	20031	1505	59.9%	42	40.1%
19972	1820	94.8%	3	5.2%	20032	1554	56.9%	54	43.1%
19973	1915	94.5%	5	5.5%	20033	1583	61.8%	43	38.2%
19974	1923	94.6%	4	5.4%	20034	1608	61.9%	47	38.1%
19981	1901	99.0%	2	1.0%	20041	1638	63.1%	52	36.9%
19982	1905	95.3%	6	4.7%	20042	1569	61.3%	83	38.7%
19983	1876	91.3%	11	8.7%	20043	1507	66.6%	40	33.4%
19984	1876	91.9%	9	8.1%	20044	1625	58.2%	65	41.8%

Table II
Occurrence Frequency of Boundary Solutions

This table reports summary statistics on the occurrence frequency of boundary solutions in every quarter between 1993 and 2004. YYYYQ represents year and quarter. For all stocks with PIN estimates, we report the number of stocks (N), the mean (Mean) and median (Median) of the number of boundary solutions. N_0 (N_{15}) represents the number of stocks that have no (more than 15) boundary solutions out of the 125 runs. For stocks in these two categories, we report the proportion of each category accounting for the total number of stocks (%), the average market capitalization in billions of dollar (Market Cap) and the average number of daily trades (# Trades).

YYYYQ	All stocks with PIN estimates			Stocks with no boundary solutions				Stocks with more than 15 boundary solutions			
	N	Mean	Median	N_0	%	Market Cap	# Trades	N_{15}	%	Market Cap	# Trades
19931	1486	15.9	8	243	16.4%	4.97	213	431	29.0%	2.01	66
19932	1481	17.5	9	212	14.3%	4.01	165	481	32.5%	2.33	66
19933	1516	19.9	9	200	13.2%	4.94	181	538	35.5%	2.17	60
19934	1592	17.9	9	185	11.6%	5.54	206	493	31.0%	2.16	61
19941	1610	18.4	10	206	12.8%	6.02	230	577	35.8%	1.74	66
19942	1490	20.3	11	152	10.2%	6.90	240	578	38.8%	1.91	63
19943	1483	21.7	12	153	10.3%	6.17	202	618	41.7%	1.95	60
19944	1542	17.1	10	197	12.8%	6.31	226	491	31.8%	1.75	61
19951	1557	22.2	12	175	11.2%	6.02	214	637	40.9%	1.92	61
19952	1596	20.5	9	204	12.8%	7.53	261	561	35.2%	1.72	58
19953	1635	17.3	8	239	14.6%	7.09	224	480	29.4%	1.84	59
19954	1637	17.2	8	259	15.8%	7.72	229	515	31.5%	2.10	61
19961	1694	18.7	9	259	15.3%	9.21	286	547	32.3%	1.69	63
19962	1733	17.8	9	260	15.0%	8.31	253	574	33.1%	2.34	71
19963	1689	15.6	8	248	14.7%	8.45	224	472	27.9%	2.09	64
19964	1777	17.2	8	275	15.5%	9.30	284	530	29.8%	2.08	65
19971	1800	20.0	10	250	13.9%	11.37	348	672	37.3%	1.99	76
19972	1820	19.1	9	271	14.9%	10.97	326	589	32.4%	2.18	79
19973	1915	19.3	9	257	13.4%	11.36	397	668	34.9%	2.27	86
19974	1923	14.1	7	326	17.0%	11.59	388	481	25.0%	1.85	74
19981	1901	18.1	10	296	15.6%	14.59	433	691	36.3%	2.30	88
19982	1905	17.8	9	277	14.5%	15.49	444	658	34.5%	2.53	96
19983	1876	13.5	8	348	18.6%	12.25	457	516	27.5%	1.47	84
19984	1876	13.3	6	382	20.4%	12.08	441	445	23.7%	1.76	92
19991	1826	15.8	8	347	19.0%	14.24	542	564	30.9%	1.27	92
19992	1824	12.9	5	439	24.1%	12.61	452	401	22.0%	1.48	90
19993	1778	12.5	6	403	22.7%	12.34	439	419	23.6%	1.47	84
19994	1823	12.6	5	417	22.9%	11.15	499	403	22.1%	1.16	87

Table II (continued)
Occurrence Frequency of Boundary Solutions

YYYYQ	All stocks with PIN estimates			Stocks with no boundary solutions				Stocks with more than 15 boundary solutions			
	N	Mean	Median	N_1	%	Market Cap	# Trades	N_2	%	Market Cap	# Trades
20001	1824	13.8	5	491	26.9%	10.04	551	439	24.1%	1.25	102
20002	1706	11.6	4	451	26.4%	11.97	534	351	20.6%	1.49	113
20003	1615	13.0	5	422	26.1%	12.36	527	388	24.0%	1.91	113
20004	1644	10.8	3	498	30.3%	12.49	564	294	17.9%	1.68	129
20011	1571	7.3	2	600	38.2%	10.55	654	207	13.2%	0.84	93
20012	1565	5.4	1	722	46.1%	9.53	640	113	7.2%	0.70	97
20013	1525	5.3	1	679	44.5%	7.82	648	118	7.7%	0.70	108
20014	1535	5.3	1	734	47.8%	8.02	694	94	6.1%	0.93	132
20021	1525	5.5	1	722	47.3%	8.64	773	119	7.8%	0.93	117
20022	1568	3.6	0	847	54.0%	6.69	768	66	4.2%	0.52	105
20023	1526	3.0	0	932	61.1%	4.46	815	43	2.8%	0.44	76
20024	1534	3.4	0	852	55.5%	4.79	860	60	3.9%	0.57	96
20031	1505	3.3	0	887	58.9%	4.16	911	65	4.3%	0.46	136
20032	1554	3.4	0	982	63.2%	4.20	891	58	3.7%	0.48	107
20033	1583	2.6	0	1034	65.3%	4.20	854	48	3.0%	0.36	108
20034	1608	2.5	0	1045	65.0%	4.72	851	41	2.6%	0.38	100
20041	1638	3.1	0	1047	63.9%	4.51	913	64	3.9%	0.46	123
20042	1569	2.5	0	1027	65.5%	4.69	950	49	3.1%	2.87	205
20043	1507	2.6	0	942	62.5%	4.89	958	45	3.0%	2.97	154
20044	1625	3.5	0	1001	61.6%	4.58	981	70	4.3%	0.76	184

Table III
Logistic Regression Models

We study three logistic regression models. The dependent variables of all three models equals 1 for stocks having more than 15 boundary solutions and 0 for stocks having zero boundary solutions. The three models have different explanatory variables; Model 1 has a single variable equal to the logarithm of the market capitalization (i.e. log(mktcap)), Model 2 has a single variable equal to the logarithm of the number of trades in a day (i.e. log(trade)), and Model 3 include both variables. We report the coefficient estimates and the residual deviation. We omit the intercepts of the three models because they have no meaningful implications for our current analysis. YYYYQ represents year and quarter. N_0 is the number of stocks that have no boundary solutions. N_{15} is the number of stocks that have more that 15 boundary solutions. RD stands for residual deviance. The superscripts indicate the results from the t-test for the coefficient estimate and the chi-square test for the residual deviance. The superscripts, ^{a}, ^{b}, and ^{c}, represent statistical significance at 1%, 5%, and 10%, respectively.

			Model 1		Model 2		Model 3		
YYYYQ	N_0	N_{15}	log(mktcap)	RD	log(trade)	RD	log(mktcap)	log(trade)	RD
19931	243	431	0.137^a	872.0^a	-0.341^a	857.0^a	1.103^a	-1.794^a	702.8
19932	211	481	0.207^a	832.2^a	-0.263^a	838.7^a	1.107^a	-1.748^a	686.7
19933	199	537	0.065	857.3^a	-0.380^a	832.9^a	0.771^a	-1.368^a	749.8
19934	185	493	-0.007	794.7^a	-0.486^a	752.6^b	0.696^a	-1.380^a	683.7
19941	206	577	-0.049	901.4^a	-0.475^a	858.6^b	0.719^a	-1.406^a	793.7
19942	152	578	-0.150^a	739.9	-0.677^a	677.9	0.843^a	-1.729^a	616.8
19943	152	618	-0.042	764.5	-0.554^a	721.2	0.864^a	-1.651^a	651.6
19944	196	491	-0.051	820.5^a	-0.612^a	759.4^b	0.888^a	-1.762^a	672.0
19951	175	634	0.032	844.5	-0.477^a	807.3	0.843^a	-1.575^a	719.3
19952	204	561	-0.135^a	879.5^a	-0.683^a	796.1	0.753^a	-1.622^a	723.3
19953	239	479	-0.104^a	908.5^a	-0.601^a	839.7^a	0.574^a	-1.299^a	786.2^b
19954	259	514	-0.076^c	983.0^a	-0.624^a	903.9^a	0.782^a	-1.620^a	814.8
19961	259	546	-0.190^a	991.9^a	-0.691^a	896.6^a	0.764^a	-1.685^a	820.0
19962	260	574	-0.031	1034.5^a	-0.534^a	969.0^a	0.727^a	-1.472^a	877.9
19963	247	472	-0.068	923.0^a	-0.552^a	862.2^a	0.750^a	-1.503^a	787.4^b
19964	274	529	-0.141^a	1019.7^a	-0.699^a	915.4^a	0.762^a	-1.684^a	826.7
19971	250	671	-0.199^a	1056.1^a	-0.651^a	969.8	0.654^a	-1.480^a	909.7
19972	270	586	-0.164^a	1052.8^a	-0.526^a	993.9^a	0.641^a	-1.332^a	942.9^b
19973	254	668	-0.212^a	1061.7^a	-0.644^a	975.6^c	0.709^a	-1.542^a	912.0
19974	325	476	-0.179^a	1061.1^a	-0.735^a	931.4^a	0.852^a	-1.853^a	824.6
19981	295	690	-0.195^a	1180.1^a	-0.657^a	1081.0^b	0.733^a	-1.587^a	1003.6
19982	276	654	-0.281^a	1087.2^a	-0.766^a	981.7	0.610^a	-1.550^a	931.7
19983	347	513	-0.386^a	1074.7^a	-1.056^a	894.4	0.687^a	-1.916^a	834.2
19984	382	445	-0.230^a	1104.6^a	-0.804^a	949.7^a	0.710^a	-1.703^a	865.2

Table III (Continued)
Logistic Regression Models

YYYYQ	N_0	N_{15}	log(mktcap)	RD	log(trade)	RD	log(mktcap)	log(trade)	RD
19991	344	563	-0.361[a]	1122.8[a]	-0.983[a]	945.1	0.763[a]	-1.968[a]	871.5
19992	436	400	-0.347[a]	1076.1[a]	-0.864[a]	943.5[a]	0.639[a]	-1.696[a]	891.6[c]
19993	400	418	-0.311[a]	1073.6[a]	-0.832[a]	935.7[a]	0.797[a]	-1.833[a]	861.3
19994	414	402	-0.411[a]	1023.6[a]	-0.910[a]	889.5[b]	0.537[a]	-1.579[a]	853.3
20001	488	437	-0.329[a]	1193.6[a]	-0.848[a]	1037.2[a]	0.632[a]	-1.664[a]	975.1
20002	448	348	-0.360[a]	1009.5[a]	-0.776[a]	912.9[a]	0.480[a]	-1.378[a]	884.5[a]
20003	421	388	-0.366[a]	1038.0[a]	-0.755[a]	952.1[a]	0.367[a]	-1.215[a]	934.5[a]
20004	496	292	-0.256[a]	987.7[a]	-0.639[a]	908.2[a]	0.373[a]	-1.133[a]	883.5[a]
20011	597	207	-0.457[a]	806.2	-0.986[a]	688.1	0.646[a]	-1.824[a]	652.2
20012	720	113	-0.478[a]	582.5	-0.855[a]	528.8	0.401[a]	-1.333[a]	519.3
20013	679	118	-0.560[a]	574.2	-1.004[a]	507.6	0.340[a]	-1.398[a]	500.6
20014	732	93	-0.409[a]	529.6	-0.800[a]	474.5	0.419[a]	-1.272[a]	462.5
20021	719	119	-0.547[a]	583.8	-0.974[a]	504.3	0.580[a]	-1.650[a]	485.5
20022	846	66	-0.552[a]	407.4	-0.869[a]	372.5	0.076	-0.946[a]	372.2
20023	932	43	-0.656[a]	295.8	-1.058[a]	257.6	0.218	-1.273[a]	256.0
20024	850	60	-0.589[a]	374.2	-1.008[a]	313.3	0.297[a]	-1.295[a]	308.5
20031	883	65	-0.649[a]	398.9	-1.016[a]	350.7	0.226	-1.243[a]	348.5
20032	982	58	-0.729[a]	364.2	-1.116[a]	309.0	0.429[a]	-1.543[a]	303.1
20033	1030	48	-0.722[a]	320.0	-1.001[a]	288.4	0.007	-1.007[a]	288.4
20034	1042	41	-0.588[a]	300.3	-0.985[a]	267.9	0.221	-1.216[a]	266.1
20041	1046	62	-0.689[a]	394.1	-1.044[a]	342.3	0.159	-1.196[a]	341.1
20042	1025	49	-0.493[a]	363.0	-0.769[a]	337.4	0.117	-0.880[a]	336.7
20043	941	45	-0.623[a]	318.2	-0.944[a]	276.8	0.310[c]	-1.231[a]	273.2
20044	998	69	-0.588[a]	443.7	-0.869[a]	400.2	0.212	-1.073[a]	397.9

Table IV
Seasonal Variation of PIN

This table reports the mean and median of quarterly percentage changes of six variables: the probability of informed trading (PIN), the probability of an information event α, the daily arrival rate of informed trades μ, the daily arrival rates of uninformed buy trades ε_b, the daily arrival rates of uninformed sell trades ε_s, and the ratio of the uninformed trades to the informed trades $(\varepsilon_b + \varepsilon_s)/\mu$. The quarterly percentage change of a variable V is defined as $\Delta V_t = (V_t - V_{t-1})/V_{t-1}$, where t indexes calendar quarters. We group quarterly percentage changes under the four quarters of a year. The mean and median are calculated after removing the most extreme 1% observations at both tails. N represents the number of observations. The time period extends from 1993Q2 to 2004Q4. The t-statistic and the sign statistic are reported in parentheses. The superscripts, [a], [b], and [c], represent statistical significance at 1%, 5%, and 10%, respectively.

	Mean quarterly change % (t-statistic)				Median quarterly change % (sign statistic)			
Quarter	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
N	16957	18277	18075	18280	16957	18277	18075	18280
$\Delta\mu$	14.8 (37.8)[a]	13.2 (34.6)[a]	7.8 (22.4)[a]	15.2 (41.7)[a]	4.8 (892.5)[a]	3.2 (653.0)[a]	-0.7 (-143.5)[a]	5.1 (1088.0)[a]
$\Delta\varepsilon_b$	18.4 (62.5)[a]	6.8 (25.0)[a]	5.0 (19.4)[a]	13.1 (52.2)[a]	13.0 (2846.5)[a]	0.9 (284.5)[a]	-0.4 (-124.0)[c]	7.8 (2192.5)[a]
$\Delta\varepsilon_s$	16.7 (63.4)[a]	4.9 (20.3)[a]	4.5 (19.8)[a]	13.7 (60.2)[a]	12.5 (2891.5)[a]	-0.6 (-171.5)[b]	-0.4 (-115.5)[c]	8.6 (2616.5)[a]
$\Delta\frac{(\varepsilon_b+\varepsilon_s)}{\mu}$	12.1 (44.4)[a]	2.7 (10.9)[a]	6.0 (24.4)[a]	7.7 (30.9)[a]	7.2 (1578.5)[a]	-2.2 (-574.0)[a]	1.2 (284.5)[a]	2.8 (720.5)[a]
ΔPIN	-1.6 (-7.2)[a]	4.9 (21.6)[a]	4.2 (18.7)[a]	4.3 (19.3)[a]	-5.9 (-1380.5)[a]	0.3 (90.5)	0.1 (17.5)	-0.1 (-25.0)
$\Delta\alpha$	9.5 (24.6)[a]	8.1 (22.2)[a]	10.8 (28.8)[a]	12.4 (33.6)[a]	0.0 (-8.5)	-1.2 (-262.0)[a]	1.3 (253.0)[a]	2.9 (630.5)[a]

Table V
Year-End Tax-Loss Selling and Seasonal Variation of PIN

We calculate the cumulative 10-month return from February to November in each year and classify a stock as winner if its cumulative return is greater than 10% or loser if the return is less than -10%. We study the quarterly percentage changes of PIN and other trading-related parameters in the fourth quarter and the first quarter of the following year. The quarterly percentage change of a variable V is defined as $\Delta V_t = (V_t - V_{t-1})/V_{t-1}$, where t indexes calendar quarters. This table reports the mean and median of the quarterly percentage changes by quarter and type of stock. The mean and median are calculated after removing the most extreme 1% observations at both tails. N represents the number of observations. The t-statistic and sign statistic are reported in parentheses. The two-sample t-test and Wilcoxon rank sum test are reported in parentheses under the column "L vs W". The superscripts, ^{a}, ^{b}, and ^{c}, represent statistical significance at 1%, 5%, and 10%, respectively.

		Fourth Quarter			Subsequent First Quarter		
		Losers	Winners	L vs W	Losers	Winners	L vs W
N		5117	8349		4763	7331	
$\Delta\mu$	Mean %	19.4	14.5	$(5.24)^a$	9.4	16.5	$(-7.72)^a$
	(t-stat)	$(26.1)^a$	$(27.5)^a$		$(12.7)^a$	$(29.3)^a$	
	Median %	8.0	5.0	$(3.62)^a$	-1.1	7.4	$(-10.94)^a$
	(sign)	$(431.0)^a$	$(505.5)^a$		$(-67.0)^b$	$(651.0)^a$	
$\Delta\varepsilon_s$	Mean %	18.6	12.6	$(10.12)^a$	4.0	24.3	$(-30.89)^a$
	(t-stat)	$(40.0)^a$	$(37.4)^a$		$(8.7)^a$	$(60.2)^a$	
	Median %	12.7	7.4	$(10.24)^a$	-0.2	19.5	$(-33.43)^a$
	(sign)	$(995.0)^a$	$(1030.0)^a$		(-10.5)	$(1963.0)^a$	
$\Delta\varepsilon_b$	Mean %	14.9	14.1	(1.15)	6.9	25.8	$(-25.00)^a$
	(t-stat)	$(30.3)^a$	$(36.4)^a$		$(13.2)^a$	$(56.0)^a$	
	Median %	9.2	8.4	(1.21)	1.6	19.5	$(-27.72)^a$
	(sign)	$(664.0)^a$	$(1092.0)^a$		$(89.5)^a$	$(1843.0)^a$	
$\Delta\frac{(\varepsilon_b+\varepsilon_s)}{\mu}$	Mean %	7.8	7.9	(-0.61)	6.3	16.3	$(-13.88)^a$
	(t-stat)	$(15.8)^a$	$(21.5)^a$		$(12.8)^a$	$(39.7)^a$	
	Median %	2.2	3.0	(-1.36)	2.1	10.9	$(-14.92)^a$
	(sign)	$(166.5)^a$	$(347.0)^a$		$(117.0)^a$	$(1060.5)^a$	
ΔPIN	Mean %	6.2	3.1	$(5.73)^a$	0.2	-3.7	$(7.09)^a$
	(t-stat)	$(14.5)^a$	$(9.5)^a$		(0.5)	$(-11.6)^a$	
	Median %	1.9	-1.2	$(5.65)^a$	-3.9	-7.7	$(6.86)^a$
	(sign)	$(146.5)^a$	$(-150.5)^a$		$(-253.5)^a$	$(-810.5)^a$	
$\Delta\alpha$	Mean %	14.9	10.7	$(4.24)^a$	6.9	10.5	$(-2.64)^a$
	(t-stat)	$(20.4)^a$	$(20.5)^a$		$(9.3)^a$	$(18.5)^a$	
	Median %	5.0	1.6	$(3.74)^a$	-2.4	1.0	$(-4.70)^a$
	(sign)	$(267.5)^a$	$(178.0)^a$		$(4763)^a$	$(83.0)^b$	

Table VI
Comparison between our and Brown's PIN Estimates

This table reports the mean, median and standard deviation of Brown's PIN estimates and ours in every quarter between 1993 and 2003. It also reports the mean, median and standard deviation of the same-stock difference between ours and Brown's PIN estimates. The statistics are calculated over all stocks that have PIN estimates in both our and Brown's sets. We match stocks in our sample with those in Brown's by eight-digit CUSIP. N equals the number of stocks. YYYYQ represents year and quarter.

		Our PIN			Brown's PIN			Ours – Brown's		
YYYYQ	N	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
19931	1408	0.173	0.169	0.060	0.186	0.180	0.065	-0.013	-0.012	0.052
19932	1402	0.169	0.165	0.061	0.185	0.178	0.068	-0.016	-0.014	0.056
19933	1436	0.166	0.161	0.060	0.181	0.173	0.065	-0.015	-0.011	0.055
19934	1520	0.170	0.168	0.061	0.186	0.180	0.065	-0.016	-0.015	0.056
19941	1535	0.167	0.159	0.063	0.180	0.175	0.065	-0.014	-0.012	0.054
19942	1428	0.164	0.157	0.060	0.181	0.170	0.066	-0.017	-0.014	0.055
19943	1404	0.165	0.159	0.064	0.184	0.175	0.069	-0.019	-0.015	0.057
19944	1453	0.167	0.163	0.052	0.185	0.177	0.062	-0.018	-0.017	0.051
19951	1456	0.161	0.156	0.061	0.188	0.179	0.067	-0.027	-0.023	0.060
19952	1478	0.162	0.156	0.059	0.186	0.177	0.066	-0.025	-0.021	0.055
19953	1500	0.166	0.160	0.057	0.188	0.179	0.068	-0.023	-0.017	0.054
19954	1520	0.168	0.163	0.062	0.190	0.182	0.067	-0.022	-0.016	0.058
19961	1601	0.166	0.159	0.060	0.180	0.170	0.068	-0.014	-0.012	0.056
19962	1641	0.164	0.157	0.062	0.178	0.168	0.066	-0.014	-0.012	0.053
19963	1605	0.171	0.166	0.059	0.186	0.179	0.065	-0.015	-0.012	0.048
19964	1690	0.169	0.163	0.060	0.185	0.176	0.064	-0.016	-.013	0.059
19971	1718	0.159	0.152	0.061	0.176	0.168	0.067	-0.018	-0.013	0.052
19972	1740	0.163	0.153	0.065	0.177	0.168	0.068	-0.015	-0.010	0.057
19973	1828	0.152	0.146	0.061	0.159	0.148	0.063	-0.006	-0.002	0.046
19974	1838	0.158	0.153	0.059	0.165	0.157	0.065	-0.007	-0.003	0.045
19981	1809	0.147	0.139	0.059	0.152	0.141	0.063	-0.005	-0.002	0.042
19982	1829	0.150	0.144	0.058	0.158	0.150	0.063	-0.008	-0.004	0.045
19983	1799	0.154	0.149	0.058	0.160	0.153	0.064	-0.006	-0.002	0.046
19984	1790	0.155	0.152	0.057	0.160	0.151	0.065	-0.004	0.000	0.045

Table VI (Continued)
Comparison between our and Brown's PIN Estimates

		Our PIN			Brown's PIN			Ours – Brown's		
YYYYQ	N	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.	Mean	Median	Std. Dev.
19991	1621	0.143	0.136	0.057	0.172	0.146	0.101	-0.029	-0.005	0.100
19992	1596	0.154	0.148	0.055	0.172	0.153	0.090	-0.017	-0.002	0.090
19993	1553	0.152	0.146	0.056	0.174	0.150	0.097	-0.022	-0.004	0.098
19994	1599	0.159	0.151	0.060	0.179	0.158	0.092	-0.020	-0.004	0.089
20001	1564	0.143	0.132	0.063	0.169	0.142	0.103	-0.026	-0.005	0.103
20002	1509	0.145	0.135	0.061	0.181	0.151	0.107	-0.036	-0.008	0.112
20003	1437	0.147	0.138	0.061	0.180	0.154	0.100	-0.032	-0.011	0.103
20004	1482	0.150	0.143	0.060	0.181	0.156	0.102	-0.031	-0.006	0.103
20011	1445	0.153	0.141	0.062	0.161	0.140	0.085	-0.008	0.006	0.077
20012	1459	0.159	0.146	0.064	0.165	0.144	0.086	-0.006	0.006	0.069
20013	1441	0.152	0.142	0.061	0.160	0.139	0.083	-0.008	0.004	0.069
20014	1439	0.153	0.139	0.066	0.164	0.142	0.089	-0.011	0.003	0.074
20021	1460	0.148	0.134	0.066	0.147	0.127	0.076	0.002	0.007	0.038
20022	1504	0.145	0.132	0.065	0.142	0.124	0.073	0.003	0.006	0.035
20023	1478	0.145	0.134	0.065	0.139	0.122	0.075	0.006	0.011	0.036
20024	1448	0.141	0.126	0.065	0.137	0.118	0.073	0.004	0.008	0.039
20031	1419	0.134	0.120	0.063	0.128	0.110	0.067	0.006	0.009	0.032
20032	1441	0.139	0.122	0.068	0.134	0.114	0.074	0.005	0.009	0.039
20033	1470	0.139	0.125	0.066	0.132	0.111	0.069	0.007	0.009	0.036
20034	1497	0.137	0.127	0.064	0.134	0.117	0.072	0.002	0.008	0.039

Table VII
PIN Difference for Two Special Groups of Stocks

This table reports the mean, median and standard deviation of the same-stock difference between our and Brown's PIN estimates for two special groups of stocks, in every quarter between 1993 and 2003. Stocks in the first group have no boundary solutions in the 125 maximizations we ran for each of them, whereas stocks in the second group have boundary solutions for more than 15 of the 125 maximizations. The statistics are calculated over all stocks that have PIN estimates in both our and Brown's sets. We match stocks in our sample with those in Brown's by eight-digit CUSIP. N equals the number of stocks. YYYYQ represents year and quarter.

	Stocks with no boundary solutions				Stocks with more than 15 boundary solutions			
YYYYQ	N	Mean	Median	Std. Dev.	N	Mean	Median	Std. Dev.
19931	225	-0.013	-0.013	0.042	412	-0.018	-0.016	0.060
19932	199	-0.013	-0.013	0.043	459	-0.023	-0.018	0.066
19933	189	-0.008	-0.008	0.045	514	-0.020	-0.015	0.063
19934	172	-0.006	-0.007	0.048	477	-0.022	-0.020	0.065
19941	188	-0.011	-0.010	0.037	553	-0.018	-0.015	0.066
19942	146	-0.023	-0.017	0.048	555	-0.018	-0.016	0.064
19943	143	-0.017	-0.013	0.050	586	-0.021	-0.017	0.064
19944	179	-0.017	-0.015	0.043	465	-0.021	-0.017	0.052
19951	156	-0.030	-0.016	0.062	605	-0.028	-0.025	0.069
19952	173	-0.012	-0.009	0.046	529	-0.031	-0.026	0.064
19953	208	-0.012	-0.007	0.038	440	-0.030	-0.025	0.064
19954	247	-0.014	-0.010	0.052	466	-0.025	-0.018	0.075
19961	243	-0.009	-0.010	0.040	518	-0.017	-0.016	0.070
19962	243	-0.012	-0.010	0.043	545	-0.019	-0.014	0.065
19963	236	-0.019	-0.012	0.045	451	-0.018	-0.014	0.056
19964	256	-0.021	-0.018	0.042	510	-0.019	-0.014	0.085
19971	232	-0.008	-0.007	0.043	654	-0.021	-0.016	0.065
19972	256	-0.010	-0.003	0.048	573	-0.014	-0.011	0.071
19973	245	0.000	0.002	0.041	641	-0.010	-0.007	0.056
19974	306	0.002	0.004	0.037	458	-0.012	-0.005	0.055
19981	280	-0.001	0.001	0.043	663	-0.008	-0.005	0.040
19982	263	-0.003	0.003	0.038	637	-0.008	-0.004	0.051
19983	333	-0.001	0.002	0.036	502	-0.007	-0.006	0.054
19984	355	0.002	0.007	0.039	429	-0.009	-0.002	0.054

Table VII (continued)
PIN Difference for Two Special Groups of Stocks

	Stocks with no boundary solutions				Stocks with more than 15 boundary solutions			
YYYYQ	N	Mean	Median	Std. Dev.	N	Mean	Median	Std. Dev.
19991	292	-0.034	-0.005	0.104	518	-0.038	-0.006	0.113
19992	363	-0.022	-0.005	0.096	358	-0.022	-0.002	0.096
19993	332	-0.036	-0.008	0.104	383	-0.023	-0.002	0.115
19994	353	-0.033	-0.005	0.107	366	-0.019	-0.002	0.095
20001	399	-0.037	-0.003	0.112	392	-0.029	-0.006	0.110
20002	386	-0.054	-0.006	0.131	319	-0.031	-0.005	0.117
20003	361	-0.060	-0.010	0.135	357	-0.032	-0.013	0.097
20004	445	-0.057	-0.005	0.135	273	-0.025	-0.010	0.085
20011	549	-0.009	0.010	0.084	197	-0.015	0.000	0.095
20012	668	-0.005	0.008	0.073	104	-0.022	-0.006	0.083
20013	647	-0.009	0.007	0.078	112	-0.001	0.000	0.045
20014	685	-0.014	0.006	0.087	89	-0.015	0.003	0.074
20021	699	0.008	0.011	0.035	108	-0.002	-0.003	0.051
20022	822	0.006	0.008	0.031	62	-0.001	0.004	0.045
20023	908	0.010	0.014	0.033	37	-0.003	0.002	0.038
20024	820	0.007	0.010	0.036	51	-0.008	0.005	0.080
20031	846	0.007	0.010	0.029	61	0.014	0.013	0.035
20032	928	0.008	0.011	0.031	51	-0.015	-0.002	0.086
20033	975	0.008	0.010	0.032	41	0.002	0.003	0.073
20034	979	0.004	0.009	0.036	37	0.008	0.014	0.061

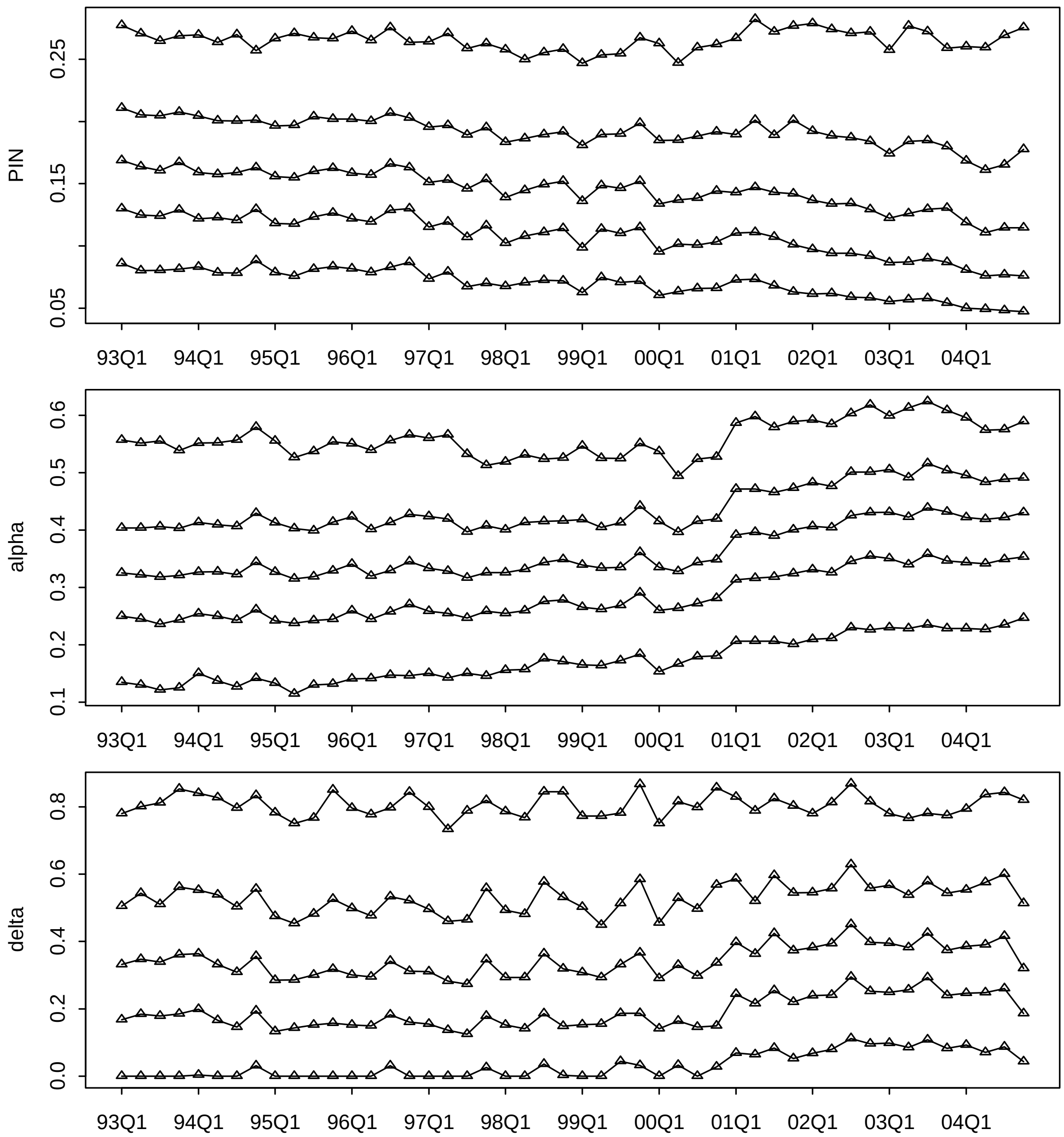

Figure 1. Time-series plots in every quarter between 1993Q1 and 2004Q4 for the 5th, 25th, 50th (i.e., median), 75th, and 95th percentiles of the probability of informed trading (PIN), the probability of an information event α, and the probability of an information event being bad news δ.

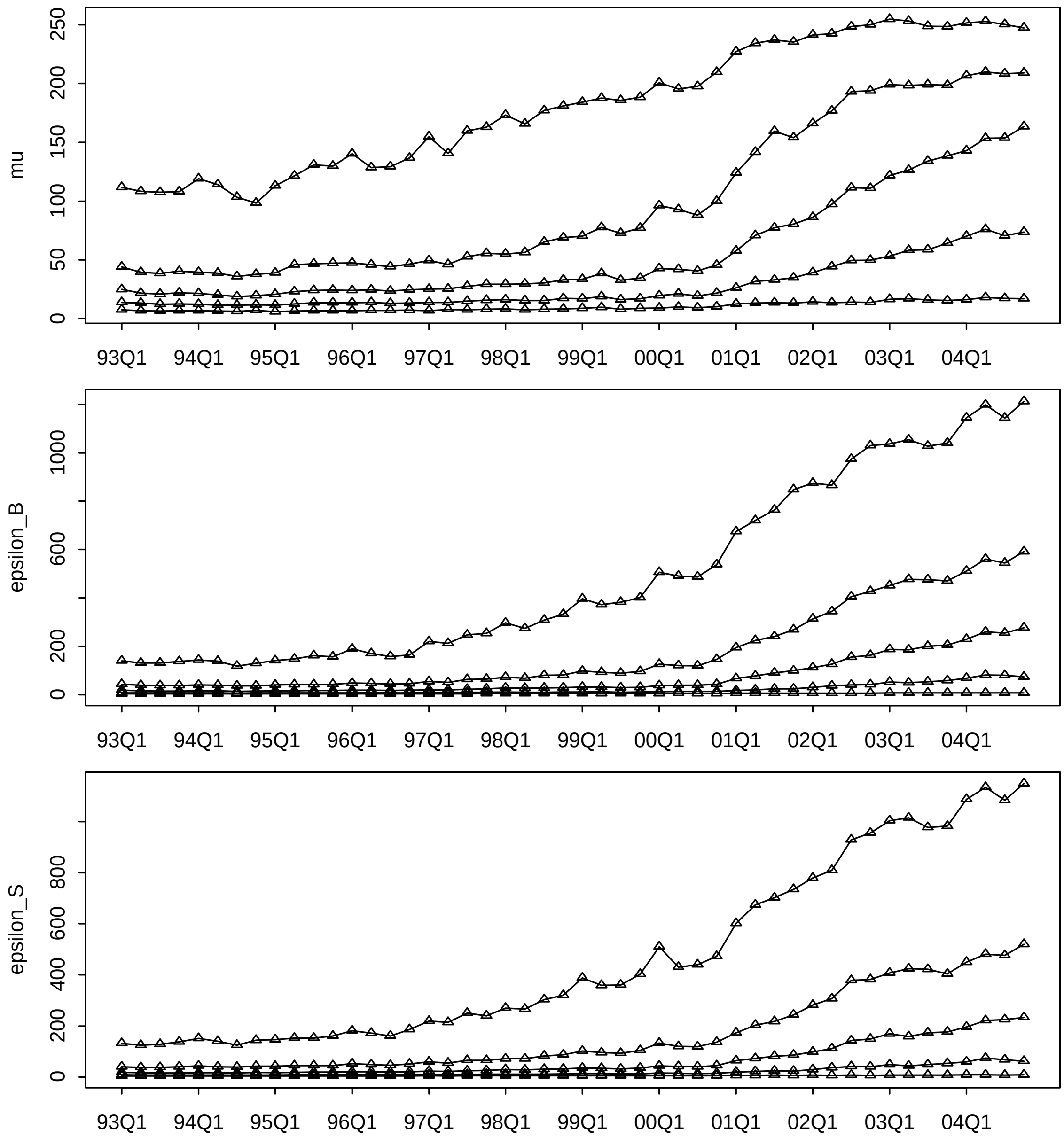

Figure 2. Time-series plots in every quarter between 1993Q1 and 2004Q4 for the 5th, 25th, 50th (i.e., median), 75th, and 95th percentiles of the arrival rate of informed trades per day μ, and the arrival rates of uninformed buy and sell trades per day ε_b and ε_s.

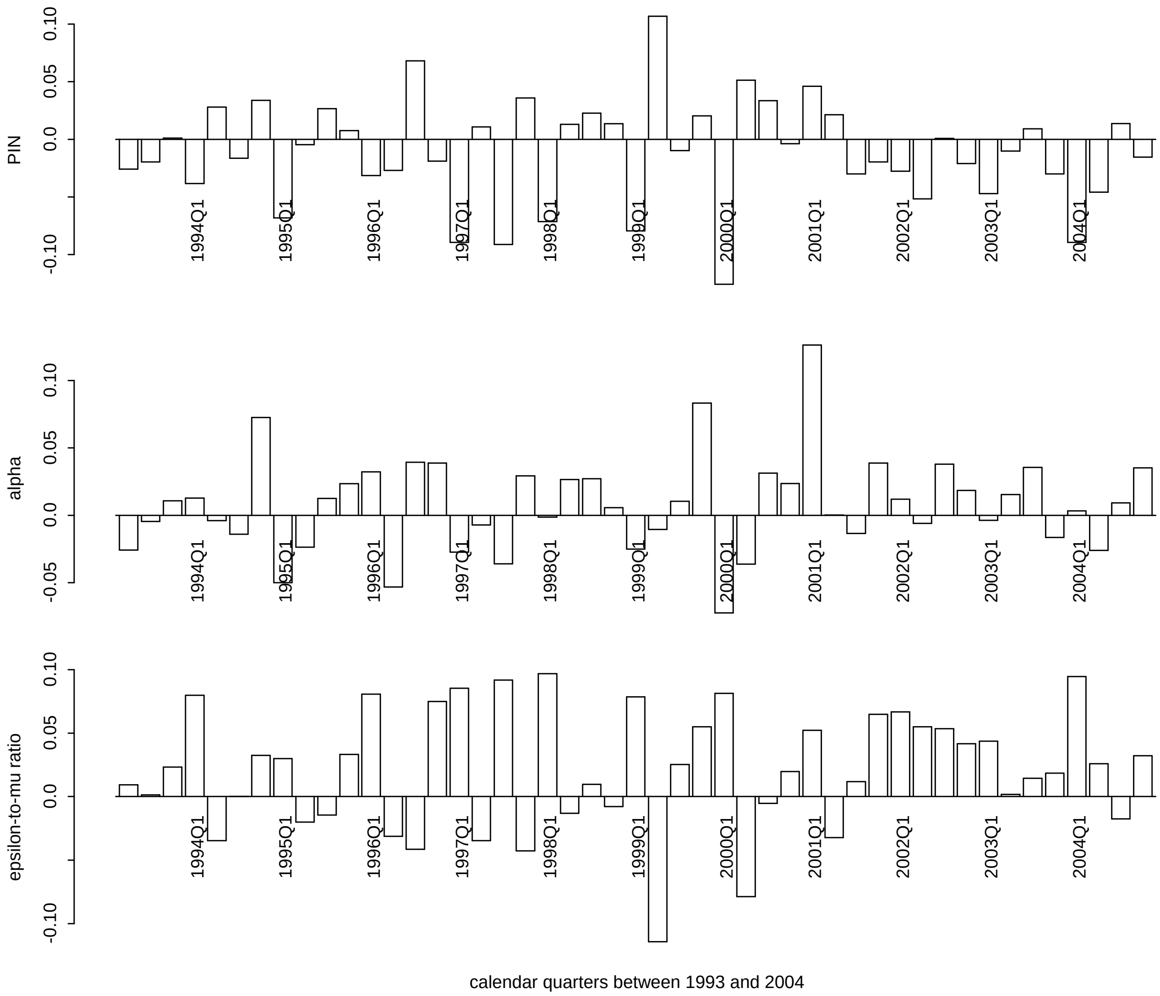

Figure 3. Bar plots in every quarter between 1993Q2 and 2004Q4 for the median quarterly percentage changes of the probability of informed trading (PIN), the probability of an information event α, and the ratio of the daily arrival rate of uninformed trades over the daily arrival rate of informed trades $\left(\varepsilon_b + \varepsilon_s\right)/\mu$. The quarterly percentage change of a variable V is defined as $\Delta V_t = (V_t - V_{t-1})/V_{t-1}$, where t indexes calendar quarters.