Journal of Financial Econometrics, 2005, Vol. 3, No. 4, 500–524

Inferring Information Frequency and Quality

JOHN OWENS
Victoria University of Wellington

DOUGLAS G. STEIGERWALD
University of California, Santa Barbara

ABSTRACT
We develop a microstructure model that, in contrast to previous models, allows one to estimate the frequency and quality of private information. In addition, the model produces stationary asset price and trading volume series. We find evidence that information arrives frequently within a day and that this information is of high quality. The frequent arrival of information, while in contrast to previous microstructure model estimates, accords with nonmodel-based estimates and the related literature testing the mixture-of-distributions hypothesis. To determine if the estimates are correctly reflecting the arrival of latent information, we estimate the parameters over half-hour intervals within the day. Comparison of the parameter estimates with measures of persistent price changes reveals that the estimates reflect the arrival of latent information.

KEYWORDS: asymmetric information, high-frequency econometrics

An attraction of market microstructure models is that they allow one to assess the empirical importance of private information in security markets. The asymmetric information model developed by Glosten and Milgrom (1985) to explain the presence of bid-ask spreads over a single trading period was extended to multiple trading periods by Easley and O'Hara (1992). With a model of multiple trading periods, succeeding articles estimated the impact of privately informed traders on price determination [see Easley et al. (1996), Easley, Kiefer, and O'Hara (1997), Easley, O'Hara and Saar (2001), Hanousek and Podpiera (2002), and Kelly and Steigerwald (2004)]. While certain features of informed traders are estimable from the model, both the frequency of (private) information arrival and the accuracy of information must be assumed. As the above mentioned empirical papers assume that information arrives at most once per day, while estimates obtained directly from buy and sell order flows [Hasbrouck (1999)] find that information arrives many times within a day, misspecification bias of the microstructure estimates is a very real possibility. We address these issues

doi:10.1093/jjfinec/nbi024
Advance Access publication August 12, 2005

with a model that allows us to estimate both the frequency and the quality of private information.

The microstructure model of Easley and O'Hara is inherently nonstationary, as the potential arrival of information occurs at fixed (and known) points in time. Estimation of the model requires that the researcher specify the fixed intervals that correspond to the arrival of information. Typically researchers assume that information potentially arrives at the beginning of each trading day, thereby limiting the frequency of information arrival to once per day. Such an assumption is in contrast not only to the findings of Hasbrouck, but also to the assumption that underpins research about the mixture-of-distributions hypothesis (MDH). Empirical assessment of the MDH, which describes the distribution of security prices arising from the presence of informed traders, is typically based on the assumption of many information arrivals within a day [e.g., Andersen (1996)]. We address the issue by constructing a model (of the type in Easley and O'Hara) in which information arrives randomly throughout each trading day (in so doing, the model is stationary as well). As a result, the length of time over which information persists is random, in accordance with the different types of information that enter asset markets.

In addition, we allow for information of varying accuracy. In the model of Easley and O'Hara, information is perfect by assumption. Yet it may be the case that not all information is created equally. We address this issue by explicitly modeling the belief a trader has in the accuracy of information.[1] By allowing the quality of information to vary, we also address another issue. In Easley and O'Hara, the group of informed traders all receive (perfect) information simultaneously. The potential for strategic behavior by the informed is eliminated through the random arrival of traders.[2] While eliminating strategic behavior, the mechanism imparts perfect correlation at the microstructure level, as the information received by one informed trader is the same as the information received by the next informed trader. By allowing the quality of news to vary, we are able to estimate the correlation of private information and so more accurately gauge the impact of information on stochastic volatility in asset prices.

In Section 1, we present the asymmetric information microstructure model for a security market. A period of asymmetric information ends with the arrival of public news. (Not all news about the asset is privately revealed, with positive probability public news will not have been previously revealed to informed traders.) Following each public news arrival is the possible arrival of (private) information. The parameter governing the public news arrival process determines the average length of time over which informed traders exploit their information and contributes to the frequency of information arrivals. We capture the varying accuracy of information through a Markov transition matrix that governs the probability that a given information signal to a trader will be publicly revealed (and hence, accurate). The model is flexible enough to allow information accuracy

[1] Damodaran (1985) shows how the accuracy of information affects the variance of returns.
[2] Kyle (1985) considers the strategic behavior of a single informed trader.

to depend on whether the information reflects positively or negatively on the asset price.

We detail how to estimate the model in Section 2. As the fundamental data are the latent individual trade decisions, we first describe how to construct an observable sequence of decisions. To do so one must specify how frequently traders arrive to the market. Easley, Kiefer, and O'Hara (1997) assume that the arrival frequency is fixed over time (with a trader arriving every five minutes). Unfortunately such a specification does not accommodate the fact that trading intensity varies in predictable ways over the course of a day. To allow for these periodic effects, and so distinguish episodes of trading that follow from information arrival, we vary the arrival rate of traders over the course of a day. We also study the bias that arises from misspecification of the arrival rate and establish that the bias vanishes as the assumed arrival frequency grows. Given the sequence of trade decisions, the likelihood function is formed from the probabilities of each trade as governed by the model. We note that maximum-likelihood estimation of the model of Easley and O'Hara (1992) implicitly conditions on the assumed frequency of information arrival and then detail the construction of the likelihood function for the model with the estimated frequency of information arrival. With the additional parameters, we are careful to distinguish between the information set of the market specialist and that of the econometrician, because the specialist has the additional knowledge of the timing of public news arrivals.

An empirical investigation of the impact of informed traders on a security market is contained in Section 3. We first focus on 2001 data for three firms of varying liquidity that trade on the New York Stock Exchange (NYSE). We estimate the frequency and accuracy of private information and find evidence that private information potentially arrives many times within a day, in accord with the findings of Hasbrouck (1988) and empirical analysis of the MDH. To link our work to previous work by Foster and Viswanathan (1993) and Madhavan, Richardson, and Roomans (1997), who study intraday effects, we estimate the model separately for each half hour of the trading day. With an expanded dataset of Dow Jones firms for all of 2002, we find that informed trading is most pronounced early in the day, in accord with earlier findings. To check if the model is capturing information arrivals rather than correlated arrivals of trades for other reasons, we check for comovement between weighted-price contributions and the estimates of informed trade. Weighted-price contributions, devised by Barclay and Hendershott (2003), measure the persistent impact of price changes within a period. As persistent price changes provide evidence of informed trading, our finding of substantial comovement reveals that the informed trade estimates capture the arrival of latent information.

1 MICROSTRUCTURE MODEL

Trade in a market for a single stock is coordinated through a market specialist. (The market is a dealership market in that the specialist does not act as a broker, thus all orders are market orders.) There are an arbitrarily large number of

potential (risk-neutral) traders, from which traders are randomly selected to meet with the specialist. Because there are an uncountable number of traders, who have only countably many opportunities to be selected to trade with the specialist, almost surely a selected trader has only one opportunity to increase utility through trade in the market. As the specialist meets with only one trader at a time, we index traders by their order of arrival, *i*. (We imagine that the market began at some time in the arbitrarily distant past, so i is an element of the integers Z.) Concordant with the arrival of traders is the generation of a signal S_i of the intrinsic stock value.

Trade occurs over a sequence of information periods. An information period captures the interval over which private information potentially exists, so the end of an information period is characterized by agreement over all participants on the value of the stock. The end of an information period thus corresponds to public revelation of the signal, which occurs with probability $\delta \in (0, 1)$ on any arrival. If we index information periods by m, then the value of the stock at the end of period m is

$$V_m = V_{m-1} + S_{i_m},$$

where i_m denotes the last arrival in period m (thus $\Delta V_m = S_{i_m}$). Because

$$E(S_{i_m}|S_{i_{m-1}}) = 0,$$

the expected price at the next public signal conditional on the current public signal always equals the price at the current public signal.

Private information is captured through the signals generated within an information period. In detail, the signal takes one of three values: $S_i \in \{-1, 0, 1\}$. (Setting the increment amount to $k \in \mathbb{R}$, rather than 1, would simply rescale the market.) At the beginning of an information period, the price-changing signals ($S_i = 1$ and $S_i = -1$) are equally likely. Only price-changing signals correspond to news privately entering the market, so we refer to $S_i \neq 0$ as (private) news. For all arrivals during information period m, $i_{m-1} < i \leq i_m$, the arriving trader observes the signal, and so is informed, with probability α. Inclusion of $S_i = 0$ is important in determination of α. Without $S_i = 0$, α must be small for an illiquid stock, otherwise one could not generate long segments without trade. With $S_i = 0$, such sequences can be generated by less frequent trade by uninformed traders, which allows α to reflect the increase in trading when information is present.

The signal that informed traders observe may differ from the publicly revealed value, and so is potentially imperfect. The evolution of the signal over trader arrivals within an information period is governed by the transition probabilities

$$M = \begin{bmatrix} \theta_1 & \theta_2 & 1-\theta_1-\theta_2 \\ \frac{1-\theta_3}{2} & \theta_3 & \frac{1-\theta_3}{2} \\ 1-\theta_1-\theta_2 & \theta_2 & \theta_1 \end{bmatrix},$$

with $P(S_i = 1 \mid S_{i-1} = 1) = \theta_1$, $P(S_i = 0 \mid S_{i-1} = 1) = \theta_2$, and $P(S_i = -1 \mid S_{i-1} = 1) = 1 - \theta_1 - \theta_2$. The parameters θ_1 and θ_2 measure state persistence, and so capture the belief that informed traders attach to private information. If $\theta_1 = \theta_3 = 1$, then the signal is perfect, as in Easley and O'Hara (1992). To understand how the two persistence parameters are identified, consider the effect of altering each parameter. As θ_1 increases, the likelihood of news $P(S_i \neq 0)$ increases. In addition, the quality of news increases, as it becomes more likely that the privately revealed value will be the future publicly revealed value. As θ_3 increases, the likelihood of news decreases, but the quality of news in unchanged. It is this asymmetry in the behavior that identifies the two persistence parameters.

While the transition matrix M describes the evolution of the signal within an information period, we must look across information periods to determine the statistical properties of the stationary series. Each information period begins with a signal that is drawn from $\bar{\pi}$, which is the stationary (unconditional) distribution of S_i. We assume that each element of $\bar{\pi} \in (0, 1)$, so that S_i is generated by a Markov process with transition probabilities

$$\begin{bmatrix} (1-\delta)M & \delta M \\ (1-\delta)\Pi & \delta\Pi \end{bmatrix},$$

where the (3×3) submatricies capture the transition if neither signal (S_i, S_{i-1}) is publicly revealed (the submatrix $(1 - \delta)M$), only S_i is publicly revealed (δM), only S_{i-1} is publicly revealed ($(1 - \delta)\Pi$, so $\Pi' = \bar{\pi} \otimes [1, 1, 1]$ reflects the fact that S_i is governed by $\bar{\pi}$ regardless of the value of S_{i-1}), both (S_i, S_{i-1}) are publicly revealed ($\delta\Pi$).

To determine the likelihood of private information, in the appendix we derive $\bar{\pi}$. Good news ($S_i = 1$) and bad news ($S_i = -1$) are (unconditionally) equally likely with total probability

$$v \equiv P(S_i \neq 0) = \frac{1 - \theta_3}{\theta_2 + (1 - \theta_3)}.$$

Although the model can accommodate very general dynamics, we concentrate on the realistic case in which the signal cannot immediately switch between the good and bad states. So, in what follows, we assume that $\theta_2 = 1 - \theta_1$, and we have

$$v = \frac{1 - \theta_3}{(1 - \theta_1) + (1 - \theta_3)}. \tag{1}$$

For this case, if $\theta_1 = \theta_3$, then private information (unconditionally) arrives half the time, $v = \frac{1}{2}$.

In order to test whether private information is perfect (that the signal remains constant until publicly revealed) we must test the hypothesis **H**o: $\theta_1 = \theta_3 = 0$. Unfortunately, under this restriction $\bar{\pi}$ becomes arbitrary as the transition probabilities no longer form an ergodic Markov chain. Nevertheless, we can nest the perfect information hypothesis by recasting the model to include v as a parameter instead of θ_3. Notice that Equation (1) is equivalent to

$$\theta_3 = 1 - \frac{v}{1-v}(1-\theta_1)$$

for any $\theta_3 < 1$ and any $0 \leq v \leq 1/(2-\theta_1)$. Moreover, if $\theta_1 = 1$, then $\theta_3 = 1$ and v varies freely over [0,1]. When we refer to the perfect information model, we refer to the model parameterized as $\theta_1 = \theta_3 = 1$, with v arbitrary on the interval [0,1] and $\bar{\pi} = \left(\frac{v}{2}, 1-v, \frac{v}{2}\right)$.

Prior to the ith arrival, the specialist sets an ask (A_i) and a bid (B_i) for one share of the stock. Let D_i represent the random variable corresponding to the decision of the ith trader, taking on d_A, d_B, or d_N if the decision is to trade at the ask, the bid, or not to trade, respectively. After each arrival, $i_{m-1} < i \leq i_m$, the specialist is aware of the entire history of trades, $\{D_j\}_{j=-\infty}^{i}$, the entire history of the public signals, $\{s_{i_k}\}_{k=-\infty}^{m-1}$, and the structure of the market. However, because of the Markovian setting and preference assumptions, only a small subset of this information is relevant to the specialist. All relevant information, other than the structure of the market, is summarized in the public information set

$$Z_i \equiv \{V_{m-1}, D_i, D_{i-1}, \ldots, D_{i_{m-1}}\}.$$

The information set of an informed trader includes the private signal and so is a finer partition of the event space at the next public signal. Because the intra-information period trade history is used by the specialist only to predict S_i, an informed trader's information set is

$$\{V_{m-1}, S_i\}.$$

The specialist does not observe the private signal and so must form beliefs about the value of the signal. Bayes' rule governs the method by which the specialist (and uninformed traders) learn through observing transactions. If trader i buys the stock, then[3]

$$P(S_i = 1 | Z_{i-1}, D_i = d_A) = P(S_i = 1 | Z_{i-1}) \frac{P(D_i = d_A | S_i = 1)}{P(D_i = d_A | Z_{i-1})}.$$

With updated beliefs, the Markov transition matrix M guides prediction of the signal at the next arrival,

$$P(S_{i+1} = 1 | Z_i) = \theta_1 P(S_i = 1 | Z_i) + \frac{1-\theta_3}{2} P(S_i = 0 | Z_i) + (1-\theta_1-\theta_2) P(S_i = -1 | Z_i).$$

For all future values, the specialist's beliefs are contained in the vector

$$\pi_{j,i} = \begin{bmatrix} P\ (S_{i+j} = 1 | Z_i) \\ P\ (S_{i+k} = 0 | Z_i) \\ P\ (S_{i+j} = -1 | Z_i) \end{bmatrix},$$

where $\pi'_{j,i} = \pi'_{0,i} M^j$.

[3] The specialist's recursion begins with the stationary probabilities.

The central problem for all agents is to determine the value of the stock at the public revelation, V_m. In predicting V_m, there are two interrelated sources of uncertainty. First, the agent must predict when the public information will arrive. To model prediction of uncertain future public news, we define a subsequence of arrivals, $\{i_m\}_{m=1,}^{\infty}$ corresponding to public news. The (random) number of arrivals until a public signal is $T(i) \equiv \#\{j{:}i < j \leq i_m\}$ for $i_{m-1} < i \leq i_m$, where # is the number of elements in the set. Because the arrival of public news occurs randomly with probability δ, the $T(i)$ are i.i.d. geometric random variables with common distribution equivalent to T, where

$$P[T = t] = (1 - \delta)^t \delta \qquad t = 0, 1, 2, \ldots.$$

Note that $T(i) = 0$ corresponds to public revelation of the signal (potentially) received by the ith trader. Second, given the expected time of public information arrival, the agent must predict the value of the stock when the value is made public. Following the arrival of trader i, the specialist's valuation $E(\Delta V_m \mid Z_i)$ is

$$\sum_{j=0}^{\infty} P[T(i) = j][P(S_{i+j} = 1|Z_i) - P(S_{i+j} = -1|Z_i)].$$

Solution of the infinite series yields (details are in the appendix)

$$E(S_i|Z_i) \cdot \frac{\delta}{1 - (1 - \delta)[\theta_1 - (1 - \theta_1 - \theta_2)]},$$

which equals the expectation of the current signal multiplied by a factor that captures the likelihood that the signal is publicly revealed. The factor equals one only if θ_1 equals one, in which case the current signal is perfect and so is revealed with certainty. We also see that the factor is an increasing function of δ and θ_1. Increasing δ tends to shorten the information period, while increasing θ_1 makes the current signal more informative, both of which imply that the current signal is more likely to be publicly revealed.

An informed trader receives S_i, which supersedes the public information. As S_i does not provide information about the timing of public news, an informed trader's valuation differs from the specialist's only in the prediction of the revealed signal

$$\sum_{j=0}^{\infty} P[T(i) = j][P(S_{i+j} = 1|S_i) - P(S_{i+j} = -1|S_i)].$$

Solution of the infinite series yields (details are in the appendix)

$$S_i \cdot \frac{\delta}{1 - (1 - \delta)[\theta_1 - (1 - \theta_1 - \theta_2)]},$$

which equals the current signal multiplied by the factor that captures the likelihood of public revelation of the signal. If $\theta_1 = 1$, then the valuation of an informed trader is simply

$$E(\Delta V_m|S_i) = S_i.$$

To complete the specification of the market microstructure, we define equilibrium by a sequence of bid-ask pairs that result in zero expected profits for the specialist; formally, at any given arrival an equilibrium obeys

$$E[V_m - A_i|Z_{i-1}, D_i = d_A] = E[B_i - V_m|Z_{i-1}, D_i = d_B] = 0,$$

where $i_{m-1} < i \leq i_m$ for all $m \in Z$. We consider this an equilibrium condition obtaining from the potential free entry of additional market specialists should the bid and ask lead to positive expected profits.

The zero-profit equilibrium imposes constraints on the quotes. First, the quotes always satisfy

$$V_{m-1} - \frac{\delta}{1-(1-\delta)[\theta_1-(1-\theta_1-\theta_2)]} < B_i \leq A_i < V_{m-1} + \frac{\delta}{1-(1-\delta)[\theta_1-(1-\theta_1-\theta_2)]},$$

where the lower and upper bounds are the "reservation prices" for an informed trader with $S_i = -1$ and $S_i = 1$, respectively. For example, if the ask exceeded the upper bound, then informed traders would never trade at the ask and the specialist could ensure positive profit from exclusive trade with uninformed traders. Because the quotes are bounded by the reservation prices, the decision of the informed is summarized by the following simple rule: Buy if $S_i = 1$, sell if $S_i = -1$.

We do not directly model the preferences of the uninformed, as the uninformed are assumed to trade for liquidity reasons rather than speculation. Because a trader who receives the signal $S_i = 0$ does not trade, we must allow for uninformed traders to elect not to trade, to avoid immediate revelation of a private signal. Uninformed traders elect to trade with probability ε. Of the proportion of uninformed traders who trade, half buy at the ask and half sell at the bid.

To see the specific form of the quotes, note that the ask for the stock is determined from

$$\alpha P(S_i = 1|Z_{i-1}) \cdot [E(V_m|S_i = 1) - A_i] = \frac{1}{2}(1-\alpha)\varepsilon \cdot [A_i - E(V_m|Z_{i-1})].$$

The left side is the expected loss the specialist incurs from trade with the informed at the ask, the right side is the expected gain the specialist receives from trade with the uninformed. The corresponding equilibrium ask is

$$A_i = \frac{\alpha P(S_i = 1|Z_{i-1})E(V_m|S_i = 1) + \frac{1}{2}(1-\alpha)\varepsilon E(V_m|Z_{i-1})}{\alpha P(S_i = 1|Z_{i-1}) + \frac{1}{2}(1-\alpha)\varepsilon},$$

where $E(V_m \mid Z_{i-1})$ equals

$$P(S_i = -1|Z_{i-1})E(V_m|S_i = -1) + P(S_i = 0|Z_{i-1})E(V_m|S_i = 0) + P(S_i = 1|Z_{i-1})E(V_m|S_i = 1).$$

To summarize, trading evolves as follows. After a public signal, the next signal is selected according to the dynamics of the signal process. Assuming that this signal is not public, only the informed are aware of the signal value. A trader is randomly selected to trade with the specialist and the signal is potentially revealed to the trader. The trader observes the bid and ask, and decides whether to buy or sell. After the decision of the trader, the signal is publicly revealed with probability δ. Upon observing the decision of the trader, and the possible public revelation of the signal, the specialist must set the bid and ask that will be in effect at the next arrival. The signal is then updated again, and the previously described process continues until another public signal occurs.

2 ECONOMETRIC ESTIMATION

The microstructure model yields the likelihood of each trade decision D_i as a function of the parameters $\Phi = (\alpha, \delta, \varepsilon, \theta_1, \theta_3)$. Thus the observed trade decisions are used to form the likelihood (function) without need of further distributional assumptions. As the actual sequence of trade decisions is unobserved, two transformations of the data are needed to construct the sequence.

The first transformation concerns the information content of elapsed time with out trade. Within the model, the frequency of trader arrivals determines the amount of time without trade that corresponds to a no-trade decision. To form a sequence of trade decisions, we must specify the frequency of trader arrivals. In doing so we must account for the fact that, as Harris (1986), Jain and Joh (1988), and McInish and Wood (1992) verify, trade volume exhibits significant cyclic patterns both within a day and across days of the week. As the predictable patterns in trade activity are likely due to the many factors affecting trade that are not captured by the model, we must allow the arrival rate of traders to vary over time. To do so, we construct an average number of trades for each hour of the week (allowing for both day-of-week and hour-of-day effects).[4] For each hour of the week, the number of arrivals is assumed to be a multiple, K, of the average number of trades. (The value of K must be large enough so that the number of arrivals always exceeds the actual number of trades, and small enough so that traders do not arrive more frequently than one per second.) Thus the length of time corresponding to a no-trade interval (the time between arrivals) varies over hours.

While our ability to allow for varying arrival rates is an improvement over a fixed arrival rate of traders, any specification of arrival rates likely introduces misspecification bias. To determine the bias, we focus on the ratio of the assumed number of arrivals to the recorded number of trades, which directly affects the estimator of Φ. For a given interval of time, the ratio depends on the (assumed)

[4] We use hourly effects, rather than a quadratic function of hours, to account for the additional effect of the lunch hour on trade activity for the NYSE.

frequency of arrivals and on the timing of recorded trades. Increasing the frequency of arrivals increases the ratio of arrivals to trades, as the number of recorded trades is unaffected by the assumed frequency of arrivals. As estimation is affected by this ratio, we account for this by reporting the invariant measures that are scaled by the arrival frequency.

The timing of recorded trades also affects the number of arrivals. If trades are recorded at times other than integer multiples of the arrival frequency, then the number of arrivals is increased (e.g., because two trades are recorded more closely than the assumed arrival frequency). The following theorem details the effects. Let L be the length of the time interval and let $f \leq 1$ be the assumed arrival frequency in seconds ($f = 0.1$ indicates a trader arrives every 10 seconds). Let T be the number of trades in the interval and A be the number of constructed arrivals, so the bias is $B = T^{-1}(A - Lf)$.

Theorem 1 The bias induced by the assumed arrival frequency f is

$$0 \leq B \leq 1.$$

Further,

$$B \to 0 \text{ as } f \to 1.$$

Proof See the appendix.

Theorem 1 is quite intuitive. As trades are recorded to the nearest second, the assumption that a trader arrives every second eliminates bias from the misalignment between recorded trades and the arrival frequency.

While an arrival frequency of once per second eliminates bias, much of the bias can be eliminated at less frequent arrival rates. This is useful because the entire sequence of trade decisions is needed for estimation. To determine the arrival frequency, and the associated length of the decision sequence, we measure the bias relative to K. The relative bias is $K^{-1}B$, which upon noting that $KT = Lf$ is simply given from the formula of B as $K^{-1}\left(\frac{A}{T} - K\right)$. To link the bias to estimates of the parameters, note that $\frac{T}{A}$ is the constructed probability of a trade, which is $\alpha\ \nu + (1 - \alpha)\varepsilon$. The implied relative bias is obtained by replacing $\frac{T}{A}$ with this probability, evaluated at the maximum-likelihood estimates, which yields $K^{-1}\left((\hat{\alpha}\hat{\nu} + (1 - \hat{\alpha})\hat{\varepsilon})^{-1} - K\right)$. We analyze the measures of relative bias to determine the value of K that balances bias reduction and computation efficiency.

The second transformation of the data arises because all trades are assumed cleared through the specialist, so each trade is classified as buyer initiated (a trade at the ask) or seller initiated (a trade at the bid). Because transaction records do not indicate who initiates a trade, a classification rule must be employed. We use a rule proposed by Lee and Ready (1991), who use the midpoint of the bid-ask spread to classify trades. A trade above the midpoint is (classified as) buyer initiated, a trade below the midpoint is seller initiated, and a trade at the midpoint depends on the preceding price movement. For example, if the price of the

preceding trade is above the midpoint, then the midpoint trade represents a price decline and is seller initiated. As consecutive midpoint trades are classified identically (if there is no intervening price movement), such a rule can produce artificial runs of trades on one side of the market. In fact, Lee and Radhakrishna (2000) found that while the rule correctly classified 93% of transactions in their test sample, only 60% of consecutive midpoint transactions were correctly classified. To mitigate this type of missclassification, we also consider both random assignment of consecutive midquote trades and removal of midquote trades.

For the model of Section 1, in which the frequency and accuracy of private information are unknown, the sufficient statistics for Φ are the entire sequence of trade decisions and public news arrivals. [If, as in Easley, Kiefer, and O'Hara (1997), one assumes that private information is perfect and can arrive only at fixed points in time, then the likelihood is considerably simplified. For this case, the sufficient statistics reduce to the number of trade decisions of each type within an information period.] From a sequence of n trader arrivals (over a span of m information periods), the likelihood from the model with unknown information frequency and accuracy is

$$L\left(\Phi|D_1 = d_1, \ldots, D_n = d_n, \{i_j\}_{j=1}^{m-1}\right) = \Pi_{i=1}^{n} P(D_i = d_i|Z_{i-1}; \Phi),$$

where $P(D_1 \mid Z_0;\Phi)$ is the stationary probability distribution for the first decision in the information period. In detail for $D_i = d_A$ we have

$$\begin{aligned} P(D_i = d_A|Z_{i-1}; \Phi) = {} & P(S_i = 1|Z_{i-1}, D_i = d_A; \Phi)\left[\alpha + (1-\alpha)\frac{\varepsilon}{2}\right] \\ & + P(S_i = 0|Z_{i-1}, D_i = d_A; \Phi)[(1-\alpha)(1-\varepsilon)] \\ & + P(S_i = -1|Z_{i-1}, D_i = d_A; \Phi)\left[(1-\alpha)\frac{\varepsilon}{2}\right]. \end{aligned}$$

The conditional probabilities for S_i are obtained directly from the learning rules for the specialist described in Section 2. We have

$$P(D_i = d_i|Z_{i-1}; \Phi) = \lambda_i' \pi_{1,i-1},$$

where the trade frequencies are captured by

$$\lambda_i = \begin{bmatrix} (1-\alpha) & \frac{\varepsilon}{2} \\ (1-\alpha) & (1-\varepsilon) \\ (1-\alpha) & \frac{\varepsilon}{2} \end{bmatrix} + \alpha \begin{bmatrix} 1(D_i = d_A) \\ 1(D_i = d_N) \\ 1(D_i = d_B) \end{bmatrix}.$$

Much of the motivation for carefully extending the sequential arrival microstructure model to a stationary setting is that it is typically not possible to identify episodes of information asymmetry. Instead, the best that can be hoped for is that a process governing the ongoing evolution of information asymmetries can be identified. As a result, the econometrician has the reduced information set $\tilde{Z}_{i-1}$, which consists of only trade decisions without knowledge of the subsequence of arrivals corresponding to public news. Burdened by the lack of knowledge about the public news state variable, the econometrician is concerned with a state space

that is twice as large as the state space confronting the specialist. The state space, $\{0, 1\} \times \{1, 0, -1\}$ underpins an ergodic bivariate Markov chain. The first element, $\tilde{S}_i$, determines whether the signal is private (1) or public (0), the second element is S_i. (The stationary distribution is $(\delta\bar{\pi}, (1-\delta)\bar{\pi})$, with $\bar{\pi}$ the stationary distribution for S_i.) The likelihood becomes

$$L(\Phi|D_1 = d_1, \ldots, D_n = d_n) = \Pi_{j=1}^{n} P\big(D_i = d_i|\tilde{Z}_{i-1}; \Phi\big),$$

where $P\big(D_1|\tilde{Z}_0; \Phi\big)$ is the stationary probability distribution for the first trade in the sample. We have

$$P\big(D_i = d_i|\tilde{Z}_{i-1}; \Phi\big) = P\big(\tilde{S}_{i-1} = 0\big)\lambda_i'\bar{\pi} + P\big(\tilde{S}_{i-1} = 1\big)\lambda_i'\tilde{\pi}_{1,i-1},$$

where

$$\tilde{\pi}_{1,i-1} = \begin{bmatrix} P\big(S_i = 1|\tilde{Z}_{i-1}\big) \\ P\big(S_i = 0|\tilde{Z}_{i-1}\big) \\ P\big(S_i = -1|\tilde{Z}_{i-1}\big) \end{bmatrix}.$$

To understand the compact expression, if public news arrives with trader $i-1$, then the conditional probabilities for S_i are reset to the stationary values $\bar{\pi}$. If public news does not arrive with trader $i-1$, then the conditional probabilities for S_i follow from the learning rules of Section 1, with the restricted information set $\tilde{Z}_i$. Because the public news process is i.i.d., $P\big(\tilde{S}_{i-1} = 1\big) = (1-\delta)$ and $P\big(\tilde{S}_{i-1} = 0\big) = \delta$, so

$$P\big(D_i = d_i|\tilde{Z}_{i-1}; \Phi\big) = \delta\lambda_i'\bar{\pi} + (1-\delta)\lambda_i'\tilde{\pi}_{1,i-1}.$$

The score function is

$$\frac{\partial}{\partial\Phi}\ \ln\ L = \sum_{t=1}^{T} \frac{\frac{\partial}{\partial\Phi}\{\delta\lambda_i'\bar{\pi} + (1-\delta)\lambda_i'\bar{\pi}_{1,i-1}\}}{L_t},$$

where

$$\frac{\partial}{\partial\Phi}\{\delta\lambda_i'\bar{\pi} + (1-\delta)\lambda_i'\tilde{\pi}_{1,i-1}\} = \frac{\partial\delta\lambda_i}{\partial\Phi}\bar{\pi} + \frac{\partial(1-\delta\lambda_i)}{\partial\Phi}\bar{\pi}_{1,i-1} + \left(\frac{\partial}{\partial\Phi}\bar{\pi}\right)\delta\lambda_i + \left(\frac{\partial}{\partial\Phi}\tilde{\pi}_{1,i-1}\right)(1-\delta)\lambda_i.$$

While the terms $\frac{\partial(1-\delta)\lambda_i}{\partial\Phi}$ and $\frac{\partial\delta\lambda_i}{\partial\Phi}$ depend only on the parameter values and trade decisions, the term $\frac{\partial}{\partial\Phi}\bar{\pi}_{1,i-1}$ must be calculated recursively.

3 EMPIRICAL IMPLEMENTATION

We examine three interesting questions. First how long is an information period? Many empirical specifications of microstructure models assume that an information

period is one trading day. This is in stark contrast to much of the MDH literature, which assumes significantly shorter information periods. Second, how much trade is information based? Easley, Kiefer, and O'Hara (1997) find that informed traders account for less than one-third of the trades in their samples of NYSE stocks. Third, how precise is private information? Signals are often assumed noiseless in the microstructure literature, raising the question of how well this assumption conforms to the data when signal quality is estimated instead of assumed.

To answer these questions, we analyze three NYSE stocks chosen according to their liquidity characteristics. We select International Business Machines (ticker symbol IBM) to represent a highly liquid stock. We select Ashland (ticker symbol ASH) to represent a moderately liquid stock.[5] Finally, we select the Commerce Insurance Group (ticker symbol CGI), a property and casualty insurer, to represent a relatively illiquid stock.

For each stock, we extract data from the first 30 trading days of 2001 (January 2, 2001 to February 13, 2001) from the NYSE Trades and Quotes (TAQ) dataset. The TAQ dataset contain a record of every trade and quote posted on the NYSE, the American Stock Exchange (AMEX) and the NASDAQ National Market System for all NYSE-listed securities. We filter the trade data to remove trades that were recorded out of sequence, canceled, executed with special conditions, or recorded with some other anomaly. We use quotes only from the NYSE [Blume and Goldstein (1997) find that the NYSE quote determines or matches the national best quote about 95% of the time]. We also filter the quote data to remove recording anomalies.

Because of certain institutional details, occasionally large trades are broken up into a sequence of smaller trades, all at the same price [see Hasbrouck (1988)]. In order to avoid misidentifying these sequences of same-sided trades as bursts of informed trades, we aggregate all trades recorded within five seconds of each other without an intervening price change or quote revision.

The data are further filtered to remove time stamps outside of the official trading hours of the NYSE (9:30 A.M. to 4:00 P.M.). Finally, the first half hour of each trading day is removed in order to avoid modeling the market opening of the NYSE, which is characterized by heavy activity following the morning call auction. As Harris (1986), Engle and Russell (1998), and many other authors have noted, the first half hour of trade exhibits substantially different properties than the rest of the day.

We investigate the potential bias from sequences of mid quote trades that are assigned the same initiator type. Fortunately, relatively few of the trades in our dataset are part of mid quote sequences. For ASH, CGI, and IBM, 0.32%, 0.26%, and 0.88% were consecutive mid quote trades without a price change, respectively. Neither randomly assigning these few mid quote trades nor removing them from the sample had a significant effect on estimated parameters. Consequently we report estimates based on the unmodified Lee and Ready approach.

[5] Ashland Oil Incorporated, also studied by Easley, Kiefer, and O'Hara, changed its name to Ashland Incorporated in 1995.

After filtering the data and assigning trader initiation, we remove periodic features from the data. In detail, we first regress the number of trades for each hour on day-of-week and hour-of-day indicators. The parameter estimates are given in Table 1, where starred items are significant at the 5% level and double-starred items are significant at the 1% level. (Hour-of-day indicators are labeled according to the beginning of the hour they correspond to, so 11 A.M. takes on 1 over the interval 11 A.M. to 12 P.M.) We determine the arrival frequency for each hour of the week by dividing the predicted number of arrivals in the hour (the predicted number of trades multiplied by K) by 3600 seconds. The arrival frequency, in turn, determines the length of time between arrivals, and whenever this length of time elapses without a trade, we record a no-trade.

Once the periodic features of the data are removed, we record the effective number of arrivals by adding the number of trades to the number of constructed no-trades. The realized relative bias is then given by $K^{-1}\left(\frac{A}{T} - K\right)$. With the maximum-likelihood estimates constructed from the sequence of trader decisions, the implied relative bias is $K^{-1}\left[\hat{\alpha}\hat{\nu} + (1-\hat{\alpha})\hat{\varepsilon}\right)^{-1} - K\Big]$.

In Figure 1, we plot the realized and implied relative bias as K varies from 2 to 20 for ASH. It is clear that the realized and implied relative biases are essentially identical, demonstrating that parameters are biased by incorrectly classifying no-trade decisions. As the figure indicates, increasing K from 2 to 10 decreases

Table 1 Regression of the hourly number of trades against hour-of-week regressors (SEs in parentheses).

Coefficient	ASH	CGI	IBM
INTERCEPT	52.406**	9.614**	755.106**
	(3.343)	(1.358)	(44.886)
TUE	−0.229	0.419	−15.395
	(3.233)	(1.312)	(43.404)
WED	−0.022	0.161	60.367
	(3.343)	(1.357)	(44.886)
THR	1.978	−0.439	159.894**
	(3.343)	(1.366)	(44.886)
FRI	2.617	−0.283	78.672
	(3.343)	(1.357)	(44.886)
11AM	−14.200**	−2.867*	−113.900*
	(3.492)	(1.417)	(46.882)
12PM	−21.200**	−2.433	−295.767**
	(3.492)	(1.417)	(46.882)
1PM	−17.367**	−1.787	−258.467**
	(3.492)	(1.430)	(46.882)
2PM	−13.433**	−0.933	−145.233*
	(3.492)	(1.417)	(46.882)
3PM	5.567	6.533**	21.533
	(3.492)	(1.417)	(46.882)
R^2	0.351	0.264	0.367

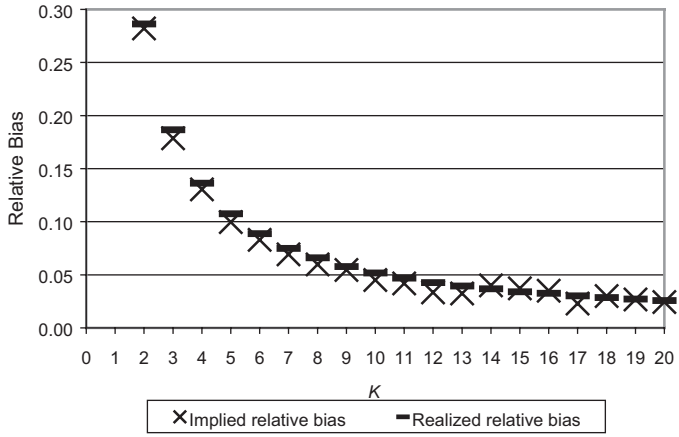

Figure 1 Realized and implied relative biases as functions of *K*.

the relative bias from 25% to 5%. Further increases in *K* have little effect on the bias. Consequently we choose $K = 10$ for the results reported on ASH and the CGI. For IBM, we choose $K = 2$ to avoid having effective arrivals too close together (on Thursdays between 3 P.M. and the close, trades occur, on average, less than 4 seconds apart).

Because we are constrained in the choice of *K* for IBM reported results tend to overestimate the time between events. For example, the estimated time between trades and the time between information arrivals are biased upward. However, many interesting estimates are largely invariant to the choice of *K*. Figure 2 shows that, for ASH, the estimated probability of informed trade as measured by $\hat{\alpha}\hat{\nu}/(\hat{\alpha}\hat{\nu} + (1 - \hat{\alpha})\hat{\varepsilon})$ and the expected probability of informed trade given the presence of private information, as measured by $\hat{\alpha}/(\hat{\alpha} + (1 - \hat{\alpha})\hat{\varepsilon})$, where the hats indicate maximum-likelihood estimates, change little as *K* varies. This is an intuitive result; the biases of the estimated probability of informed trade and of uninformed trade tend to offset each other in measurements that include their ratios.

Having settled the implementation issues, we next turn to estimating the model and testing information quality. Table 2 lists the parameter estimates for ASH.[6] The first item that stands out is that information quality is estimated to be very high. The chi-square statistic for the likelihood ratio test of H_0: $\theta_1 = \theta_3 = 1$ has one degree of freedom and is not significant at the 10% level. Moreover, the shared parameter estimates are almost identical between the two models (SEs in parantheses). From the pattern that emerges, information periods correspond to short bursts of $7(\hat{\delta}^{-1} = 6.8)$ arrivals. Roughly 40% of the bursts are associated with private information $(\hat{\nu} = 0.41)$ and within a burst 14% $(\hat{\alpha} = .14)$ of the traders are

[6] As described in Section 1, for the unrestricted model, ν is determined by $\nu = \frac{1-\theta_3}{(1-\theta_1)+(1-\theta_3)}$. The standard errors ν reported in Table 2 are derived by the delta method.

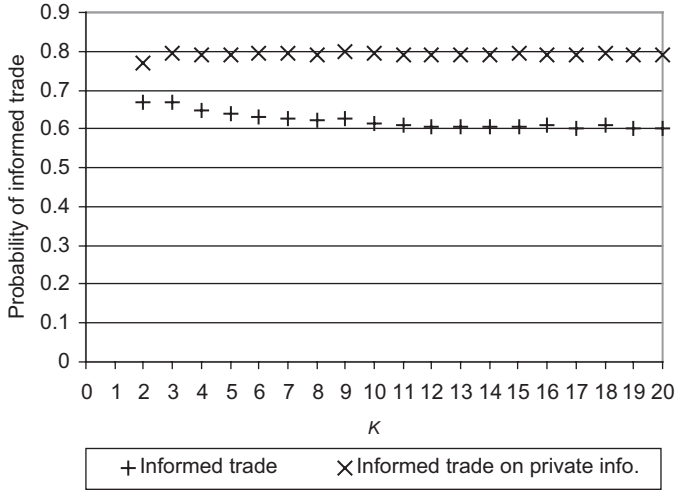

Figure 2 Probability of informed trade.

Table 2 Maximum-likelihood estimates of the restricted and unrestricted models.

	ASH	
Coefficient	Restricted	Unrestricted
α	.143097	.143133
	(.005008)	(.005370)
ν	.411310	.412034
	(.021685)	(.017239)
ε	.042826	.042823
	(.002314)	(.002453)
δ	.146922	.146738
	(.008310)	(.018375)
θ_1	1	.999601
	(N/A)	(.016371)
θ_3	1	.999721
	(N/A)	(.011461)
Log likelihood	−30,343.47	−30, 342.59
χ^2		0.88

informed. While the remaining 86% of arrivals are associated with uninformed traders, such trade is infrequent as uninformed arrivals result in trade less than 5% of the time ($(\hat{\varepsilon} = 0.04)$. As a result, price movements during bursts are heavily influenced by informed traders, with 80% of trade attributed to the informed $\left(\frac{0.14}{0.14+0.86.004} = 0.80\right)$. Even though bursts occur less than half the time, the remaining periods see very little trading (as the uninformed trade infrequently). The

Table 3 Maximum-likelihood estimates of restricted and unrestricted model for CGI.

	CGI	
Coefficient	Restricted	Unrestricted
α	.312000	.312002
	(.029530)	(.029559)
ν	.173258	.173158
	(.027062)	(.019422)
ε	.055401	.055402
	(.008384)	(.008387)
δ	.361355	.361358
	(.033811)	(.034561)
θ_1	1	.999805
	(N/A)	(.017933)
θ_3	1	.999959
	(N/A)	(.003754)
Log likelihood	−6506.96	−6505.30
χ^2		1.66

probability of informed trade (PIN), which is the overall share of trading attributed to the informed, is slightly more than 60% $\left(\frac{0.41 \cdot 0.14}{0.41 \cdot 0.14 + 0.86 \cdot 0.04} = 0.63\right)$. Finally, translating the bursts into clock time, we find that during active times in the market (such as the first hour, 10–11 A.M., on Monday) a trader arrives roughly every 7 seconds, so a burst lasts only 50 seconds. As a result, a burst occurs (i.e., private information potentially arrives) roughly every 2 minutes $\left(\frac{49\,\text{sec}}{0.41}\right)$.

Many of the results for ASH carry over to the case of a far less liquid stock. Estimated information quality for the CGI is also very high (see Table 3). Trade bursts now correspond to only three arrivals and are far less prevalent (occurring only 17% of the time). Given the less frequent arrival of bursts, it is perhaps not surprising that only 30% of the traders are informed. As the uninformed again trade only 5% of the time, the frequency of informed trade within a burst is nearly 90%. Because bursts are rarer, the larger probability of informed trade within a burst plays a smaller role and the overall probability of informed trade remains at roughly 60%. During an active market hour, a trader arrives every 40 seconds, so a burst lasts 2 minutes and bursts occur every 10 minutes. Thus CGI is characterized by relatively rare, but quite potent, bursts of trade.

In contrast, the liquid stock IBM has bursts that arrive quite frequently, 75% of the time, and last for 14 arrivals. Within a burst, roughly 20% of the traders are informed. Yet, because the uninformed trade with much greater frequency (30%), the overall impact of informed trade is substantially lower: 50% within a burst and 35% overall. Finally, during an active market hour a trader arrives every 2 seconds, so bursts last just under 30 seconds and occur (on average) every 40 seconds.

Our results differ from previous empirical work by Easley, Kiefer, and O'Hara (1997), who assume that information arrives at most once per day and

Table 4 Maximum-likelihood estimates of restricted and unrestricted model.

	IBM	
Coefficient	Restricted	Unrestricted
α	.233794	.259330
	(.001777)	(.002337)
ν	.759209	.545455
	(.006302)	(.004844)
ε	.294319	.338974
	(.001787)	(.002311)
δ	.069861	.103751
	(.001607)	(.002891)
θ_1	1	.999999
	(N/A)	(.00001)
θ_3	1	.999999
	(N/A)	(.00001)
Log likelihood	−277,946.38	−277,547.97
χ^2		398.41

takes a full day for the market to absorb. In contrast, we find that information potentially arrives many times within an hour. Moreover, information is quickly absorbed by the market, often in a matter of minutes. In an appendix to the Web version of this article [Owens and Steigerwald (2004)], we detail how misspecification of the information arrival frequency leads to biased parameter estimates.

One fact that emerges clearly is that, regardless of liquidity, stock trades are characterized by frequent bursts of trade activity. As these bursts are short, trade activity is typically one-sided (i.e., clustered at either the ask or bid), so information is always estimated to be very accurate. As a consequence, we set $\theta_1 = \theta_3 = 1$ in what follows.

3.1 Time-Varying Parameters

In the above analysis, the length of time between trader arrivals and the expected time to a burst of trading vary over the course of the day. Yet other interesting features, such as the probability of informed trade, are not allowed to vary. Recent work by Madhavan, Richardson, and Roomans (1997) and Foster and Viswanathan (1993) find interesting variation in the information content of trades within the day.

To capture some of this richer detail, we allow the parameters to vary by time of day (we also allow the parameters to vary by day-of-week).[7] In addition, we expand the scope of our investigation, both by including more stocks and more trading days. We reestimate the model for 27 Dow Jones Industrial stocks traded on the NYSE for

[7] In estimating the model for separate time periods, the sample consists of noncontinguous segments (e.g., the first half hour of a given Monday is followed by the first half hour of the following Monday). We initialize the prior distribution of the state variable at the beginning of each segment.

Table 5 Overall distribution of parameter estimates.

Coefficient	Lower quartile	Mean	Upper quartile
α	.0589	.0726	.0834
ν	.6790	.7398	.7926
ε	.0345	.0489	.0607
δ	.0264	.0319	.0369
PIN	.4903	.5510	.6068

all full trading days in 2002.[8] We filter the data as reported above, with two exceptions. First, as we allow the parameters to vary over the course of the day we do not need to remove a cyclic pattern in trade. Second, as all 27 stocks are liquid, we assume that a trader arrives every second, which in essence is the limiting value of *K*.

We divide each trading day into 13 half-hour periods (the opening half-hour is now included) and estimate the model separately for each of these periods.[9] In Table 5 we present the overall distribution of the parameter estimates. (For each parameter there are 27 estimates for each of the 65 periods. For each of the following tables, there is little evidence of asymmetry as the median virtually equals the mean and is not reported.) The estimated values of δ and ν are quite similar to the above estimates, indicating that the frequency and length of bursts of one-sided trading is relatively constant. As one would expect, the estimates of the probability of trade (either by the informed, α, or the uninformed, ε) have declined from those reported earlier for IBM. This follows from the more frequent arrival of traders, as the parameters now represent the probability of trade for one-second intervals. We find the probability of informed trade to be between 49% and 61%.

In columns 2–4 of Table 6 we display the results of the probability of informed trade over the course of Monday. Column 3 contains the mean estimate for all 48 Mondays in the sample. For the opening period, if we subtract 0.05 from the mean estimate we obtain the lower quartile of the estimates. In similar fashion, if we add 0.06 to the mean estimate we obtain the upper quartile. The pattern across the periods of the day indicates that the distribution of estimates is roughly symmetric about the mean. (The remaining days of the week exhibit a similar pattern.) The first half hour contains the highest probability of informed trade. There is a continued decline in informed trade through 1:30 P.M. followed by a slight increase through 3:30 P.M., with a substantive drop in the last half hour. (With less finely aggregated periods, Madhavan, Richardson and Roomans (1997) find that a measure of the information content of trades is highest early in the day, declines through 2:00 P.M. and is relatively constant through the end of the day. Similarly, Foster and Viswanathan (1993) find

[8] There are 30 stocks in the Dow Jones Industrial Average. Two (Microsoft and Intel) do not trade on the NYSE. Hewlett-Packard merged with Compaq during the sample and so is also excluded. The 27 firms we include are listed in the appendix.

[9] We discard the first trade of the day, as this trade alone arises from an auction. All stocks opened within 22 minutes of 9:30 A.M. Excluding the 10 days on which a stock took more than 15 minutes to open made no material difference to the results. Therefore no days are removed from the analysis.

Table 6 Measures of informed trade over Monday. (LQ = -0.05 - subtract 0.05. to obtain lower quartile, UQ = +0.06 - Add 0.06 to obtain upper quartile.)

	PIN			WPC		
Period	LQ	Mean	UQ	LQ	Mean	UQ
9:30-10.00	−.05	.5880	+.06	−.12	.2346	+.14
10:00-10:30	−.05	.5806	+.06	−.11	.1614	+.15
10:30-11:00	−.06	.5801	+.06	−.09	.1066	+.07
11:00-11:30	−.08	.5567	+.05	−.06	.0638	+.05
11:30-12:00	−.06	.5466	+.04	−.05	.0607	+.09
12:00-12:30	−.06	.5337	+.07	−.06	.0456	+.05
12:30-1:00	−.06	.5324	+.07	−.07	.0204	+.07
1:00-1:30	−.07	.5280	+.04	−.07	.0207	+.06
1:30-2:00	−.06	.5338	+.05	−.08	.0155	+.08
2:00-2:30	−.07	.5393	+.07	−.07	.0667	+.06
2:30-3:00	−.06	.5445	+.07	−.08	.0676	+.06
3:00-3:30	−.07	.5359	+.04	−.08	.0764	+.08
3:30-4:00	−.05	.4577	+.04	−.07	.0600	+.06

that the information content of trades is high early in the day, declines through midday and then increases in the afternoon.) The estimated diurnal pattern in informed trade provides evidence that informed trading is most pronounced early in the trading day. To determine if we are correctly attributing bursts of trade to informed traders, we study the price impact associated with each period.

3.2 Weighted-Price Impact

If a trade really is due to information, then the price impact of that trade should persist. Thus if the estimates of PIN are capturing information, then periods with a large PIN estimate should have persistent price impacts. To measure the persistence of price impacts, we follow Barclay and Hendershott (2003) and construct the weighted-price contribution (WPC) of the return for each period. In detail, for period i on trading day j,

$$WPC_{ij} = \sum_{s=1}^{27} \left(\frac{|r_{js}|}{\sum_{s=1}^{27} |r_{js}|} \right) \left(\frac{r_{ijs}}{r_{js}} \right),$$

where s indexes each stock. The first component reflects the contribution of stock j to the daily return r_{js} (which equals $\sum_{i=1}^{13} r_{ijs}$). The second component reflects the contribution of period i to the daily return for stock s. Large values of WPC_{ij} refer to periods with persistent price impact. In columns 5–7 of Table 6 we present estimates of the WPC for Monday. The largest price contribution, which is the period with the most persistent innovation to price, is the opening half hour. In parallel to the estimates of informed trade probabilities, the price contribution declines sharply through 2:00 P.M., before climbing slightly to a rather constant level for the remainder of the day.

While the measures confirm that the estimates of informed trade reflect persistent price impacts, there is still the question of the level of informed trade. Because the estimates of informed trade are driven by bursts of one-sided trade, any forces at work that lead to trade clustering influence the estimates. In practice, forces such as the division of large block trades into a sequence of smaller trades invariably create upward bias in the estimates of informed trade.

To gauge the magnitude of the bias, we use changes in the weighted-price contribution over the day. We first decompose the PIN estimate into a component that reflects informed trade (I) and a noise component arising from other forces (N), $PIN_i = I_i + N_i$. If the noise component is roughly constant over the course of the day, then changes in PIN reflect changes in information,

$$PIN_1 - PIN_{13} = I_1 - I_{13}.$$

Information changes are also reflected in changes in WPC. From Table 6, the weighted-price impact in the closing period is only 26% $\left(\frac{0.06}{0.2346} = 0.2558\right)$ of the impact for the opening period. If WPC and PIN are linearly related, then a change in WPC should lead to a proportionate change in the trade arising from information. With the average value of each of the 65 periods, we find the fitted value of WPC to be

$$\underset{(0.11)}{-0.46} + \underset{(0.20)}{0.97}\, PIN.$$

Given the linear relation in the two information measures, $I_{13} = 0.2558 I_1$. As $PIN_1 - PIN_{13} = 0.1303$ from Table 6, it follows that $I_1 = 0.1751$ and $N = 0.4129$. This rough bias adjustment leads to estimates of informed trade contribution that are substantially smaller: 18% for the opening period and 4% for the closing period.

4 CONCLUSION

In this article we develop and estimate a microstructure model that allows for estimation of both the frequency and quality of private information. The frequency of private information is captured through the potential revelation of the information at each arrival. The length of time prior to revelation follows a geometric distribution and the parameter of this distribution characterizes the frequency of information. The quality of private information is captured with a Markov transition matrix; information that is likely to be publicly revealed is of high quality.

Construction of the trade decision sequence depends on two classification algorithms. We show the potential bias that results from each of the algorithms. For the no-trade classification algorithm, we prove that the bias shrinks as the arrival frequency approaches the time scale on which trade is resolved. The ability to estimate the frequency and quality of information is not without cost. The sufficient statistic for the microstructure parameters is the entire sequence of trade decisions, in contrast to the case in which the frequency and quality are assumed known, for which the sufficient statistics are simply the number of trades of each type.

With recent data from the NYSE, we examine the characteristics of private information. For stocks of widely varying liquidity, the answer is surprisingly robust, high-quality information potentially arrives frequently and trading reveals the information quickly. These results pose a challenge to previous microstructure studies in which information is assumed to arrive at most once per day. Our finding of frequent information arrival allows one to rigorously underpin tests of the mixture-of-distributions hypothesis, in which information is assumed to arrive frequently, with a microstructure model.

APPENDIX

A.1 Calculation of the Stationary Distribution

The stationary probabilities for the private signal are the values of $\bar{\pi}$ such that $M'\bar{\pi} = \bar{\pi}$. If θ_1 and θ_3 do not equal one, then the stationary probability vector is the eigenvector of M′ that corresponds to an eigenvalue of one. In detail, we first note that $|M' - I_3| = 0$, where I_3 is the 3×3 identity matrix, so that one is an eigenvalue of M'. The corresponding eigenvector, $\bar{\pi}$, satisfies

$$(M' - I_3)\bar{\pi} = 0.$$

From the first and third equations in the system, it follows that $\bar{\pi}_1 = \bar{\pi}_3 = c$. From the first equation,

$$\bar{\pi}_2 = \frac{2\theta_2}{1-\theta_3}c.$$

Because the three stationary probabilities sum to one,

$$2c + \frac{2\theta_2}{1-\theta_3}c = 1, \text{ so } c = \frac{1-\theta_3}{2[\theta_2 + (1-\theta_3)]}.$$

Hence $\bar{\pi}_2 = \frac{\theta_2}{\theta_2+(1-\theta_3)}$. Because $\bar{\pi}$ derived in the preceding displayed equation is a stationary distribution if $\theta_1 = \theta_3 = 1$ (although the stationary distribution is not unique for this case), we assume the derived $\bar{\pi}$ forms the initial distribution following public news for all values (θ_1, θ_3). The formula displayed in the text is arrived at by noting $\theta_2 = 1 - \theta_1$.

A.2 Stock Valuation

The specialist's valuation $E(\Delta V_m | Z_i)$ is

$$\sum_{j=0}^{\infty} P[T\ (i) = j]\pi'_{j,i}s$$

with s = [1, 0, −1]′. As noted in the text, $P[T\ (i) = j] = \delta(1-\delta)^j$ and $\pi'_{j,i} = \pi'_{0,i}M^j$. Thus

$$E(\Delta V_m|Z_i) = \sum_{j=0}^{\infty} \delta\ (1-\delta)^j\ \pi'_{0,i} M^j s$$
$$= \delta\pi'_{0,i} \sum_{j=0}^{\infty} [(1-\delta)\ M]^j \cdot s$$
$$= \delta\pi'_{0,i} (I - (1-\delta)\ M)^{-1}\ s.$$

Now,

$$(I - (1-\delta)\ M) = \begin{pmatrix} a & b & \delta - a - b \\ c & \delta - 2c & c \\ \delta - a - b & b & a \end{pmatrix}$$

where $a = 1 - (1-\delta)\ \theta_1$, $b = -(1-\delta)\ \theta_2$, and $c = -\frac{1}{2}(1-\delta)\ (1-\theta_3)$. From matrix algebra $(I - (1-\delta)\ M)^{-1}\ s = \frac{(1,0,-1)'}{(2a-\delta+b)}$, so

$$\delta\pi'_{0,i}\ (I - (1-\delta)\ M)^{-1}\ s = \delta \cdot \frac{P(S_i = 1|Z_i) - P(S_i = -1|Z_i)}{1 - (1-\delta)[\theta_1 - (1 - \theta_1 - \theta_2)]}$$
$$= E(S_i|Z_i) \frac{\delta}{1 - (1-\delta)[\theta_1 - (1 - \theta_1 - \theta_2)]}.$$

To obtain the valuation of an informed trader, $E\ (\Delta V_m|S_i)$, simply replace the specialist's information set with the signal. Thus

$$E\ (\Delta V_m|S_i) = \delta \cdot \frac{P\ (S_i = 1|S_i) - P\ (S_i = -1|S_i)}{1 - (1-\delta)[\theta_1 - (1 - \theta_1 - \theta_2)]}$$
$$= S_i \cdot \frac{\delta}{1 - (1-\delta)[\theta_1 - (1 - \theta_1 - \theta_2)]}.$$

Proof of Theorem 1 The arrival time of trader i is given by

$$\tau_i = \tau_{i-1} + \min\ (f^{-1}, g),$$

where g is the elapsed time to the first recorded trade since τ_{i-1} and τ_0 is the beginning of the interval. If all recorded trades are at integer multiples of f^{-1}, then $A = Lf$. If some recorded trades are not at integer multiples, then for at most T arrivals, $\tau_i - \tau_{i-1} < f$. As a result, $A \leq Lf + T$, so

$$B = \frac{A}{T} - \frac{Lf}{T} \leq \frac{T}{T}.$$

As trades cannot be less than one second apart, if $f = 1$, then $\tau_i - \tau_{i-1} = f^{-1}$ for all arrivals and $B = 0$.

A.3 Firms Included in the 2002 Analysis

The Dow Jones Industrial firms (together with their ticker symbols) used for the analysis of data from 2002 are

American Express (AXP)	General Motors (GM)	3M (MMM)
Alcoa (AA)	Home Depot (HD)	Philip Morris (MO)
Boeing (BA)	Honeywell (HON)	Merck (MRK)
Citigroup (C)	IBM (IBM)	Procter Gamble (PG)
Caterpillar (CAT)	International Paper (IP)	SBC Communications (SBC)
Du Pont (DD)	Johnson and Johnson (JNJ)	ATT (T)
Walt Disney (DIS)	JP Morgan Chase (JPM)	United Technologies (UTX)
Eastman Kodak (EK)	Coca-Cola (KO)	Wal Mart (WMT)
General Electric (GE)	McDonalds (MCD)	Exxon Mobil (XOM)

Received March 19, 2004; revised April 19, 2005; accepted May 27, 2005

REFERENCES

Andersen, T. (1996). "Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility." *Journal of Finance 51*, 169–204.

Barclay, M., and T. Hendershott. (2003). "Price Discovery and Trading After Hours." *Review of Financial Studies 16*, 1041–1073.

Blume, M., and M. Goldstein. (1997). "Quotes, Order Flow, and Price Discovery." *Journal of Finance 52*, 221–244.

Damodaran, A. (1985). "Economic Events, Information Structure and the Return-Generating Process." *Journal of Financial and Quantitative Analysis 20*, 423–434.

Easley, D., N. Kiefer, and M. O'Hara. (1997). "One Day in the Life of a Very Common Stock." *Review of Financial Studies 10*, 805–835.

Easley, D., N. Kiefer, M. O'Hara, and J. Paperman. (1996). "Liquidity, Information, and Infrequently Traded Stocks." *Journal of Finance 51*, 1405–1436.

Easley, D., and M. O'Hara. (1992). "Time and the Process of Security Price Adjustment." *Journal of Finance 47*, 577–605.

Easley, D., M. O'Hara, and G. Saar. (2001). "How Stock Splits Affect Trading: A Microstructure Approach." *Journal of Financial and Quantitative Analysis 36*, 25–51.

Engle, R., and J. Russell. (1998). "Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data." *Econometrica 66*, 1127–1162.

Foster, A., and S. Viswanathan. (1993). "Variations in Trading Volume, Return Volatility and Trading Costs: Evidence on Recent Price Formation Models." *Journal of Finance 48*, 187–211.

Glosten, L., and P. Milgrom. (1985). "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders." *Journal of Financial Economics 14*, 71–100.

Hanousek, J., and R. Podpiera. (2002). "Information-Driven Trading at the Prague Stock Exchange: Evidence from Intra-day Data." *Economics of Transition 10*, 747–759.

Harris, L. (1986). "A Transaction Data Study of Weekly and Intradaily Patterns in Stock Returns." *Journal of Financial Economics 16*, 99–117.

Hasbrouck, J. (1988). "Trades, Quotes, Inventory, and Information." *Journal of Financial Economics 22*, 229–252.

Hasbrouck, J. (1999). "Trading Fast and Slow: Security Market Events in Real Time." Salmon Center Working Paper S/99/17, New York University.

Jain, P., and G. Joh. (1988). "The Dependence Between Hourly Prices and Trading Volume." *Journal of Financial and Quantitative Analysis 23*, 269–283.

Kelly, D., and D. Steigerwald. (2004). "Private Information and High-Frequency Stochastic Volatility" *Studies in Nonlinear Dynamics and Econometrics* 8, 1–28.

Kyle, A., 1985, "Continuous Auctions and Insider Trading" *Econometrica* 53, 1315–1335.

Lee, C. and R. Radhakrishna, 2000, "Inferring Investor Behavior: Evidence from the TORQ Data" *Journal of Financial Markets 3*, 83–112.

Lee, C. and M. Ready, 1991, "Inferring Trade Direction from Intraday Data" *Journal of Finance* 46, 733–746.

Madhavan, A., M. Richardson, and M. Roomans, (1997) "Why Do Security Prices Change? A Transaction-Level Analysis of NYSE Stocks." *Review of Financial Studies 10*, 1035–1064.

McInish, T., and R. Wood. (1992). "An Analysis of Intraday Patterns in Bid/Ask Spreads for NYSE Stocks." *Journal of Finance 47*, 753–764.

Owens, J., and D. Steigerwald. (2004). "Inferring Information Frequency and Quality." Available at http://www.departments.bucknell.edu/management/apfa/Stockholm%20Papers/Owens.pdf.