MODELING TRADE DIRECTION

DALE W.R. ROSENTHAL

ABSTRACT. I propose a modeling approach to classifying trades as buys or sells. Modeled classifications consider information strengths, microstructure effects, and classification correlations. I also propose estimators for quotes prevailing at trade time. Comparisons using 2,800 US stocks show modeled classifications are 1–2% more accurate than current methods across dates, sectors, and the spread. For Nasdaq and NYSE stocks, 1% and 1.3% of improvement comes from using information strengths; 0.9% and 0.7% of improvement comes from estimating quotes. I find evidence past studies used unclean data and indications of short-term price predictability. The method may help detect destabilizing order flow. (*JEL*: C53, D82, G14)
Keywords: trade classification, delay models, trade publishing delays, prevailing quotes, trade direction, flash crash detection

Studies of market dynamics often require classifying the aggressor in a trade as buyer or seller. Researchers often infer this classification because it is missing from most available databases of trades and quotes. While many different methods have been proposed, this paper is the first to model trade classifications and trade reporting delays. We then show this approach is more accurate than previously proposed methods.

Trade classification is used in many different areas and is thus important to many researchers. For example, some event studies seek to determine the balance of buying and selling in response to a release of information or a governmental action. Odders-White (2000) notes that classification inaccuracies have yield spurious inferences, such as "buys" following negative earnings news. She also has shown that misclassification results in overestimating effective spreads by about \$0.01. This is troubling since effective spreads are used to study the effects of changes in market structure or policy and to infer the portion of the spread due to adverse selection.

University of Illinois at Chicago, Department of Finance, Chicago, IL 60607, email: daler@uic.edu.
I thank Per Mykland, ex-colleagues from Morgan Stanley's Equity Trading Lab, Vanja Dukic, David Modest, Stephen Stigler, Torben Andersen, and two referees. Helen Barounis at NYSE Arca provided data help; John Zekos and the Stevanovich Center for Financial Mathematics at the University of Chicago provided computing resources. Financial support from the National Science Foundation under grants DMS 06-04758 and SES 06-31605 is also gratefully acknowledged. Address correspondence to: Dale W.R. Rosenthal, University of Illinois at Chicago, Department of Finance, Chicago, IL 60607 or email: daler@uic.edu.

Boehmer *et al.* (2007) document that classification inaccuracy downwardly biases estimates of the probability of informed trading (aka PIN). Others use trade classifications to study stealth trading and to measure liquidity. These suggest a novel use for trade classification methods that measure signal strength (as done here): regulators and market authorities may use them to detect destabilizing order flow. Such a method might have detected the 6 May 2010 "flash crash" and allowed it to be stopped long before markets were severely affected.

Further afield, trade classifications have been used to study accounting issues, such as Boone and Raman's (2001) study of off-balance-sheet assets and liquidity and Danielsen *et al.*'s (2007) study showing auditors price firm opacity into fees. Corporate finance researchers also use trade classifications, as in Schultz and Zaman's (1994) and Benveniste *et al.*'s (1998) studies of secondary-market price supports for IPOs.

Inaccurate trade classifications used in a regression model create an errors-in-variables problem which biases coefficient estimates. Aigner (1973) shows this is worse for binary variables (*e.g.* trade classification). This bias also causes trouble for researchers who seek to model the trade generation process by determining if trade classifications are autocorrelated. Tanggaard (2004) documents these problems and explores the effects in depth.

This bias also affects researchers who use trade classifications to estimate price impact models. These models measure the effects of liquidity demands and are used to assess market efficiency, transactions costs, and risk (since liquidity has been shown to be a priced risk factor). Since estimating price impact models can be noisy, even a modest improvement in classification accuracy is worthwhile. More accurate price impact models enable better trading of customer orders and more accurate return predictions based on inferred price impact (for investment funds). A 1%–2% accuracy improvement would increase estimates of price impact by 2%–4% and result in more careful and inexpensive trading of illiquid orders. Given that a large investment bank may trade $2 billion per day, a cost savings of 0.01% would be worth $100 million annually to one such bank.

A few methods exist for classifying trades, most comparing trade prices to prevailing price quotes. Since these methods all have sensible economic rationales, choosing which is "superior" can be difficult and may vary with the data. A further complication is the delays between publishing times (*i.e.* timestamps) of trades and quotes; however, classification methods without rigorous delay assumptions are often compared.

To improve classification accuracy, I incorporate different methods into a model for the likelihood a trade was buyer-initiated. I also allow for joint estimation of a (latent) delay model. Modeling trade classifications is a new approach and one of the unique contributions of this work. A modeling approach lets us incorporate richer information, such as: information from multiple tests; the strengths of those test results; effects of microstructure

peculiarities (*e.g.* short sales rules)[1]; autocorrelations and cross-correlations in buys/sells; and, the likelihood our classification is correct.

Previous work has mainly focused on the time lag between publishing trades and contemporaneous bid and ask quotes. Recent work implies that delays are now small; however, a commensurate increase in quote updates means these small delays are still not ignorable. Also, even if such issues were resolved today, researchers studying liquidity would still need to classify older trades. Further, the increasing fragmentation of markets (and trading in dark pools) suggests that trades are reported with varying delays. Finally, attempting to pick correct quotes is likely to add more noise to inferences than estimating quotes. For these reasons, I use a model to estimate contemporaneous quotes.

Thus I explore two improvements: the use of estimated quotes and a modeling approach to trade classification. The goal of these improvements is to increase the accuracy of trade classification.

I begin by discussing different classification methods and previous analyses. I next discuss the two improvements: delay models for estimating quotes and the modeling approach to trade classification. I then describe the data used to demonstrate these improvements followed by the model estimation. I analyze the estimated model's out-of-sample performance in depth; and, finally I conclude and suggest areas for further study.

1. Classification Methods and Previous Analyses

Trade classification infers which trade participant initiated a trade by being the aggressor, consistent with Odders-White (2000) defining the later-arriving order as the trade initiator. Three approaches to trade classification dominate the literature: tick tests, midpoint tests, and bid/ask tests.

Finucane (2000) recommends a tick test: comparing a trade price to the previous (differing) trade price for that stock. A lower previous trade price is taken as evidence the current trade was buy-initiated.

Lee and Ready (1991) suggested a midpoint test: comparing a trade price to the lagged midpoint (*i.e.* average of best bid and ask quote). Trades at prices above (below) the midpoint are classified as buy- (sell-) initiated. Trades at the midpoint are resolved with a tick test.

Ellis *et al.* (2000) suggested a bid/ask test for Nasdaq stocks; Peterson and Sirri (2003) then suggested it for NYSE stocks. Trades at the lagged ask (bid) are classified as buy- (sell-) initiated; other trades are resolved with a tick test.

Trades are published with delay relative to quotes. Therefore, midpoint and bid/ask methods require delay assumptions. For midpoint methods, Lee and Ready (1991) use a delay of five seconds; Vergote (2005) suggests

[1]Asquith *et al.* (2010) notes the difficulties some classification tests have if short-sales are only allowed on zero-plus ticks.

two seconds; and, Henker and Wang (2006) suggest a one-second delay. For bid/ask methods, all previous analyses used unlagged quotes.

The datasets used in previous analyses also bear consideration. Computing power and data availability limited past analyses to a small number of stocks (often a small sub-section of the market) for only one listing exchange. Past analyses also used older data although this study does not totally escape that problem (our data are from December 2004). Table 1 summarizes datasets used in previous analyses.

Study	Stocks (#, Type)		Market	Year
Lee and Ready (1991)	150	Large cap	NYSE	1987
Hasbrouck (1992)	144	Large cap	NYSE	1990
Ellis *et al.* (2000)	313	Post-IPO internet	Nasdaq	1997
Peterson and Sirri (2003)	N/A	Round-lot SuperDOT	NYSE	1997
Henker and Wang (2006)	401	Large cap	NYSE	1999
Chakrabarty *et al.* (2007)	750	ARCA+INET trades	Nasdaq	2005

TABLE 1. Summary of datasets used in past analyses. None look at both NYSE and Nasdaq stocks simultaneously. Apart from Chakrabarty *et al.* (2007), all look at a small sample of stocks that do not represent the entire market.

2. Delays Between Trades and Quotes

Many studies note that trades are published with non-ignorable delays. Lee and Ready (1991) first suggested a five-second delay (now commonly used) for 1988 data, two seconds for 1987 data, and "a different delay ... for other time periods". Ellis *et al.* (2000) note (Section IV.C) that quotes are updated almost immediately while trades are published with delay[2]. Therefore, determining the quote prevailing at trade time requires finding quotes preceding the trade by some (unknown) delay.

Important sources of this delay include time to notify traders of their executions, time to update quotes, and time to publish the executions. For example, an aggressive buy order may trade against sell orders and change the inventory (and quotes) available at one or more prices. Notice is then sent to the buyer and sellers; quotes are updated; and, the trade is made public. This final publishing timestamp is what researchers see in non-proprietary transaction databases.

Erlang's (1909) study of information delays forms the theory for modeling delays. Bessembinder (2003) and Vergote (2005) are probably the best prior studies on delays between trades and quotes.

Trades collected from all markets come from differing venues with different distributions for the preceding delays (and no formal clock synchronization).

[2]Interestingly, they note this but use no delay between trades and quotes.

This further complicates the use of quotes and might make delay models more important than the analysis here suggests.

2.1. **Delay Models for Quotes.** Since the total delay is determined by a chain of events, it is a sum of constituent delays. While market participants prefer short data paths ("fresher data"), delay constituents may be correlated[3]. Using the Central Limit Theorem to approximate the delay distribution is thus unwise and why I explore small-sample approximations. These approximations require that (1) delay constituents are exponentially distributed; (2) observations of total delay are independent; and, (3) there are at least two delay constituents. To express these ideas mathematically, we require some notation:

$$\begin{aligned}
Y &= \text{delay between trade timestamps and quotes used by initiators;}\\
\kappa_r &= r\text{-th cumulant of total delay } Y;\\
\nu, \hat{\nu} &= \text{number, estimated number of delay constituents;}\\
\hat{\lambda} &= \text{estimated iid-delay rate parameter of a gamma distribution;}\\
\tilde{\kappa}_r &= r\text{-th pseudocumulant of total delay } Y;\\
f_Y(y) &= \text{density of total delay } Y \ (y \geq 0);\\
b_t, \hat{b}_t &= \text{bid, estimated bid prevailing at time } t;\\
a_t, \hat{a}_t &= \text{ask, estimated ask prevailing at time } t;\\
m_t, \hat{m}_t &= \text{midpoint, estimated midpoint prevailing at time } t,\ \hat{m}_t = (\hat{b}_t + \hat{a}_t)/2;\\
p_t &= \text{price of a trade reported at time } t; \text{ and,}\\
\mathcal{F}_t &= \text{past prices; } \sigma < a_s, b_s, p_s : s \leq t >.
\end{aligned}$$

We assume the delay Y_i associated with the i-th trade (recorded at time t) is independent of $\mathcal{F}_t$[4]. Gamma distribution parameters are chosen to match the sample mean and variance (κ_1, κ_2). This yields $\hat{\nu} = \kappa_1^2/\kappa_2$ and $\hat{\lambda} = \kappa_1/\kappa_2$. Pseudocumulants $\tilde{\kappa}_r$'s are as in McCullagh (1987): differences between sample cumulants (κ_r's) and those of the Gamma($\hat{\nu}, \hat{\lambda}$) distribution.

2.1.1. *Small-Sample Approximations.* I use a gamma-based Edgeworth approximation to the total delay density. This form is simple; low-order approximations are likely to fit well; and, it puts no probability mass on negative delays. The regularity conditions may preclude some or all of the correction terms[5], however the base gamma density alone often fits well.

[3]Heavy information flow at the start of a data path implies heavy information flow throughout the path and induces correlations.

[4]The independence assumption seems reasonable since trade publishing delays are affected by many unobservable variables such as the publishing system, processing capacity, contemporaneous system load, and other processing happening at the same time.

[5]See Rosenthal (2008) for discussion of the regularity conditions.

$$
\begin{aligned}
f_Y(y) = \gamma_{\hat{\nu},\hat{\lambda}}(y) &+ \frac{\tilde{\kappa}_3\hat{\lambda}^3}{6}\sum_{j=0}^{3}(-1)^{3-j}\binom{3}{j}\gamma_{\hat{\nu}-j,\hat{\lambda}}(y) \\
&+ \frac{\tilde{\kappa}_4\hat{\lambda}^4}{24}\sum_{j=0}^{4}(-1)^{4-j}\binom{4}{j}\gamma_{\hat{\nu}-j,\hat{\lambda}}(y) \\
&+ \frac{\tilde{\kappa}_3^2\hat{\lambda}^6}{72}\sum_{j=0}^{6}(-1)^{6-j}\binom{6}{j}\gamma_{\hat{\nu}-j,\hat{\lambda}}(y) + O(\nu^{-3/2}),
\end{aligned} \tag{1}
$$

where $\gamma_{\nu,\lambda}(y)$ is the Gamma(ν,λ) pdf if $\nu > 0$, 0 otherwise.

2.2. **Prevailing Quotes.** Since there are delays between quotes and trades, we use the approximate f_Y to estimate prevailing quotes. If the delay density is $f_Y(y)$, the expected value of the ask for a trade recorded at time t is:

$$\tilde{a}_t = E(a_t|\mathcal{F}_t) = \int_0^\infty a_{t-z} f_Y(z)dz, \tag{2}$$

since positive delays ($z > 0$) correspond to quotes further in the past[6].

We truncate the integration at T and estimate the ask price $\tilde{a}_t$ by $\hat{a}_t$. Since a_t is bounded and $f_Y(z)$ dies off quickly (like e^{-z}) as z increases, we can choose T so that $\tilde{a}_t$ and $\hat{a}_t$ are arbitrarily close. Since quotes are simple processes, this simplifies to a sum involving the delay CDF $F_Y(s)$:

$$\hat{a}_t = \frac{\int_0^T a_{t-z} f_Y(z)dz}{\int_0^T f_Y(z)dz} = \sum_{i=1}^{n} a_{t-s_i}\frac{F_Y(s_i;\kappa) - F_Y(s_{i-1};\kappa)}{F_Y(s_n;\kappa)} \tag{3}$$

where $t - s_i$ are observed quote times with $s_0 = 0$ and $s_n = T$. F_Y depends on unknown κ's estimated jointly with the classification model.

3. Trade Classification Model

The idea of a trade classification model is simple: instead of using one classification method, use many of them and let them "vote."

3.1. **Information Strength.** It seems sensible to consider the strength of information used to classify trades. However, previous studies note *lower* accuracy for trades outside the prevailing spread. An online addendum discusses how this may result from past studies using contaminated data.

To measure information strength relative to midpoints or previous trades, I use a log-return function $g = \log(p_t) - \log(\hat{m}_t)$. While not explored here, it might be more informative to divide g by the stock-specific volatility or mean spread.

[6]This and the following argument also apply to the expected bid $\tilde{b}_t$ and midpoint $\tilde{m}_t$.

To measure information strength relative to the bid/ask, we consider proximity since the estimated bid and ask may not be decimal prices. The proximity function J is approximately -1 and +1 for trade prices near the expected bid and ask:

$$J(p_t, \hat{b}_t, \hat{a}_t; \tau) = \exp\left(-\left(\frac{p_t - \hat{a}_t}{\tau}\right)^2\right) - \exp\left(-\left(\frac{p_t - \hat{b}_t}{\tau}\right)^2\right) \tag{4}$$

where τ controls the J function width near the estimated quotes. (A picture of J for various values of τ is in Figure 1.) To reinforce when we work with information strength, I call these measures (g and J) "metrics." Boolean results (*e.g.* $I(p_t > \hat{m}_t)$) are referred to as "tests."

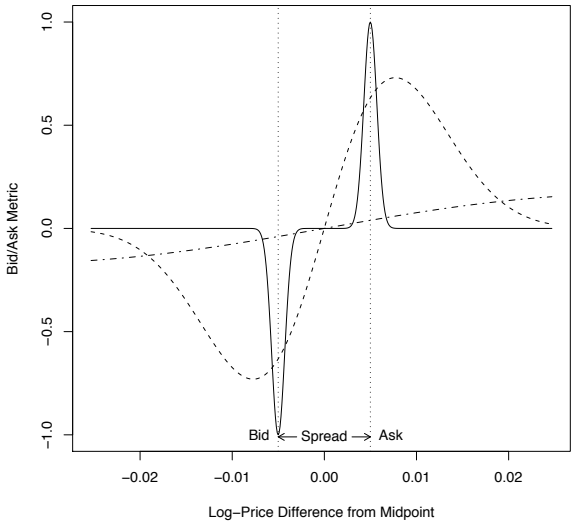

FIGURE 1. Example of bid/ask metric $J(\cdot; \tau)$ for a 1% bid-ask spread and $\tau = 0.001$, 0.01, and 0.05. As τ grows, J becomes nearly linear (like g) over a wide range. For large τ, J may be collinear with midpoint metric g.

3.2. **Model.** To express the model mathematically, we define:

t = time at which a trade is published;
p_{t-} = price of most recent trade reported before time t;
p'_{t-} = price of most recent differing-price trade reported before time t;
B_t = initiating side of trade reported at time t, (1=buy, 0=sell);
$\hat{B}_t$ = predicted initiating side of trade reported at time t;
π_t = P(trade at time t was a buy) = $P(B_t = 1|\mathcal{F}_t)$; and,
η_t = linear model prediction at time t (log-odds of $B_t = 1|\mathcal{F}_t$).

Since the response we predict is binary (buyer- or seller-initiated), I use a logistic-link generalized linear model (GLM). Errors are assumed to be independent and to have a Bernoulli variance (*i.e.* $\text{Var}(\pi_t) = \pi_t(1-\pi_t)$)[7].

In a panel of multiple stocks, classifications may be serially- and cross-correlated. For example, the classification of a trade in General Motors stock might help infer the classification of a contemporaneous trade in Ford stock or in *any* stock. We model these classification correlations in order to get more accurate standard errors and inferences.

Correlation modeling is done separately (i) across time and (ii) cross-sectionally within sectors. Lagged (*i.e.* preceding trade in that stock) midpoint, tick, and bid/ask metrics were used to capture autoregressive behavior. Random effect terms capture cross-correlations within contiguous time periods ("bins") by imposing a covariance structure. Time bins are used since stocks rarely trade simultaneously.

Correlated classifications among all trades in a time bin are handled by a time random effect: all trades within that time bin are assumed to have a common bias toward being buyer- or seller-initiated. This corrects for unpredictable momentum across all stocks and also allows for higher volatility during some portions of a trading day. Correlated classifications among stocks in the same sector and time bin are handled by a sector$\times$time random effect: all trades for a given sector within that time bin are assumed to have a common bias toward being buyer- or seller-initiated. This corrects for unpredictable momentum across stocks in a sector. While correlations across industries and individual stock pairs may also be significant, this would be unwieldy for large numbers of stocks.

A misspecified dependence structure clouds insight into classification correlations. However, Heagerty and Zeger (2000) suggest this is better than using no correlation model. While bootstrap or Huber-White standard errors may help, Kauermann and Carroll (2001) show they may greatly underperform explicit correlation modeling.

3.3. **Partial Likelihood.** The classification model is formally valid when formulated as a partial likelihood as in Cox (1975) and Wong (1986). Since we are classifying a sequence of trades and conditioning on $\mathcal{F}_t$, t is not random. The randomness in the (conditional) classification model is due only to (i) the unknown amount of time to look backwards for a quote; and, (ii) the unknown trade classification. Were this not so, we would need to condition on the likelihood of each trade happening at its observed time.

If t_i is the i-th trade time and $\mathcal{G}_{i-1}$ is a sigma-field encapsulating trade classifications $1, \ldots, i-1$, the full likelihood ratio can be decomposed as:

$$\mathcal{L}(\text{all data}) = \prod_{i=1}^{n} \mathcal{L}(B_{t_i} | \mathcal{F}_{t_i}, \mathcal{G}_{i-1}) \times \prod_{i=1}^{n} \mathcal{L}(\mathcal{F}_{t_i}, \mathcal{G}_{i-1} | \mathcal{F}_{t_{i-1}}, \mathcal{G}_{i-1}). \tag{5}$$

[7]Error assumptions for GLMs are detailed in McCullagh and Nelder (1989).

For inference we only use the first factor, making this a partial likelihood. We assume B_{t_i} is conditionally independent of $\mathcal{G}_{i-1}$ given $\mathcal{F}_{t_i}$, yielding $\mathcal{L}(B_{t_i}|\mathcal{F}_{t_i}, \mathcal{G}_{i-1}) = \mathcal{L}(B_{t_i}|\mathcal{F}_{t_i})$.

3.4. **Model Statement.** Thus we can now state the (conditional) model:

$$
\begin{aligned}
&P(B_{jt} = \text{Buy}|\mathcal{F}_t, c_k, d_{k\ell}; \theta_o, \kappa_o) = \pi_{jt};\\
&\pi_{jt} = \text{logit}(\eta_{jt}); \quad \text{and,}\\
&\eta_{jt} = \underbrace{\beta_0}_{\substack{\text{bias}\\ =0?}} + \underbrace{\beta_{o1} g(p_{jt}, \hat{m}_{jt})}_{\text{midpoint metric}} + \underbrace{\beta_{o2} g(p_{jt}, p'_{jt-})}_{\text{tick metric}} + \underbrace{\beta_{o3} J(p_{jt}, \hat{b}_{jt}, \hat{a}_{jt}; \tau)}_{\text{bid/ask metric}} +\\
&\qquad \underbrace{\beta_{o4} g(p_{jt-}, \hat{m}_{jt-})}_{\text{lag-1 midpoint metric}} + \underbrace{\beta_{o5} g(p_{jt-}, p'_{jt-})}_{\text{lag-1 tick metric}} + \underbrace{\beta_{o6} J(p_{jt-}, \hat{b}_{jt-}, \hat{a}_{jt-}; \tau)}_{\text{lag-1 bid/ask metric}} +\\
&\qquad \underbrace{c_k}_{\substack{\text{overall}\\ \text{effect}}} + \underbrace{d_{k\ell}}_{\substack{\text{within-}\\ \text{sector}\\ \text{effect}}}
\end{aligned}
\tag{6}
$$

where j indexes stocks, k indexes time bins; ℓ indexes sectors; and, o indexes primary exchanges (*e.g.* NYSE, Nasdaq). A stock j thus implies a sector ℓ and primary exchange o. The parameters of f_Y are estimated jointly with model coefficients. The random effects are assumed to be $c_k \overset{\text{iid}}{\sim} N(0, \sigma_c^2)$ for all bins k and $d_{k\ell} \overset{\text{iid}}{\sim} N(0, \sigma_d^2)$ for all bins k and sectors ℓ. Further, c_k and $d_{k\ell}$ are assumed to be independent of the sigma-field $\mathcal{F}_t$.

3.5. **Multi-Stock Model Coefficients.** Stocks are listed on a primary exchange; and, as shown by Stoll (2006), this largely determines where trading takes place[8]. Therefore, different parameters are estimated for stocks with different primary exchanges. This implies that there are (differing) stock-specific coefficients and that the estimated model coefficients are averages of these stock-specific coefficients weighted by the number of transactions.

The main reason to use these "population average" coefficients is for precise comparison. Current classification methods use the equivalent of population average parameters: parameters chosen for overall classification performance. Thus a fair comparison between these methods and a modeling approach requires using population average coefficients.

To classify trades for just one stock or to improve classification performance, we would use random model coefficients or normalize the information strength functions g and J by stock-specific measures like volatility or spread[9]. This is briefly explored in an online addendum.

[8]The Nasdaq is a system of market centers which I abstract as a monolithic market.

[9]We might also examine other characteristics: trade size, typical volume, liquidity risk, index memberships, and intraday patterns (diurnals) in these measures.

4. Data and Description

To explore the modeling approach, I use the ArcaTrade dataset which gives the non-initiating (first arriving) trade classification and has all trades on the Archipelago ECN and Exchange for December 2004. Archipelago Holdings, Inc. (2005b) reports their share of traded volume as 22.5% and 2.3% for Nasdaq- and NYSE-listed stocks. For inside quotes, I used the ArcaSIP consolidated NBBO dataset for the same month.

4.1. **Data Synchronization.** The lagged tick test in the model forced a choice of data sources: use preceding trades from ArcaTrade or try to get them from a market-wide source like TAQ? Using another data source would require finding the Arca trade and then the preceding trade. This matching would be less accurate for more common (smaller) trades and could induce serious bias. This would also mix datasets timestamped by different clocks. Since time is a crucial part of the analysis, that is a troubling prospect.

Using the preceding Arca-executed trade avoids these issues. While that trade might not be the preceding trade market-wide (*e.g.* the preceding trade might have happened on the NYSE), I assume this merely adds noise to the tick test covariates. If the *location* of trading is autocorrelated, this assumption would not hold.

4.2. **Data Cleaning and Augmentation.** ArcaTrade data includes pre- and post-market trades. Unlike TAQ, ArcaTrade data lacks negotiated and auction trades. To eliminate opening and closing auction effects, I exclude all trades before 10:00 AM and after 3:30 PM. To handle the one-second resolution of trade and quote timestamps, I assume messages in a file are correctly ordered and uniformly distributed within a second. Thus two messages at "9:35:01" are assumed to have occurred one- and two-thirds of a second after "9:35:01".

I restrict attention to stocks in the Russell 1000 large–mid-cap and 2000 small-cap indices (the "Russell 3000") as of the 2004 annual June rebalance.

4.3. **Summary Statistics.** The resulting dataset covered two days: 1–2 December 2004; 2,178,307 transactions across the 2,836 stocks in the Russell 3000 still active under the same ticker as on the rebalance date[10]. These stocks represented all three primary US markets (AMEX, Nasdaq, and NYSE) and 13 sectors. AMEX stocks were a fraction of the data (2,797 trades), so AMEX-specific analysis is omitted. Characteristics of NYSE and Nasdaq stocks are shown in Tables 2 and 3.

Table 2 shows that the average Nasdaq spread is about half that of the NYSE, but the average Nasdaq trade size is about three-quarters that of the NYSE. Also, while most trades are for Nasdaq stocks, we still have a large sample of NYSE trades.

[10]While more days for in-sample estimation would have been preferrable, matrix inversions needed to fit random effects and the time needed to fit nonlinear parameters ($\nu, \lambda, \tau, \kappa$'s) limited estimation to using two days of data.

	Number of		Trade-Weighted Average			
Market	Stocks	Trades	Shares	Price	Mkt Cap	Spread
Nasdaq	1,391	2,014,236	319.4	27.59	6,252MM	0.07%
NYSE	1,420	161,274	406.2	38.98	6,785MM	0.15%
Total	2,836	2,178,307	326.1	$28.51	$6,285MM	0.14%

TABLE 2. Characteristics by market of the stocks analyzed. All were members of the Russell 1000 or 2000 as of July 2004. Total includes 2,797 trades in 25 AMEX stocks. While the trade-weighted spread is larger for NYSE stocks, the average trade size is smaller on the Nasdaq.

Table 3 (a sector breakdown) shows that most stocks in the dataset are service-related companies while most trades are for technology-related companies. The unusually small industrial goods sector is an artifact of the data-gathering process and changing sector names over time[11].

	Number of		Trade-Weighted Average			
Sector	Stocks	Trades	Shares	Price	Mkt Cap	Spread
Capital Goods	159	24,976	187.4	$39.78	$5,732MM	0.13%
Conglomerates	19	3,728	307.2	54.63	1,328MM	0.03%
Cons. Cyclical	121	32,754	229.9	31.92	2,580MM	0.13%
Energy	110	40,542	251.5	34.79	2,984MM	0.09%
Financial	468	126,337	193.7	35.80	5,194MM	0.12%
Healthcare	338	295,327	227.4	28.42	6,222MM	0.12%
Indust. Goods	2	827	355.2	12.82	1,012MM	0.14%
Materials	149	38,605	228.9	36.88	9,150MM	0.10%
Non-Cyclical	95	20,262	221.0	33.96	1,783MM	0.13%
Services	657	433,999	349.0	36.26	3,809MM	0.09%
Technology	573	1,107,925	372.3	23.59	7,687MM	0.18%
Transportation	60	43,065	319.1	28.82	5,789MM	0.10%
Utilities	95	9,960	379.5	30.76	3,495MM	0.10%

TABLE 3. Characteristics by sector of the stocks analyzed. All were members of the Russell 1000 or 2000 as of July 2004. The small industrial goods sector is an artifact of data collection. The spread is largest for technology stocks and smallest for conglomerates; however, this does not affect classification accuracy.

Summary statistics (Table 4) for the tick, midpoint, and bid/ask metrics by market. The means confirm there were no major imbalances between

[11]The net effect for the model is minor: The random effects had slightly more freedom to account for covariances. Coefficient estimates and out-of-sample prediction are unaffected.

buying and selling. The extremes and standard deviations show the order of magnitude for the covariates. Thus Nasdaq stocks often trade within 10bp[12] and 16bp of the midpoint and preceding trade; NYSE stocks often trade within 10bp and 24bp of the midpoint and preceding trade.

Market	Metric	Min	Mean	Max	St. Dev.
Nasdaq	Tick	-0.11	3.8×10^{-5}	0.12	1.6×10^{-3}
	Midpoint	-0.17	3.6×10^{-5}	0.06	1.0×10^{-3}
	Bid/Ask	-1.00	4.8×10^{-2}	1.00	0.79
NYSE	Tick	-0.05	3.5×10^{-5}	0.05	2.4×10^{-3}
	Midpoint	-0.05	4.8×10^{-5}	0.02	9.6×10^{-4}
	Bid/Ask	-1.00	0.10	1.00	0.79

TABLE 4. Summary statistics by market of the covariates used in the classification model. The midpoint and tick metrics compare prices in percentage (not absolute) terms. The bid/ask metric is approximately +1 and -1 for trade prices near the estimated ask and bid.

Correlation matrices for covariates and their lagged values (Table 5) show that current-trade metrics are strongly correlated with preceding-trade metrics. Also, the bid/ask metric is more strongly correlated with other metrics on the (decentralized) Nasdaq than on the (specialist-driven) NYSE. This suggests more trading near the bid/ask for Nasdaq versus NYSE stocks.

	Nasdaq			NYSE		
	Bid/Ask	Midpt	Tick	Bid/Ask	Midpt	Tick
Midpoint	0.47			0.43		
Tick	0.36	0.55		0.18	0.35	
Pr. Bid/Ask	0.58	0.30	0.20	0.38	0.22	0.07
Pr. Midpt	0.29	0.63	0.20	0.21	0.64	0.15
Pr. Tick	0.22	0.33	0.55	0.08	0.20	0.57

TABLE 5. Correlations by market between covariates used in the classification model. Correlations between lagged covariates are omitted since they are identical to correlations between unlagged covariates.

Plots comparing the bid/ask, midpoint, and tick metrics suggest they are correlated, but that some of the correlation is due to extreme observations. Plotting the lagged versus unlagged metrics suggests serial correlations and cross-correlations. Finally, a plot of the midpoint versus tick metrics for NYSE stocks shows secondary modes which are absent from the Nasdaq plot and may be related to trading by the NYSE specialist.

[12]A basis point (1bp) is 1/100-th of 1%.

5. Model Estimation

5.1. **Estimation Caveats.** Since delay parameters are used to estimate quotes used in (6), the model cannot be linearized nor is there a tractable way to get closed-form gradients or Hessians for the log-likelihood. Further, the τ used in the $J(\cdot;\tau)$ bid/ask metric must be positive and small relative to a typical bid-ask spread to avoid collinearity issues between J and g (see Figure 1). A weak restriction, $\tau < 0.05$ (*i.e.* a 5% spread) preserves identifiability and model interpretation.

The nonlinearity, lack of closed-form derivatives, and constraints were handled with a modified conjugate direction method to find optimal non-linear (delay) parameters[13]. The approximation of gradients and inverse Hessians yielded inflated estimates of standard errors for delay parameters. Standard errors for classification parameters were estimated from summing an MCMC resampling variance and the variance in coefficient estimates from dividing the dataset into days[14].

5.2. **Estimation.** Estimating the model with the ArcaTrade dataset and performing model selection via backward elimination yielded the coefficient estimates shown in Table 6. The fixed effect parameter estimates are highly significant and roughly in line with results from various analyses of deviance. In addition, the intercept is not large enough to be troubling. Since the model coefficients are "population average" coefficients (for comparison with other classification methods), I assume the coefficients are averages of stock-specific coefficients weighted by how often each stock is traded. This means standard errors reflect both (i) differences between population betas and sample estimates and (ii) variation in the population betas due to the averaging of stock-specific betas.

The estimated parameters imply delays of 5.0 seconds and 0.8 seconds for Nasdaq and NYSE trades versus quotes. While these estimates are nearly the opposite of those used as accepted practice[15], they are not surprising since the distributed nature of the Nasdaq market should yield longer delays in trade reporting. The sign difference for the Nasdaq versus NYSE tick metric is likely due to differing short-sales rules in force when these trades occurred. This suggests modeled trade classifications avoid the problems Asquith *et al.* (2010) find with classifying short sales. Negative coefficients for previous bid/ask metrics are consistent with negative autocorrelations from bid-ask bounce, as in Roll (1984). The random effects suggest a 2% correlation of buying or selling across same-sector stocks in a ten-minute period and a 0.2% correlation of buying or selling across all stocks in a ten-minute period. The significance of lagged metrics suggests even shorter-term predictability of buying and selling.

[13]Chapter 9 of Nocedal and Wright (2006) has more on conjugate direction methods.

[14]The combination of these variances is similar to an analysis of variance.

[15]The LR method uses a delay of 5 seconds for NYSE stocks; the EMO method uses a delay of 0 seconds for Nasdaq stocks.

Fixed Effect	Nasdaq	NYSE
J width (τ)	Overall: 2.1×10^{-4} (0.3)	
Delay count (ν)	1.65 (0.65)	0.62 (0.47)
Delay rate (λ)	0.33 (0.40)	0.78 (0.35)
Excess skew ($\tilde{\kappa}_3$)	—	—
Excess kurtosis ($\tilde{\kappa}_4$)	—	—
Intercept	Overall: 0.06 (0.02)	
Midpoint	209 (11)	122 (13)
Tick	29.4 (8.4)	-20.5 (8.5)
Bid/Ask	1.41 (0.02)	2.04 (0.20)
Prev. Midpoint	—	—
Prev. Tick	—	—
Prev. Bid/Ask	-0.14 (0.01)	-0.17 (0.05)
Random Effect	Std. Dev.	
Time Bin	0.08 (0.01)	
Sector × Time Bin	0.27 (0.03)	
Residual Deviance: 2,390,436		

TABLE 6. Estimated parameters and standard errors for trade direction model in equation (6). For numerical reasons, standard errors for the nonlinear parameters (τ, ν, λ) are overstated. The sign difference for tick metric coefficients is likely due to differing short sale rules for NYSE versus Nasdaq stocks.

6. Out-of-Sample Performance

6.1. **Using the Model to Classify Trades.** Having fit model coefficients, classifying trades with the model is straightforward. If we want an easy, computationally-light method, we can just use metrics instead of tests and forego quote estimation. Thus we would calculate differences of log-prices (or the J function) instead of comparing those prices; we then sum these metrics multiplied by their model coefficients. If the result is positive, we classify the trade as a buy.

For example, suppose a NYSE stock trades at \$21.13 when the unlagged quotes are \$21.10–\$21.15, the 5-second-lagged quotes are \$21.09–\$21.15, and the previous trade was at \$21.09 with a then-unlagged quote of \$21.07–\$21.10. We would calculate log-odds of $122\log(\frac{21.13}{21.12}) - 20.5\log(\frac{21.13}{21.09}) + 2.04(e^{-(\frac{21.15-21.13}{0.00021})^2} - e^{-(\frac{21.13-21.09}{0.00021})^2}) - 0.17(e^{-(\frac{21.10-21.09}{0.00021})^2} - e^{-(\frac{21.09-21.07}{0.00021})^2}) = 0.0189$. Since the log-odds are positive, we would classify the trade as a buy. The probability the trade is a buy is $\exp(0.0189)/(1+\exp(0.0189)) = 0.505$.

For those willing to pay the computational price for even higher accuracy, we can use the delay model-estimated quotes for computing the metrics.

This more involved approach is used here; however, foregoing this step would still be more accurate than any of the current methods.

6.2. **Competing Schemes.** The last 20 days in December 2004 were used for out-of-sample evaluation. Modeled trade classifications were compared to those from other methods.

The Lee and Ready test uses quotes from five seconds earlier; however, quotes in this dataset are only published with one-second resolution. Since multiple quotes might be found in the $(-6s, -5s]$ window, the Lee and Ready test using the oldest "five seconds prior" quote is referred to as "LR.old" while the test using the newest such quote is referred to as "LR.new".

The methods evaluated are henceforth referred to as: Modeled (the model in (6)), EMO (Ellis, Michaely, and O'Hara's method), LR.new, LR.old, and Tick (a tick test).

6.3. **Trade Classification Accuracy.** Across markets, sectors, and dates, modeled trade classifications are generally the most accurate followed by the EMO bid/ask test, LR.new midpoint test, LR.old midpoint test, and the tick test. We can see this in Tables 7 and 8 and Figure 2.

		Percent of Trades Correctly Classified				
Market	N	Modeled	EMO	LR.new	LR.old	Tick
Nasdaq	15,220,579	**74.3%**	72.3%	71.8%	71.4%	66.7%
NYSE	1,264,866	**80.7%**	79.6%	76.1%	75.6%	60.7%
Total	16,504,880	**74.7%**	72.8%	72.1%	71.7%	66.2%

TABLE 7. Percent of trades correctly classified across all US markets for 3–31 December 2004. Excluding December 31 would lower accuracies by about 2% for the NYSE. The totals reflect 19,435 AMEX trades (not shown).

Modeled classifications are more accurate than all other methods. The model accuracy versus the next best method is 2% higher for Nasdaq stocks and 1.1% higher for NYSE stocks. Ten to twenty years ago, the LR, EMO, and tick methods were more accurate: 75% to 85%. More recent studies have typically noted lower accuracy.

If we excluded December 31, all methods would be about 2% less accurate for NYSE stocks (and only slightly worse for Nasdaq stocks). Why were NYSE trades on December 31 easier to classify? Trades on that day might be more aggressive than normal due to "window dressing" by investment funds, sellers wanting to file tax losses; or, lower volume. Also unclear is why this effect is concentrated on the specialist-driven NYSE.

Table 8 indicates that the overall accuracy is not due to superior performance in a particular sector. Modeled classifications are superior to the next best method, often by 1–2%, except for the (very small) industrial goods sector where the EMO method is 1.3% more accurate. Figure 2 shows that

		Percent of Trades Correctly Classified				
Sector	N	Modeled	EMO	LR.new	LR.old	Tick
Capital Goods	216,800	**74.7%**	73.0%	72.1%	71.8%	61.6%
Conglomerates	33,863	**84.7%**	83.4%	79.5%	78.9%	63.7%
Cons. Cyclical	236,193	**73.4%**	72.1%	71.7%	71.4%	62.9%
Energy	228,978	**77.3%**	76.1%	73.3%	72.9%	62.5%
Financial	1,014,479	**74.2%**	72.4%	72.4%	72.2%	63.3%
Healthcare	2,314,251	**72.2%**	71.5%	69.7%	69.4%	63.8%
Ind. Goods	4,917	62.5%	**63.8%**	60.4%	60.3%	54.3%
Materials	247,166	**74.7%**	73.4%	71.6%	71.3%	62.1%
Non-Cyclical	149,270	**73.8%**	72.5%	71.6%	71.3%	60.5%
Services	3,278,245	**73.2%**	71.9%	70.5%	70.1%	64.7%
Technology	8,440,206	**76.0%**	73.4%	73.2%	72.8%	68.5%
Transportation	279,582	**75.1%**	73.5%	72.6%	72.3%	64.3%
Utilities	60,930	**81.2%**	79.7%	78.1%	77.7%	61.5%

TABLE 8. Percent of trades correctly classified across sectors. The large and small average spreads for technology and conglomerate stocks does not appear to affect classification accuracy.

while accuracy varies, modeled classifications are superior to the next best method for all out-of-sample dates and across all ten-minute bins during the trading day. Classification accuracy is lower in the morning and higher in the evening, perhaps due to spreads decreasing across the trading day.

6.4. **Accuracy Across the Spread.** Many studies note lower classification accuracy for trades outside the spread. An online addendum discusses how negotiated trades might be responsible for some of this. (Briefly: A large sell (buy) order executed over time with a final aggregate "print" is likely to have an average price above (below) the end-of-trading spread.) Since our dataset excludes negotiated trades, we can examine how the performance of these methods relative to the prevailing quote differs from past studies.

Defining the "prevailing quote" requires taking a view on the delay between quotes and trades. Therefore, I examine classification accuracy across three versions of the "prevailing quote:" EMO (0s delay), LR.new[16] (5s delay), and delay-modeled (mean of 5.0s, 0.8s for Nasdaq, NYSE).

For the EMO (0s lag) quote, only 7% of trading occurs outside the spread. About 21% of trading occurs outside the LR-defined spread. About 27% of trading occurs outside the model-estimated spread. It is unclear if this much trading really takes place outside the spread or if the model-estimated quotes also capture some persistence in quote changes.

[16]The results for LR.old are not materially different.

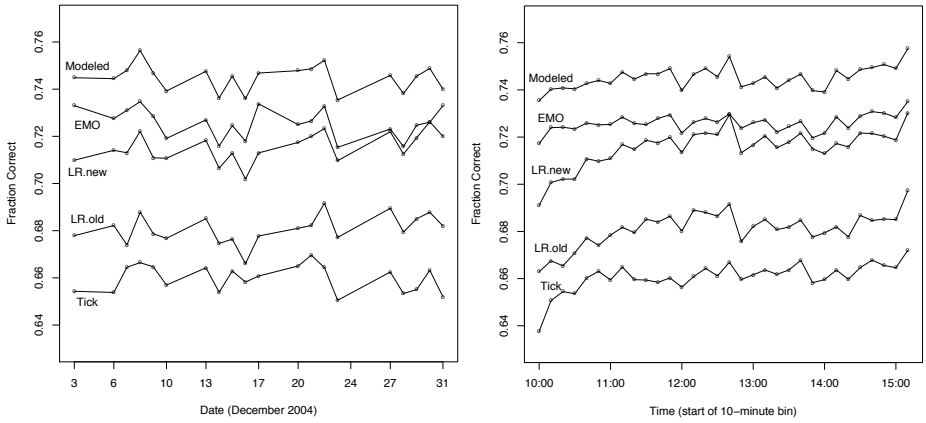

FIGURE 2. Fraction of trades correctly classified across time. Classification methods used: Modeled, EMO, LR.new, LR.old, and Tick. Modeled classifications are more accurate across all days and all time bins. All methods appear to be more accurate near the end of the trading day.

Trade Price	EMO (0s)	LR (5s)	Modeled
[0, Bid)	545,877	1,701,011	2,187,956
Bid	5,476,692	5,429,370	3,917,366
(Bid, Mid)	695,830	370,777	1,870,881
Midpoint	490,690	284,534	30,316
(Mid, Ask)	705,964	361,013	1,879,244
Ask	7,951,786	6,629,664	4,387,127
(Ask, ∞)	638,041	1,728,511	2,231,990

TABLE 9. Trade counts relative to various estimates of the prevailing quote. Note that where trades appear to take place relative to the prevailing quote varies with the delay between trades and quotes.

Classification accuracy across these three versions of prevailing quotes are shown in Figure 3. Bars are ordered by classification method: Modeled, EMO, LR.new, LR.old, and Tick. Modeled trade classifications are superior with only a few exceptions: the EMO method is 0.1% more accurate for prices at the modeled ask; LR and tick methods are 3.8% more accurate for trades at the modeled midpoint; the LR.new method is 0.1% more accurate for trade prices in the EMO (Bid, Mid); and, the EMO method is 0.3% and 1.0% more accurate for prices in the LR (Bid, Mid) and (Mid, Ask). Modeled classifications do poorly for the few trades at the model-estimated midpoint: these would be better classified by flipping a coin.

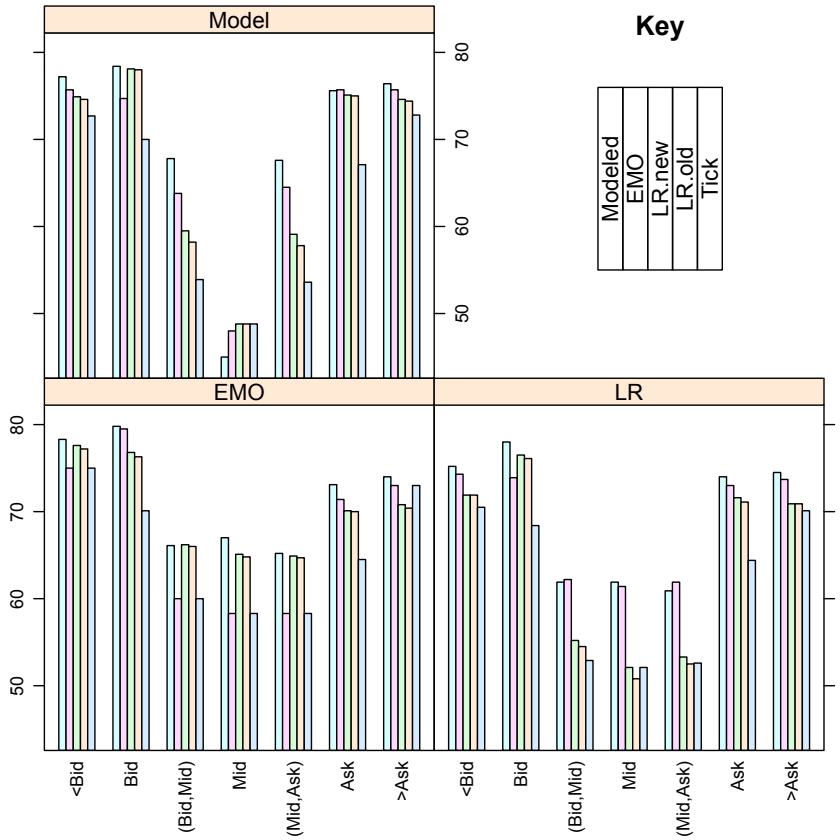

FIGURE 3. Classification accuracy for different methods across various ideas of the prevailing quote: Modeled, EMO (0s lag), and LR (5s lag). Bars show accuracy for each classification method: Modeled, EMO, LR.new, LR.old, and Tick. Modeled classifications are usually more accurate than all other methods. Accuracy outside the spread is also higher than inside, perhaps due to omitting negotiated trades.

For all three versions of the prevailing quote, modeled trade classifications are more accurate than other methods. Classification is generally more accurate outside the spread than at or inside the quote; this differs from Ellis *et al.*'s (2000) and Peterson and Sirri's (2003) finding lower accuracy outside the spread. As mentioned previously, this difference may be due to omitting negotiated trades. (See the online addendum for details.)

Figure 3 also suggests that a better method might be to use the midpoint method for trades not at the EMO bid or ask — an interaction effect which further endorses a modeled approach. Similar to this, Chakrabarty *et al.* (2007) suggest using the LR (midpoint) method for trades at the bid, ask, or more than ±20% of the spread away from the midpoint and the tick method

for other trades. They report that their method for Nasdaq stocks is 0.7% more accurate than the next-best, EMO method. (The model here is 2.0% more accurate than the next-best, EMO method.)

Accuracy for a few subsets of data are poor. Odd-lot orders are not protected by Reg NMS order handling rules: they need not be displayed in order books nor are they protected from specialists trading in front of them or other markets trading though their limit price. Classifying odd-lot trades is more difficult, as seen in Table 10. The performance of the LR.old method suggests that odd-lot buys (sells) at the midpoint are almost always preceded 5–6 seconds earlier by the midpoint increasing (decreasing).

Trade Price vs.		Percent of Trades Correctly Classified				
Estimated Quote	N	Modeled	EMO	LR.new	LR.old	Tick
[0, Bid)	154,761	73.8%	**74.1%**	73.4%	72.4%	72.8%
Bid	354,350	73.2%	**73.3%**	73.1%	69.1%	66.7%
(Bid, Mid)	211,416	57.0%	56.0%	**60.8%**	56.1%	56.6%
Midpoint	11,095	39.8%	47.0%	53.2%	0.6%	**53.3%**
(Mid, Ask)	210,457	60.9%	57.3%	**63.4%**	58.9%	57.4%
Ask	363,550	**79.0%**	75.9%	**79.0%**	77.5%	71.5%
(Ask, ∞)	156,123	**76.8%**	75.6%	76.0%	75.3%	74.6%

TABLE 10. Classification accuracy of odd-lot trades for different methods across the model-estimated quote. Accuracy inside the spread is worse than for outside the spread (in contrast with prior studies). Accuracy for the few trades at the midpoint is very poor.

6.5. **Multiple Comparisons.** In looking at out-of-sample performance across various factors, we are making multiple comparisons. Given enough subsets, even a mediocre method will have apparent areas of excellence due to sheer randomness. One elegant way to handle this is to use the model confidence set of Hansen *et al.* (2011). In this case, we can use a cruder and much more conservative method: a Bonferroni correction. The preceding tables and figure compared five models, suggesting a null hypothesis that the modeled approach "wins" with probability 1/5. Ninety-five comparisons were made with the modeled approach winning 85[17]. The probability of the modeled approach winning this many or more of the 95 comparisons is a simple binomial sum yielding 4.3×10^{-38}. This exceeds even the 1% (Bonferroni-corrected to $\frac{1}{95}\%$) level of significance.

[17]The ninety-five comparisons come from comparing performance across 2 markets, 13 industries, 20 dates, 32 time bins, 7 locations relative to the spread $\times$ 3 quote definitions, and 7 locations relative to the spread for odd-lot trades.

6.6. **Performance Attribution.** Since the model performs well, we might wonder how much of the superior performance is due to various model features. We can examine this with a sequence of models from simple to the full model of the previous section. To properly attribute performance, the models are all nested. I compare the models using out-of-sample classification accuracy by market. Six models are compared, each designed to isolate the effect of a particular aspect of the full model. The models are:

1. **Tests** A simple GLM using tick, midpoint, and bid/ask tests[18]. The tests use conventional quote delays: 5s for midpoint, 0s for bid/ask. No delay model nor random effects are used.
2. **Metrics** The preceding model except using tick, midpoint, and bid/ask *metrics* (g and J from section 3.1).
3. **AR Effect** The preceding model plus the lagged bid/ask metric.
4. **Random Effects** The preceding model plus time and time-sector random effects.
5. **Ad-hoc Delay** The preceding model plus a simple *ad hoc* universal delay distribution: Gamma$(3, 2)$.
6. **Full Model** The preceding model with estimated delay distributions for each market.

Table 11 shows the estimated coefficients for these models. Intercepts are estimated in-sample but not reported since they are nuisance parameters and thus not used out of sample. Table 12 shows the out-of-sample performance of these models. Since the models are nested, changes in accuracy (moving from left to right) are additive[19].

From Tables 7 and 12 we see that the initial model (Tests) is less accurate than the EMO method for Nasdaq stocks — and only marginally (0.2%) better for NYSE stocks. Compared to the (conventional) LR.new method, the Tests model is less accurate for Nasdaq stocks by 1.5% and more accurate for NYSE stocks by 3.7%. Incorporating multiple sources of information seems helpful but insufficient to achieve superior performance.

The Metrics model is the same as Tests except that it also accounts for information strength. That increases accuracy by 3% and 1.1% for Nasdaq and NYSE stocks. This increased accuracy makes the Metrics model 1% and 1.3% more accurate than the EMO method for Nasdaq and NYSE stocks.

The AR Effect model adds a lagged bid/ask metric, an autoregressive term, which decreases out-of-sample accuracy 0.1% and 0.6% for Nasdaq stocks and NYSE stocks. As expected, the random effects in model Random Effects have no effect on the out-of-sample classification accuracy[20]. The

[18]The tests return -1 and +1 to indicate likely sells and buys and 0 otherwise.

[19]For example, the accuracy of the Ad-hoc Delay model includes the effects of (i) using multiple sources of information; (ii) using the strength of that information; (iii) allowing for autocorrelated information; (iv) using random effects; and, (v) adding an ad-hoc universal delay model.

[20]The random effects should mostly affect the estimated coefficient standard errors.

	Coefficients	Tests	Metrics	AR Effect	Random Effects	Ad-hoc Delay	Full
	J width (τ) $\times 10^{-4}$	—	2.09	2.09	2.09	2.09	2.09
Nasdaq	Delay Count (ν)	—	—	—	—	3	1.65
Nasdaq	Delay Rate (λ)	—	—	—	—	2	0.33
Nasdaq	Bid/Ask	0.70	1.21	1.19	1.19	1.19	1.41
Nasdaq	Midpoint	0.52	160	155	157	214	209
Nasdaq	Tick	0.18	80	82	83	67	67
Nasdaq	Prev. Bid/Ask	—	—	0.05	0.06	-0.20	-0.20
NYSE	Delay Count (ν)	—	—	—	—	3	0.62
NYSE	Delay Rate (λ)	—	—	—	—	2	0.78
NYSE	Bid/Ask	1.19	1.87	1.89	1.91	1.72	2.04
NYSE	Midpoint	0.52	197	205	211	344	122
NYSE	Tick	0.02	-17.5	-19.0	-18.9	-27.6	-20.5
NYSE	Prev. Bid/Ask	—	—	-0.05	-0.05	-0.22	-0.17

TABLE 11. Estimated model coefficients for the nested performance attribution models. The consistently negative NYSE tick metric coefficient contrasts with the positive Nasdaq tick metric coefficient. This is likely due to different short sale constraints for the two markets.

		Percent of Trades Correctly Classified					
Market	N	Tests	Metrics	AR Effect	Random Effects	Ad-hoc Delay	Full
Nasdaq	15,220,579	70.3%	73.3%	73.2%	73.2%	74.1%	74.3%
NYSE	1,264,866	79.8%	80.9%	80.3%	80.3%	81.0%	80.7%
Total	16,504,880	71.1%	73.8%	73.7%	73.7%	74.6%	74.7%

TABLE 12. Percent of trades correctly classified across US markets for the nested performance attribution models. The nesting allows us to attribute the full model's performance to various improvements. The model using just classification metrics is 1% and 1.3% more accurate than the EMO method for Nasdaq and NYSE stocks. Adding a delay model increases accuracy by 0.9% and 0.7% for Nasdaq and NYSE stocks. Total includes 19,435 AMEX trades.

Ad-hoc Delay model shows improvement of 0.9% and 0.7% for Nasdaq and NYSE stocks. The Full model adds market-specific delay parameters with mixed results: a 0.2% accuracy improvement for Nasdaq stocks; and, a 0.3% accuracy loss for NYSE stocks.

Summarizing, two improvements stand out. First, including the strength of information (converting the EMO, LR, and tick tests to *metrics*) yields

a 3% and 1.1% improvement in accuracy for Nasdaq and NYSE stocks. Second, adding a basic delay model is responsible for a 0.9% and 0.7% improvement in classification accuracy for Nasdaq and NYSE stocks. Some improvements did not work: The lagged bid/ask metric (a basic autoregressive effect) was significant in-sample; out of sample it reduced accuracy for both Nasdaq and NYSE stocks.

6.7. **Resorting to the Tick Test.** Stoll and Schenzler's (2006) finding more trading outside the (unlagged) spread indicates that bid/ask methods will increasingly resort to the tick test. The original Ellis *et al.* analysis of Nasdaq trades resorted to the tick test for classifying 25% of trades. Peterson and Sirri's (2003) analysis of NYSE trades resorted to the tick test for classifying 11–20% and 19–30% of trades for tick sizes of \$1/8 and \$1/16. In our data, the EMO test resorted to the tick test 18.4% and 14.2% of the time for Nasdaq and NYSE trades and anywhere from 16.0% to 19.4% of the time on a given day. Across all markets, the frequency and interday pattern of resorting to the tick test are almost the same as the overall numbers. These numbers might be higher had we looked at trades across all venues.

Overall, the EMO test is not lucky about when it resorts to the tick test. The accuracy when the EMO test resorts to the tick test is 64.6% and 60.6% on the Nasdaq and NYSE — versus overall tick test accuracy (Table 7) of 66.7% and 60.7%. However, there are times when the EMO test benefits from resorting to the tick test. Table 9 shows that for trades outside the EMO-defined spread, the tick test is much more accurate than a midpoint test. The overall accuracy when using the tick test is dragged down by the poor performance of the tick test inside the EMO spread.

The daily numbers reveal a small but significant relationship between higher volumes and increased use of the tick test. A basic linear model for this relationship has an intercept of 0.152 (*i.e.* 15.2%, s.e. 0.009) and a slope of 0.032 (s.e. 0.010) for volume in millions. This suggests exploring volume or trade-count interactions in our model.

7. Conclusion

Modeled trade classifications and estimated quotes yield more accurate trade classifications across many different strata. We decompose accuracy improvements to find that using classification metrics yields a 3% and 1.1% improvement in accuracy for Nasdaq and NYSE stocks; and, using (at least) a basic delay model is responsible for a 0.9% and 0.7% improvement in classification accuracy for Nasdaq and NYSE stocks. Also, we note that the differing signs of the tick metric correlate with the difference in short-sales constraints between the Nasdaq and NYSE pre-Reg SHO.

A strong benefit of the modeled approach is the ease of improvement. For example, an interaction with volume might give more weight to some tests at low-volume times versus high-volume times. An interaction with trade size could capture differences in order handling for small and very large orders

or odd-lot versus round-lot and mixed-lot orders. We could also try scaling covariates by the stock-specific mean spread, volatility, trade size, or some measure of daily volume. Finding these transformations could yield further economic insights.

Market observers cannot use random effects for out-of-sample prediction; but, market participants who know their own trade classifications can use them to discern an increased tendency toward buying or selling for a given sector and time bin. Market makers could use this to better control inventory risk. One could even place random trades, compare their known classifications to the model and estimate nonparametric (model-free) short-term alpha.

Another area for further study is the time trend of classification accuracy. The increased use of ECNs, automated trading, and microstructure-based speculation might affect the accuracy of the methods studied here. The best-efforts classification accuracy might even be seen as a measure of market fragmentation and data quality.

Finally, the method here is the only trade classification method which also produces a form of "signal strength." Therefore, this method would be particularly useful at detecting the presence of highly-aggressive or destabilizing order flow. The linear predictor η (log-odds) can be inspected across time to find outliers and see if they share certain time patterns or come from the same source. This might help regulators and market authorities detect the presence of destabilizing flow and could help prevent an event similar to the 6 May 2010 "flash crash" from occurring.

References

Aigner, D. J. (1973)., "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics*, 1, 49–60.

Archipelago Holdings, Inc. (2005a)., *ArcaBook and ArcaTrade Historical: For the Archipelago Exchange and ArcaEdge, Version 1.1.* Chicago.

Archipelago Holdings, Inc. (2005b)., "ArcaEx Releases December 2004 Transaction Volume Data." Retrieved on 11 June 2008 from `http://www.archipelago-exchange.com/inside/news/news_20050111.asp`.

Asquith, P., R. Oman, and C. Safaya. (2010)., "Short Sales and Trade Classification Algorithms." *Journal of Financial Markets*, 13, 157–173.

Benveniste, L. M., S. M. Erdal, and J. William J. Wilhelm. (1998)., "Who Benefits From Secondary Market Price Stabilization of IPOs?" *Journal of Banking and Finance*, 22, 741–767.

Bessembinder, H. (2003)., "Issues in Assessing Trade Execution Costs." *Journal of Financial Markets*, 233–257.

Boehmer, E., J. Grammig, and E. Theissen. (2007)., "Estimating the Probability of Informed Trading: Does Trade Misclassification Matter?" *Journal of Financial Markets*, 10, 26–47.

Boone, J. P. and K. K. Raman. (2001)., "Off-Balance Sheet R&D Assets and Market Liquidity." *Journal of Accounting and Public Policy*, 20, 97–128.

Chakrabarty, B., B. Li, V. Nguyen, and R. A. V. Ness. (2007)., "Trade Classification Algorithms for Electronic Communications Network Trades." *Journal of Banking and Finance*, 31, 3806–3821.

Cox, D. R. (1975)., "Partial Likelihood." *Biometrika*, 62, 269–276.

Danielsen, B. R., R. A. V. Ness, and R. S. Warr. (2007)., "Auditor Fees, Market Microstructure, and Firm Transparency." *Journal of Business Finance and Accounting*, 34, 202–221.

Ellis, K., R. Michaely, and M. O'Hara. (2000)., "The Accuracy of Trade Classification Rules: Evidence from Nasdaq." *Journal of Financial and Quantitative Analysis*, 35, 529–551.

Erlang, A. K. (1909)., "The Theory of Probabilities and Telephone Conversations." *Nyt Tidsskrift for Matematik*, B, 33–39.

Finucane, T. J. (2000)., "A Direct Test of Methods for Inferring Trade Direction from Intra-Day Data." *Journal of Financial and Quantitative Analysis*, 35, 553–576.

Hansen, P. R., A. Lunde, and J. M. Nason. (2011)., "The Model Confidence Set." *Econometrica*, 79, 453–497.

Hasbrouck, J. (1992)., "Using the TORQ Database." Working Paper 92-05, NYSE.

Heagerty, P. J. and S. L. Zeger. (2000)., "Marginalized Multilevel Models and Likelihood Inference." *Statistical Science*, 15, 1–19.

Henker, T. and J. Wang. (2006)., "On the Importance of Timing Specifications in Market Microstructure Research." *Journal of Financial Markets*, 9, 162–179.

Kauermann, G. and R. J. Carroll. (2001)., "A Note on the Efficiency of Sandwich Covariance Matrix Estimation." *Journal of the American Statistical Association*, 96, 1387–1396.

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

McCullagh, P. (1987)., *Tensor Methods in Statistics*. London: Chapman and Hall.

McCullagh, P. and J. A. Nelder (1989)., *Generalized Linear Models*. 2nd edn., London: Chapman and Hall.

Nocedal, J. and S. J. Wright (2006)., *Numerical Optimization*. 2nd edn., New York: Springer.

Odders-White, E. R. (2000)., "On the Occurrence and Consequences of Inaccurate Trade Classification." *Journal of Financial Markets*, 3, 259–286.

Peterson, M. and E. Sirri. (2003)., "Evaluation of the Biases in Execution Cost Estimation Using Trade and Quote Data." *Journal of Financial Markets*, 6, 259–280.

Roll, R. (1984)., "A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market." *Journal of Finance*, 39, 1127–1139.

Rosenthal, D. W. R. (2008)., "Trade Classification and Nearly-Gamma Random Variables." PhD dissertation, University of Chicago, Department of Statistics.

Schultz, P. H. and M. A. Zaman. (1994)., "Aftermarket Support and Underpricing of Initial Public Offerings." *Journal of Financial Economics*, 35, 199–219.

Stoll, H. R. (2006)., "Electronic Trading in Stock Markets." *Journal of Economic Perspectives*, 20, 153–174.

Stoll, H. R. and C. Schenzler. (2006)., "Trades Outside the Quotes: Reporting Delay, Trading Option, or Trade Size?" *Journal of Financial Economics*, 79, 615–653.

Tanggaard, C. (2004)., "Errors in Trade Classification: Consequences and Remedies." Working Paper 420680, SSRN.

Vergote, O. (2005)., "How to Match Trades and Quotes for NYSE Stocks?" Working paper, Katholieke Universiteit Leuven.

Wong, W. H. (1986)., "Theory of Partial Likelihood." *Annals of Statistics*, 14, 88–123.