Modeling Trade Direction

Rosenthal, Dale W.R.

University of Illinois at Chicago

August 2008

MODELING TRADE DIRECTION

DALE W.R. ROSENTHAL

Abstract. The problem of classifying trades as buys or sells is examined. I propose estimated quotes for midpoint and bid/ask tests and a modeling approach to classification. Prevailing quotes are estimated using flexible approximations to the distribution for delays of quotes relative to trade timestamps. Classification is done by a generalized linear model which includes improved versions of midpoint, tick, and bid/ask tests. The model also considers the relative strengths of these tests, can account for market microstructure peculiarities, and allows for autocorrelations and cross-correlations in trade direction. The correlation modeling corrects for pseudoreplication, yielding more accurate standard errors and fixed effect estimates. Further, the model estimates probabilities of correct classification. The model is compared to various trade classification methods using a sample of 2,836 domestic US stocks from an unexplored, recent, and readily-available dataset. Out of sample, modeled classifications are 1–2% more accurate overall than current methods; this improvement is consistent across dates, sectors, and locations relative to the inside quote. For Nasdaq and NYSE stocks, 1% and 1.3% of the improvement comes from using relative strengths of the various tests; 0.9% and 0.7% of the improvement, respectively, comes from using some form of estimated quotes. For AMEX stocks, a 0.4% improvement is attributed to using a lagged version of the bid/ask test. I also find indications of short- and ultra-short-term alpha. (*JEL*: C53, D82, G14)

1. Introduction

Over the past forty years, econometricians have discussed how to determine if the aggressor in a trade was the buyer or seller.

Solutions to this problem have been primarily algorithmic: Should one classify the trade using a:

Date: August 6, 2008.

This article is based on my PhD dissertation at the University of Chicago. Suggestions and assistance were provided by my advisor, Per Mykland; ex-colleagues from LTCM and Morgan Stanley's Equity Trading Lab; my committee members: Vanja Dukic, David Modest, and Stephen Stigler; Helen Barounis at NYSE Arca; and, John Zekos. 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.

1. prevailing midpoint test, such as Lee and Ready (1991);
2. tick test, as recommended by Finucane (2000); or,
3. a prevailing bid/ask test, such as Ellis, Michaely, and O'Hara (2000)?

Scholarship about prevailing midpoint and bid/ask methods has mainly focused on the time lag between a trade's time of publication and the prevailing bid and ask quotes. Many articles imply that if this lag were known, we would know the prevailing market price used by the trade initiator.

But is this so? Are we even asking the right question? I suggest we are not.

I believe that trying to pick the prevailing midpoint (or bid and ask) is suboptimal: that it is "less correct" and introduces too much volatility into the classification process. Instead, I propose that a weighted average of the midpoint process should yield an estimate that has more predictive power.

If changes in the midpoint process are not serially correlated we should estimate a number that is close to the prevailing midpoint. However, if changes to the midpoint process are autocorrelated, we might want to refer to quotes older than that prevailing at order submission time. This latter approach essentially uses ultra-short-term alpha[1] to improve our classification.

I also believe we should model the likelihood a trade was buyer-initiated. This allows for richer models which:

1. include information from midpoint, tick, and bid/ask tests;
2. consider the strengths of those test results;
3. account for microstructure peculiarities (*e.g.* short sales rules)[2];
4. allow for autocorrelations and cross-correlations in buys/sells; and,
5. indicate the likelihood our classification is correct.

Thus I propose two improvements: better quote estimates and a better model. I explore these improvements with a relatively unknown dataset which is readily available, easy to work with, and current. This dataset lets us see glimmerings of the ultra-short-term alpha mentioned previously.

2. Trade Classification

Determining which trade participant "caused" a trade to occur is a problem which goes by various names:

- determining the *initiator* (*i.e.* who came later, buyer or seller?);
- guessing which *side* was the aggressor;

[1] I use "ultra-short-term alpha" to mean return predictability over a small number of trades and beyond bid-ask bounce. This parallels the use of "ultra-high frequency" for data including all trades.

[2] A working paper by Asquith, Oman, and Safaya (2007) notes the difficulties some classification tests have if short-sales are only allowed on zero-plus ticks.

- *classifying* trades;
- inferring *trade direction*; and,
- *signing* trades or volume: attaching a sign to traded volume (*e.g.* "−" for sell- and "+" for buy-initiated trades).

Hasbrouck and Schwartz (1987) define the trade initiator as the order which incurs execution costs. However, any limit order increases the cost of trading through that order's limit price. Thus all orders must impart some bias to market prices. Instead, I use Odders-White's (2000) definition of the later-arriving order as the trade initiator.

Trade signing may seem esoteric but Hasbrouck (1991) argues that signed volumes are themselves important. Signed trades are critical for inferring the probability of informed trading, effective spreads, and the market impact of trading. Sums of signed volumes may be thought of as "net order flow".

How worthwhile is a small improvement? Even a one-half percent gain in classification accuracy could result in better estimation of market impact. A more accurate market impact model would likely result in more efficient trading of customer orders (for an investment bank) or more accurate return predictions based on inferred market impact (for investment funds). Thus a 1%–2% accuracy improvement could easily be worth millions of dollars to either of these market participants.

The side of a trade can be inferred by some established methods. Most of these have focused on stock trades; however, these methods are based on standard economic concepts and thus are generally useful. Further, reporting an execution to a buyer or seller is distinct from, and probably more urgent than, publishing that trade to the public. Thus the modeling concepts I develop may be similarly useful across markets and asset classes.

Early work on trade signing is illustrated by Osborne (1965), Niederhoffer and Osborne (1966) and Garman (1976). One could even claim trade signing is discussed in Lefèvre (1923). However, none of these considers the delays between quotes and trade reports. (Lefèvre briefly mentions market data delays in bucket shops.)

Erlang's (1909) study of information delays forms the foundation of thought on delays. Forrester (1980) examined delayed data in macroeconomic models. Lee and Ready (1991) first considered the delay in trade reporting. The assumptions and models used herein were developed in Rosenthal (2008).

2.1. **Approaches.** Currently, three approaches to trade classification dominate the literature. These three approaches can be thought of as competing families of tests: midpoint tests, tick tests, and bid/ask tests.

2.1.1. *Midpoint Tests.* Lee and Ready (1991) suggested a midpoint test with delay (the "LR method"): compare the trade price to a midpoint which lags the trade publishing time; resolve midpoint trades with a tick test. Further, they noted reasonable lags: five seconds (now commonly used) for 1988 data, two seconds for 1987 data, and that "a different delay may be appropriate for other time periods".

Vergote (2005) suggests using the LR method with a two-second lag; Henker and Wang (2006) suggest a one-second delay for NYSE TAQ data.

2.1.2. *Tick Tests.* Tick tests have received comparatively little attention: Finucane (2000) recommends a tick test (*i.e.* comparing a trade price to the previous trade price for that stock). Most literature refers to tick tests only insofar as to use them to resolve inconclusive midpoint or bid/ask tests.

2.1.3. *Bid/Ask Tests.* Bid/ask tests were explored by Ellis, Michaely, and O'Hara (2000) for Nasdaq stocks. Their "EMO method" compares trade prices to prevailing bids and asks and resolves indeterminacies with a tick test. Peterson and Sirri (2003) suggested the EMO method for NYSE stocks.

2.1.4. *Modeling.* Caudill, Marshall, and Garner (2004) was the only (unsuccessful) attempt to find a trade-classifying generalized linear model (GLM).

2.2. **Previous Analyses.** While previous analyses were advanced for their time, electronic data on financial markets has increased tremendously. In the context of currently available data, the data used in past analyses are:

- Old: dating from 1987 (LR), 1990 (TORQ database), 1997 (EMO, Peterson and Sirri), and 1999 (Henker and Wang);
- Narrow: consisting of trades in 144 stocks (TORQ), 150 stocks (LR), 313 stocks (EMO), and 401 stocks (Henker and Wang)[3];
- Biased: composed of solely large-cap stocks or internet boom post-IPO stocks (EMO, 1996–1997); and,
- Time-skewed: lacking contemporaneous trades for both Nasdaq and NYSE stocks.

3. Microstructure for Quotes

Better quotes require examining what quotes are and how they might be delayed relative to trades.

[3]The number of stocks Peterson and Sirri analyze is not stated.

3.1. **Bid and Ask Processes.** Bid and ask quotes constitute two simple processes[4] for the prices at which somebody else would buy or sell. Averaging these we get another simple process: the midpoint process.

Discerning the midpoint prevailing when an initiating order was sent seems to be the problem. But we can think of a more general question: What is a good measure of the state of the market preceding order submission?

This question admits that an observed price might have been spurious or an ephemeral "blip"; allows for market participants who receive quotes with differing delays; and, embraces the possibility of predicting the future market state encountered by an order.

3.2. **Market Data Delays.** The first sort of delay we consider is the delay in transmitting market data from various sources to users.

Suppose a trader examines the prevailing market prices before placing an order — either to assign a limit price to the order or to guess at the price a market order would get. Unfortunately, we know neither who the trader was nor the bid and ask prices seen pre-order.

There is more we do not know. Market data originates from many market centers (exchanges, ECNs, market makers). Intermediaries combine data sources or add data of their own creation. Consumers buy data from both original sources and intermediaries and may disseminate that data inside their organization. Eventually the data reaches a trader or trading agent.

From these details, we can consider some example market data pathways (Figure 1). Each arrow and node represents a possible delay in receiving a quote. Not shown is that each node may differ in data processing speed.

3.3. **Order Submission Delays.** When a quote is received, a trader may decide to send an order at a limit price to a market center. The time to decide adds a small delay; transmitting the order adds another small delay.

By definition, the transmitted initiating order is marketable[5]. If the transmitted order were not marketable, it would not be the initiator and we would not care about its submission delay[6].

3.4. **Trade Reporting Delays.** A marketable order reaching a market center trades against orders in the order book (or the specialist). That trade changes the inventory available at one or more prices.

[4]Piecewise-constant processes jumping finitely often in a finite compact timespan.

[5]A marketable order satisfies the far side of the inside quote or is a market order.

[6]The non-initiating (liquidity providing) order is in the domain of limit order models and is explored by Chacko, Jurek, and Stafford (2008).

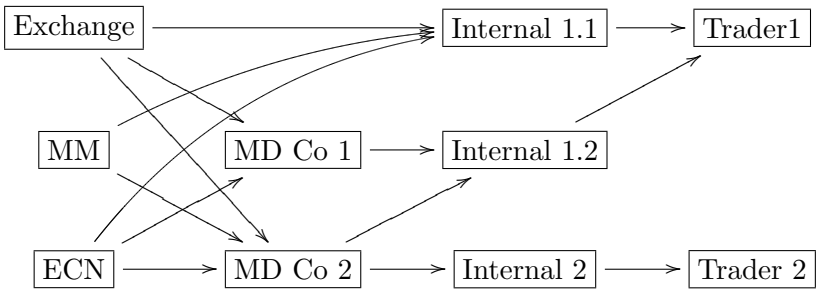

FIGURE 1. Example market data pathways from sources to consumers. Internal 1.1 is a high-performance disseminator of direct data feeds; Internal 1.2 aggregates data from market data companies; and, Internal 2 takes a simple approach.

A trade execution ("fill") must then be sent to involved customers; quotes are updated; and, the trade must be made public. This final publishing timestamp is what researchers see in non-proprietary transaction databases.

3.4.1. *Reporting Fills.* Fills must be reported to the customer within 90 seconds. Boehmer (2006) found average delay was 10–20 seconds in 2004.

In the US, average execution times and other metrics appear in execution quality reports mandated by the Securities and Exchange Commission[7]. These surveys are meant to inform customers so they may preference market centers. Recent reports indicate the median execution time for a market or marketable-limit order is now often fractions of a second.

3.4.2. *Updating Quotes.* Market centers have a strong incentive to keep their quotes current: Any market participant not honoring their quotes is subject to fines and censure. SEC and FINRA rules, as well as securities training materials, repeatedly emphasize this fact[8].

For Nasdaq market makers, penalties for "backing away" from a quote can be swift: FINRA may block a firm from making a market in the related stock for one or more days starting the same day as a violation.

[7]This first took effect in 2001 as an amendment to the Securities Exchange Act (see SEC Release 34-43590). It is often quoted in reference to the Exchange Act as "11Ac1-5" and "11Ac1-6". In 2005, this was incorporated into Regulation NMS as Rule 605.

[8]Securities licensing preparation materials, such as Securities Training Corp (2006), offer useful summaries of suggested and proscribed practices as well as relevant laws.

3.4.3. *Publishing Trades.* Reporting trades to the public must also be done within 90 seconds of execution. However, the time to report trades to the public is not measured for surveys of execution quality nor is there a market mechanism to reward fast public reporting.

In many markets, trading and receiving a fill happens on a sub-second timeframe. Compared to that, 90 seconds is a long time. Trades may even be published more than 90 seconds late without disciplinary action — if there is a reasonable explanation for the delay. For example: the heavy processing needed near market close at month- or quarter-end would likely be such an exceptional (albeit predictable) situation.

3.4.4. *Differing Priorities.* Thus publishing quotes versus trades have different priorities: keeping quotes current and notifying customers of a fill are higher priorities than reporting an execution to the public. Given this difference in priorities, we should not be surprised that the (well-documented) publishing delay is on the order of seconds.

This is more than just theorizing; Ellis, Michaely, and O'Hara (2000) note (Section IV.C) that quotes are updated with little-to-no delay whereas trades are published with delay.

3.4.5. *Putting It All Together.* We can put the various constituents of delay together to see the net delay observed by the public (and microstructure researchers). An illustration of these constituents and the net delay is shown in Figure 2. As we can see, estimating the quotes prevailing when a marketable order was sent requires looking at quotes before the trade timestamp.

3.5. **Challenges for the Central Limit Theorem.** Since the total delay is a sum of constituent delays (sub-delays), we might be tempted to use the Central Limit Theorem to approximate the delay distribution. However, two likely possibilities make the CLT poorly-suited to this situation.

3.5.1. *Short Data Paths.* Some market participants have direct connections to market centers to get the freshest data available. Other participants are less time-sensitive and buy their data from a single market data company. Both approaches yield data paths of only a few segments. Thus the number of sub-delays is neither growing nor close to "asymptopia".

3.5.2. *Correlated Delay Constituents.* Delay constituents may be correlated if portions of the data path are shared. Additionally, the structure of a data path ensures that heavy information flow at the start yields heavy information flow throughout the path — thereby inducing correlations. Using the CLT would be less accurate for positively-correlated delay constituents.

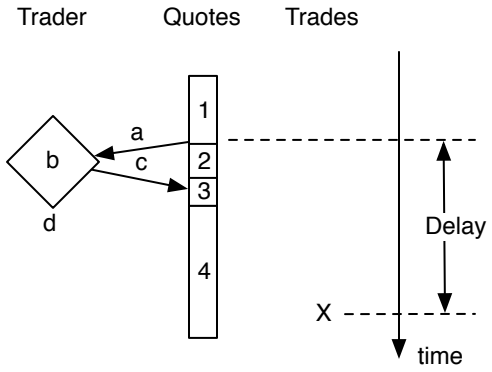

FIGURE 2. An example of delay between quote and trade timestamps. 1) Transmission a: Quote #1 received by trader. 2) Decision b: Trader decides to trade; assembles order. 3) Transmission c: Order transmitted to market center. 4) Trade occurs; quote updated quickly (quote #3 → #4). 5) Private message d: Trader notified of fill. 6) Trade ("X") published (and timestamped) later.

4. Delay Models for Quotes

Since the constituent delays may be few and correlated, I explore small-sample approximations based on the gamma distribution. The theory behind these expansions is found in Rosenthal (2008).

4.1. **Setup.** The approximations I use require a few assumptions:

1. Delay constituents (sub-delays) are exponentially distributed;
2. Observations of total delay are independent; and,
3. There are at least two delay constituents.

Note that I do not assume independence of the delay constituents.

To express these ideas mathematically, we need to introduce some notation:

Y = the delay between trade timestamps and quotes used by initiators;
κ_r = the r-th cumulant of total delay Y;
ν = the number of delay constituents (aka sub-delays);
$\hat{\nu}$ = the estimated scale parameter of a gamma distribution;
$\hat{\lambda}$ = the estimated rate parameter of a gamma distribution;
$\tilde{\kappa}_r$ = the r-th pseudocumulant of total delay Y;
$f_Y(y)$ = the density of total delay Y ($y \geq 0$);
$\hat{b}_t$ = the estimated bid prevailing at time t;
$\hat{a}_t$ = the estimated ask prevailing at time t; and,

$\hat{m}_t$ = the estimated midpoint prevailing at time t, $\hat{m}_t = (\hat{b}_t + \hat{a}_t)/2$.

The estimated gamma distribution parameters are chosen to match the first two sample cumulants κ_1 and κ_2. This yields $\hat{\nu} = \kappa_1^2/\kappa_2$ and $\hat{\lambda} = \kappa_1/\kappa_2$. The pseudocumulants $\tilde{\kappa}_r$'s are as in McCullagh (1987): differences between sample cumulants (κ_r's) and cumulants of the Gamma$(\hat{\nu}, \hat{\lambda})$ distribution.

4.2. **Small-Sample Approximations.** I use a gamma-based Edgeworth approximation to the total delay density[9]. The gamma-based Edgeworth density approximation is attractive for a number of reasons:

1. it is fairly simple;
2. even low-order approximations are likely to fit well; and,
3. it puts no probability mass on negative delays.

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

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

The regularity conditions for this approximation may preclude some or all of the correction terms[10]. However, the base gamma density alone has been shown to fit well in many circumstances.

4.3. **The Prevailing Quote.** Given the background information, the aforementioned literature seems unduly concerned with finding the correct prevailing quote. Instead of trying to pick the correct quote, we seek estimates which characterize the prevailing market conditions.

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

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

[9]A mélange Edgeworth approximation, as in Rosenthal (2008) coould also be explored.
[10]See Rosenthal (2008) for discussion of the regularity conditions.

since positive delays ($z > 0$) correspond to times further in the past. This, and the following, also applies to the expected bid price $\tilde{b}_t$.

I assume nobody would trade on "old" quotes[11]. Thus we truncate the above integration at T and estimate the expected ask price $\tilde{a}_t$ by $\hat{a}_t$:

$$\hat{a}_t = \frac{\int_0^T a_{t-z} f_Y(z) dz}{\int_0^T f_Y(z) dz}. \tag{3}$$

Since forms of $f_Y(z)$ considered die off sufficiently quickly (like e^{-z}) as z increases, we can choose T such that $\tilde{a}_t$ and $\hat{a}_t$ are arbitrarily close (since a_t is, naturally, a bounded process). Because quotes are simple processes, (3) simplifies to a sum involving the delay CDF $F_Y(s)$:

$$\hat{a}_t = \frac{1}{F_Y(s_n; \kappa)} \sum_{i=1}^{n} a_{t-s_i} (F_Y(s_i; \kappa) - F_Y(s_{i-1}; \kappa)) \tag{4}$$

where $t - s_i$ are the observed quote times in the data with $s_0 = 0$ and $s_n = T$. F_Y depends on unknown κ's which are estimated jointly with the classification model parameters. The work to calculate (4) is less than for (2) or (3) — but greater than current methods which "pick" an a_t.

I assume the data are sufficiently frequent that the estimation error $\hat{a}_t - \tilde{a}_t$ is negligible. Formally, this can be motivated by high-frequency asymptotics with the right-continuity of the a_t process. Under such an asymptotic regime, $\hat{a}_t$ is consistent for $\tilde{a}_t$.

Since we condition on $\mathcal{F}_t$, t is not random. The randomness in the (conditional) classification model is due to (i) the unknown amount of time to look backwards for a quote; and, (ii) the unknown trade classification.

5. Microstructure for Trade Signing

Before modeling trade direction, we should consider the microstructure of the trading processes we observe. We should also think about how to normalize the strength of modeling information.

To understand why these issues are important, we begin by examining the price tests currently used for trade signing.

[11]Since traders can monitor the time disseminated by market-data providers, they may have an idea when quotes are old.

5.1. **Price Tests.** Trade signing has typically been inferred using a hierarchy of tests. The methods can be classified by their dominant test:

Midpoint Tests: include the test of Lee and Ready (LR):

Condition	Classification
Trade price > Prevailing midpoint price	Trade is a buy;
Trade price < Prevailing midpoint price	Trade is a sell;
Trade price = Prevailing midpoint price	Use tick test.

Tick Tests: are recommended by Finucane (2000):

Condition	Classification
Trade price > Prior differing trade price	Trade is a buy;
Trade price < Prior differing trade price	Trade is a sell.

Bid/Ask Tests: include the test of Ellis, Michaely, and O'Hara (EMO):

Condition	Classification
Trade price = Prevailing inside ask price	Buy;
Trade price = Prevailing inside bid price	Sell;
Otherwise	Use tick test.

The original 2000 EMO analysis of Nasdaq trades resorted to the tick test for classifying 25% of trades. Peterson and Sirri's (2003) EMO 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. Stoll and Schenzler (2006) suggest trading at the bid and ask is decreasing, implying EMO tests may increasingly resort to the tick test.

5.2. **Is Information Strength Misleading?** It might seem sensible to consider the strength of the information we use to classify trades. However, almost all studies of trade signing note that midpoint, tick, and bid/ask tests appear to be less accurate for trades outside the prevailing spread.

Therefore, it would seem to be misleading to consider the strength of the information we use. As we will see, this conclusion may be premature.

5.3. **Negotiated Trades.** The trades Hasbrouck (2007) refers to as derivatively priced trades are, in general, negotiated trades. SEC (2005) Rule 611 of Regulation NMS covers the flagging of these executions

5.3.1. *Description.* Negotiated trades are trades whose prices are agreed upon in light of the order size and, possibly, adverse selection risk. Since small trades can be executed without negotiation in the continuous trading market, negotiated trades tend to be large.

Examples of negotiated trades include block trades, auction trades, prior reference price trades, and volume-weighted average price trades[12].

Block trades: have historically traded monolithically, *i.e.* the entire order amount moved from one holder to another; the trade was not executed in pieces via the continuous market. From personal experience, I can say this is no longer necessarily true[13].

Auction trades: are often used to start and end continuous trading — or to resume trading after a halt or excess price volatility. Multiple orders, possibly on both sides of the market, are executed together at one time and at a clearing price: the price causing the most shares to trade. If the clearing price is not unique, the price used is the one closest to the previous "normal hours" trade price[14].

Prior reference price (PRP) trades: include trades benchmarked to (i) the bid or ask when the order arrives at the broker-dealer, and (ii) some opening-price trades. The price recorded may be the benchmark price or may include a fee (aka "markdown" or "markup").

Volume-weighted average price (VWAP): and other average price trades are often traded across time to target some average price benchmark. The aggregate quantity is then printed at the benchmark or realized average price, perhaps with a markup or markdown[15].

5.3.2. *Decoupling of Trading and Recording.* A side effect of negotiated trades is the decoupling of trading and recording times. For VWAP and PRP trades, sub-executions are recorded as they happen. After trading is done, an aggregate trade is recorded at the average price and aggregate volume.

Some block trading desks now print the block trade and take it into inventory instead of waiting to find a buyer for the position. They may then trade out of the position or manage a portfolio of "internalized" block trades.

In all of these cases, an analysis not filtering out these aggregate prints will double-count some large trades. Further, the aggregate trade may be incorrectly signed as we see next.

5.3.3. *Why Some Trades Print Outside the Spread.* A simple thought experiment illustrates the problem this decoupling can cause if the data are not properly filtered. Note that the same effect can result from a PRP order.

[12]In TAQ, PRP and VWAP/average price trades are flagged with a "P" and "W" (respectively) in the COND field.

[13]Keim and Madhavan (1996) note that most block trades are seller-initiated and trade below the spread.

[14]FINRA (2007), SEC (2005), and NASD (1999) have more information on Nasdaq opening price trades. NYSE (2006), Securities Training Corporation (2006), and Nasdaq (2006a, 2006b) have more information on closing auctions.

[15]Usually this is driven by whether the order is a principal or agency order, respectively.

Suppose a customer sends a dealer an order to sell 100,000 shares at the four-hour VWAP for 11:00am–3:00pm. (Figure 3 shows the price trajectory.)

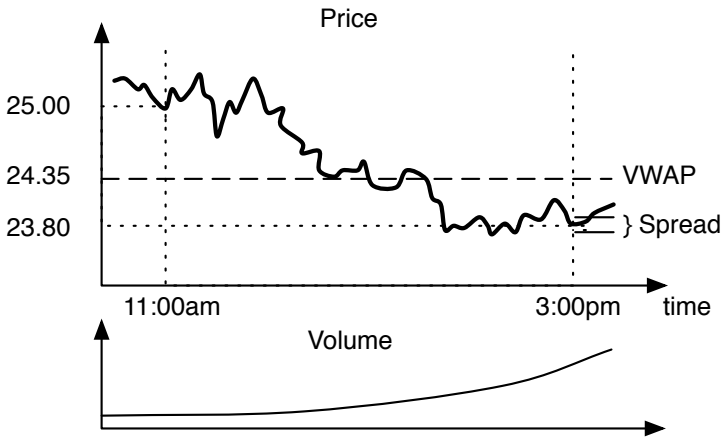

FIGURE 3. An example price and volume trajectory for a four-hour VWAP trade showing how a sell order can appear to be a buy. The trade pushes down the price such that the aggregate trade print is above the contemporaneous spread.

The stock price starts at \$25.00 with a national best bid and offer (NBBO) of \$24.95–\$25.03 at 11:00am. As the trade is split into pieces and traded throughout the four-hour window, the price impact[16] biases the price processes downward. When trading ends at 3:00pm, the stock price is at \$23.80 with a NBBO of \$23.78–\$23.89; the VWAP for 11:00am–3:00pm is \$24.35.

Since the ask price at 3:00pm is below \$24.35, the VWAP print appears to be far above the spread. However, no single trade for that aggregate volume ever occurred; and, that (negotiated) trade was not buyer-initiated.

5.3.4. *Contamination of Previous Studies.* These aggregate prints should be removed. Most studies using TAQ or TORQ data refer to Hasbrouck (1992) or Hasbrouck, Sofianos, and Sosebee (1993) for which trades to remove.

While block trades are mentioned in these sources, VWAP and PRP trades are not mentioned. Since information for how to publish these trades came later (*e.g.* NASD (1999)), this is no sin of omission. Rather, this indicates an evolution of market practice.

How serious is this problem? Ellis, Michaely, and O'Hara (2000) find trades outside the spread for 1996–1997 Nasdaq stocks less than 5% of the time. Peterson and Sirri (2003) find trades outside the spread rose from 2–3%

[16]Not to mention the adverse selection bias inherent to receiving a large sell order.

to 4.5–7% for mid-1997 NYSE stocks when tick sizes decreased from \$1/8 to \$1/16. Stoll and Schenzler (2006) show that trading outside the spread increased during 1999–2002. They also show that in 2002, trades of more than 10,000 shares occurred outside the spread over 50% of the time for Nasdaq stocks versus 6% of the time for NYSE stocks.

Thus large trades outside the spread seem to be a growing problem, particular for Nasdaq-listed stocks. Studies which put more weight on larger transactions will be adversely affected by incorrect trade classification. Most notably, this includes studies of price impact.

Whether negotiated trades account for many of the prints outside the spread is unclear. However, the failure to remove VWAP and PRP trades is a possible contaminant of previous and future studies.

5.4. **Return to the Strength of Information.** If inaccuracy outside the spread is due to data contamination, we can use distances from the midpoint, tick, and bid/ask — the strength of our information.

5.4.1. *Form of Information Strength.* Different stocks have different prices, liquidities, and volatilities. A trade \$0.10 above the midpoint (or spread) might be very informative for a low volatility stock trading at \$20 and nearly uninformative for a highly volatile stock trading at \$200.

My approach is simple: log-proportions. If our information strength function is g and we compare the trade price to the midpoint, this means $g = \log(\text{trade price}) - \log(\text{midpoint})$. While not explored here, it might be more informative to divide g by the stock-specific volatility or mean spread.

5.4.2. *Form of Signed Indicator-like Function.* Since the expected bid and ask may not live on the price lattice, a trade occurring at the expected bid or ask might be a measure-zero event. Therefore, I create a signed indicator-like function J which is approximately -1 and +1 for trade prices near the expected bid and ask[17]. If p_t is the trade price:

$$J(p_t, \hat{b}_t, \hat{a}_t; \tau) = \exp\left(-\left(\frac{p_t - \hat{a}_t}{\tau}\right)^2\right) - \exp\left(-\left(\frac{p_t - \hat{b}_t}{\tau}\right)^2\right) \tag{5}$$

5.4.3. *Terminology.* I reinforce the change from a Boolean test to a signed distance-like measure by a change in terminology. I refer to Boolean tests as "tests" and to signed measures as "metrics". This terminology shift emphasizes when we are working with the strength of our information.

[17]Recall that the estimated prevailing quotes $(\hat{b}_t, \hat{a}_t, \hat{m}_t)$ are functions of the κ_r's. This has computational implications mentioned later.

6. Trade Signing Models

Modeling the sides of a sequence of trades requires two pieces: estimates of prevailing quotes; and, a model using these estimates (and other data) to infer the sequence of sides.

A subtle point here bears emphasis: We are classifying a sequence of trades, not a subset of observable trades in a given stock nor trades which happened to occur at some chosen set of times. Were this not so, we would need to condition on the likelihood of each trade happening at its observed time.

This might seem to be a pedantic difference; and, in a way, it is. However, these two slightly differing statements imply very different models. Formally, our model has a partial likelihood interpretation given in Appendix A.

6.1. **Notation.** To express these ideas mathematically, we define:

t = the time at which a trade is published;
p_t = the price of a trade reported at time t;
p_{t_-} = the price of the most recent trade reported before time t;
p'_{t_-} = the price of the most recent differing-price trade reported before time t;
$\mathcal{F}_t$ = the information known up to time t;
B_t = the initiating side of the trade reported at time t, (1=buy, 0=sell);
$\hat{B}_t$ = the predicted initiating side of the 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 $B_t = 1|\mathcal{F}_t$).

Formally, the bid, ask, and transaction prices at time s (b_s, a_s, p_s) are observed if a quote change or transaction takes place at time s; otherwise they are unobserved. We assume, however, that a_s, b_s are càdlàg. The sigma-field $\mathcal{F}_t$ is based on these processes: $\mathcal{F}_t = \sigma < a_s, b_s, p_s : s \leq t >$.

Further, we suppose there is a delay between each transaction and the quote used by the initiating order. This delay is a random variable Y with cdf $F(y) = \int_0^y f_Y(z)dz$. (Recall section 4.3.) However, the parameters of f_Y are taken to be freely varying: They are effectively model coefficients.

If we abuse our notation slightly: for Y_i associated with the i-th transaction (recorded at time t) we assume Y_i is independent of $\mathcal{F}_t$.

Is this reasonable? The delay Y_i could be related to the size of $\mathcal{F}_t$, *i.e.* how many quote changes and transactions occurred "recently". However, a market might spread trade processing across multiple computers to reduce this effect. Further, economics suggests preferred markets would have sufficient capacity to handle most processing in a timely manner.

Without knowing how an exchange processes trades, we cannot make any inferences about these relationships. For markets that are increasingly competitive, the independence assumption may also be increasingly reasonable.

6.2. **Multi-Stock Model for Trade Side.** Since the response we are predicting is dichotomous, I use a logistic-link GLM. An error term is not included since that is specified by the GLM form: errors are assumed to be independent and to follow a Bernoulli variance (*i.e.* $\text{Var}(\pi_t) = \pi_t(1-\pi_t)$)[18].

Whether the error variance would increase (*i.e.* exhibit overdispersion) on high or low volatility days is unclear. Lower volatility might increase the amount of (strategic) trading within the spread — making classification more difficult; on the other hand, higher volatility might increase the noise in differences of trades and lagged quotes.

Since this model is for multiple stocks, we need to worry about pseudoreplication. To correct for this, we model the side correlations for trades executed at nearby times. This should yield more accurate standard errors and inferences. The effect is similar to Zellner's (1962) SUR and changes the GLM into a generalized linear mixed model (GLMM).

6.2.1. *Correlations and Pseudoreplication.* Trade and quote data may be serially- and cross-correlated. For example, the side of a trade in General Motors stock might help infer the side of a subsequent trade in Ford stock. Accounting for these correlations reduces pseudoreplication (ignorantly treating the data as independent)[19].

Thus any multi-stock model should allow for correlations:

- across time (*e.g.* more buyer-initiated trades in all stocks after a positive news announcement); and,
- across sectors (*e.g.* more buyer-initiated trades in a sector's stocks after a positive sector-related news announcement).

Correlations across industries and individual stock pairs are also likely to be significant. Ideally, we would model the full covariance matrix. However, estimating the entire covariance matrix would be too unwieldy: 3000 stocks would require over 4.5 million parameters. A covariance matrix for 50 industries would still require estimating 1275 parameters.

Econometrics has relied heavily on robust Huber-White standard errors. However, Kauermann and Carroll (2001) show the Huber-White sandwich estimator can substantially underperform explicit correlation modeling.

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

[19]Page 113, point (iii) of Mead (1988) discusses pseudoreplication in a scenario similar to many time series analyses. Page 108 of Van Belle (2002) concisely defines pseudoreplication and explains its toxicity.

If the dependence structure is misspecified, our intuition into the correlation between trade sides will be incomplete. However, Heagerty and Zeger (2000) indicate that misspecification is much less dangerous than omitting a correlation model altogether.

In other words: a wrong or incomplete correlation model is no worse than mindless use of Huber-White standard errors. Our standard errors should be robust to misspecification; and, explicit modeling of the correlation structure may yield insights we would otherwise have lacked.

We also have further recourse. Correcting the standard errors for overdispersion (if the estimated residuals are overdispersed) would probably be more conservative than using Huber-White standard errors with a correlation model. Bootstrap standard errors would be more accurate than any of these methods but might be computationally demanding.

6.2.2. *Different Markets.* In using multiple stocks we may be working with delays originating from multiple market centers. Stocks are listed on a primary exchange; and, as shown by Stoll (2006), where trading takes place is largely segregated by primary exchange[20]. Therefore, different parameters are estimated for stocks with different primary exchanges.

6.2.3. *Indices.* These augmentations to our model require indices:

- j indexes stocks;
- k indexes contiguous time periods ("bins");
- ℓ indexes sectors; and,
- o indexes primary exchanges (*e.g.* NYSE, Nasdaq).

To be clear: a given j implies a value for ℓ and o.

6.2.4. *The Model (Almost).* The classification model may then be written:

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

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

where g is an information strength function and J is a signed indicator-like function as in section 5.4.

6.3. **Correlation Modeling.** Correlation modeling is done separately (i) across time and (ii) cross-sectionally within sectors.

6.3.1. *Autoregression Form.* If we thought autocorrelations diminished with time, we could specify an effect like $\phi_o e^{-\lambda_o(t-t_-)}\eta_{jt_-}$. Unfortunately, a distance-decaying autoregression might be computationally intractable.

Further, autoregression on the preceding log-odds may not be stationary. I avoid this problem and favor a more interpretable model. Therefore, lagged[21] values of the midpoint, tick, and bid/ask metrics were used to capture any autoregressive behavior.

6.3.2. *Random Effects.* The c_k and $d_{k\ell}$ terms are random effects, a statistical technique largely lacking from the econometric time series literature[22].

The random effects capture cross-correlations at the time binning granularity and help account for pseudoreplication. The binning of time is inelegant but handles the reality that stocks rarely trade simultaneously.

The first random effect is a time effect, implying cross-correlations for the initiating side among all stocks transacting during a time bin. This corrects for unpredictable momentum across all stocks and also allows for higher volatility during some portions of a trading day.

The second random effect is a sector effect, implying initiatiing side cross-correlations among all stocks in the same sector and the same time bin. This corrects for unpredictable momentum across all stocks in a sector.

Statistically, the random effects are defined as $c_k \overset{\text{iid}}{\sim} N(0, \sigma_c^2)$ for all bins k and $d_{k\ell} \overset{\text{iid}}{\sim} N(0, \sigma_d^2)$ for all bins k and sectors ℓ. The random effects c_k and $d_{k\ell}$ are assumed to be independent of the sigma-field $\mathcal{F}_t$.

6.4. **Maximum-Likelihood Edgeworth Parameters.** One detail in the above models is more involved than it might seem. The preceding models all use estimates of the prevailing quotes.

The estimated prevailing quote depends on the delay distribution for a primary market center. That distribution is characterized by parameter tuples $(\nu_o, \lambda_o, \tilde{\kappa}_{o3}, \tilde{\kappa}_{o4})$ implied by cumulants ($\kappa_{o,r}$'s) having unknown values[23].

[21] By "lagged" I mean from the preceding trade in that stock.

[22] Pesaran (2007) is a nice example of the burgeoning interest in random effects models.

[23] We do not observe the actual delays incurred by market participants.

I estimate delay distribution and GLMM parameters together. This should be harmless from a modeling perspective and may yield insights into the actual delay distribution or ultra-short-term price predictability.

However, estimating the delay parameters greatly increases the time needed for model fitting. This is another reason for computationally lighter approximations as in equation (4).

6.5. **Multi-Stock Model Coefficients.** The coefficients in the multi-stock model are what biostatisticians call "population average" estimates. However, datasets of trades are unlike most biostatistical longitudinal data analyses in a key way: the number of observations per "individual" (*i.e.* per stock) can vary widely.

I say these are population average estimates, because I believe there are (differing) stock-specific coefficients per stock and that the estimated model coefficients would be weighted averages of these stock-specific coefficients — where the weighting is by the number of transactions.

6.5.1. *Why We Want "Population Average" Coefficients.* There are three reasons we want population average estimates. I believe the first reason is what should drive us to accept the use of these coefficients.

The first reason is driven by the desire for precise comparison. I am comparing my model to other methods based on population average parameters. While those parameters might not have been determined statistically, their purpose was to provide good classification performance overall for the trades in those panels. Thus to make a fair comparison between these methods and a modeling approach, my model should use population average coefficients.

The other two reasons are less relevant but are presented for consideration.

We could view weighting the stock-specific coefficients by trade frequency as assigning importance: Each trade is a vote for the importance of a stock and our coefficients give greater representation to more important stocks.

Alternatively, we could assume each trade contains some small amount of information about the changing state of the world. Then stocks which are traded more receive greater weight since their trades contain more information. This reason says nothing about the relative importance of stocks and implicitly assumes commensurate information in each trade.

6.5.2. *Why We Don't Want "Population Average" Coefficients.* If we want to classify trades for just one stock or improve classification performance by finding better coefficients, the preceding reasons are irrelevant and even dilatory. Longitudinal data analysis suggests two possible remedies.

Random Coefficients. A common longitudinal approach is to assume each coefficient has a mean and a stock-specific random deviation from that mean. This approach is flexible but has a few drawbacks.

The random effects have no explicit relationship to stock characteristics. This makes the model flexible but eliminates any insight about how tests perform as stock characteristics vary.

Further, we can determine the stock-specific coefficients (mean+BLUPs) for stocks in our estimation panel, but not for stocks outside the estimation set. In that case, our best estimate for the stock-specific coefficient is the (population average) mean coefficient. Again, we lack insight.

Transforming Covariates. I previously suggested we might want to normalize the information strength function g by some stock-specific measure like volatility or spread[24]. Since these characteristics exhibit a diurnal, we might even use the average volatility or spread at that time of day.

This should yield coefficients which are identically distributed across all stocks listed on a given market. These market-wide coefficients might give us greater insight into the microstructure of a market. Thus the transformation approach is clearly superior to the random coefficient approach.

Whither Transformations. Comparing a model involving transformed covariates to the standard LR or EMO methods would be unfair: If modeling is a superior approach, that should be evident even with population average coefficient estimates.

However, higher classification accuracy is the ultimate goal. We should certainly search for such transformations. This search should also involve other characteristics: trade size, typical volume, liquidity risk, index memberships, and more. Such a search cannot be resolved in one article.

7. Empirical Analysis: Data

7.1. **Data Source.** To explore modeling trade direction, I used a dataset containing the non-initiating side. (Thus the initiating side is the opposite of the side in the dataset. See Archipelago (2005a) for more information.)

The ArcaTrade dataset from Archipelago (now NYSE Arca) has all trades occurring on the Archipelago ECN and Exchange for a given month[25]. For December 2004, Archipelago (2005b) reports their share of traded volume as 23.2%, 22.5%, and 2.3% for AMEX-, Nasdaq-, and NYSE-listed stocks.

[24]This could also be done for the bid/ask metric J or its parameter τ.

[25]Previously-studied datasets with the initiating side are much older and not openly-accessible. The ArcaTrade dataset, by comparison, is updated monthly and is available from the NYSE's `www.nysedata.com` website.

For inside quotes, I used the ArcaSIP consolidated NBBO dataset for the same month; that dataset has since been retired. Future researchers must use quotes from the NYSE's TAQ or from the ArcaBook dataset.

7.2. **Time Resolution Augmentation.** One problem with the dataset: trades and quotes are only timestamped up to one-second resolution. This problem is not unique to the ArcaTrade dataset; many datasets may have multiple trades resolving to the same time.

My approach is to assume messages in a file are uniformly distributed within their second. Thus two trades at "9:35:01" are assumed to have occurred one- and two-thirds of a second after "9:35:01". Since all messages within a file must be counted, data cleaning cannot happen before time resolution.

7.3. **Data Synchronization.** The lagged tick test in the model forced a choice of data sources. I could get the preceding trades from TAQ or I could use the preceding Arca-executed trades.

Using TAQ would require finding the Arca trade so as to then find the preceding tick. This matching would be less accurate for more common (smaller) trades. Thus matching errors could induce serious bias. I would also have to mix datasets from possibly differing clocks. Since time is a crucial part of the analysis, that is a troubling prospect[26].

To avoid these issues, I used the preceding tick in the ArcaTrade dataset. Sometimes that might be the tick from two or more trades prior. The assumption I have made is that this merely adds noise to the tick test covariates. If the *location* of trading is autocorrelated, this assumption would not hold[27]. However, I have no reason to believe this is so.

7.4. **Data Cleaning and Augmentation.** ArcaTrade data includes pre- and post-market trades. However, unlike TAQ, ArcaTrade data does not include negotiated trades and auction trades.

The microstructure of the market post-opening auction and pre-closing auction is the subject of other studies; here, we are interested in the microstructure of continuous trading. I exclude all trades occurring before 10:00 AM to eliminate pre-market trades and trades affected by opening auctions (including those of related stocks). I also exclude trades after 3:30 PM to eliminate post-market trades and trades affected by closing auctions[28].

[26]For future researchers, this suggests preferring the ArcaBook dataset to TAQ.

[27]Since the dataset is a subset of all transactions, we also need to consider that ArcaTrade transactions might not be representative of the overall market.

[28]Orders for AMEX and NYSE closing auctions are due at the specialist by 3:40 PM; orders for the Nasdaq closing cross are due by 3:50 PM. After these times, specialists and market makers may be transacting to hedge the liquidity risks of these orders and speculators may be trading based on published estimates of auction order imbalances.

For estimating random effects, I assign trades to sectors[29] and contiguous ten-minute bins. I also restrict my attention to stocks in the Russell 1000 large–mid-cap and 2000 small-cap indices. (Together, the "Russell 3000".) Index membership was as of the 2004 annual June rebalance[30].

7.5. **Summary Statistics.** The resulting dataset covered two days: 1 and 2 December 2004. This was composed of 2,178,307 transactions across the 2,836 stocks in the Russell 3000 which were still active under their ticker on the rebalance data.

7.5.1. *Stock Characteristics.* These stocks represented all three primary US markets (AMEX, Nasdaq, and NYSE) and 13 sectors. Characteristics of those stocks are shown in Tables 1 and 2.

Worth noting in Table 1 is 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, the bulk of the transactions are for Nasdaq stocks, probably because most trading of NYSE-listed stocks happens on the NYSE[31].

	Number of		Trade-Weighted Average			
Market	Stocks	Trades	Shares	Price	Mkt Cap	Spread
AMEX	35	2,797	489.7	$36.42	$1,664MM	0.13%
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%
All	2,836	2,178,307	326.1	$28.51	$6,285MM	0.14%

TABLE 1. Characteristics by market of the stocks analyzed. All were members of the Russell 1000 or 2000 as of July 2004.

In Table 2, we can note that a plurality of stocks in the dataset are service-related companies; however, the bulk of the trades in this dataset are of technology-related companies. The industrial goods sector appears unusually small — an artifact of the data-gathering process and the changing of sector names over time[32].

7.5.2. *Covariate Characteristics.* The midpoint and tick metrics are differences of log-prices; this compares prices in percentage rather than absolute

[29]Sectors are available from historical Yahoo web pages stored at `www.archive.org`.

[30]Active futures exist on these indices so the data likely include index arbitrage transactions.

[31]Recall Archipelago's 2.3% market share of NYSE trading from section 7.1.

[32]The net effect for the model is minor at best: The random effects had slightly more freedom to counteract pseudoreplication. Coefficient estimates and out-of-sample prediction are completely unaffected.

	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 2. Characteristics by sector of the stocks analyzed. All were members of the Russell 1000 or 2000 as of July 2004.

terms. The bid/ask metric is approximately +1 for trade prices near the estimated ask and approximately -1 for trade prices near the estimated bid.

Table 3 lists summary statistics for these covariates by market. The means confirm that on most of these markets there was not a major imbalance between buying and selling. The extremes and the standard deviation show the order of magnitude for the covariates. Thus Nasdaq stocks often trade within 10bp[33] and 16bp of the midpoint and preceding trade; NYSE stocks often trade within 10bp and 24bp of the midpoint and preceding trade.

Correlation matrices for covariates and their lagged values (Table 4) show that current-trade metrics are strongly correlated with preceding-trade metrics. Also, the bid/ask metric is more strongly correlated with the tick metric on the (decentralized) Nasdaq than on the two (specialist-driven) exchanges.

7.5.3. *Covariate Plots.* We can also try to visualize these relationships. However, plain scatterplots will not work since the Nasdaq and NYSE observations are too numerous: The overlap of points presents us from seeing the variation in point densities.

One solution is to create a contour plot of the scattering of data points: plot contour lines where point densities are above visual resolution and individual points where the point density is lower. These "scatter contour plots" take some getting used to; but, they offer a way to visualize millions of overlapping data points.

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

Market	Stat	Metrics Bid/Ask	Midpoint	Tick
AMEX	Max	1	0.03	0.06
	Mean	0.29	2.7×10^{-4}	3.2×10^{-4}
	Min	-1	-0.02	-0.02
	StDev	(0.79)	(1.9×10^{-3})	(4.2×10^{-3})
Nasdaq	Max	1	0.06	0.12
	Mean	4.8×10^{-2}	3.6×10^{-5}	3.8×10^{-5}
	Min	-1	-0.17	-0.11
	StDev	(0.79)	(1.0×10^{-3})	(1.6×10^{-3})
NYSE	Max	1	0.02	0.05
	Mean	0.10	4.8×10^{-5}	3.5×10^{-5}
	Min	-1	-0.05	-0.05
	StDev	(0.79)	(9.6×10^{-4})	(2.4×10^{-3})

TABLE 3. Summary statistics by market of the covariates used in the classification model.

	AMEX			Nasdaq			NYSE		
	B/A	Mid	Tick	B/A	Mid	Tick	B/A	Mid	Tick
Midpt	0.37			0.47			0.43		
Tick	0.17	0.43		0.36	0.55		0.18	0.35	
Pr. Bid/Ask	0.47	0.17	0.09	0.58	0.30	0.20	0.38	0.22	0.07
Pr. Midpt	0.16	0.46	0.16	0.29	0.63	0.20	0.21	0.64	0.15
Pr. Tick	0.08	0.20	0.59	0.22	0.33	0.55	0.08	0.20	0.57

TABLE 4. Correlations by market between covariates used in the classification model. Correlations among lagged covariates are omitted since they are identical to correlations among unlagged covariates. "B/A" and "Mid" denote the bid/ask and midpoint metrics.

Scatter plots (AMEX stock trades) and scatter contour plots (Nasdaq and NYSE) of the relationships between the bid/ask, midpoint, and tick metrics are shown in Appendix B. The plots suggest the metrics are indeed correlated — but that some of the correlation is due to extreme observations.

Most unusual is the plot of the midpoint versus tick metrics for AMEX and NYSE stocks (Figure 4). The plot for NYSE stocks shows secondary modes. These secondary modes imply a clustering of trade prices about $\pm0.7\%$ away from the midpoint and about twice that distance from the previous trade. Part of this effect is faintly observable in the AMEX plot; however, the effect is completely absent from the Nasdaq plot.

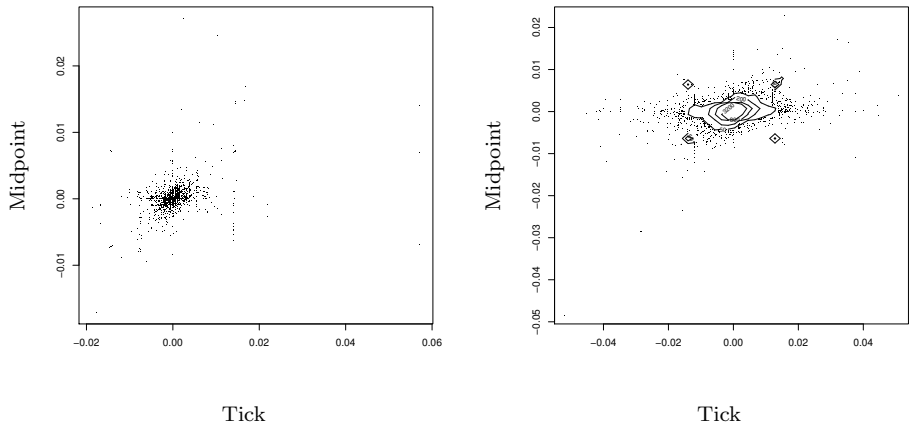

FIGURE 4. Scatter plots of midpoint versus tick metrics for AMEX (left) and NYSE (right) stocks. Secondary modes are clearly visible on the NYSE plot and faintly visible on the AMEX plot, suggesting some trades can be characterized as bid-ask bounce or occur at successive bids/asks.

Since the tick metric compares successive trades, some elevated number of successive NYSE stock trades occur at twice their distance from the midpoint. The mode tick coordinates ($\pm 1.5\%$) are about ten times the average NYSE spread of 0.15% in Table 1.

One explanation for these secondary modes would be that some market participants maintain wide quotes (presumably of a large number of shares) to provide liquidity to the market when price jumps occur. Another explanation is that NYSE stocks trade on Arca for quotes which are relatively wide. This explanation would suggest Arca supplies liquidity to the market (i) at times of wide spreads, or (ii) for stocks with wide spreads.

Either way, a noticeable amount of Arca trading in NYSE stocks happens at some quoted or effective spread. The NYSE midpoint versus tick metric plot also indicates those stocks were subject to both bid-ask bounce (secondary modes in the upper left and lower right quadrants) and trades at successive bids or asks (lower left and upper right quadrants).

Why are these secondary modes absent from the Nasdaq plot? Arca's higher market share of Nasdaq trading could be a reason; but, then the effect should also be absent from the AMEX plot. The absence could also be due to the competitive nature of the Nasdaq market versus the centralized nature of specialist markets.

7.5.4. *Lagged Covariate Plots.* Scatter (AMEX) and scatter contour plots (Nasdaq, NYSE) of the relationships between the lagged and unlagged metrics are shown in Appendix C. The plots suggest serial correlation of the various metrics as well as serial cross-correlations. The secondary modes seen on a NYSE midpoint versus tick metric plot are also visible on these lagged covariate plots involving the midpoint and tick metrics.

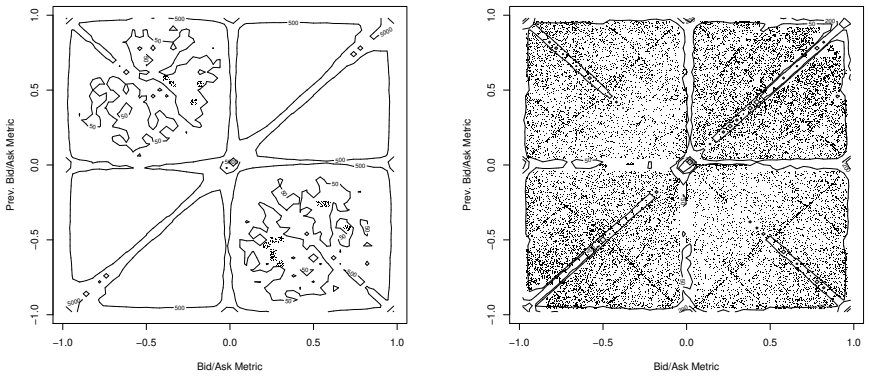

FIGURE 5. Scatter plots of lagged versus current bid/ask metrics for Nasdaq (left) and NYSE (right) stocks. The clustering along the 45-degree line implies persistence of buying/selling; the secondary clustering along the -45-degree line implies reversals of buying/selling at the inside quote.

Unusual among these plots are those of the previous versus current bid/ask metric for Nasdaq and NYSE stocks (Figure 5). The clustering along the 45-degree line in both these plots implies a persistence of buying and selling. The secondary clustering along the -45-degree line implies reversals of buying/selling at the inside quote — including bid-ask bounce[34].

These plots also exhibit two lesser patterns of clustering. Lesser clustering in the upper left and lower right quadrants but not on the -45-degree line indicates partial bid-ask-like bounce. Other lesser clustering creates a faint diamond pattern joining the middle points of each side of the plot. This could be due to (i) orders "sweeping" beyond the spread into the order book and (ii) the inside quote shifting around some equilibrium trading price.

[34]The clusters along the zero bid/ask (vertical) axis are 0+ and 0- ticks; the clusters along the zero previous bid/ask (horizontal) axis are "-0" and "+0" ticks.

8. Modeling Issues

The previously-stated model seems to cleanly extend the various tests traditionally used to classify trades. However, estimating coefficients for these models faces difficulties.

The model cannot be nicely linearized nor is there a tractable way to get closed-form gradients or Hessians for the log-likelihood. Thankfully, this difficulty is only encountered in fitting the model; using a fitted model to classify trades is computationally simple.

The model also has a potential identifiability problem. A modified conjugate direction method resolves this through bound constraints. Using truncation as in (4) to estimate the prevailing quotes requires other constraints.

8.1. **Computation.**

I estimate the nonlinear parameters; generate covariates which use these parameters; and then, fit a GLMM using these (conditional) covariates. The optimization iterates these three steps to find the best parameters and their information matrix. The control structure and hardware details are given in Appendix D[35]

This process is computationally intensive. Computing tricks and smart optimization are needed for reasonable performance. Exploring the nonlinear parameter space was done with a conjugate direction (CD) method — an analog of the conjugate gradient method for derivative-free optimization[36]. Details are in Chapter 9 of Nocedal and Wright (2006).

The CD method was modified to enforce bound constraints were enforced on the base gamma distribution parameters[37] and on the bid/ask (τ) parameter. Quadratic interpolation was used for line searches if the suggested minimizer decreased the model log-likelihood.

Penalized quasi-likelihood (PQL) was used for estimating the GLMM parameters. PQL can fail if the estimated model is nearly- or fully-separable (*i.e.* some predictions are numerically zero or one)[38]. Unbalanced analysis of deviance, Laplace approximation, and adaptive Gaussian quadrature were not used due to their memory requirements and/or their sloth.

Despite performance tuning, the task was formidable. Fitting one model took over 43 hours of "wall clock" time. I also tried estimating models with

[35]Perl and C code for data manipulation and model fitting is available online.

[36]This method allows us to find optimal nonlinear parameters; however, the approximation of gradients and inverse Hessians may result in inflated estimates of standard errors for the nonlinear parameters.

[37]The base gamma distribution parameters (ν, λ) must be positive.

[38]PQL also tends to have problems when there are a small number of observations per realization of a random effect; however, that is certainly not of concern here.

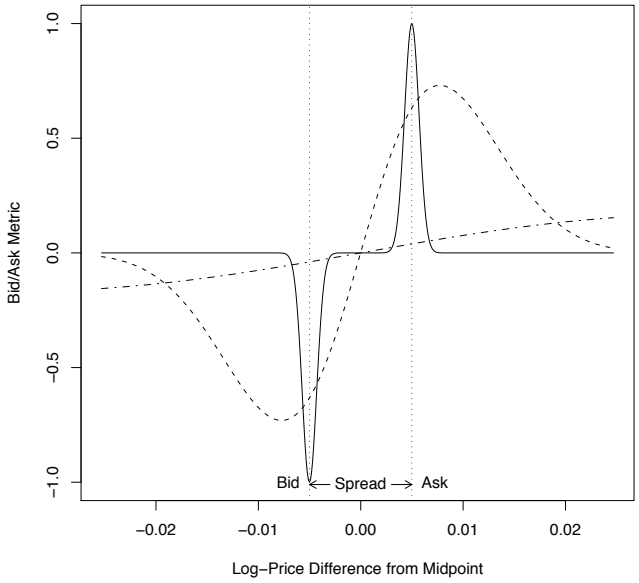

FIGURE 6. Example of how the bid/ask metric $J(;\tau)$ may be collinear with the midpoint metric. The graphs show J for a 1% bid-ask spread when $\tau = 0.001$ (solid line), $\tau = 0.01$ (dashed line), and $\tau = 0.05$ (dash-dotted line). As τ grows, the bid/ask metric becomes nearly linear over a wide range. Since the midpoint metric is linear and most trades occur within the spread, this can cause an identifiability problem.

random samples of a limited number of stocks, but these attempts either did not converge or could not estimate the random effects.

8.2. **Identifiability: Bid/Ask Metric versus Midpoint Metric.** The $J(;\tau)$ function which measures the strength of the bid/ask-like metric requires τ to be positive, but τ must also be restricted from growing large relative to a typical bid-ask spread.

Suppose we let τ grow beyond the fractional spread (*e.g.* 0.01 for a 1% bid-ask spread). As τ grows J begins to resemble a midpoint test (see Figure 6). Since most trading occurs within the spread, this can introduce an identifiability problem.

To resolve this problem, I restrict τ to be less than 0.05. This restriction is very mild (a 5% spread is large) and is sufficient to preserve identifiability and interpretation of the model.

Does this bias the τ parameter? It should not; and, if it did, we should not be alarmed. Bid/ask tests (*e.g.* EMO) are equivalent to bid/ask metrics with

τ forced to be very small[39]. Further, we should not be afraid of introducing bias if it allows us to perform model selection. The utility of the LASSO and ridge regression are testaments to this perspective.

8.3. **A Truncation Caveat.** I use truncation as in (4) to estimate the prevailing quotes. Specifically, I truncate quotes beyond 90 seconds. Given finite memory and enforcement of the 90-second reporting rule, truncation seems sensible. I also require that the truncated probability mass be trivially small. This requires further delay parameter bounds[40].

9. Model Estimation

9.1. **Model Statement.** We can now formally state a model and estimate it for the data. The model has six market-specific terms as well as an overall intercept and cross-market random effects:

$$
\begin{aligned}
&P(B_{jt} = \text{Buy} | \mathcal{F}_t, c_k, d_{k\ell}; \theta_o, \kappa_o) = \pi_{jt};\\
&\pi_{jt} = \text{logit}(\eta_{jt}); \quad \text{and,}\\
&\eta_{jt} = \underbrace{\beta_0}_{\substack{\text{bias}\\ =0?}} + \underbrace{\beta_{o1} g(p_{jt}, \hat{m}_{jt})}_{\text{midpoint test}} + \underbrace{\beta_{o2} g(p_{jt}, p'_{jt-})}_{\text{tick test}} + \underbrace{\beta_{o3} J(p_{jt}, \hat{b}_{jt}, \hat{a}_{jt})}_{\text{bid/ask test}} + \\
&\qquad \underbrace{\beta_{o4} g(p_{jt-}, \hat{m}_{jt-})}_{\text{lag-1 midpoint test}} + \underbrace{\beta_{o5} g(p_{jt-}, p'_{jt--})}_{\text{lag-1 tick test}} + \underbrace{\beta_{o6} J(p_{jt-}, \hat{b}_{jt-}, \hat{a}_{jt-})}_{\text{lag-1 bid/ask test}} + \\
&\qquad \underbrace{c_k}_{\substack{\text{overall}\\ \text{effect}}} + \underbrace{d_{k\ell}}_{\substack{\text{within-}\\ \text{sector}\\ \text{effect}}} .
\end{aligned}
\tag{7}
$$

9.2. **Model Estimation.** Estimating the model in (7) with the ArcaTrade dataset and performing some basic model selection yielded the model coefficient estimates shown in Table 5.

Standard errors were estimated using an efficient MCMC resampling technique plus dividing the dataset into days and computing the variance of those coefficient estimates[41]. The resulting standard errors apply to coefficients estimated across multiple days for Russell 3000 stocks.

The fixed effect parameter estimates are highly significant and roughly in line with the results from various analyses of deviance. In addition, the intercept (a nuisance parameter) is not large enough to be troubling.

[39] J is only +1 or -1 for strict equality with the bid or ask and 0 otherwise.

[40] The mean implied delay (ν_o/λ_o) was restricted to be less than five seconds; and, the truncated probability was restricted to be less than 0.01.

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

Fixed Effect	AMEX	Nasdaq	NYSE
τ	Overall: 2.1×10^{-4} (0.3)		
ν	1.66 (0.58)	1.65 (0.65)	0.62 (0.47)
λ	0.35 (3.7)	0.33 (0.40)	0.78 (0.35)
$\tilde{\kappa}_3$	—	—	—
$\tilde{\kappa}_4$	—	—	—
Intercept	Overall: 0.06 (0.02)		
Midpoint	—	209 (11)	122 (13)
Tick	—	29.4 (8.4)	-20.5 (8.5)
Bid/Ask	1.20 (0.25)	1.41 (0.02)	2.04 (0.20)
Prev. Midpoint	—	—	—
Prev. Tick	—	—	—
Prev. Bid/Ask	0.33 (0.31)	-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 5. Estimated parameters for trade direction model in equation (7). Standard errors are computed by a combination of bootstrap resampling and subsampling by time. Due to numerical reasons, the standard errors for the nonlinear parameters (Greek leters) are overstated.

Recall that the model coefficients in Table 5 are "population average" coefficients — suitable for comparison with the LR and EMO methods. I assume the coefficients are related to market microstructure, but are averages of stock-specific coefficients weighted by how often each stock is traded. This means the standard errors reflect both (i) the difference between the population betas and sample estimates and (ii) variation in the population betas due to the averaging of stock-specific betas.

9.3. **Analysis Commentary.** A handful of features should be noted from these estimates:

1. Trading at the bid or ask seems to be very frequent on the AMEX since bid/ask metrics alone were sufficient for modeling;
2. Trading near the bid or ask on the two major markets is highly informative and statistically significant for classifying trades;
3. Midpoint metrics are less informative than trading at the bid or ask but are still important for modeling;
4. Trade classifications seem to show strong autocorrelations: classification metrics from a preceding trade were significant for all markets;

5. The sign difference for the Nasdaq versus NYSE tick metric may be due to a difference in short-sales rules. NYSE short sales are governed by SEC Rule 10a-1 (the zero-plus tick test); Nasdaq short sales are governed by FINRA Rule 3350 (the bid test rule)[42]
6. Negative coefficients for preceding bid/ask metrics agrees with other studies' observations of bid-ask bounce;
7. The sector × time random effect indicates about a 2% correlation of buying or selling across same-sector stocks in a ten-minute period; and,
8. The time random effect indicates about a 0.2% correlation of buying or selling across all stocks in a ten-minute period.

The tick metric appears less informative in our models. However, we introduced noise (see section 7.3) into the tick metric by using the preceding trade in the ArcaTrade dataset for the preceding trade across all markets. For the AMEX and Nasdaq stocks, Arca trades a large share of total volume; thus those tick metrics should be less noisy[43].

The negative tick metric coefficient for NYSE stocks indicates some correction for short sales in that market. Not correcting for how and when short sales can take place is considered by Asquith, Oman, and Safaya (2007). Thus the problems they highlight may be less problematic in these modeled trade classifications.

Had the ArcaTrade dataset noted which trades were short sales, we might have been able to better develop and assess the model's performance. Unfortunately, this information is not in the ArcaTrade dataset currently.

9.4. **Correlated Trade Classifications.** Autocorrelations in trade direction seem plausible *a priori*. And for various models considered, all Nasdaq and NYSE preceding-trade metrics had significant coefficients.

However, the aggressive model selection PQL requires eliminated all but the preceding bid/ask metrics. These can be thought of as either modeling autocorrelations or as nuisance parameters to correct for bid-ask bounce. The random effects indicate strong (for financial markets) cross-correlations of buying and selling within a sector and short time period.

9.5. **Varying Coefficients.** Transforming covariates by stock characteristics might yield coefficients applicable to any stock[44]. I explored one such characteristic: the average delay-model-estimated spread. I computed the

[42]The bid test rule states that short sales cannot occur at or below the inside bid if that inside bid is lower than the previous inside bid.

[43]Nonetheless, model selection removed the tick metric from the AMEX model.

[44]Johnson (2008) indicates likely characteristics: measures of typical volatility, volume, and spreads.

average model-estimated spread for each stock, computed average spread deciles, and fit a model with a different coefficient for each spread decile.

The plots in Figure 7 show how the four metric coefficients vary with average spread for the three US markets: Some model coefficients vary with the average spread — or with something related to the average spread.

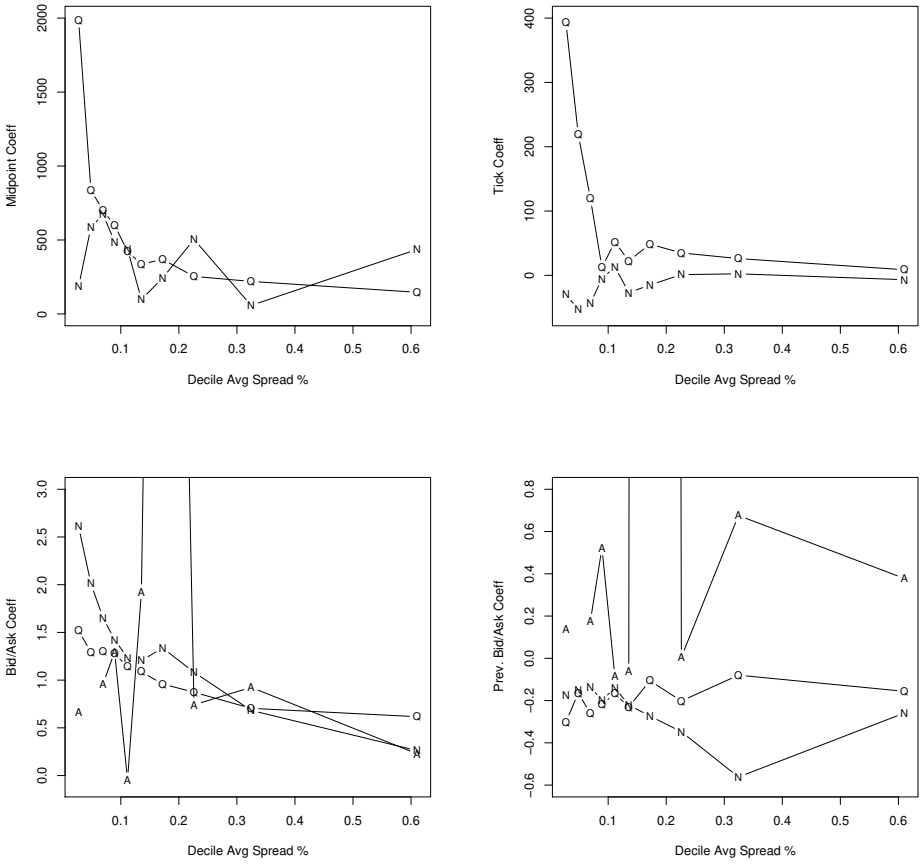

FIGURE 7. Model coefficients estimated for deciles of average model-estimated spread. The midpoint coefficients are shown in the upper left plot; the tick coefficients are in the upper right plot; the bid/ask coefficients are in the lower left plot; and, the previous bid/ask coefficients are in the lower right plot. Data for the AMEX, Nasdaq, and NYSE are marked with "A", "Q", and "N". The off-plot AMEX bid/ask coefficient is almost 17; the off-plot AMEX previous bid/ask coefficient is almost 73.

Midpoint metric coefficients (used only for Nasdaq and NYSE stocks) behave differently for the two markets. The Nasdaq coefficients decrease with increasing average spreads — indicating that normalizing the midpoint metric by the average spread might be useful for Nasdaq stocks. The NYSE coefficients, however, appear to be increasing then decreasing. Whether this is spurious or a result of the NYSE microstructure is not clear.

Tick metric coefficients (again only Nasdaq and NYSE) show a similar but more pronounced pattern. The Nasdaq coefficients decline sharply and then more gradually as spread increases above the 4-th decile. The NYSE coefficients may be increasing for the first five deciles but are relatively constant for the last five deciles.

Bid/ask metric coefficients are clearly declining with increasing average spread for Nasdaq and NYSE stocks. While the AMEX coefficients are very noisy, one could perhaps say they also decline with spread.

Preceding bid/ask metric coefficients exhibit no clear patterns. Whether this is due to the model term being spurious or unrelated to spreads is unclear.

We could conduct hypothesis tests to check if the coefficients are all constant across various stock characteristics. However, the results for some of these plots are strong enough that even a back-of-the-envelope nonparametric test is very significant.

9.6. **Short- and Ultra-Short-Term Alpha.** The cross-correlations implied by the random effects suggest there is some short-term[45] predictability of buying and selling. This is nonparametric evidence of short-term alpha — alpha available to those who know some trade classifications and estimate the random effect BLUPs.

The significance of the lagged metrics suggests the presence of ultra-short-term alpha. Further confirmation of this suggestion comes from a curious accident in the modeling process.

Unconstrained fitting of the delay parameters yielded a mean quote delay much greater than suggested by Table 5. Mean "delays" implied by these unconstrained parameters were 106s and 34s for the AMEX and NYSE; Nasdaq parameters were unchanged.

This does not indicate delays are greater than previously expected since we do not observe delays. Rather, these parameters may be due to ultra-short-term autocorrelations in returns and quote changes: Stronger metrics preference older quotes. This preference could also be related to the speed at which the AMEX and NYSE (both specialist-driven markets) publish

[45]On the order of ten minutes.

trades or update quotes; or, this preference could be due to autocorrelations in specialist trading.

10. Out-of-Sample Performance

10.1. **Competing Schemes.** The last 20 days in December 2004 were used for out-of-sample evaluation. Modeled trade classifications were compared to those from a Lee and Ready-style midpoint test; an Ellis, Michaely, and O'Hara-style bid/ask test; and, a tick test.

The Lee and Ready test uses quotes from five seconds earlier; however, quotes in this dataset are only published with one-second resolution. Multiple quotes might be found in the one-second window of "five seconds prior".

Should one use the oldest quote in that window or the newest (most recent) quote in that window? I have done both. The Lee and Ready test using the oldest "five seconds prior" quote is referred to as "LR.old"; the Lee and Ready test using the newest such quote is referred to as "LR.new"[46].

10.2. **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 6 and 7 and Figure 8.

		Percent of Trades Correctly Classified				
Market	N	Modeled	EMO	LR.new	LR.old	Tick
AMEX	19,435	69.8%	70.3%	59.2%	58.9%	52.5%
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%
Overall	16,504,880	74.7%	72.8%	72.1%	71.7%	66.2%

Table 6. Percent of trades correctly classified as buyer- or seller-initiated across three US markets for 3–31 December 2004. Classification methods used: modeled; Ellis, Michaely, and O'Hara's bid/ask test; Lee and Ready's midpoint test using the newest five-seconds-prior quote; Lee and Ready's test using the oldest five-seconds-prior quote; and, the tick test. Excluding December 31 would lower accuracies by about 0.5% for the AMEX and about 2% for the NYSE.

Modeled classifications are more accurate than all other methods for Nasdaq and NYSE stocks; the EMO test is slightly more accurate than the model

[46]Researchers who use the WRDS database should be aware that WRDS uses "LR.new" for trade classification.

		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 7. Percent of trades correctly classified as buyer- or seller-initiated across sectors. Classification methods used: modeled; Ellis, Michaely, and O'Hara's bid/ask test; Lee and Ready's midpoint test using the newest five-seconds-prior quote; Lee and Ready's test using the oldest five-seconds-prior quote; and, the tick test.

for AMEX stocks. Since these constitute a small part of this sample (and of the Russell 3000 universe), this underperformance is a minor concern.

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.

Excluding December 31, the accuracy across all methods would be about 0.5% lower for AMEX stocks and about 2% lower for NYSE stock; accuracy for Nasdaq stocks would be roughly unchanged.

Why were 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 markets.

Does estimating random effects involving sectors affect performance in a particular sector? Table 7 indicates it does not. Modeled classifications are superior to the next best method, often by 1–2% — except for the Industrial Goods sector where the EMO method is 1.3% more accurate. This sector has little representation in our universe and is of minor concern.

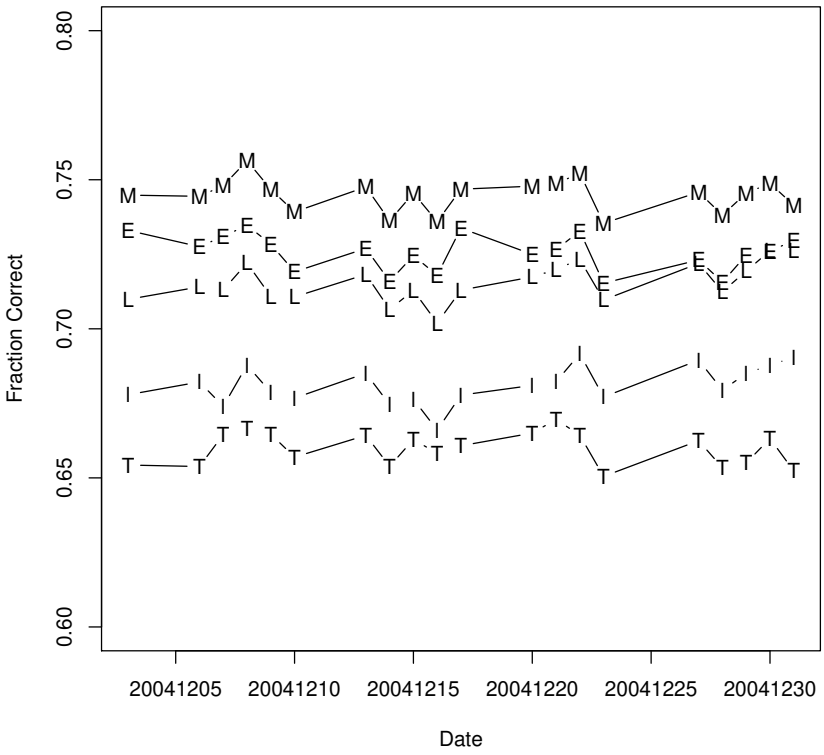

FIGURE 8. Fraction of trades correctly classified as buyer- or seller-initiated across time. Classification methods used: modeled (M); Ellis, Michaely, and O'Hara's bid/ask test (E); Lee and Ready's midpoint test using the newest five-seconds-prior quote (L); Lee and Ready's test using the oldest five-seconds-prior quote (l); and, the tick test (T).

Finally, we might wonder about the variation across dates. Figure 8 shows that while there is variation in accuracy, modeled classifications are superior to the next best method for all of our out-of-sample dates.

10.3. **Accuracy Across the Spread.** Many studies note lower classification accuracy for trades occurring outside the spread. We have already discussed how negotiated trades might be responsible for some of this; and, the ArcaTrade dataset excludes negotiated trades. However, we should still examine the performance of these methods relative to the prevailing quote.

Since the "prevailing quote" is not universally agreed-upon, I examine classification accuracy across the three versions of the "prevailing quote": EMO, LR, and delay-modeled. [47]

10.3.1. *EMO Prevailing Quote.* Classification accuracy across the Ellis, Michaely, and O'Hara version (0s lag) of the bid-ask spread is shown in Table 8. For an unlagged quote[48], only 7% of trading occurs outside the spread.

Location vs.		Percent of Trades Correctly Classified				
EMO Quote	N	Modeled	EMO	LR.new	LR.old	Tick
Above Ask	638,041	74.0%	73.0%	70.8%	70.4%	73.0%
At Ask	7,951,786	73.1%	71.4%	70.1%	70.0%	64.5%
>Mid, <Ask	705,964	65.2%	58.3%	64.9%	64.7%	58.3%
At Midpoint	490,690	67.0%	58.3%	65.1%	64.8%	58.3%
>Bid, <Mid	695,830	66.1%	60.0%	66.2%	66.0%	60.0%
At Bid	5,476,692	79.8%	79.5%	76.8%	76.3%	70.1%
Below Bid	545,877	78.3%	75.0%	77.6%	77.2%	75.0%

TABLE 8. Classification accuracy relative to EMO prevailing quote (0s lag). Note that accuracy outside the spread is higher, perhaps because the ArcaTrade dataset eliminates negotiated trades.

Note that classification is not less accurate outside the spread; rather, it is more accurate. This contrasts with Ellis, Michealy, and O'Hara's (2000) and Peterson and Sirri's (2003) findings of lower accuracy outside the spread than at the bid or ask. Whether the difference is due to a more recent dataset, the elimination of negotiated trades, or both is not clear.

We can also note that the modeled trade classifications beat all other methods across all parts of the EMO-defined spread with one exception: the LR.new method is 0.1% more accurate for trades above the bid but less than the midpoint.

The EMO method resorts to the tick test for 18.6% of the trades in Table 8. Table 8 also suggests that a better method would be to use the midpoint method for trades not at the EMO bid or ask.

10.3.2. *LR Prevailing Quote.* The classification accuracy across the Lee and Ready version (5s lag) of the prevailing bid-ask spread is shown in Table 9. About 21% of trading occurs outside the LR-defined spread.

[47]Here I refer to the newest quote five seconds prior to a trade report. This is the "LR.new" method in the preceding tables and is what is returned by WRDS.

[48]We should be skeptical about using unlagged quotes since trades are often published with delay. A quote contemporaneous with the trade print could easily be directly affected by the printed trade. This possible endogeneity seems to have garnered little attention.

Location vs.		Percent of Trades Correctly Classified				
LR Quote	N	Modeled	EMO	LR.new	LR.old	Tick
Above Ask	1,728,511	74.5%	73.7%	70.9%	70.9%	70.1%
At Ask	6,629,664	74.0%	73.0%	71.6%	71.1%	64.4%
>Mid, <Ask	361,013	60.9%	61.9%	53.3%	52.5%	52.6%
At Midpoint	284,534	61.9%	61.4%	52.1%	50.8%	52.1%
>Bid, <Mid	370,777	61.9%	62.2%	55.2%	54.5%	52.9%
At Bid	5,429,370	78.0%	73.9%	76.5%	76.1%	68.4%
Below Bid	1,701,011	75.2%	74.3%	71.9%	71.9%	70.5%

TABLE 9. Classification accuracy relative to LR prevailing quote (5s lag). Accuracy is higher outside than inside the spread, perhaps due to eliminating negotiated trades.

Note that classification is again more accurate outside than inside the spread. Less clear is whether classification outside the LR-defined quote is more accurate than classification at the LR quote.

However, modeled trade classifications are not universally better. For trades inside the LR-defined spread but not at the midpoint, the EMO method is between 0.3% and 1% more accurate than the model. For all other parts of the LR-defined spread, modeled trade classifications are more accurate.

10.3.3. *Delay-Modeled Prevailing Quote.* The classification accuracy across the delay model-estimated prevailing bid-ask spread is shown in Table 10. 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.

Location vs.		Percent of Trades Correctly Classified				
Modeled Quote	N	Modeled	EMO	LR.new	LR.old	Tick
Above Ask	2,231,990	76.4%	75.7%	74.6%	74.4%	72.8%
At Ask	4,387,127	75.6%	75.7%	75.1%	75.0%	67.1%
>Mid, <Ask	1,879,244	67.6%	64.5%	59.1%	57.8%	53.6%
At Midpoint	30,316	45.0%	48.0%	48.8%	48.8%	48.8%
>Bid, <Mid	1,870,881	67.8%	63.8%	59.5%	58.2%	53.9%
At Bid	3,917,366	78.4%	74.7%	78.1%	78.0%	70.0%
Below Bid	2,187,956	77.2%	75.7%	74.9%	74.6%	72.7%

TABLE 10. Classification accuracy relative to delay model-estimated prevailing quote. Note that accuracy outside the spread is higher, perhaps due to eliminating negotiated trades. Also, accuracy at the estimated midpoint is abysmal.

In Table 10, performance for trades at the model-estimated midpoint is abysmal: trades at the model-estimated midpoint would be better classified by flipping a coin — an alarming prospect. Thankfully, trades at the model-estimated midpoint are a tiny fraction of all trades.

The model is also slightly less accurate (0.1%) than the EMO method for trades at the model-estimated ask. Across the rest of the model-estimated spread, modeled trade classifications are more accurate than all other methods. Again, classification is generally more accurate outside the spread than at or inside the quote.

10.4. **Performance Attribution.** Since the model performs well, we might wonder how much of the superior performance is due to various aspects of the model. We can examine this with a series of models; the models range from simple to the full model of the previous section.

To properly attribute the 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[49]. If the midpoint or bid/ask test is inconclusive, the test contributes no information to the model. The tests use conventional quote delays: 5s for midpoint, 0s for bid/ask. AMEX stocks use only the bid/ask test to mirror the full model. No delay model nor random effects used.
2. **Metrics** The preceding model except using tick, midpoint, and bid/ask *metrics* (g and J from section 5.4). AMEX stocks use only the bid/ask metric.
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 in are nested, changes in accuracy (moving from left to right) are additive[50].

[49]The tests are modified to return -1 and +1 to indicate likely sells and buys.

[50]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)

		Models					
Coefficients		Tests	Metrics	AR Effect	Random Effects	Ad-hoc Delay	Full
Bid/Ask	τ	—	0.000209	0.000209	0.000209	0.000209	0.000209
AMEX	ν	—	—	—	—	3	1.66
	λ	—	—	—	—	2	0.35
	Bid/Ask	1.27	1.30	1.14	1.17	1.06	1.21
	Pr. B/A	—	—	0.45	0.48	0.31	0.33
Nasdaq	ν	—	—	—	—	3	1.65
	λ	—	—	—	—	2	0.33
	Bid/Ask	0.70	1.21	1.19	1.19	1.19	1.41
	Midpt	0.52	160	155	157	214	209
	Tick	0.18	80	82	83	67	67
	Pr. B/A	—	—	0.05	0.06	-0.20	-0.20
NYSE	ν	—	—	—	—	3	0.62
	λ	—	—	—	—	2	0.78
	Bid/Ask	1.19	1.87	1.89	1.91	1.72	2.04
	Midpt	0.52	197	205	211	344	122
	Tick	0.02	-17.5	-19.0	-18.9	-27.6	-20.5
	Pr. B/A	—	—	-0.05	-0.05	-0.22	-0.17

TABLE 11. Estimated model coefficients for the nested performance attribution models. "Pr. B/A" denotes coefficients for the preceding-trade bid/ask metric.

		Percent of Trades Correctly Classified					
Market	N	Tests	Metrics	AR Effect	Random Effects	Ad-hoc Delay	Full
AMEX	19,435	67.7%	70.2%	70.6%	70.6%	69.8%	69.8%
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%
Overall	16,504,880	71.1%	73.8%	73.7%	73.7%	74.6%	74.7%

TABLE 12. Percent of trades correctly classified as buyer- or seller-initiated across US markets for various models. This lets us attribute the full model's performance to various improvements. Models increase in complexity from left to right and are defined in section 10.4; the names indicate the added feature compared to the preceding simpler model. The "Full" column is the same as the "Modeled" column in Table 6.

From Tables 6 and 12 we see that the initial model (Tests) is less accurate than the EMO method for AMEX and 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 AMEX and NYSE stocks (by 8.5% and 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 2.5%, 3%, and 1.1% for AMEX, Nasdaq, and NYSE stocks. This increased accuracy makes the Metrics model compare favorably to the EMO method; the Metrics model is 0.1% less accurate for AMEX stocks and 1% and 1.3% more accurate for Nasdaq and NYSE stocks.

The AR Effect model adds a lagged bid/ask metric, an autoregressive term. This increases accuracy 0.4% for AMEX stocks, but decreases accuracy 0.1% for Nasdaq stocks and 0.6% for NYSE stocks. The AR Effect model is, however, more accurate than the EMO and LR methods across all markets.

As expected, the random effects in model Random Effects have no effect on the out-of-sample classification accuracy[51].

The Ad-hoc Delay model shows varying improvement: 0.8% worse for AMEX stocks, 0.9% better for Nasdaq stocks, and 0.7% for NYSE stocks.

The Full model adds market-specific delay parameters. The effects of these also vary: no improvement for AMEX stocks; a 0.2% accuracy improvement for Nasdaq stocks; and, a 0.3% accuracy loss for NYSE stocks.

Summarizing, three improvements stand out:

1. Including the strength of information (converting the EMO, LR, and tick tests to *metrics*) yields a 2.5%–3% improvement in accuracy for NYSE and Nasdaq and a 1.1% improvement for AMEX stocks.
2. Adding a basic delay model is responsible for a 0.7%–0.9% improvement in classification accuracy for NYSE and Nasdaq stocks.
3. For AMEX stocks only, the addition of the autoregressive effect improved accuracy by 0.6%. This might indicate that trading direction on the AMEX truly is autocorrelated; or, this could be an artifact of the base AMEX model simplicity.

---

allowing for autocorrelated information; (iv) using random effects; and, (v) adding an ad-hoc universal delay model.

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

Term	Estimate	Std Error
Intercept	0.152	0.009
Volume (millions)	0.032	0.010

TABLE 13. Linear model for EMO method resorting to the tick test versus volume in millions of shares. The scale (0.152 = 15.2%) and significant positive relationship indicate a possible interaction between higher volume and lower accuracy.

Some improvements did not work: The lagged bid/ask metric (a basic autoregressive effect) was significant in-sample; out of sample it reduced accuracy by about 0.1% and 0.6% for Nasdaq and NYSE stocks. Also, for AMEX stocks, delay models yielded 0.8% lower classification accuracy.

10.5. **Resorting to the Tick Test.** Stoll and Schenzler's (2006) findings that more trading is occurring outside the (unlagged) spread indicates that bid/ask methods (*e.g.* the EMO test) will increasingly resort to the tick test.

The EMO test resorted to using the tick test 16.5%, 18.4%, and 14.2% of the time for AMEX, Nasdaq, and NYSE trades. Across our test period, the EMO test used the tick test to classify anywhere from 16.0% to 19.4% of trades on a given day.

Across all three markets, the frequency and interday pattern of resorting to the tick test are almost the same as the overall numbers. These numbers are less than the 25% tick test use found by Ellis, Michaely, and O'Hara in their sample of Nasdaq stocks. Despite Stoll and Schenzler's findings, the EMO test does not appear to be increasingly resorting to the tick test.

The interday numbers reveal a small but significant relationship between higher volumes and increased use of the tick test. A basic linear model for this relationship is given in Table 14. This suggests exploring volume or trade-count interactions in our model.

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 54.3%, 64.6%, and 60.6% on the AMEX, Nasdaq, and NYSE — versus overall tick test accuracy (Table 6) of 52.5%, 66.7%, and 60.7%.

However, there are some times at which the EMO test intelligently resorts to the tick test. Table 8 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.

11. Further Questions

Modeled trade classifications are more accurate across many different strata. The preceding has demonstrated the power of using the strength of classification metrics; using a basic delay model; and, allowing for autocorrelations in trade direction on some markets. How can we do even better?

One possibility is to include interactions. This plays to the strength of a modeled approach. The most obvious interaction is the one just mentioned: an interaction with volume. This might yield a model which gives more weight to some tests at low-volume times versus high-volume times.

Another possible interaction is with share size. This could capture differences in order handling for small versus large orders or odd-lot versus round-lot and mixed-lot orders[52].

Exploring covariate transformations could be fruitful. We might scale the covariates by the stock-specific mean spread (as briefly examined here), volatility, trade size, or some measure of typical daily volume. Finding these transformations could yield insight for theoretical microstructure models.

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.

Mere observers cannot use random effects to help out-of-sample prediction; but, market participants can use random effects to compute BLUPs[53] for a given sector and time bin. This would let them discern an increased tendency toward buying or selling and better control inventory.

More intriguing is the possibility of placing random trades and comparing their known classifications to the model. This would allow us to estimate the random effects for time and sector bins and can be thought of as a way to calculate nonparametric (model-free) short-term alpha.

Models including the preceding-trade tick and midpoint metrics could not be estimated with PQL. However, Laplace approximation or unbalanced analysis of deviance might allow models with these terms to be estimated. Whether this would be overfitting for all markets is unclear.

Also of interest: if we detect cross-correlations at a ten-minute timespan, how strong are cross-correlations at other tenors? Would using one-minute bins eliminate the need to restrict the model from using much older quotes?

[52]Large orders are often traded on a best efforts basis, avoiding Reg NMS stipulation of "best execution". Odd-lot orders are explicitly excluded by the Manning and Display Rules (IM-2110-2 and 11ac1-4).

[53]Best Linear Unbiased Predictors estimate particular realizations of a random effect.

We could also explore correlations across asset classes. Buying on futures markets might help predict buying or selling in the "cash" markets — whether from index arbitrage or co-movement.

Finally, we could explore market features suggested by the covariate plots. The NYSE trades exhibit clear secondary modes at tick metrics twice the spread metrics. The AMEX trades faintly exhibit similar secondary modes. These suggest an elevated amount of trading at effective bids and asks.

Trades at successive bids/asks imply some adverse selection for quote providers. Since Arca later[54] merged with the NYSE, we might wonder if these secondary nodes have disappeared for NYSE trades. Post-merger data would show if this pattern has changed on the NYSE alone — indicating a reduction in adverse selection for quoters.

Appendix A. Partial Likelihood Interpretation

The trade classification model is formally valid when formulated as a partial likelihood as in Cox (1975) and Wong (1986).

Let t_i be the time of the i-th trade and $\mathcal{G}_i$ be a sigma-field encapsulating the information on whether trades $1, \ldots, i$ were buys or sells. Then the full likelihood ratio (Radon-Nikodỳm derivative) can be decomposed as:

$$\mathcal{L}(\text{all data}) = \prod_{i=1}^{n} \mathcal{L}(B_{t_i} | \mathcal{F}_{t_i}, \mathcal{G}_{i-1}) \times \prod_{i=1}^{n} \mathcal{L}(\mathcal{F}_{t_i}, \mathcal{G}_{i-1} | \mathcal{F}_{t_{i-1}}, \mathcal{G}_{i-1}). \tag{8}$$

For inference we only use the first of these two factors — thus making this a partial likelihood. We assume B_{t_i} is conditionally independent of $\mathcal{G}_{i-1}$ given $\mathcal{F}_{t_i}$, yielding

$$\mathcal{L}(B_{t_i} | \mathcal{F}_{t_i}, \mathcal{G}_{i-1}) = \mathcal{L}(B_{t_i} | \mathcal{F}_{t_i}). \tag{9}$$

Appendix B. Model Covariate Plots

[54]The merger of the NYSE and Archipelago Holdings closed on 7 March 2006.

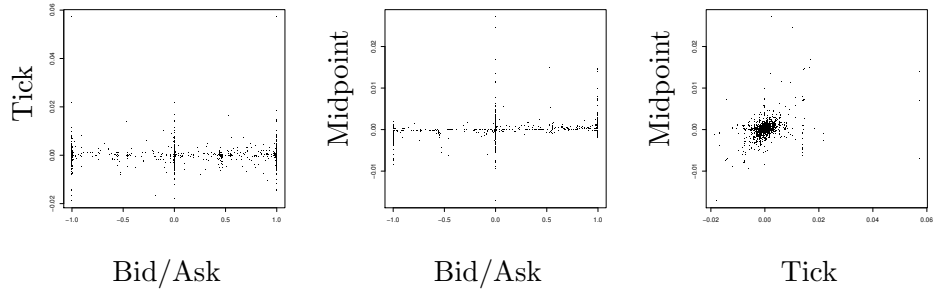

FIGURE 9. Scatter plots for AMEX stocks of metrics versus one another.

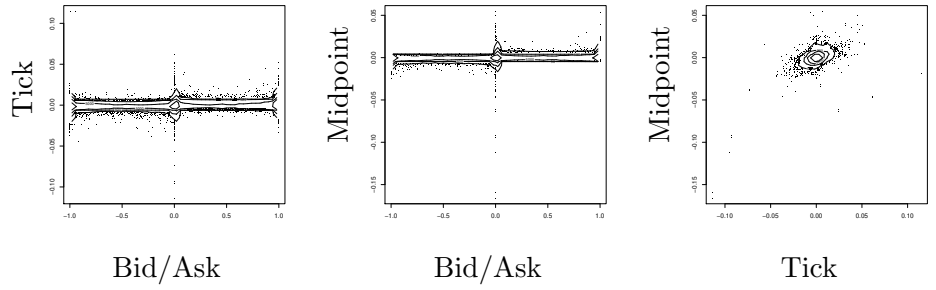

FIGURE 10. Scatter plots for Nasdaq stocks of metrics versus one another.

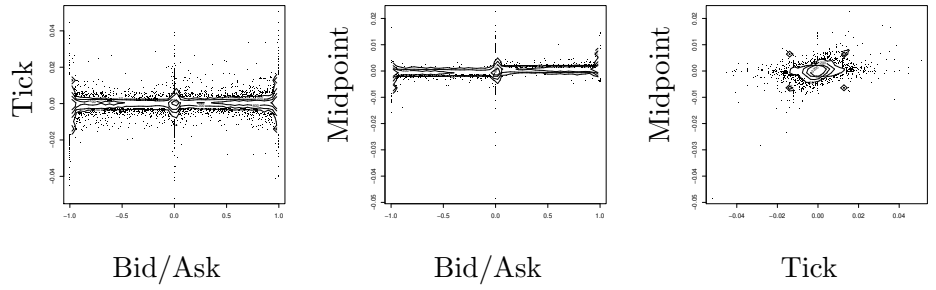

FIGURE 11. Scatter plots for NYSE stocks of metrics versus one another. Note the secondary modes on the midpoint vs. tick plot suggesting some NYSE trades can be characterized as bid-ask bounce or trades at successive bids/asks.

Appendix C. Lagged Covariate Plots

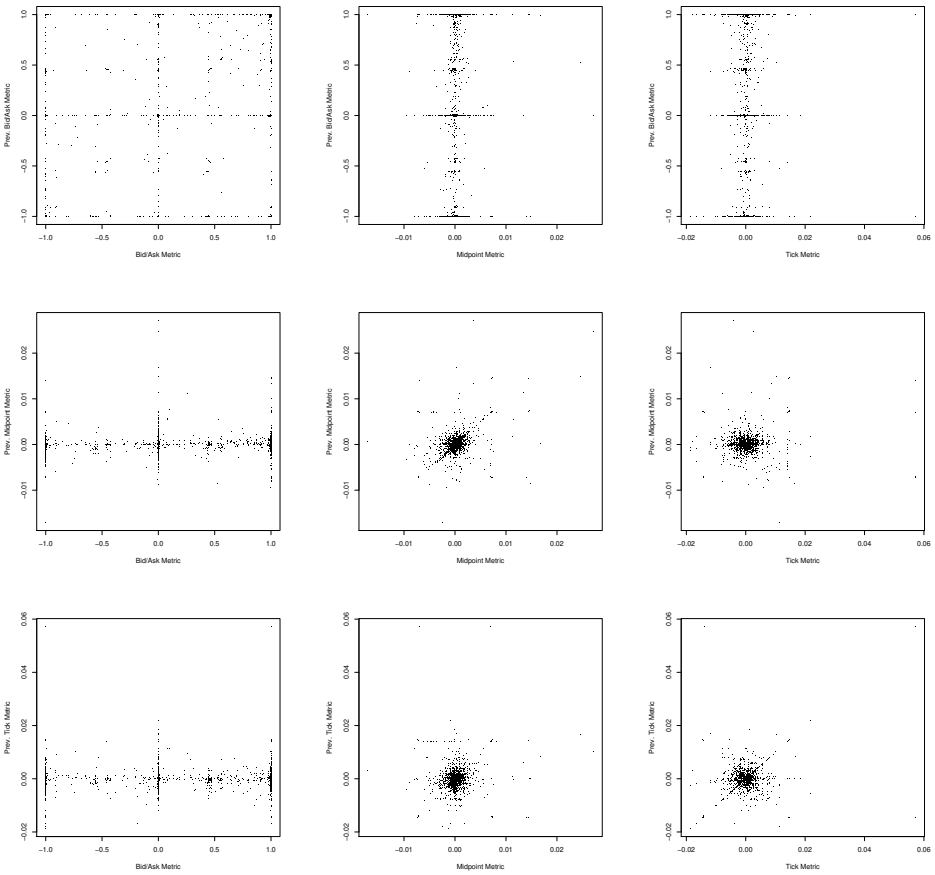

Figure 12. Scatter plots of lagged versus current metrics for AMEX stocks.

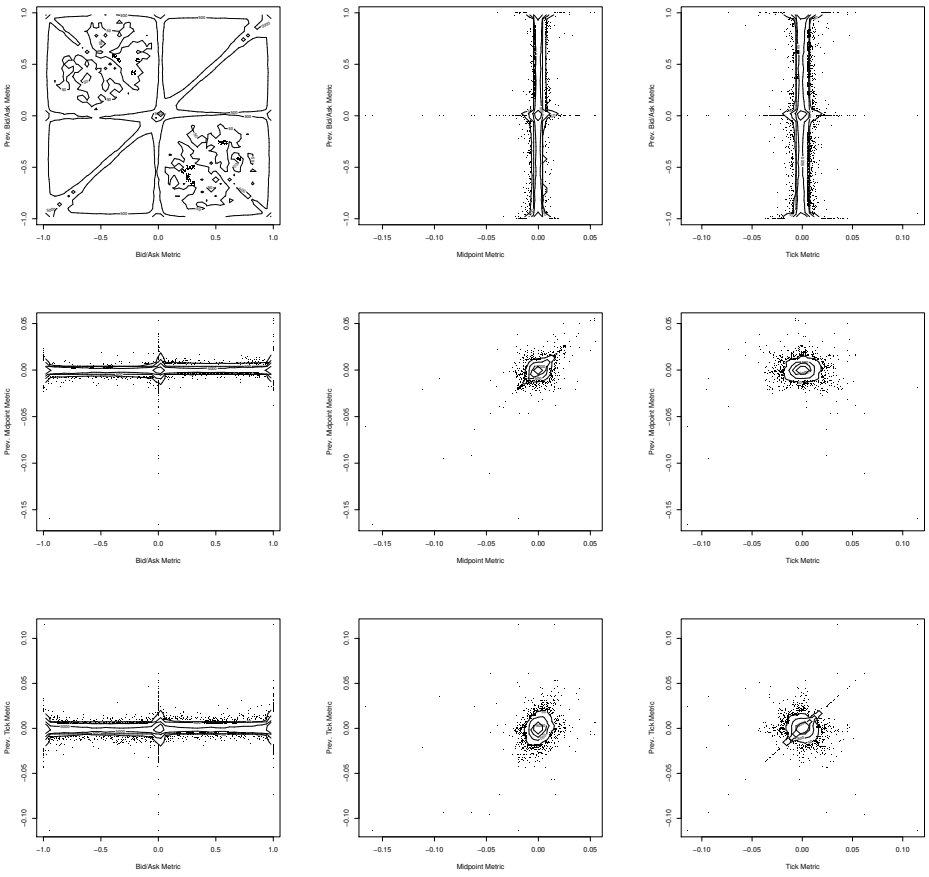

FIGURE 13. Scatter plots of lagged versus current metrics for Nasdaq stocks.

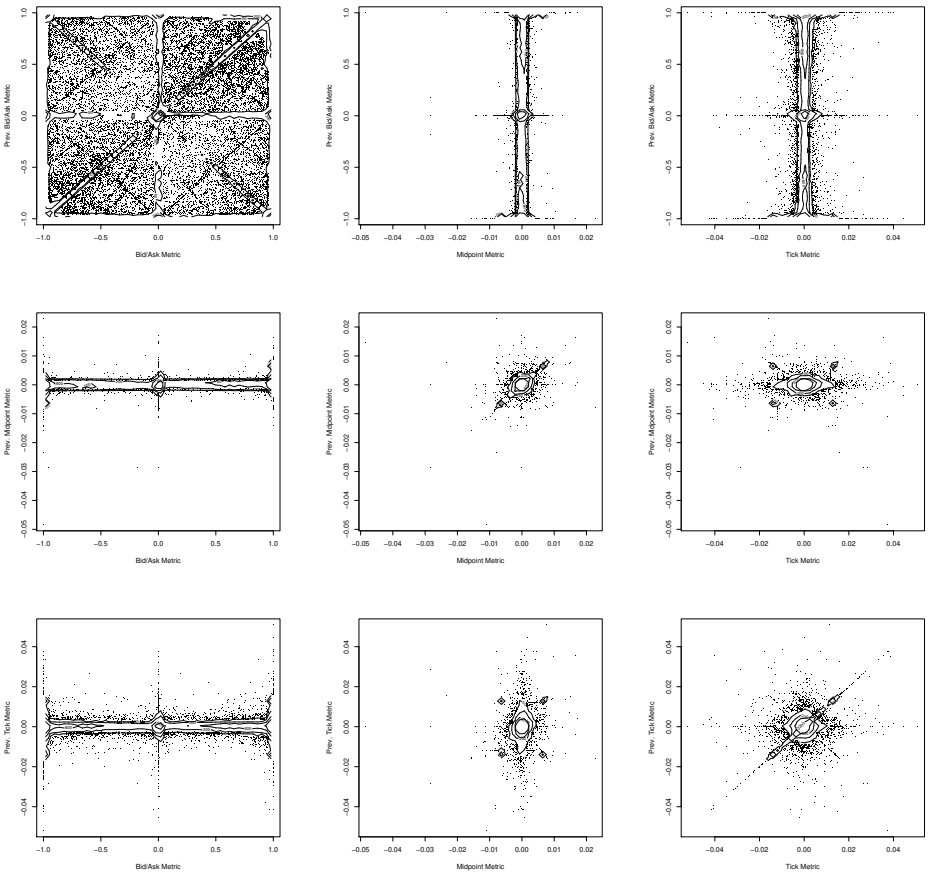

FIGURE 14. Scatter plots of lagged versus current metrics for NYSE stocks.

Appendix D. Computation Structure

The only dataset preprocessing was done on quotes by a Perl script. The script used the daily ArcaSIP consolidated NBBO data files to create daily streamlined quote files. These streamlined files had vectors the preceding 90 seconds of bid and ask changes and times for each trade.

The driver script was written in Perl augmented with the `Math::MatrixReal` package to implement conjugate direction linear algebra. C was used for the calculation of expected quotes for a given day. This allowed multiple processes to run simultaneously, each for a separate day.

A modified version of *R* used the `lme4` package to fit the GLMM. *R* was compiled to use the GotoBLAS threaded linear algebra package. This allowed some linear algebra calculations to use all of the machine's processors and generated a speed increase of about 50%.

The combination of Perl, distributed C, and multi-threaded R seems baroque. However, experimenting with other approaches showed this to be simplest, most robust high-performance approach available. Estimation was done on a 64-bit Linux machine with 32 GB of memory and 8 CPUs. Each CPU ran at 2.33 GHz and accessed data on a 1.07 GHz bus. Fitting the model took 262 iterations which took 43 hours and 18 minutes of "wall clock" time.

The interplay of these processes and data is illustrated in Figure 15. Since this structure was complicated, the driver script was built to allow recovery from crashes by re-reading saved output files.

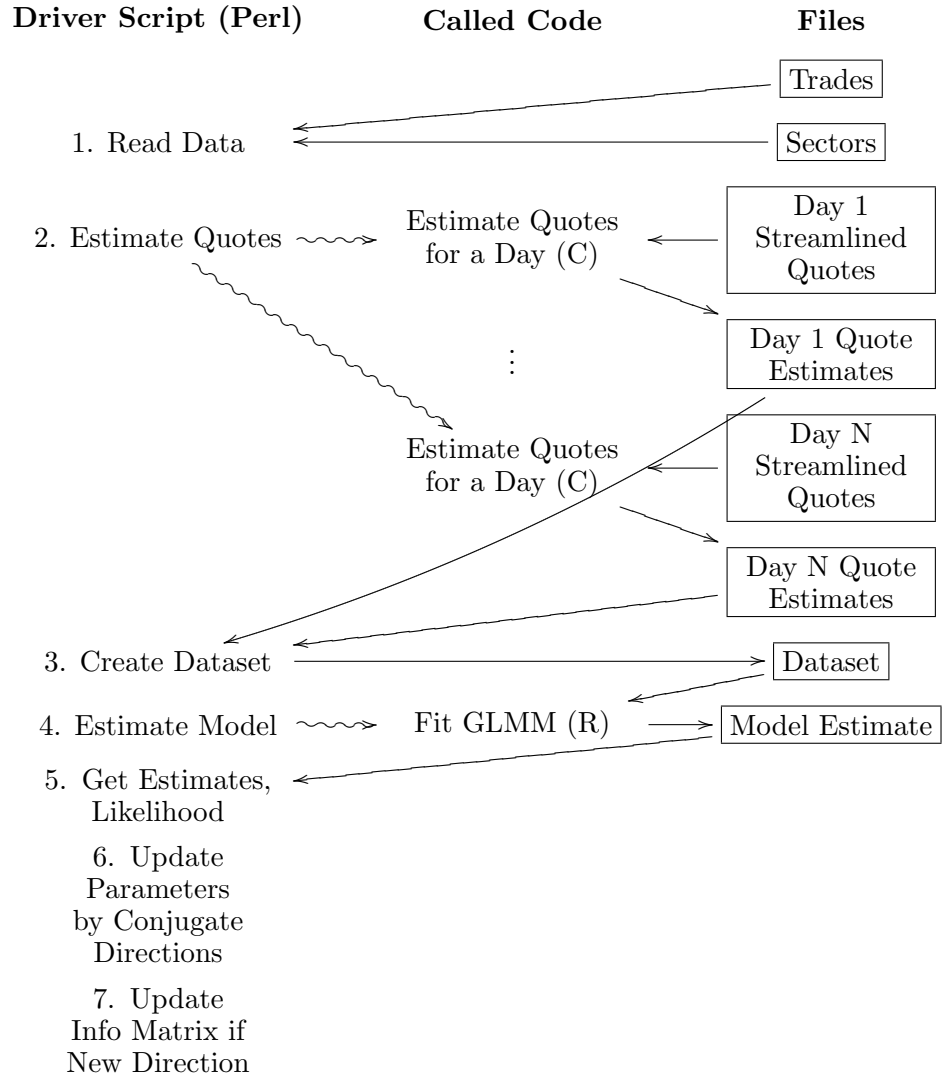

FIGURE 15. Control structure and data flow for model estimation. Steps 2–7 are repeated until convergence. Straight arrows (→) indicate data flow; squiggly arrows (⇝) indicate control flow.

References

[1] Archipelago Holdings, Inc. *ArcaBook and ArcaTrade Historical: For the Archipelago Exchange and ArcaEdge, Version 1.1*, February 2005. Archipelago Holdings, Inc.: Chicago.

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

[3] Asquith, P.; R. Oman and C. Safaya, "Short Sales and Trade Classification Algorithms" SSRN Working Paper (July 2, 2007). Retrieved on 15 June 2008 from `http://ssrn.com/abstract=951420`.

[4] Bartlett, M. S. "Approximate Confidence Intervals", *Biometrika*, 40:1/2(1953a), 12–19.

[5] Bartlett, M. S. "Approximate Confidence Intervals: II. More Than One Unknown Parameter", *Biometrika*, 40:3/4(1953), 306–317.

[6] Bennett, P. and Wei, L. "Market Structure, Fragmentation, and Market Quality", *Journal of Financial Markets*, 9:1(2006), 49–78.

[7] Boehmer, E. "Dimensions of Execution Quality: Recent Evidence for US Equity Markets", *Journal of Financial Economics*, 78:3(2005), 553-582.

[8] Box, G. E. P. "A General Distribution Theory for a Class of Likelihood Criteria", *Biometrika*, 36:3/4(1949), 317–346.

[9] Caudill, S. B.; B. B. Marshall and J. Garner. "Improved Trade Classification Rules: Estimates Using a Logit Model Based on Misclassified Data", *Atlantic Economic Journal*, 32:3(2004), 256.

[10] Chacko, G. C.; J. W. Jurek and E. Stafford. "The Price of Immediacy", *Journal of Finance*, 63:3(2008), 1253–1290.

[11] Cox, D. R. "Partial Likelihood", *Biometrika*, 62:2(1975), 269–276.

[12] Ellis, K.; R. Michaely and M. O'Hara. "The Accuracy of Trade Classification Rules: Evidence from Nasdaq", *Journal of Financial and Quantitative Analysis* 35:4(2000), 529–551.

[13] Erlang, A. K. "The Theory of Probabilities and Telephone Conversations", *Nyt Tidsskrift for Matematik*, B:20(1909), 33–39.

[14] Financial Industry Regulatory Authority. Notice to Members 07-23: NASD Trade Reporting Requirements, 11 May 2007. Retrieved on 30 December 2007 from `http://www.finra.org/RulesRegulation/NoticestoMembers/2007NoticestoMembers/P019150`.

[15] Finucane, T. J. "A Direct Test of Methods for Inferring Trade Direction from Intra-Day Data", *Journal of Financial and Quantitative Analysis* 35:4(2000), 553–576.

[16] Fitzmaurice, G.M.; N. M. Laird and J. H. Ware. *Applied Longitudinal Analysis*, 2004. Wiley: New York.

[17] Forrester, J. W. "Information Sources for Modeling the National Economy", *Journal of the American Statistical Association*, 75:371(1980), 555–566.

[18] Garman, M. B. "Market Microstructure", *Journal of Financial Economics*, 3:3(1976), 257–275.

[19] Hasbrouck, J. "Measuring the Information Content of Stock Trades", *Journal of Finance*, 46:1(1991), 179–207.

[20] Hasbrouck, J. *Empirical Market Microstructure*, 2007. Oxford University Press: New York.

[21] Hasbrouck, J. "Using the TORQ Database." *NYSE Working Paper #92-05*. NYSE (1992).

[22] Hasbrouck, J. and Schwartz, R. A. "Liquidity and Execution Costs in Equity Markets", *Journal of Portfolio Management*, 14:3(1988), 10–16.

[23] Hasbrouck, J.; G. Sofianos and D. Sosebee. "New York Stock Exchange Systems and Trading Procedures." *NYSE Working Paper #93-01*. NYSE (1993).
[24] Heagerty, P. J. and Zeger, S. L. "Marginalized Multilevel Models and Likelihood Inference", *Statistical Science* 15:1(2000), 1–19.
[25] Henker, T. and Wang, J. "On the Importance of Timing Specifications in Market Microstructure Research", *Journal of Financial Markets* 9(2006), 162–179.
[26] Johnson, T. C. "Volume, Liquidity, and Liquidity Risk", *Journal of Financial Economics*, 87:2(2008), 388–417.
[27] Kauermann, G. and Carroll, R. J. "A Note on the Efficiency of Sandwich Covariance Matrix Estimation", *Journal of the American Statistical Association* 96:456(2001), 1387–1396.
[28] Keim, D. B. and Madhavan, A. "The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects", *Review of Financial Studies* 9:1(1996), 1–36.
[29] Lee, C. M. C. and Ready, M. J. "Inferring Trade Direction From Intraday Data", *Journal of Finance* 46:2(1991), 733–746.
[30] Lefèvre, E. *Reminiscences of a Stock Operator*, 1923. Doubleday, Doran & Company: New York.
[31] McCullagh, P. and Nelder, J. A. *Generalized Linear Models, 2nd Edition*, 1989. Chapman and Hall: London.
[32] McCullagh, P. *Tensor Methods in Statistics*, 1987. Chapman and Hall: London.
[33] McCulloch, C. E. and Searle, S. R. *Generalized, Linear, and Mixed Models*, 2001. Wiley: New York.
[34] Mead, R. *The Design of Experiments*, 1988. Cambridge University Press.
[35] National Association of Securities Dealers. Notice to Members 99-66: NASD Trade Reporting Requirements, 10 August 1999. Retrieved on 31 December 2007 from `http://www.finra.org/RulesRegulation/NoticestoMembers/1999NoticestoMembers/P004186`.
[36] The Nasdaq Stock Market LLC. *The Nasdaq Closing Cross Fact Sheet*, November 2006. Retrieved on 3 January 2007 from `http://www.nasdaqtrader.com/trader/openclose/ccfactsheet.pdf`.
[37] The Nasdaq Stock Market LLC. *The Nasdaq Opening Cross Fact Sheet*, November 2006. Retrieved on 3 January 2007 from `http://www.nasdaqtrader.com/trader/openclose/openfactsheet.pdf`.
[38] New York Stock Exchange, Inc. *TAQ 3 User's Guide, Version 1.0*, 3 January 2005. New York Stock Exchange, Inc.: New York.
[39] New York Stock Exchange, Inc. New Product: NYSE Liquidity Replenishment Points, 7 September 2006. Retrieved on 7 December 2006 from `http://www.nysedata.com/nysedata/default.aspx?tabid=155\&id=114`.
[40] Niederhoffer, V. and Osborne, M. F. M. "Market Making and Reversal on the Stock Exchange", *Journal of the American Statistical Association* 61:316(1966), 897–916.
[41] Nocedal, J. and Wright, S. J. *Numerical Optimization, 2nd Edition*, 2006. Springer: New York.
[42] O'Hara, M. *Market Microstructure Theory*, 1997. Blackwell: London.
[43] Odders-White, E. R. "On the Occurrence and Consequences of Inaccurate Trade Classification", *Journal of Financial Markets* 3(2000), 259–286.
[44] Osborne, M. F. M. "The Dynamics of Stock Trading", *Econometrica* 33:1(1965), 88–113.
[45] Pesaran, M. H. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure", *Econometrica* 74:4(2006), 967–1012.
[46] Peterson, M. and Sirri, E. "Evaluation of the Biases in Execution Cost Estimation Using Trade and Quote Data", *Journal of Financial Markets* 6(2003), 259–280.

[47] Rosenthal, D. W. R. "Trade Classification and Nearly-Gamma Random Variables" (Ph.D. dissertation, University of Chicago, 2008).
[48] Searle, Shayle R.; G. Casella and C. E. McCulloch. *Variance Components*, 1992. Wiley: New York.
[49] Securities Training Corporation. *Series 55: The Equity Trader Examination Study Manual*, 6 November 2006. Securities Training Corporation: New York.
[50] Stoll, H. R. "Electronic Trading in Stock Markets", *Journal of Economic Perspectives* 20:1(2006), 153–174.
[51] Stoll, H.R. and Schenzler, C. "Trades Outside the Quotes: Reporting Delay, Trading Option, or Trade Size?", *Journal of Financial Economics* 79(2006), 615–653.
[52] U.S. Securities and Exchange Commission. Final Rule: Disclosure of Order Execution and Routing Practices, 17 November 2000. Retrieved 20 December 2007 from `http://www.sec.gov/rules/final/34-43590.pdf`.
[53] U.S. Securities and Exchange Commission. Regulation NMS, 9 June 2005. Retrieved 20 December 2007 from `http://www.sec.gov/rules/final/34-51808.pdf`.
[54] van Belle, G. *Statistical Rules of Thumb*, 2002. Wiley: New York.
[55] Vergote, O. "How to Match Trades and Quotes for NYSE Stocks?" KU Working Paper. Katholieke Universiteit Leuven (2005).
[56] Wong, W. H. "Theory of Partial Likelihood", *Annals of Statistics*, 14:1(1986), 88–123.
[57] Zellner, A. "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association* 57:298(1962), 348–368.

rosenthal@galton.uchicago.edu, 5734 S. University Ave, Chicago, IL, 60637