Optimal Execution Size in Pro-Rata Markets

IRENE ALDRIDGE

IRENE ALDRIDGE
is a managing partner at ABLE Alpha Trading, LTD. in New York, NY.
ialdridge@ablealpha.com

Financial markets in the U.S. are home to many trading venues with several competing market structures. At the time this research was written, in the U.S. equity markets alone, there were 13 exchanges and multiple dark pools. Due to the competitive nature of the trading landscape, exchanges attempt to differentiate their offerings by deploying different pricing and order matching systems. As discussed by Aldridge [2013a], some equity exchanges, known as "normal," offer traders monetary incentives to provide liquidity in the Demsetz [1968] sense: as immediacy of execution. Such immediacy of execution translates into posting limit orders. Thus, normal exchanges offer "rebates" for providing liquidity (posting limit orders) and charge for taking liquidity (placing market orders). Other exchanges, known as "inverted," do the opposite, charging for limit orders and paying for market orders. The NYSE is an example of a normal exchange, while the Boston OMX is an inverted exchange. Many exchange firms have offerings in each category: BATS, for example, has normal and inverted exchanges.

In addition to competing on price, however, exchanges may also compete on their matching processes. Most U.S. equity exchanges deploy so-called price-time priority schedule, illustrated in Exhibit 1. Under the price-time schedule, the orders in the best-price queue are matched with incoming market orders in the sequence in which they arrived. Thus, the "oldest" limit order found in the top-of-the-book queue will be matched with the incoming opposing market order first. If the size of the market order exceeds the size of the oldest limit order in the top-of-the-book queue, the market order will proceed to match with the next oldest limit order in the top-of-the-book queue, until the market order is filled.[1] In most cases, the price-time matching will grant the first-come, first-served priority to limit order arrivals.

By contrast, under pro-rata, first modeled by Glosten [1994] and most recently by Field and Large [2012] and Aldridge [2013b], all limit orders at the top-of-the-book queue are matched simultaneously with the incoming market order, as shown in Exhibit 2. The proportion of each limit order executed under pro-rata depends on the size of the arriving market order as well as on the cumulative size of limit orders at the top-of-the-book queue: all orders in the top-of-the-book queue will be executed in equal proportion equal to the size of the incoming market divided by the total size of all limit orders in the top-of-the-book queue. In U.S. cash equities, the Philadelphia Stock Exchange was operating pro-rata until May 2013, when it was converted to the price-time priority model. The pro-rata exchanges,

however, remain popular in other U.S. markets, such as options and commodities.

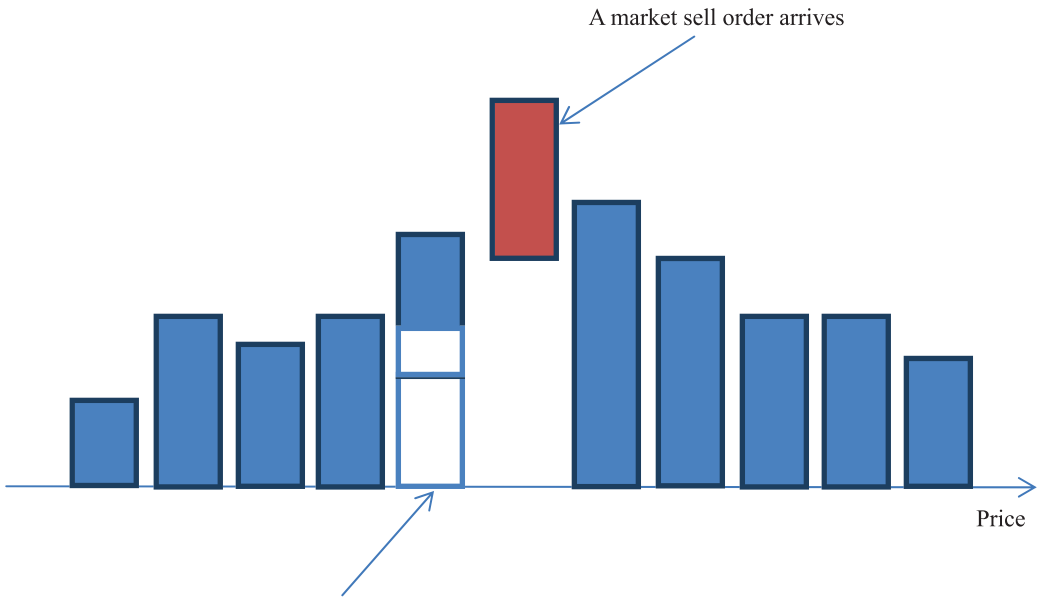
In price-time priority markets, execution or fill of a limit order is not certain and depends on the number and size of limit orders pending execution ahead of the given limit order. For example, a limit order placed close to the market price has a higher probability of execution than does a limit order placed further away from the market price behind many other limit orders. Even in the "top-of-the-book" price queue under the price-time priority, the execution of each limit order further depends on the number and the size of limit orders that arrived prior to the given limit order. The more orders there are in a queue prior to the given order's arrival, the higher the probability of non-execution of that given order.

To combat this uncertainty, many brokers today offer a service whereby they place multiple limit orders in the limit order book, thereby securing the price-time priority for their trading clients. If the broker's orders are about to be executed, but no relevant client flow has come in, the broker simply cancels the excessive limit orders. This widespread practice is known as "layering" and is primarily responsible for excessive order cancellation rates in today's markets. According to Hautsch and Huang [2011], for example, 95% of all limit orders placed on NASDAQ are canceled within just one minute of their placement.

The order cancellation message rate in inverted markets happens to be smaller than in the normal markets due to the built-in incentive structures in either market. In normal markets, limit order traders are rewarded for executing their limit orders. Therefore, a broker who fails to cancel an unwanted limit order fast enough to prevent the execution of the order is granted a rebate. Such a rebate will, at least in part, offset the

EXHIBIT 1

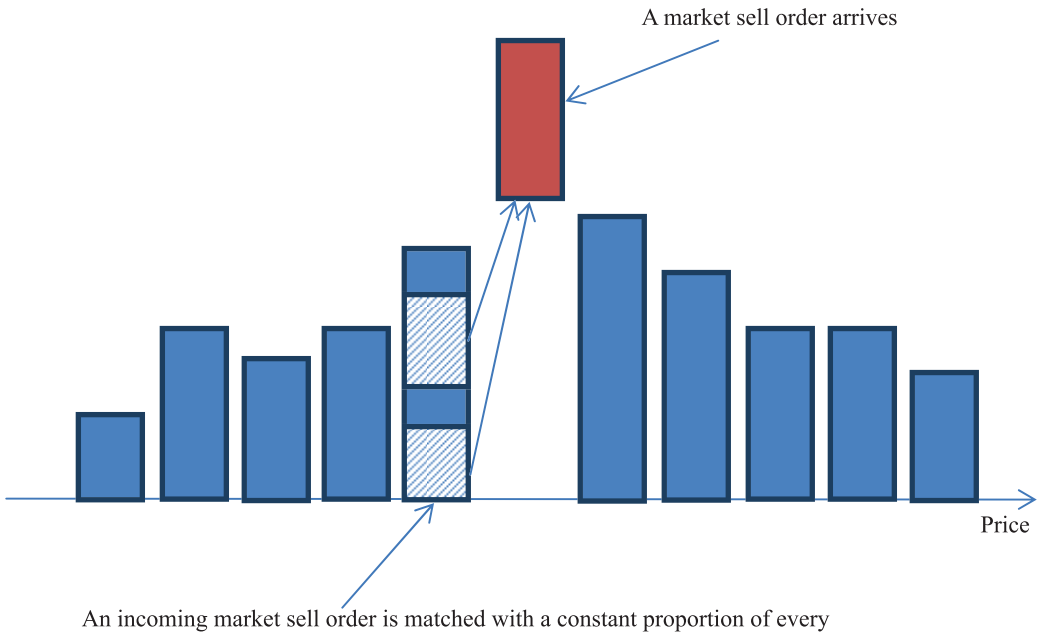
Price-Time Priority Matching, an Illustration



The "oldest" limit buy order placed at the best bid earliest is executed first; if the market sell is not filled completely with the oldest limit order, the market sell is next matched with the second oldest limit buy order placed at the best bid. The process continues until the market sell order is matched in full or the best bid queue is exhausted.

Source: Aldridge (2013a).

EXHIBIT 2
Pro-Rata Matching, an Illustration



Source: Aldridge (2013a).

costs associated with the risk borne by the broker of carrying an unwanted position on his books. In contrast, in inverted markets, most limit order traders are charged for execution of their limit orders, forcing layering brokers to be more careful with limit order placement and cancellation in order to avoid the costs associated with inadvertent order execution.

In comparison, in pro-rata markets, the price-time priority simply does not exist, as all orders in the top-of-the-book queue are executed simultaneously. As a result, the brokers' layering practice is invalid in pro-rata markets by the market-design. Having said that, in pro-rata markets, limit order traders may need to place larger orders to ensure that the fractional fill of their orders adds up to the desired order size. The surplus portion of their orders is subsequently canceled, also resulting in order cancellation traffic. As shown in Aldridge [2013b], however, the order cancellation traffic in pro-rata markets is considerably smaller than that in price-time priority markets, both of the normal and the inverted variety.

As discussed in Aldridge [2013b], the absence of layering brokers in pro-rata markets increases transparency of the markets and allows traders to spot and weed out high-frequency market manipulators much more effectively. As a result, the quality of pro-rata markets can be thought to be superior to that of inverted markets, which, in turn, is higher than the quality of normal markets.

Despite the advantages of pro-rata markets, however, the uncertainty of order size execution is daunting for many pro-rata limit order traders. Based on the findings of Aldridge [2013b], the current paper derives optimal limit order size in pro-rata and shows its performance with various latency assumptions. This article is organized as follows: The first section presents the derivation of the optimal order execution size for orders placed at the top of the book in a pro-rata market, the second section shows empirical tests of the optimal order size formula performance in the pro-rata equities Philadelphia Stock Exchange under various latency conditions, and the final section provides a conclusion.

DERIVATION OF THE OPTIMAL LIMIT ORDER SIZE

In pro-rata markets, the expected trading size of the limit order of size N is proportional to the expected aggregate number of limit order units in the limit order book queue at the time of the order execution:

$$\mathbb{E}[N_{executed}] = N\frac{N}{N+\mathbb{E}[z]} \tag{1}$$

To execute the desired order size N in full, a pro-rata trader needs, therefore, to place a larger-than-warranted order, and then quickly cancel the "surplus" order once the desired execution size has been realized in the market. The expected number of orders in a top-of-the-book queue, $\mathbb{E}[z]$, can be estimated as the expected value of all order sizes:

$$\mathbb{E}[z] = \sum_{z>0} zP(I_\infty = z) \tag{2}$$

where $P(I_\infty = z)$ is the steady-state probability of observing exactly z orders in the top-of-the-book queue. As shown in Aldridge [2013b], this probability $P(I_\infty = z)$ is

$$P(I_\infty = z) = \left(1+\sum_{z=1}^{\infty}\left(\frac{\lambda}{\mu+\delta}\right)^z\right)^{-1}\left(\frac{\lambda}{\mu+\delta}\right)^z \tag{3}$$

where λ is the rate of limit order arrivals, μ is the rate of arrival of opposing market orders, and δ is the rate of limit order cancellations. Practical step-by-step estimation of rates λ, μ and δ is described in the appendix.

The notations imply that:

- A trading unit of limit orders (say, one futures contract or 100 equity shares) in the top-of-the-book queue in a given pro-rata market arrives on average every $1/\lambda$ units of time (seconds, milliseconds, and so on).
- A trading unit of the matching (opposing) market orders arrives every $1/\mu$ units of time.
- A trading unit of the limit orders at the top of the book is canceled every $1/\delta$ time units.

The order arrival and cancellation rates depend on the order flow and may be different for buy and sell orders. Furthermore, since the order flow depends on market conditions, the order arrival and cancellation rates may vary throughout the day.

Solving Equation (1) for the optimal order size N^* to ensure that the executed portion of the order is, in fact, the desired order size, $\mathbb{E}[N_{executed}]$, we obtain the following quadratic equation:

$$(N^*)^2 - \mathbb{E}[N_{executed}]N^* - \mathbb{E}[N_{executed}]\mathbb{E}[z_k] = 0 \tag{4}$$

Solving for N^*, we obtain

$$N^* = \frac{\mathbb{E}[N_{executed}] + \sqrt{\mathbb{E}[N_{executed}]^2 + 4\mathbb{E}[N_{executed}]\mathbb{E}[z_k]}}{2} \tag{5}$$

since $N^* > \mathbb{E}[N_{executed}]$ as the trader will necessarily place a larger order N^* than the required size N. It follows that

$$N^* = \frac{\mathbb{E}[N_{executed}] + \sqrt{\mathbb{E}[N_{executed}]^2 + 4\mathbb{E}[N_{executed}]\sum_{z>0} zP(I_\infty = z)}}{2} \tag{6}$$

By equations (2) and (3),

$$N^* = \frac{\mathbb{E}[N_{executed}] + \sqrt{\mathbb{E}[N_{executed}]^2 + 4\mathbb{E}[N_{executed}]\sum_{z>0} z\left(1+\sum_{z=1}^{\infty}\left(\frac{\lambda}{\mu+\delta}\right)^z\right)^{-1}\left(\frac{\lambda}{\mu+\delta}\right)^z}}{2} \tag{7}$$

Subsequently, the trader on a pro-rata exchange is optimally expected to cancel the unexecuted portion of his order equal to

$$N^* - \mathbb{E}[N_{executed}] = \frac{\sqrt{\mathbb{E}[N_{executed}]^2 + 4\mathbb{E}[N_{executed}]\sum_{z>0} z\left(1+\sum_{z=1}^{\infty}\left(\frac{\lambda}{\mu+\delta}\right)^z\right)^{-1}\left(\frac{\lambda}{\mu+\delta}\right)^z}}{2} - \frac{\mathbb{E}[N_{executed}]}{2} \tag{8}$$

While the actual order cancellation of the surplus may cost nothing, the pro-rata trader's expected cost is strictly positive, as he always faces the positive probability of execution of his surplus order in the top-of-the-book queue: an opposing market order may arrive between the time the pro-rata limit order trader receives acknowledgement of execution of his order and the time his subsequent excess-cancellation order reaches the exchange. As such, pro-rata microstructure naturally discourages traders from placing limit orders that are larger than prescribed by Equation (8) for fear of bearing the cost for limit orders not canceled in time.

CONCLUSION

Different market microstructures of equity exchanges lead to ex ante divergent properties of the markets. As shown in Aldridge [2013b], pro-rata matching leads to lower volatility, lower probability of crashes, and higher liquidity than price-time priority matching. In addition, pro-rata markets are least susceptible to spoofing, while normal exchanges are most susceptible to such market manipulation. Untill now, however, traders and brokers have not had access to a reliable methodology for optimal sizing of limit orders in pro-rata markets. This research fills the void.

APPENDIX

PRACTICAL ESTIMATION OF ORDER ARRIVAL AND CANCELLATION RATES

A trader desiring to execute N contracts would begin by estimating the arrival rates of limit orders, market orders, and limit order cancellations in the top-of-the-book queue of interest. A trader seeking to execute N limit buy orders will count the number of limit buy orders, λ, the number of limit buy orders canceled, δ, and the number of market sell orders, μ, that have arrived over a certain period of time, say five minutes, immediately preceding the limit order placement. A trader looking to trade N limit sell orders will count the number of limit sell orders, λ, the number of canceled limit sell orders, δ, and the number of market buy orders, μ, over the past five minutes. The estimated variables λ, δ, and μ are ready to be used in the Equation (8).

Specifically, the required parameters may be estimated using the prevalent Level I data comprised of:

- timestamp in microseconds
- best bid price and size
- best offer price and size
- last trade price and size

To estimate the parameters, all data need to be separated into "bars" corresponding to finite time periods of 300 seconds (five minutes). Within each time bar, the number of limit order arrivals, market order arrivals, and limit order cancellations is computed following the "volume clock" first described by Easley, Lopez de Prado, and O'Hara [2010]. Under the volume clock process, each contract or trading unit is considered to be an independent arrival. As a result, an exchange order of 100 shares is considered to be an arrival of 100 instantaneously sequential orders in our model. The orders are further classified into limit, limit cancel, or market according to the following rules:

I. When the last recorded tick of data was a trade:
   a. If a trade is recorded at the most recent best bid price or a lower price, the trade was considered to be initiated by a market sell order. The market sell order arrival counter in the model was then incremented by the number of shares recorded in the trade.
   b. If a trade is recorded at the most recent best ask price or a higher price, the trade was considered to be initiated by a market buy order. The market buy order arrival counter in the model was then incremented by the number of shares recorded in the trade.
   c. If a trade is recorded at a price different than a prevailing bid or a prevailing ask, the trade was classified according to the Lee-Ready rule (Lee and Ready [1991]):
      i. A trade recorded below the midpoint of prevailing bid and ask prices was assumed to be initiated by a market sell order.
      ii. A trade recorded above the midpoint of prevailing bid and asks was assumed to be initiated by a market buy order.
      iii. A trade recorded exactly at the midpoint of the bid and ask prices prevailing at the time was accounted for by the tick rule: If the price movement accompanying the trade was positive, the trade was classified as market-buy-initiated, while the trades with negative price move were considered to be market-sell-initiated. If the price did not move on the latest trade, previous price directions were considered in a similar manner.

II. When the last recorded tick of data was a quote:

a. If the reported best bid price was the same as the previous best bid price, but the best bid size increased, a new limit buy order was considered to be received. The limit order was recorded as an increment in the limit buy order arrival counter, with the increase equal to the difference in size between the new best bid size and the previously recorded best bid size.

b. If the reported best ask price was the same as the previous best ask price, but the best ask size increased, a new limit sell order was considered to be received. The sell limit order was recorded as an increment in the limit sell order arrival counter, with the increase equal to the difference in size between the new best ask size and the previously recorded best ask size.

c. If the reported best bid was above the best bid preceding the latest quote, a new limit buy order of size of the new best bid was considered to have been received. The limit buy arrival counter was incremented by the size of the new best bid.

d. If the reported best ask was below the best ask preceding the latest quote, a new limit sell order of size of the new best ask was considered to have been received. The limit sell arrival counter was incremented by the size of the new best ask.

e. If the reported best bid price was the same as the previous best bid price, but the best bid size decreased, and no trade at the best bid preceded the latest quote, a new limit buy order cancellation was considered to be received. The canceled limit order was recorded as an increment in the limit buy order cancellation counter, with the increase equal to the difference in size between the previous best bid size and the new best bid size.

f. If the reported best ask price was the same as the previous best ask price, but the best ask size decreased, and no trade at the best ask was recorded between the previous and current asks, a new limit sell order cancellation was considered to be received. The sell limit order cancellation was recorded as an increment in the limit sell order cancellation counter, with the increase equal to the difference in size between the previous best ask size and the new ask size.

g. If the reported best bid was below the best bid preceding the latest quote, and no trade at the best bid preceded the latest quote, a new limit buy order cancellation of the size of the previous best bid was considered to have been received. The limit buy cancellation counter was incremented by the size of the previous best bid.

h. If the reported best ask was above the best ask preceding the latest quote, and no trade at the best ask preceded the latest quote, a new limit sell order cancellation of size of the previous best ask was considered to have been received. The limit sell cancellation counter was incremented by the size of the previous best ask.

Exhibit 3 shows an example of tick data records classified as limit buy, limit sell, market buy, market sell, buy cancel, or sell cancel orders. Exhibit 4 and 5 illustrate the logic diagrammatically.

EXHIBIT 3

A Stylized Sample of Historical Reuters Tick Data, Classified into Market Buys, Market Sells, Limit Buys, Limit Sells, Buy Cancellations, and Sell Cancellations

Print Type	BB	Size	BO	Size	Trade	Size	Print Classification
Quote	1.08	50	1.10	100			
Trade					1.08	30	Market sell 30
Quote	1.08	20	1.10	100			
Trade					1.08	20	Market sell 20
Trade					1.07	100	Market sell 100
Quote	1.07	50	1.09	80			Limit sell 80
Quote	1.08	100	1.09	80			Limit buy 50
Quote	1.08	200	1.09	80			Limit buy 100
Quote	1.08	200	1.09	50			Cancel sell 30
Trade					1.09	50	Market buy 50

EXHIBIT 4

Diagram of the Logic for Determining Whether a Trade Print Was Initiated by a Market Buy or a Market Sell

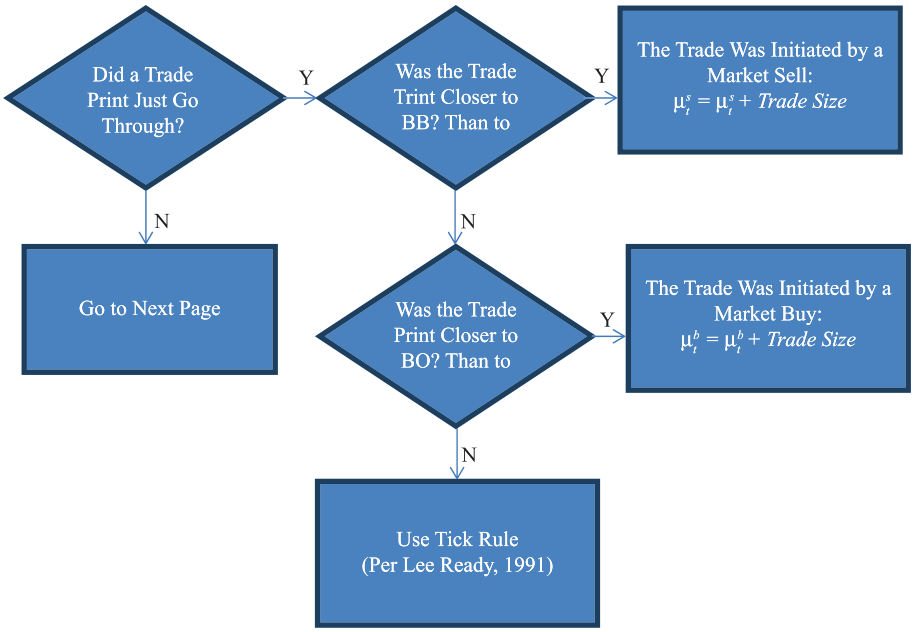
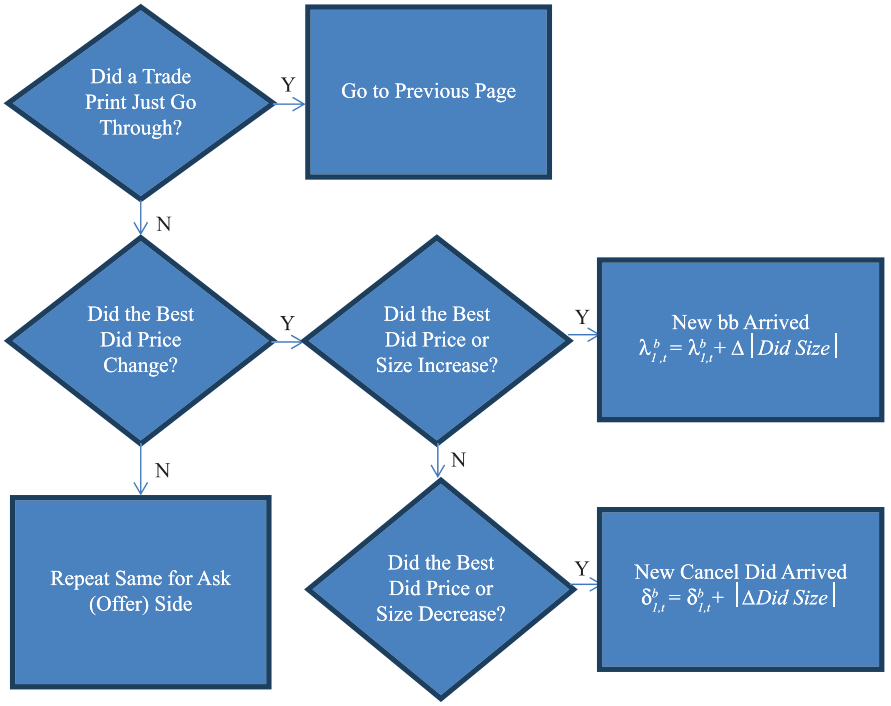


EXHIBIT 5

Diagram of the Logic for Determining Whether a Quote Revision was Driven by an Arrival of a New Limit Order or a Limit Order Cancellation



ENDNOTE

[1]Certain restrictions may apply on "sweeping the book"—in some cases, a large market order is allowed to sweep only through a certain number of queues or ticks of the limit order book, after which point the unfilled portion of the order is automatically converted into a limit order.

REFERENCES

Aldridge, I.E. *High-Frequency Trading: A Practical Guide to Algorithmic Strategies and Trading Systems*, 2nd ed. Hoboken, NJ: Wiley, 2013.

——. "Market Microstructure and the Risks of High-Frequency Trading." Working paper, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294526.

Cont, R., S. Stoikov, and R. Talreja. "A Stochastic Model for Order Book Dynamics." Working paper, http://ssrn.com/abstract=1273160 or http://dx.doi.org/10.2139/ssrn.1273160, 2008.

Demsetz, H. "The Cost of Transacting." *The Quarterly Journal of Economics*, Vol. 82, No. 1 (1968), pp. 33-53.

Easley, D., M. Lopez de Prado, and M. O'Hara. "Flow Toxicity and Liquidity in a High-Frequency World." *Review of Financial Studies*, Vol. 25, No. 5 (2012), pp. 1414-1493.

Field, J., and J. Large. "Pro-Rata Matching in One-Tick Markets." Working paper, Cass Business School, London, 2012.

Glosten, L.R. "Is the Electronic Limit Order Book Inevitable?" *Journal of Finance*, 49 (1994), pp. 1127-1161.

Hautsch, N., and R. Huang. "The Market Impact of a Limit Order." *Journal of Economic Dynamics and Control*, http://ssrn.com/abstract=1677343 or http://dx.doi.org/10.2139/ssrn.1677343, 2011.

Kitaev, M., and S. Yashkov. "Distribution of the Conditional Sojourn Time in a System with Division of Time of Servicing." *Engineering Cybernetics*, 16 (1978), pp. 162-167.

Kleinrock, L. "Analysis of a Time-Shared Processor." *Naval Research Logistics Quarterly*, 11 (1964), pp. 59-73.

Lee, C.M.C., and M.J. Ready. "Inferring Trade Direction from Intraday Data." *Journal of Finance*, Vol. 46, No. 2 (1991), pp. 733-746.

Yashkov, S. "A Derivation of Response Time Distribution for an M/G/1 Processor-Sharing Queue." *Problems of Control and Information Theory*, 12 (1983), pp. 133-148.