Recommendations for Equitable Allocation of Trades in High Frequency Trading Environments[1]

John McPartland

July 25, 2013

Executive Summary

Most industry observers and much of the academic research in this area have concluded that High Frequency Trading (HFT) is generally beneficial. Many institutional investors, however, believe that HFT places them at a competitive disadvantage. Digital computers will always have some structural (speed) advantages over human traders. This is inevitable.

This paper (1) acknowledges and summarizes much of the relevant published research[2] (2) discusses some of the HFT strategies that likely run counter to good public policy and (3) makes six recommendations that, if implemented, would not preclude any current HFT strategies, but would likely restore some competitive advantage to market participants that would be willing to expose their resting orders to market risk for more than fleeting milliseconds.

Readers should avoid the tendency to review the Working Paper only within the framework of their own nationality and market domain. The paper is meant to be global in scope. Some HFT practices that may be inappropriate (or banned) in some markets in some countries are alive and well in other markets in other countries.

[1] The author wishes to acknowledge the very significant contributions that David Marshall and Rajeev Ranjan made to this paper. He also wishes to thank the many industry professionals who helped review the many versions of the document and its recommendations prior to publication. Any opinions expressed in this paper are those of the author, and those opinions do not necessarily reflect the opinions of the Federal Reserve Bank of Chicago or the opinions of the Board of Governors of the Federal Reserve System.

[2] See, for example, Anton Golub, 2011, "Overview of high frequency trading," Manchester Business School, April 15, and Investment Industry Regulatory Organization of Canada, 2012, "The HOT Study: Phases I and II of IIROC's study of high frequency trading activity on Canadian equity marketplaces," report, Toronto, December 12.

An exceptionally abbreviated summary of the six recommendations follows.

1. Where appropriate, utilize a new trade allocation formula that is intermediate between the Pro Rata trade allocation formula and the Price/Time or FIFO trade allocation formula.
2. Create a new, optional, term limit order type that, as part of the trade allocation process, would reward traders for the time that their orders are committed to be resting in the Order Book.
3. Completely dark orders or the hidden portion of resting orders that are not fully displayed (lit) in the Order Book should go to the very end of the queue (within limit price) with respect to trade allocation.
4. Prior to trade allocation, multiple small orders from the same legal entity entered contemporaneously for the sole purpose of exploiting the rounding conventions of a trading venue should firstly be aggregated as a single order and should carry the lowest allocation priority time stamp of all of the orders so aggregated.
5. Rather than running a continuous trade match, trading venues should divide their trading sessions into periods of one half second. At a completely random time within each half second period, the trade match and trade allocation algorithms should be run once.
6. Visibility into the Order Book should be no more granular than aggregate size at each limit price. Market participants should neither be able to view the size of individual orders nor any other identifiers of any orders of others. This more granular information is not information that any market participant needs to make a fully informed economic decision as to the instantaneous value of the instrument being traded.

Background

Most industry observers seem to believe that HFT offers many benefits to organized financial markets and to society, including improved liquidity, tightened bid/offer spreads and a decrease in intraday price volatility. This Working Paper describes some of the HFT techniques that have developed in electronic markets around the world.

Different financial centers have different rules and regulations regarding the appropriateness of some HFT techniques. This Working Paper is intended to be global in its scope and in its recommendations. All of its six recommendations might not be appropriate for every electronic exchange in every financial center. Throughout the Working Paper, when discussing different trade allocation methodologies, we refer to “shares”, “futures” or “lots”; these terms are completely interchangeable.

Review of the Academic Literature[3]

Brogaard, Hendershott and Riordan (2012) analyzed NASDAQ and NYSE high frequency trading data[4] that show high frequency traders increase price efficiency by trading in the same direction of permanent price changes and trading in the opposite direction of transitory pricing errors on normal trading days and on days with the highest price volatility. In contrast, HFT liquidity-supplying non-marketable orders are adversely selected in terms of the permanent and transitory components as these trades are in the direction opposite to permanent price changes and in the same direction as transitory pricing errors. HFT predicts price changes in the overall market over short horizons measured in seconds. HFT is correlated with public information, such as macro news announcements, market-wide price movements, and limit order book imbalances.[5]

Jones (2013) notes that the volume of HFT has increased sharply over the past several years, has reduced trading costs and has steadily improved liquidity. The main positive is that HFT can intermediate trades at lower cost. However, HFT speed could disadvantage other investors, and the resulting adverse selection could reduce market quality. Ideally research in this area should attempt to determine the incremental effect of HFT beyond other structural and technological changes in equity markets. The best papers for this purpose attempt to isolate market structure changes that facilitate HFT. Virtually every time a market structure change results in more HFT, liquidity and market quality have improved because liquidity suppliers are better able to adjust their

[3] See Investment Industry Regulatory Organization of Canada (2012, appendix A, pp. 51-56).

[4] The HFT data represent a sample of 120 randomly selected stocks listed on NASDAQ and NYSE for all of 2008 and 2009. Trades are time-stamped to the millisecond and identify the liquidity demander and supplier as a high frequency trader or non-high frequency trader.

[5] Jonathan Brogaard, Terrence Hendershott and Ryan Riordan, 2013, “High frequency trading and price discovery,” University of Washington, University of California, Berkeley and University of Ontario Institute of Technology, working paper, April 22.

quotes in response to new information. Jones cites the concern that HFT may not help to stabilize prices during unusually volatile periods and notes that there is a potential for an unproductive arms race among HFT firms for speed.[6]

Cartea and Panalva (2012) conclude that the presence of high frequency traders increases the price impact of liquidity trades and that this price impact increases as the size of the trades increase. High frequency traders increase microstructure noise of prices and increase trading volume. High frequency traders and non-high frequency professional traders coexist as competition drives down profits for new HFT entrants while the presence of HFTs does not drive out traditional professional traders. Finally, the paper concludes that high frequency traders clearly generate costs, but they also generate benefits, and that the net effect requires more precise empirical analysis.[7]

The Litzenberger, Castura and Gorelick (2010) paper concludes that overall market quality has improved significantly, including bid-ask spreads, liquidity, and transitory price impacts (measured by short term variance ratios). Studies using proprietary exchange provided data that identify the trades of high frequency trading firms show that HFT firms contributed directly to: narrowing bid-ask spreads, increasing liquidity, and reducing intra-day transitory pricing errors and intra-day volatility.[8]

Questionable HFT Techniques

Notwithstanding the evident benefits of HFT in electronic markets, many market participants believe that some HFT practitioners utilize trading techniques that are detrimental to the well-functioning of financial markets.[9]

[6] Charles M. Jones, 2013, "What do we know about high-frequency trading?," Columbia Business School, research paper, No. 13-11, March 20.

[7] Álvaro Cartea and José Penalva, 2012, "Where is the value in high frequency trading?," University College London and Universidad Carlos III, Madrid, working paper, February 17.

[8] Robert Litzenberger, Jeff Castura, Richard Gorelick and Yogesh Dwivedi, 2010, "Market efficiency and microstructure evolution in U.S. equity markets: A high-frequency perspective," RGM Advisors LLC, working paper, August 30.

[9] See, for example, German Federal Ministry of Finance, "Speed limit for high-frequency trading–federal government adopts legislation to avoid risks and prevent abuse in high-frequency trading," press release, Berlin, September 26, available at www.bundesfinanzministerium.de/Content/EN/Pressemitteilungen/2012/2012-09-26-speed-limit-for-high-frequency-trading.html, and Australian Securities & Investments Commission (ASIC), 2012, "Australian market structure: Draft market integrity rules and guidance on automated trading," consultation paper, No. 84, Victoria, Australia, available at www.asic.gov.au/asic/pdflib.nsf/LookupByFileName/cp184-published-13-August-2012.pdf/$file/cp184-published-13-August-2012.pdf.

Some of the trading techniques generally considered to be detrimental and not capital formative are *spoofing, layering* and *quote stuffing.*[10]

Spoofing and *layering* are not at all unique to HFT. Both almost always involve feigning to be a buyer when you are really a seller or vice versa. Algorithmic HFT has, however, allowed these two strategies to be taken to new levels.

> "Generally, spoofing is a form of market manipulation which involves placing certain non-*bona fide* order(s), usually inside the existing National Best Bid or Offer (NBBO), with the intention of triggering another market participant(s) to join or improve the NBBO, followed by canceling the non-*bona fide* order, and entering an order on the opposite side of the market."[11]
>
> "Layering involves the placement of multiple, non-*bona fide*, limit orders on one side of the market at various price levels at or away from the NBBO to create the appearance of a change in the levels of supply and demand, thereby artificially moving the price of the security. An order is then executed on the opposite side of the market at the artificially created price, and the non-*bona fide* orders are immediately canceled."[12]
>
> "Quote stuffing is a practice in which a large number of orders to buy or sell securities are placed and then canceled almost immediately. During periods of intense quoting activity stocks experience decreased liquidity, higher trading costs, and increased short term volatility."[13]

Imagine that you are bidding at an art auction, and the serious bidders are now reduced to two or three. One of the persons pretending to be an interested bidder is really the owner of the art piece currently being auctioned off. It is to their advantage to get the bona fide bidders to pay as much as possible for their art piece. Bidders indicate their willingness to bid to the auctioneer by raising the bidder numbers assigned to them by the auction house. The *spoofing* equivalent in this physical environment would be if the seller of the art piece, pretending to be a buyer, raised his or her bidder number one last time, solely to get the last remaining buyer to pay more than they otherwise would be willing to pay. Granted, the *spoofer* in this case is absolutely at risk of buying their own art piece unless their *spoofing* strategy is successful, and a bona fide

[10] See the proposed amendments to the ASIC Market Integrity Rules (ASX Market) to preclude market misconduct, manipulation or false trading in Australian Securities & Investments Commission, 2013, "Dark liquidity and high-frequency trading," report, No. 331, Victoria, Australia, March, p. 10.

[11] Financial Industry Regulatory Authority (FINRA), 2012, "FINRA joins exchanges and the SEC in fining Hold Brothers more than $5.9 million for manipulative trading, anti-money laundering, and other violations," press release, Washington, DC, available at www.finra.org/Newsroom/NewsReleases/2012/P178687.

[12] Ibid.

[13] Jared Egginton, Bonnie Van Ness and Robert Van Ness, 2012, "Quote stuffing," Louisiana Tech University and University of Mississippi, March 15.

bidder betters the *spoofer's* bid. While the practice of allowing sellers to masquerade as buyers is probably not allowed at proper art auctions, its electronic equivalent is permitted and well-practiced among some HFT practitioners. Make no mistake, HFT *spoofers'* bids and offers are exposed to market risk as much as the bids and offers of click traders, even if they are often so exposed for only milliseconds. Spoofing is intentionally designed to be deceptive and, at a minimum, frustrates fair value investors' ability to determine the true market value of the instruments that are being traded.

Layering is only a slightly different technique, designed to similarly deceive market participants' perception of the aggregate size of the bids and offers in the Order Book. By entering thousands of bids or offers, and then cancelling them virtually immediately, but only after they have been acknowledged as having been present in the Order Book, HFT practitioners can successfully create the illusion of greater size at the bid (or offer) than is realistically executable. Investment managers often refer to this phenomenon as "phantom liquidity" as the visible liquidity is often not there when one goes to hit the bid or lift the offer. Not unlike the massive white clouds in the sky, they are actually nothing more than thin water vapor that simply gives the deceptive illusion of being huge, massive objects. Frequently, high frequency traders *layer* quotes on the bid side of the market, in an attempt to attract other bidders and then hit the bid side of the market as a seller in size. *Layering* is designed to be intentionally deceptive and similarly frustrates institutional investors' ability to ascertain the fair market value of the instruments traded. It also intentionally and unduly complicates order execution.

Quote stuffing is roughly equivalent to driving a race car at 190 miles per hour, but preventing the other drivers from exceeding 160 miles per hour. By clogging a trading venue's outbound quotation system (or inbound order entry systems) with near worthless quotes, an astute HFT practitioner can execute a trade on another, or on the same trading venue with some degree of confidence that at least a plurality of market participants (including many other high frequency traders) will, at best, be reacting to delayed quotes, creating an arguably unfair trading advantage for HFT practitioners that can slow the other traders down by increasing their reaction times.[14] "The ultimate goal of many of these programs is to gum up the system so it slows down the quote feed to others and allows the computer traders (with their co-located servers at the exchanges) to gain a money-making arbitrage opportunity"[15] Price transparency is considered a public benefit of organized financial markets. It is

[14] Quote stuffing is an offensive tool that high speed traders most typically use to gain a competitive advantage over other high speed traders. Click traders would not likely be adversely affected if outbound quotations were intentionally delayed by, say, 200 milliseconds, nor would they likely even be able to detect any such delay.

[15] John Melloy, 2012, "Mysterious algorithm was 4% of trading activity last week," CNBC, October 8, available at www.cnbc.com/id/49333454.

difficult to envision that the practice of intentionally slowing down the dissemination of trade prices to the public is an activity that serves the public interest.

The thesis of this paper is that, rather than attempting to ban these techniques (which could likely be difficult to enforce in practice), one could change the character and economics of the trading environment so as to disincent these and similar undesirable trading techniques. Rather than propose solutions that preclude specific HFT strategies, we propose to simply change the economics of the trade match and trade allocation processes, to strike a more equitable balance between the high frequency trading community and the investment management community.

Recommendations

The proposal contains no "thou shalt not" recommendations.

The proposal consists of six recommendations that should be deemed as one complete set that should be considered and implemented as a whole, where appropriate. Several of the recommendations are admittedly rather complex; so also are the current electronic market structures in which we find ourselves. Our recommendations follow.

1. Trade Allocation with Cardinal Weighting of Time in the Order Book

The ideal trade allocation algorithm should be a combination of the Pro Rata trade allocation algorithm and the Price/Time or FIFO trade allocation algorithm.[16] Descriptions of the Price/Time and Pro Rata algorithms would seem to be in order.

Price/Time or FIFO

The Price/Time trade allocation algorithm is also known as the FIFO (First In, First Out) algorithm. The Price/Time trade allocation algorithm first prioritizes all bids and orders based upon price, and within price, prioritizes orders (in an ordinal ranking) based upon the time that each order was received. An order

[16] There are at least half a dozen other trade allocation algorithms currently in use but not specifically referenced in this section. While it might be quite valuable for interested market participants to have a detailed treatise on the various trade allocation algorithms currently in use, that is not the objective of this section.

can always gain priority by bettering its price, while keeping its original time stamp. Within the best price, the Price/Time algorithm attempts to completely fill the order with the oldest time stamp, (the lowest ordinal ranking) with any residual contracts or shares subsequently allocated to the next oldest bid or offer, until the appropriate contracts or shares have been fully allocated.

The Price/Time trade allocation algorithm was the first algorithm utilized when the era of electronic trading was ushered in. Some market participants erroneously believe that electronic markets still utilize the Price/Time trade allocation exclusively; that there *are* no other trade allocation algorithms. While it is equitable, some trading venues have diversified away from the Price/Time trade allocation as market participants tend to feel disconnected when they join the bid or offer, but are not senior enough to participate in any trade allocation. If there is a valid criticism of the Price/Time trade allocation algorithm, it is that it allocates trades based only on a simple ordinal ranking of bids or offers based upon their respective time stamps. Basing the allocation of trades on a cardinal weighting (ranking) of trades based upon their actual time stamps would seem to be a superior approach.

Pro Rata

In the Pro Rata trade allocation algorithm, all bids are allocated their pro rata share of the allocation of a matched trade based solely upon the lot size of their respective resting bid relative to the aggregate sum of all of the resting bids at the same price. For example, if there are a total of 2200 lots bid for at 12 and 220 offers hit the bid, each resting bid would be allocated 10% of the lot size for which they were bidding.[17]

If there is a criticism of the Pro Rata trade allocation logic, it is that many market participants are constantly bidding or offering unrealistically large quantities, often far greater than they could likely realistically absorb.

NYSE/Liffe

NYSE/Liffe has a hybrid trade allocation algorithm that assigns resting bids and offers with an ordinal ranking (based on their time stamp) and then

[17] This trade allocation process has been shown to be prone to the apparently unavoidable rounding error chicanery that occurs when dozens and dozens of one lot orders are intentionally entered by a single market participant, in the hope of being unjustly enriched by the trading system rounding of what would have been anything greater than 51/100ths of a futures contract or share to one full contract or share. See recommendation 4.

allocates trades based upon a combination of the Pro Rata approach and the ordinal ranking of the bids and offers.

In the formula below, the first bracketed expression simply says that a market participant should be allocated the lesser of (1) the full amount of the quantity of their order or (2) a lesser quantity based upon where their respective order ranks in the Order Book, based upon its time stamp, relative to the time stamps of the other orders in the Order Book. The second bracketed expression determines the pro rata quantity of any given order relative to the aggregate quantity of orders at the same price. The third bracketed expression determines the ordinal ranking (by time stamp) of any given order relative to the time stamps of all of the other orders at the same price, in the Order Book. It is this third bracketed expression that we believe could be improved.

NYSE/Liffe Time Pro-Rata algorithm[18]

$$A_n = Min\left(v_n, \frac{f_n}{\sum_{r=1}^{N} f_r} * L \right) \quad \text{where,} \quad f_n = \left[\frac{v_n}{\sum_{r=1}^{N} v_r} \right] * \left[\frac{(N+1)-n}{\sum_{r=1}^{N} r} \right] \qquad (1)$$

N - Total number of resting buy (sell) orders sorted by time, n = 1(oldest) to N (newest)
n - Individual order being considered
r - Ascending sequence, 1 to N
A_n - Allocation for resting buy (sell) order, n
v_n - Volume of resting order being considered, n
f_n - 'Time Pro Rata Factor' calculated for resting buy (sell) order being considered, n
L - Incoming sell (buy) order volume

Recommended Trade Allocation Algorithm

When allocating trades, the instant proposal places a greater weighting on the time that an order is exposed to market risk. We extrapolate from the

[18] Fractional allocations are rounded down to the nearest integer for all allocations greater than 1 and rounded up to 1 for all fractional allocations less than 1. For equally sized fractional allocations, priority is granted to the oldest order. If any volume remains unallocated following this sequence (for instance, as a result of rounding or when the calculated allocation for an order is constrained by the Min function above), then a further pass of the sequence will occur.

NYSE/Liffe model and assign a cardinal, rather than an ordinal ranking[19] to resting bids and offers based upon the actual length of time that bids and offers have been resting in the Order Book, relative to the time that all of the other orders have been resting in the Order Book. This is accomplished by raising "Time in the Order Book" (Tau) to a low but effective exponential power (α).

$$A_n = Min\left(v_n, \frac{f'_n}{\sum_{r=1}^{N} f'_r} * L \right) \quad \text{where, } f'_n = \left[\frac{v_n}{\sum_{r=1}^{N} v_r} \right] * \left[\frac{\tau_n^{\alpha}}{\sum_{r=1}^{N} \tau_r^{\alpha}} \right] \qquad (2)$$

τ - Time duration (in milliseconds) for every resting order (time difference between the time of trade match and an incoming order's time stamp)

N - Total number of resting buy (sell) orders sorted by time, n = 1(oldest) to N (newest)

n - Individual order being considered

α - A constant parameter set by the trading venue

A_n - Allocation for resting buy (sell) order, n

v_n - Volume of resting order being considered, n

f'_n - 'Proposed Time Pro Rata Factor' calculated for resting buy (sell) order being considered, n

L - Incoming sell (buy) order volume

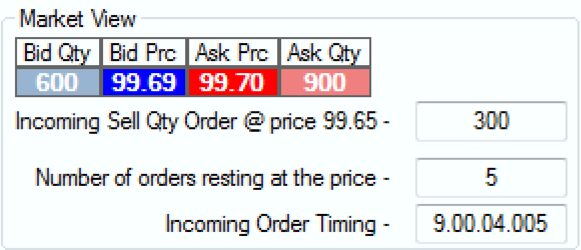
Note that equation (2) is identical to equation (1) except for the third bracketed expression. Whereas the third bracketed expression in (1) is based on a simple ordinal ranking of time in the Order Book, the third bracketed expression in (2) raises time in the Order Book (τ) to an exponential power (α). Increasing α causes a non-linear marginal increase in the number of lots a longer-duration order is allocated, compared to a shorter-duration order, based on the time

[19] Ordinal ranking of resting orders involves creating a simple ranking, not unlike athletes who finish first, second or third in an athletic contest. That is, no consideration is given to the *difference* in athletic performance between the first and second finisher and the second and third finisher. A cardinal ranking would involve assigning a numeric value to the performance of the athletes, not unlike in gymnastics, where there is a quantitative evaluation of individual performances. In the instant proposal, our recommendation is to allocate matched trades based upon the actual time that the orders have been resting in the Order Book relative to the times that other orders have been resting in the Order Book–and not based on the ordinal ranking of the respective time stamps of resting orders. Thus, in a cardinal ranking structure, an order that has been resting in the Order Book for four hours would be entitled to a far greater allocation of trades than an order that has been resting in the Order Book for only four seconds. In an ordinal ranking system, those orders would simply be ranked as #1 and #2 in priority.

that the order has been exposed to market risk. This additional weighting (resulting in a greater allocation of trades) could be set by the trading venue by increasing the parameter α.

Figure 1 illustrates the progressive trade allocation results when α is set equal to or greater than zero and less than or equal to 2.3. As α is set to increasingly higher values, the weight associated with Tau (time in the Order Book) increases exponentially and the actual time that an order has been resting in the Order Book becomes an increasingly dominant component when the algorithm allocates trades. It may be helpful to think of this recommendation as the introduction of perfect gradient shades of gray that lay between black (Pro Rata) and white (Price/Time).

Market View – Top of Book



Market View

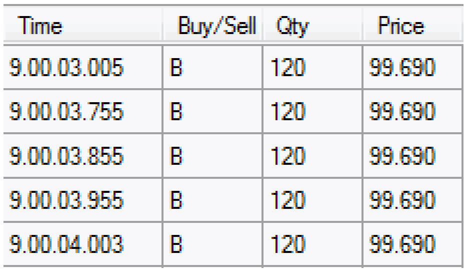
Bid Qty	Bid Prc	Ask Prc	Ask Qty
600	99.69	99.70	900

Incoming Sell Qty Order @ price 99.65 - 300

Number of orders resting at the price - 5

Incoming Order Timing - 9.00.04.005

Limit Orders – Listing all Bids at price of 99.69



Time	Buy/Sell	Qty	Price
9.00.03.005	B	120	99.690
9.00.03.755	B	120	99.690
9.00.03.855	B	120	99.690
9.00.03.955	B	120	99.690
9.00.04.003	B	120	99.690

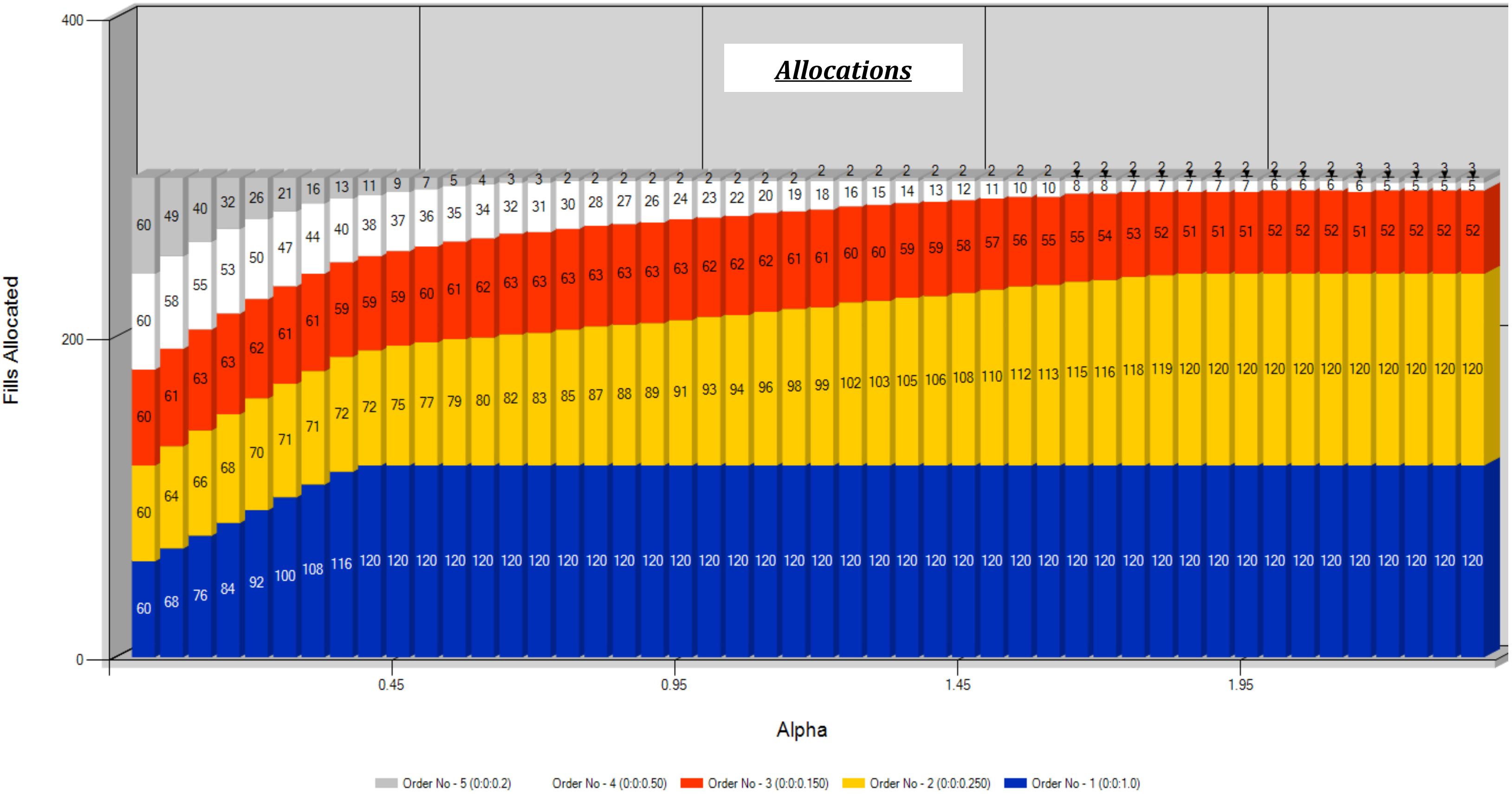


Figure – 1 (α: 0 – 2.3)

The horizontal axis reflects the value of α. The vertical axis reflects the quantities (lots) that would be allocated to resting orders based upon their respective time in the Order Book. When α is set to zero, each of the five resting bids would be allocated 60 lots, i.e., an exact Pro Rata trade allocation (where time in the Order Book means nothing). As α is set to increasing levels, time in the Order Book receives more and more weighting. If α were set to a value higher than 1.9, the recommended trade allocation algorithm closely approximates the Price/Time or FIFO trade allocation algorithm (where time in the Order Book means everything). This continuum approach would allow trading venues to select allocation outcomes of varying degrees between the Pro Rata and the Price/Time trade allocation outcomes.[20]

In the example above, there are 900 lots offered at 99.70 and 600 lots bid for at 99.69. The 600 lots on the bid side are comprised of five individual bids, each for 120 lots, all at a price of 99.69 but with different time stamps. There is a new incoming order to sell 300 lots at a price of 99.65, well below the resting bids at 99.69. The incoming order of 300 lots is therefore going to take out half of the 600 lots bid for at 99.69. The graphic demonstrates how 300 lots would be allocated to the five market participants bidding at 99.69 if α were set to various values between zero and 2.3.

Note that in this example, the oldest bid (in blue) resting in the Order Book is 750 milliseconds older than the second oldest resting bid (orange) but the remaining bids are separated by only about 10 milliseconds. The effect of this time differential can be dramatic, depending on the value of α that is selected by the trading venue.

If α were set to 0.40, the oldest resting bid would be allocated 100% of its bid quantity (120). This is largely due to the 750 millisecond time differential between the oldest bid and the second oldest bid. When α is set to 0.40, the remaining resting bids would be allocated 72, 59, 38 and 11 lots, respectively.

It is anticipated that, for every relevant instrument, a trading venue would select a fairly permanent value of α that strikes an equitable balance that rewards both liquidity providers and institutional market participants. When making such a determination, trading venues will undoubtedly consider the

[20] It is assumed that trading venues would change the value of α very infrequently, if at all, as market participants would need to have a clear understanding and expectation of the trade allocation process in order to correctly size the quantities of their bids and offers. Setting α to a value between zero and 2.3 would allow trading venues a continuum from which they could select the optimal trade allocation result for any financial instrument or product family.

current preferences of market participants and the business risks and costs of changing trade allocation algorithms for legacy products.

2. Term Limit Order Type

Create a new, entirely optional non-cancellable term limit order type, e.g. Buy at 12, good for 4 seconds or Buy at 12, good for 4/10ths of a second. The order may not be cancelled during its stated term (see footnote 21) and would be displayed in the Order Book just as any other resting bid at 12. The term of the order is the *minimum* amount of time that the order would be exposed to the market. Like any other non-term limit orders, a term limit order remains open until it is either filled (either during or after its stated term) or cancelled after its term has expired.[21] Importantly, the instant a term limit order enters the Order Book, it has the trade match and trade allocation priorities of having already been in the Order Book for the stated term of the order. For example, the instant that an order to Buy at 12, good for 4 seconds enters the Order Book, it would have the trade match priority and trade allocation priority identical to an order that has already been resting in the Order Book for 4 seconds.

This order type should provide a more equitable balance between the interests of institutional market participants and HFT practitioners, whose orders are often in the Order Book for only a few milliseconds. When combined with recommendation 1, allocation of matched trades should be directly related to how long a resting order was exposed (or committed to be exposed) to market risk. Orders that are resting (or committed to be) in the Order Book for a material amount of time are exposed to market risk, provide tangible price transparency to the public, and deserve a higher trade match priority and trade allocation priority than orders that have barely been in the Order Book for a few milliseconds and have provided only marginal (if not intentionally deceptive) pricing information to the public. The combined effect of

[21] A trading venue might elect to allow a term limit order to be amended to a price *better* than the original price of the term limit order during the period in which the order could not otherwise be cancelled. Lot size could not be amended. This should only be allowed if the original order was initially entered to become or to join the best bid or offer. Allowing such a trade (with the better price) to keep the original time stamp would seem equitable and would seem consistent with public policy objectives.

recommendations 1 and 2 should increase the likely allocation of lots[22] to orders that are exposed to market risk for a greater period of time, at the expense of orders that are exposed to market risk for only a few fleeting milliseconds.

There is no apparent reason this optional order type could not be used in a fragmented market structure where any number of trade allocation methods are in use.[23] Where financial markets are fragmented into multiple trading venues, different trade execution venues could have different trade allocation formulae. This proposed new order type would have no effect at all on trades allocated on trading venues that use the Pro Rata trade allocation method, as time in the Order Book would still be given no weight. It could, however, have a material effect on trades allocated on trading venues that use the Price/Time algorithm or which might adopt the trade allocation formula that we are recommending.

Implementing any new order type would, of necessity, involve implementation costs for trading venues, trade intermediaries and, to a lesser degree, for some end user market participants. It may be advantageous for one or more trading venues to inaugurate pilot implementation programs for a limited number of traded products to better gauge potential commercial acceptance of this concept and to make a more informed business decision regarding more universal implementation of the proposed new term limit order type.

3. Time Stamp Conventions of Dark or Unlit Orders

"Dark liquidity refers to orders that are not known to the rest of the market before the orders are matched as executed trades. Such trades, known as 'dark trades' can occur on exchange markets … and in venues other than exchange markets."[24] Orders of all types (except those noted in recommendation 2) should have a time stamp reflecting the start of the period during which that order was continuously visible in the Order Book to all market participants. Said another way, an order originally entered as an unlit

[22] Lots could mean shares, options, futures, swaps or any other specialized descriptor of traded financial instruments. Our intention is not to limit the scope of the applicability of the recommendations.

[23] In the fragmented U.S. equities market, Regulation NMS would still require that an order initially be routed to the trading venue with the best price. That trading venue may or may not be utilizing a trade allocation method that places a value on the time an order has been resting in the Order Book.

[24] Australian Securities & Investments Commission (2013, p. 12).

order should have no valid time stamp as long as that order remains unlit and not visible in the Order Book to all market participants. Without a time stamp, all unlit orders at a limit price should stand behind all lit orders at the same limit price, with respect to trade allocation.

If the offer were to go through all valid lit bids in the Order Book at the best bid limit price and should there remain unlit orders resting in the Order Book at the same limit price, the trading venue should allocate the residual amount of lots to the unlit orders on a Pro Rata basis. Doing so should reinforce the concept that dark orders should have no valid time stamp and is entirely consistent with Principle 3 of the OICU-IOSCO Principles for Dark Liquidity, Final Report[25]

> "Principle 3: In those jurisdictions where dark trading is generally permitted, regulators should take steps to support the use of transparent orders rather than dark orders executed on transparent markets or orders submitted into dark pools. Transparent orders should have priority over dark orders at the same price within a trading venue."

It is unfortunate that some trading venues allow traders to submit orders that are not visible to others and then modify the order while retaining its original time stamp. This practice runs counter to all principles of fairness. While this recommendation would not ban dark or unlit orders, it should provide an appropriate economic disincentive to utilize them to any great extent.

4. Aggregation of Consecutive Small Orders from the Same Legal Entity

Some traders have exploited trade allocation formulae that round fractional lot allocations up to the next integer. For example, a buyer of one 100 lot order might be entitled to an allocation of 62 futures or options contracts based upon the applicable trade allocation formula. If the buyer were to have entered its 100 lot order as 100 *one lot* buy orders and if the trading venue rounds fractional allocations up to the next integer, the trade allocation formula would allocate 62/100ths of a futures or options contract to each one lot order. Since the trading venue cannot allocate 62/100ths of a futures or options contract, the trader may be unjustly enriched by a rounding convention (that rounds fractions greater than ½ up) only because the trader entered a 100 lot order as 100 one lot orders. Taken to the extreme, the trader could be allocated as

[25] Technical Committee of the International Organization of Securities Commissions, "Principles for dark liquidity: Final report," Madrid, Spain, May, pp. 28-29.

many as 100 lots (one for each one lot order) while really only being entitled to 62 lots.

Prior to allocating trades, trading venues should first aggregate all matched trades submitted by the same legal entity to mitigate the potential for gamesmanship due to rounding conventions of one lot orders.[26] The aggregation routine should run once, every time that the trade allocation algorithm is run and would involve only orders that would be entitled to a fill or partial fill and only orders that appear to have been intentionally entered sequentially, by their respective time stamps. This should leave unaffected, the bona fide trades of unequal quantities entered minutes apart by the same legal entity. For example, it should be easy enough to discern between the two or three unequal resting buy orders from a grain elevator, all entered within five minutes[27] and 100 one lot orders from an algo trader, all entered three milliseconds apart.

The aggregated order should be assigned the worst time stamp of all of the multiple orders that comprise it.[28] There seems to be no reason to run the aggregation routine continuously or prior to running the trade allocation algorithm (note recommendation 5 below). This recommended procedure should eliminate much of the potential for gamesmanship, provided that market participants do not then violate other rules of trading venues by lying about their true identity.

Tangentially, maintaining the IT infrastructure to process a plethora of one and two lot fills, greatly increases the operating expenses of trading venues, clearing organizations and trade intermediaries because the *number* transactions drives the relevant operating expenses (scale), not the *number of shares or futures contracts* of those transactions. Processing a one lot order consumes just as much bandwidth and just as many IT resources as processing one 10,000 lot order that is matched and clears as one 10,000 lot order. A serendipitous result of implementing this recommendation (and recommendation 5 below) could be a material reduction in the operating expenses of trading venues, clearing organizations and trade intermediaries

[26] In the alternative, trading venues may elect to round down, rather than round up, allocating one lot trades that would otherwise be allocated a fractional lot, nothing, as at least one trading venue already does.

[27] The orders from the grain elevator should not be affected and should each severally retain their own original time stamp.

[28] This is not intended to affect all trades from the same legal entity. This recommendation is intended only to mitigate the unjust enrichment associated with multiple and sequential one lot and, perhaps, two lot trades.

that must scale their infrastructure to handle the *number* of transactions that they process or retransmit.

5. Random Timing of the Trade Match Algorithm

Trading venues should divide their trading sessions into time periods of one half of one second. At a completely random time during each one half second trading period, the trading venue should run its trade match algorithm (allocating trades utilizing the cardinal raking of resting bids and offers as described in recommendation 1 above) *once.*[29,30] This procedure has several advantages. Because high frequency traders would never know when the trade match algorithm would be run during any one half second time interval, the value of entering thousands of quotes for only a few milliseconds should be at least partially diminished. Additionally, as many of those quotes are typically not actually intended to be matched, they would occasionally, under this proposal, become swept up and executed in the trade match algorithm, due to its random timing within each half second period. Under this proposal, high frequency traders could continue the practice of entering thousands of quotes per second, only with more substantial potential financial implications.

There is some anecdotal evidence that suggests that most humans can read, recognize and process two to three numerical quantities per second.[31] Dividing

[29] The practice of trading venues to display the probable single opening price (as the market is about to open) usually comes with a policy that precludes the cancellation of orders within ½ of a second prior to the opening time. Market opening trade match algorithms determine a single price at which the maximum number of trades would optimally be matched. From a technical perspective, it will likely then be impossible to run the market opening trade match algorithm at a random time within every ½ second trading interval and display such a single price, before the fact. The instant proposal is simply to run the trade match and trade allocation algorithm once, not necessarily with all of the bells and whistles that come with the single price market opening trade match algorithm. HFT market participants and click traders should be able to deduce whether the trade match algorithm will match trades at the highest bid or the lowest offer by watching the size at the bid and offer.

[30] Alternatively, trading venues could elect to have the next subsequent one half second trading interval begin as soon as practicable after the prior trade match and trade allocation algorithms have been run for the prior trading interval. Doing so would compound the randomness of running the trade match and trade allocation algorithms. One potential downside of doing so might be that occasionally, the trade match and trade allocation algorithms could be running in excess of twice per second. Most human click traders would not be able to read, digest and respond to price and quantity information at such a fast pace.

[31] See Kimron L. Shapiro, Karen M. Arnell and Jane E. Raymond, 1997, "The attentional blink," *Trends in Cognitive Sciences*, Vol. 1, No. 8, November, pp. 291-296, and Jane E. Raymond, Kimron L. Shapiro and Karen M. Arnell, 1992, "Temporary suppression of visual processing in

the trading session into half second periods should provide human institutional traders with useful visual information on their trading screens as fast as that information can reasonably be comprehended.

Derivatives exchanges often group related product types on their trade match engines, e.g., equities, interest rates, foreign exchange, agriculture, energy and precious metals. Under this proposal, derivatives exchanges would run the trade match and trade allocation cycle, by product type (on each server) at a completely random time, once during each half second time period. Doing so should preserve the so-called "implied functionality" execution functionality e.g., the soybean crush spread, the crude oil crack spread or simple calendar spreads. Again, if performed properly, this should also randomize the sequence in which the servers run their product specific trade match and trade allocation cycles.

Equity trading venues also typically run multiple trade match engines, e.g., all stocks that begin with the letters A through F might trade on one server. It is envisioned that such a trading venue that has "n" number of servers would run one trade match and trade allocation cycle at random times during each half second period on each server. If done properly, this would also randomize the sequence in which each of the "n" number of servers would run their respective trade match and trade allocation algorithms.

The additional computational processing that would likely be required to assimilate both the new term limit order (Recommendation 2) and the allocation of lots based upon a cardinal ranking rather than a simple ordinal ranking (recommendation 1) should be partially if not totally ameliorated by only having to run the trade match and trade allocation algorithms once every half second. Doing so should allow the trading venues to conserve significantly on network bandwidth as outbound quotations would only be disseminated once every half second, and then, only in batches.[32] One firm that was invited to review and comment upon a prior draft of this paper indicated that implementing this recommendation could potentially free up much of the firm's annual IT budget to develop value added features to their client user interface, rather than spending that money on more and more servers to handle the

an RSVP task: An attentional blink?," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 18, No. 3, August, pp. 849-860.

[32] In theory, disseminating outbound quotations in batches might also thwart the practice of quote stuffing, as market venues would push quotation data out in data "packages," which might make quote stuffing a less effective strategy to intentionally clog up trading venues' outbound quotation systems.

tsunami of millisecond by millisecond data that they must retransmit to their human clients who have no possible ability to react to millisecond by millisecond price and quantity information. It is entirely possible that these potential cost savings could similarly be realized by a broader section of market participants.

Moving modern electronic markets away from continuous trade matching to discrete auction processing might also improve the technological framework within which national supervisory authorities will be held responsible for providing supervisory market oversight now, and for many years to come. Implementation of this recommendation would almost certainly materially reduce the amount of quotation and match trade data that would comprise the audit trail for today's modern electronic financial markets. Common sense would argue that the probability of achieving some success in this regulatory area might be greater if the challenges of doing so could be made less formidable.

At least two leading electronic foreign exchange trading markets have implemented or are actively considering implementing processes to slow down incoming orders. ParFX currently imposes randomized pauses on incoming orders.[33] EBS is considering submitting batches of incoming orders to its trade match engine. By doing so, many incoming orders would effectively be randomly delayed.[34]

Some lawmakers and regulators have suggested that quotations must be exposed to the market for a minimum amount of time. Fifty to 500[35] milliseconds is an eternity to a proprietary algo trader. The likely (and logical) reaction would be for marketmakers to widen their respective bid/offer spreads to compensate themselves for the additional market risk to which their quotes would be exposed under any such minimum cancellation time regimes. Our recommendation would still allow algo traders to cancel their orders at any time and should thus render moot, any potential arguments about having marketmakers' quotes unduly exposed to market risk. There is some potential that implementation of this recommendation would considerably dampen, or could even put an end to the incessant low latency arms race.

[33] Stephen Foley, 2013, "High-frequency traders face speed limits," *Financial Times*, April 28, available at www.ft.com/cms/s/0/d5b42402-aea3-11e2-8316-00144feabdc0.html?ftcamp=published_links%2Frss%2Fhome_uk%2Ffeed%2F%2Fproduct#axzz2TCwjGGsR.

[34] Ibid.

[35] Australian Securities & Investments Commission (2013, p. 10).

6. Granularity of Information in the Order Book

In general, all market participants should have access to the same information and with the same level of granularity of information in the Order Book. In general, market participants should only be able to have access to information that they legitimately need to make an informed economic decision on market depth, price and liquidity. Market participants that have the ability to query the Order Book should ideally only be able to see the aggregate size at each bid and offer levels as shown in the Order Book.

No market participants should be able to see any other identifying data in the Order Book that would reveal the identity or origin of the other market participants that have entered orders. No market participant should be able to see the time stamps of any orders in the Order Book other than their own.[36] No market participants should be able to see the individual lot sizes of orders entered, other than their own. Such granular data is not information that any market participant legitimately needs to make an informed economic trading decision.

It is our understanding that the status quo currently partially satisfies this recommendation.[37] It is also our understanding that high speed traders have already requested that trading venues begin to provide them with this more granular information, which we believe would be inappropriate. High frequency traders should continue to have the unfettered ability to attempt to reverse engineer aggregated data and reach any conclusions that they may care to reach.

Transparency into organized financial markets is beneficial and consistent with good public policy. One might attempt to argue that this recommendation goes against this principle. Dissemination of more granular data from the Order Book would assist algo traders in gaining an insight into the trading patterns of both algo traders and click traders.

[36] This is based on the premise that traders join resting *prices*, not resting *times*. It may be helpful to approach the issue from the perspective of a click trader, rather than from the perspective of an algo trader.

[37] See Sal Arnuk and Joseph Saluzzi, 2012, *Broken Markets: How High Frequency Trading and Predatory Practices on Wall Street Are Destroying Investor Confidence and Your Portfolio*, Upper Saddle River, NJ: FT Press, pp. 102-103, and Charles Duhigg, 2009, "Stock traders find speed pays, in milliseconds," *New York Times,* July 23, available at www.nytimes.com/2009/07/24/business/24trading.html.

Implications

Recommendation 5 (random trade match within half second time intervals) may have the greatest potential to disincent all three questionable behaviors, spoofing, layering and quote stuffing. If you don't know when the next trade match is going to occur, the downside risk of pretending to be a seller when you are really a buyer could leave a trader with a position exactly opposite of the desired position. This could even more so act against the interests of such a trader that engages in a combination of spoofing and layering, creating the illusion that there is size building on the bid side of the market, when the trader is really a seller. The trader could get stuck with a substantial position completely the opposite of what they want.

One important potential implication of recommendation 5 is the possible elimination of quote stuffing as a strategy to slow down other algo traders. As no one will know when the trade match and trade allocation algorithms will actually run, one would either have to abandon this strategy (as simply no longer being effective) or attempt to clog up the outbound quotation system continuously. Trading venues are reasonably adept at identifying and penalizing traders that have an exceedingly high ratio of quotes to trades (as continuous quote stuffing would undoubtedly require).

Perhaps most importantly, there is some possibility that recommendation 5 could dampen, stifle or put an end to the incessant low latency arms race. If the trade match engine only runs once every half second, and (assuming some trading venues might adopt recommendations 1 and 2) the allocation of orders would increasingly become a function of time in the Order Book, the so-called "real money" resting orders would receive increasing allocations and very short term traders would receive decreasing allocations. As very short term traders get allocated fewer and fewer lots, their respective quote to trade ratios would logically increase, which almost always carries penalties assessed by the relevant trading venues. If trading venues only matched trades and disseminated price and quantity information once every half second, there would arguably be considerably less financial incentive for all concerned to invest increasingly large sums in an effort to enable digital computers to respond to the trades of other digital computers and shave one or two milliseconds off the process.

Recommendations 3, 4 and 6 would likely only indirectly disincent spoofing, layering and quote stuffing. But those three trading strategies are not the only behaviors that should arguably be disincented. The questionable behavior addressed by recommendations 3 and 4 is obvious; using dark orders and gaming rounding conventions.

Recommendation 6 (providing only aggregated pretrade information from the Order Book) has more complex implications. Recent research by Weller[38] and by Baron, Brogaard and Kirilenko[39] indicate that the fastest high frequency traders (1) are the most profitable and (2) tend not to have their trades match opposite other fast high frequency traders. While this phenomenon has been detected in futures contracts, where such trades are completely anonymous, some equity trading venues currently provide more specific trade identifiers, more than only the aggregate quantity bid or offered at each limit price.[40] This more granular information, cannot be of much value to human click traders; it could only be of value to high frequency traders. By obtaining this more granular pre-trade information, high frequency traders could (1) more efficiently reverse engineer the trading algorithms of their competitors and (2) more effectively discriminate among the counterparties whose resting orders are resident in the Order Book. It is difficult to see how either of these activities serves the public good.

Conclusion

Term limit orders, and running the trade match algorithm at random times within half second intervals would seem to provide an equitable balance between human institutional traders and automated liquidity providers and could drastically reduce the current tsunami of data disseminated by trading venues. Allocating trades based on the actual time that orders have been exposed (or committed to be exposed) to market risk would appear to be a more equitable approach than some trade allocation algorithms currently in use.

[38] Brian Weller, 2012, "Liquidity and high frequency trading," University of Chicago Booth School of Business and University of Chicago, Department of Economics, working paper, November 10, pp. 42-43.

[39] Matthew Baron, Jonathan Brogaard and Andrei Kirilenko, 2012, "The trading profits of high frequency traders," Princeton University, University of Washington and Commodity Futures Trading Commission, working paper, November, pp. 20-21.

[40] Arnuk and Saluzzi (2012, pp. 102-103).

Implementing both term limit orders and the new trade allocation formula could return some equitability that some believe may have been lost.

Recommendations for establishing appropriate rounding conventions, appropriate treatment of invisible orders and granularity of available pre-trade information visible in the Order Book represent approaches not inconsistent with sound and defensible public policy.

National authorities and purveyors of modern electronic trading venues should consider these recommendations and the informed comments of interested market participants.