Overview of High Frequency Trading*

Anton Golub[1]

[1]Marie Curie Fellow, Manchester Business School

April 15, 2011

1 Introduction to High Frequency Trading

It seems like everyone defines "high-frequency trading" slightly differently. Ask 10 industry veterans and you will get 10 variations of similar concept. But while it may be difficult to pin down an exact definition of high-frequency trading, there are definite characteristics that help define the trading strategy.
First, high frequency trading depends on speed of execution and ultra-low latency. That generally is coupled with a very high trading turnover - many trades taking place over a short time period. Finally, high-frequency trading shops take few, if any, overnight positions as they are focused on the accumulation of small, short-term profits.
In addition, Samuelson notes, high-frequency trading generally leverages participant's proprietary money. Although some high-frequency shops offer to hedge funds the infrastructure to enable low-latency executions for high-frequency strategies, he says, high-frequency trading is not conducted on an agency basis.
According to Robert Iati, partner and global head of consulting at TABB Group, *"We define high frequency trading as fully automated trading strategies that seek to benefit from market liquidity imbalances or other short-term pricing inefficiencies. And that goes across asset classes, extending from equities and derivatives into currencies and little into fixed income"*.
Iati also makes the distinction that high frequency trading shops depend on hundreds of algorithms and he has ever heard of policies to never let an algorithm go longer than four to six weeks without being changes in some way. *"They are paying millions of dollars to coder to build a hundred variations of a particular algorithm or series of algorithms That was eye opening"*, says Iati.
Rishi Narang, founding principal of Telesis Capital, a Southern California-based alternative investment manager focused on quantitative trading strategies, attempts to define a strategy by what it is not:*If you take home overnight position, and if trading turnover is not north of 100 percent per day, then you're probably not really high-frequency trading in aggregate"*, he asserts.

*The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7-PEOPLE-ITN-2008 under grant agreement number PITN-GA-2009-237984. The funding is gratefully acknowledged.

High frequency trading requires immediate, real-time data analysis, which leads to automatic trading decisions. *"It means analyzing what is happening in the market on the spot - without the time to store the data in a database - doing automatic tick-by-tick analysis and making decisions based on that"*, he adds.

1.1 Who is Trading?

Just as it difficult to define high-frequency trading, it is a challenge to define the high-frequency traders. Telesis Capital's Narang says generally thre are two types of high-frequency trader in the market. The first is a liquidity provider or market maker, such as GETCO, Citadel and Tradebot.
The second is more serious buy-side alpha trader. *"They are not just out there trying to provide liquidity; they are trying to forecast near-term movements in instruments - not just stocks"*, Narang explains. *"They are implementing their strategies on a fast infrastructure to move in and out of positions quickly"*.
Narang also points out that while the liquidity provider-type of high-frequency trader may have an alpha component in its strategy, the alpha trader-type generally does not have a market-making component in it strategy. *"The market maker-type often employs some type of forecasting to help decide what to own or what to avoid"*, he says.
TABB Group's Iati goes further a step further, defining three types of firms that generally are high-frequency traders. First, he says, there are the traditional broker-dealer undertaking high-frequency strategies on their proprietary trading desks, separate from their client business. Second, he points out to high-frequency hedge funds. Third are proprietary trading firms that are mainly using private money.
"Not all of these [proprietary] shops are the same, though", Iati adds. *"There is a separation based on how they trade, and the clearest delineation runs between virtual market making and everything else"*. The virtual market makers, he explains, are primarily providing liquidity into the market and profiting from rebate trading. The "everything else" category of proprietary shops generally makes its money from arbitrage opportunities.
While it is tough to say with certainty how many high-frequency firms fall into each category, Iati estimates that there are between 10 and 20 broker-dealer proprietary desks and fewer than 20 active high-frequency hedge funds. The hardest category to quantify, he says, are the independent proprietary shops, which he numbers at more than 100 but less than 300.

High frequency trading has a solid hold in the UK markets at 77 percent of transactions, according to a study by Tabb Group.

Orders from long-only funds that bet stock will rise, hedge funds and retail investors account for 23 percent of activity in continuous markets, according to findings of a Tabb Group report. High frequency trading - method where trades occur over the course of seconds - accounts for the rest. The practice makes up 35 percent of 3.9 trillion-euro ($5.3 trillion) U.K. turnover, says the Tabb report.

Bloomberg reports that "Tabb's data covers what it calls continuous markets where trades occur electronically, including venues where prices are publicly displayed and dark pools, where they aren't". Over-the-counter trading, conducted away from exchanges and alternative systems, isn't included, Tabb said".

According to Tabb Group findings, the U.K. makes up about 21 percent of all European trading.

"What the study shoes is that so little of the continuous market is natural order flow". "It's critical for pension funds to have alternative strategies to achieve best execution and alternative sources of liquidity which they trust".

Rhode estimates there are between 35 and 40 independent high frequency trading firms such as Getco LLC and Optiver operating in the U.K.

1.2 HFT Wizard: *Is this really new? Part I*

1.3 High Frequency Trading and Co-location

Obsessed with high-speed trading and capturing fleeting price discrepancies between financial instruments, high frequency traders are pushing the envelope when it comes to low-latency technology. In the race to be first, firms are *co-locating* their strategies in data centers nearer to trading venues and tapping complex event processing to speed up executions.

Behaving more like technology firms than trading shops, high frequency firms are building their own proprietary models and algorithms, and back-testing them against historical tick data feeds before unleashing them on the real-time market feeds. "On of the fundamentals is an historical tick database - very large data arrays of available ticks from exchanges", says Dave Dugan, COO of Buttowood Trading Group, a proprietary trading firm in Chicago that is building high frequency trading strategies for listed futures and options on futures.

According to Dugan, Buttonwood has been back-testing its strategies against historical ticks from the Chicago Mercantile Exchange. "That was one of the first projects we put together", he says. To speed up the lunch of its algorithms across multiple markets, the global futures trading firm partnered with Frankfurt-based RTS Realtime System Group in February of this year and is using RTS's RTD Tango platform to create, test and deploy the trading strategies. "You're driving a Ferrari", Dugan notes. "What you eke out of these platforms largely depends on your skill and how you drive".

TO eke out even more speed, many of the high frequency trading firms are moving their black-box strategies into co-location facilities - data centers that are closer to the exchanges and ECN's matching engines. "With automated trading, the black-box traders develop strategies and put them into a black-box, and they host it in data centers operated by market centers such as Direct Edge, Nasdaq or Arca," explains Sang Lee, managing parter at Aite Group. "A key element of what they're going to do is get as close as possible to the venue. Firms are willing to pay a premium to get that slot".

When Atlanta-based Hyde Park Global started developing proprietary algorithms four years ago, 80 milliseconds was considered low latency, recalls Adam Afshar, president of the proprietary trading firm, which builds adaptive models for statistical arbitrage and other strategies. "Now the objective is to execute trader below 1 milliseconds, and many conversations are in the microseconds", he relates.

To shave off a few milliseconds of latency, the firm is in the process of co-locating its servers in New York. "We can co-locate our computers in New York for a very reasonable fee", says Afshar. "We don't pay for their cooling or the electricity - just the space on the rack". ....

According to Afshar, "By co-locating in New York, we are able to take 21 milliseconds off our trades. In the past 21 milliseconds was a a trivial matter. Now it's a pivotal matter"

1.4 HFT Wizard: *Is this really new? Part II*

1.5 Meeting the Co-location Demand

To capture liquidity from these high-volume trading shops, exchanges and other market centers have opened new data centers or expanded existing ones to offer co-location services. What's more, an entire industry of co-location specialists has sprung up to serve the needs of low-latency trading firms. Many of these players - including BT, Equinix, Savvis, 7Ticks, and Switch and Data - have opened up data centers in the major financial markets around the world to accommodate high frequency trading firms.

"The NYSE's impeding data center move is on the radar screen for all of these high-frequency guys", notes Panzica. As announced last year, NYSE is building a nearly 400,000-square-foot data center in Mahwah, New Jersey, and moving the equity markets (NYSE Arca and NYSE) matching engines out of its downtown Manhattan and Brooklyn facilities. "They're talking to all their customers, and they have room for some of their order flow servers". "[High frequency trading firms] need to move their operations potentially up north because that's where the new matching engines will be", he adds, noting that he's been in talks with HFT firms to house ....

1.6 Connecting the Data Center Dots

But while co-location facilities are a hot topic in high frequency trading circles, they address only one piece of the latency equation: the distance between the server and the exchange. "It doesn't really take into account how fast the provider feeds market data into the client's servers, or how fast the market data is converted from the different exchanges protocols and filtered so that it's useful for the black-box process". There is also the issues of how fast a broker-dealer and the exchange can process the orders generated by the black boxes.

1.7 Acceleration Executions with Complex Event Processing

As the demand grows for faster executions, firms are turning to complex event processing, or CEP, technology to detect patterns in real-time data. Typically the CEP engine sits on the black box server in the co-location facility.

"With high frequency trading you're analyzing flows of market data coming in against complex patterns that indicate trading opportunities, and you're placing orders in the market in real time," ... whose CEP engine is used to build algorithms that detect patterns in streaming data.

Ultimately, the push toward zero latency will drive further advances in high frequency trading technology, and firms seeking to be first will continue to invest in the latest low-latency solutions. But while faster and cheaper technology has made high-frequency trading more accessible to smaller firms, it's by no means cheap proposition. "I thing you could do it for a few hunderd thousand dollars". But as firms scale out to the multiple product types and geographies, he notes, they need more data center space, networking equipment and databases, and the cost escalates.

2 Market Microstructure

2.1 Low Latency Trading

The paper by Hasbrouck, Saar study market activity in the "millisecond environment" where computer algorithms respond to each other almost instantaneously. Using order-level Nasdaq data, they find that the millisecond environment consists of activity by some traders who respond to market events (like changes in limit order book) within roughly 2-3 milliseconds, and other who seem to cycle in wall-clock time (e.g. access the market every seconds). They define low-latency activity as strategies that respond to market events in the millisecond environment, the hallmark of proprietary trading by a variety of players including electronic market makers and statistical arbitrage desks. They reconstruct a measure of low-latency activity by identifying "strategic runs" which are linked by submissions, cancelations, and executions that are likely to be parts of a dynamic strategy. They use this measure to study the impact that low-latency activity has on market quality both during normal market conditions and during a period of declining prices and heightened economic uncertainty. Their conclusion is that increased low-latency activity improves traditional market quality measures such as short-term volatility, spreads, and displayed depth in the limit order book.

The financial environment is characterized by an ever increasing pace of both information gathering and the actions prompted by this information. Speed is important to traders in financial markets for two main reasons. First, the inherent fundamental volatility of financial securities means that rebalancing positions faster could result in higher utility. Second, irrespective of the absolute speed, being faster that other traders can create profit opportunities by enabling a prompt response to news or market-generated events. This latter considerations appears to drive an arms race where traders eploy cutting-edge technology and locate computers in close proximity to the trading venue in order to reduce the latency of their orders and gain an advantage. As a result, today's markets experience intense activity in the "millisecond environment", where computer algorithms respond to each other at a pace 100 times faster than it would take a human trader to blink.

They define the term "latency" as the time it takes to learn about an event (e.g., a change in the bid), generate a response, and have the exchange act on the response[1]. Exchanges have been investing heavily in upgrading their systems to reduce the time it takes to send information to customers as well as to accept and handle customers' orders. The have also begun to offer traders the ability to co-locate the traders' computer systems next to theirs, thereby reducing transmission times to under a millisecond (a thousandth of a second). As traders have also invested in the technology to process information faster, the entire event/analysis/action cycle has been reduced for some traders to a few milliseconds.

An important question is, who benefits from such massive investment in technology? After all, most trading is a zero sum game, and the reduction in fundamental risk mentioned above would seem very small for time intervals on the order of several milliseconds. There is a new set of traders in the market who implement low-latency strategies, which they define as strategies that respond to market events in the millisecond environment. These traders now generate most message activity in financial markets and according to some accounts also take part in the majority of trades[2]. While it appears that intermediate trading is on the rise (with these low-latency traders providing liquidity to other market participants), it is unclear whether intense low-latency activity harms of helps market quality.

The goal of the paper[3] examine the influence of these low-latency traders on the market environment. They being by studying the millisecond environment to ascertain how low-latency strategies affect the time-series properties of market activity. They then ask the following question: How does the interaction of these traders in millisecond environment impact the quality of markets that human investors can observe? In other words, they would like to know how their combined activity affects attributes such as the short-term volatility of stocks, the total price impact of trades, and the depth of the market. To investigate these questions, they utilize Nasdaq order-level data (TotalView-ITCH) that are identical to those supplied to subscribers and which provide real-time information about orders and executions on the Nasdaq system. Each entry (submission, cancelation, or execution of an order) is time-stamped in the millisecond, and hence these data provide a very detailed view of activity on the Nasdaq system.

They find that the millisecond environment shows evidence of two types of activities: one by traders who respond to market events and the other by traders who seem to op-

[1] More specifically, we define latency as the sum of three components: the time it takes for information to reach the trader, the time it takes for the trader's algorithms to analyze the information, and the time it takes for the generated action to reach the exchange and get implemented. The latencies claimed by many trading venues, however, are usually defined much more narrowly, typically as the processing delay measured from the entry of the order (at the vendor's computer) to the transmission of an acknowledgement (from the vendor's computer).

[2] See the discussion of high frequency traders in the SEC's Concept Release on Equity Market Structure

[3] *

erate according to a schedule (e.g. access the market every second). The activity of the latter creates periodicities in the time-series properties of market activity based on the wall-clock time. They believe that low-latency activity (i.e. strategies that respond to market events) is the hallmark of proprietary trading by electronic market making firms and statistical arbitrage operations conducted by hedge funds and other financial firms. On the other hand, the periodicity is more likely generated by the activity of agency algorithms employed to minimize trading costs of buy-side money managers. The interaction among different types of algorithms gives rise to intense episodes of submissions and cancellations of limit orders that start and stop abruptly, but these episodes aren't necessarily associated with the elevated execution rates. In other words, intense high frequency activity in the millisecond environment need to translate into a surge in high frequency trading.

They use the data to construct *"strategic runs"* of linked messages that describe dynamic order placement strategies. By tracking submissions, cancellations, and executions that can be associated with each other, they create a measure of low-latency activity. They use a simultaneous equation framework to examine how the intensity of low-latency activity affects market quality measure. They find that an increase in low-latency activity lowers short-term volatility, reduces quoted spreads and the total price impact of trades, and increases depth in the limit order book. If their econometric framework successfully corrects for the simultaneity between low-latency activity and market attributes, then increased activity of low-latency traders is beneficial to the traditional benchmark of market quality.

They employ two distinct sample periods to investigate whether the impact of low-latency trading on market quality (and the millisecond environment in general) differs between calm days and periods of declining prices and heightened uncertainty. They find that the millisecond environment with its various attributes is rather similar across the two sample periods. Higher low-latency activity enhances market quality in both environments, and is especially beneficial in reducing volatility for small stocks during stressful times[4]

2.2 Data and Sample - Nasdaq Order-Level Data

The Nasdaq Stock Market is a pure agency market. It operates as an electronic limit order book that utilizes the INET architecture (which was purchased by Nasdaq in 2005.)[5] All submitted orders must be price-contingent (i.e. limit orders), and the traders who seek immediate execution need to price the limit order to be marketable (e.g. a buy order priced at or above the prevailing ask price). Traders can designate their orders to display in the Nasdaq book or mark them as "non-displayed", in which case they reside in the book but are invisible to all traders. Execution priority follows price, visibility and time. All displayed quantities at a price are executed before non-displayed quantities at that price can trade.

[4] They note that this does not imply that the activity of low-latency traders would help curb volatility during extremely brief episodes such as the "flash crash" of May 2010, in which the market declined by about 7% over a 15-minute interval before partially rebounding.

[5] See Hasbrouck and Saar (2009) description of the INET market structure

The Nasdaq data they use, TotalView-ITCH, are identical to those supplied to subscribers, providing real-time information about orders and executions on the Nasdaq system. These data are comprised of time-sequenced messages that describe the history of trade and book activity. Each message is time-stamped to the millisecond, and hence these data provide a detailed picture of the trading process and the state of the Nasdaq book. They are able to observe four different types of messages in the TotalView-ITCH dataset:

- the addition of a displayed order to the book
- the cancellation of a displayed order
- the execution of a displayed order
- the execution of non-displayed order,

With respect to executions, we believe that the meaningful economic event is the arrival of the marketable order. In the data, when an incoming order executes against multiple standing orders in the book, separate messages are generated for each standing order. They view these as a single marketable order arrival, so they group as one event multiple execution messages that have all the same millisecond time stamp, are in the same direction, and occur in a sequence unbroken by any non-execution message. The component executions need not occur at the same price, and some (or all) of the executions may occur against non-displayed quantities.

2.3 Characterizing the New Trading Environment

Current market observers often comment of the rapid pace of activity. In fact, the typical average message rate is unremarkable. The sum of the median numbers of submissions cancellation, and executions for 2007 if 53,993. With 23,400 seconds in a 6.5 hour trading session, a representative average message arrival is about 2.3 messages per second.

The average, belies the intensely episodic nature of the activity. To illustrate this, we estimate the hazard rate for the inter-message durations. The hazard rate is the message arrival intensity (for a given stock), conditional on the time elapsed since the last message.

In the first millisecond (after the preceding message) the hazard rate for submissions/cancellations is 334 messages per second in 2007, and 283 messages in 2008, i.e. roughly one hundred times the average arrival intensity. These high values, however, rapidly dissipate. In 2008, the initial average drops by about 90 percent in the first ten milliseconds, and by about 98 percent in the first hundred milliseconds.

A declining hazard rate is consistent with event clustering. From an economic perspective, variation in trading intensity has long been believed to reflect variation in information intensity. While the information can be diverse in type and origin, it is often viewed as relating to the fundamental value of the stock and originating from outside the market (e.g. a news conference with the CEO ir a change in an analyst's earnings forecast.) At horizons of extreme brevity, however, there is simply not sufficient time for an agent to be reacting to anything *except* very local market information[6]. The information is about whether someone is interested in buying or selling, and it may lead to a transient price movement rather than a permanent shift.

While the hazard rate graphs are dominated by the rapid decay, they also exhibit local peaks. Over the very short run, submissions/cancellations have distinct peak in both the 2007 and 2008 samples at around 60 milliseconds. The magnitude of the peaks is rather large. For example, the peak at around 60 milliseconds in the 2007 sample implies a hazard rate that is twice as large as the hazard rate on would get by averaging the rates a few milliseconds before and after this specific duration. There are also discernible peaks at 11-12 milliseconds. These are somewhat less visible because they occur in a region dominated by the rapid decay. They are nevertheless about 30% higher than the average surrounding values. These peaks do not appear as distinctly in the execution hazard rates. The later, also peak around 2-3 milliseconds. Over a longer interval, submissions/cancellations exhibit peaks around 100 and 1,000 milliseconds.

What do does peaks represents? The peaks at 60, 100 and 1,000 milliseconds corresponds to "natural" rates (1,000 times per minute, ten times per second, and once per second), and so may reflect algorithms that access the market periodically. The peaks at shorter durations, may represent strategic responses to market events, and so serve as useful indications of effective latency.

2.4 Periodicity

They examine the level of activity in wall-clock time (the hazard rate analyzes are effectively set in event time). The time stamps in the data are milliseconds past midnight. Therefore for a given timestamp t, the quantity mod(t,1000) is the millisecond reminder, i.e. a millisecond time stamp within the second. Assuming that message arrival rates are constant of (if stochastic) well-mixes within a sample one would expect the milliseconds reminders to be uniformly distributed over the integers $\{0, 1, \ldots, 999\}$.

The distributions in both sample periods exhibit marked departures from uniformity. Both feature large peaks occurring shortly after the one-second boundary (at roughly 10-30 milliseconds), and also around 150 milliseconds. Broad elevations occur around 600 milliseconds. We believe that these peaks are indicative of automated trading systems that periodically access the market, near the second on the half-second. These intervals are substantially longer than the sub-100 milliseconds horizon that characterizes the elevated hazard rates.

In other words, unlike low-latency trader who respond to market-created events, these

[6] It is unlikely that the time it takes to process and extract the pricing-relevant implications of fundamental information (e.g. statements made by the CEO of the firm) is as low as 2-3 ms. Furthermore, the frequency of fundamental information events is so low that orders reacting to such events are unlikely to generate observable peaks in the hazard rate that are computed from tens of thousands of observations for each stock (in one month).

algorithms submit an orders and periodically revisit it. These periodic checks would also be subject to latency delays. Even if an algorithm is programmed to revisit an order exactly on the second boundary, any response would occur subsequently. The time elapsed from the one-second mark would depend on the latency of algorithm (i.e. how fast the algorithm receives information from the market, analyzes it, and responds by sending messages to the market). The observed peaks at 10-30 milliseconds or at 150 milliseconds could be generated by clustering in transmission time (due to geographic clustering of algorithmic trading firms), technology, or simply the large volume handled by particular firms.

2.5 Response Time

The definition of low-latency trading is "strategies that responds to market events in the market milliseconds environment". Although any event might be expected to affect all subsequent events, the interest here is the speed of response. It is therefore reasonable to focus on conditioning events that seem especially likely to trigger rapid reactions. One such event is the improvement of a quote. An increase in the bid may lead to an immediate trade (against the bid) as potential sellers race to hit it. Alternatively, competing buyers may race to cancel and resubmit their own bids to remain competitive and achieve or maintain time priority. We call the former response a same-side execution, and the latter response a same-side submission/cancellation. Events on the sell side of the book, subsequent to a decrease in the ask price, are defined similarly.

...These peaks are much more sharply defined in the conditional analysis, particular for executions. This suggests that the fastest responder are subject to 2-3 milliseconds latency. For comparison purposes, we note that human reaction times are generally though to be on the order of 200 milliseconds[7]. It is therefore reasonable to assume that these responses represent actions by automated agents (various types of trading algorithms).

2.6 High Frequency Episodes

Both the short-term intensity dependence and clock-time periodicity could in principle be modeled statistically with standard time series decomposition techniques. Our attempts to accomplish this (with spectral and wavelet analysis), however, were not very fruitful. Despite this, certain idiosyncrasies of the decompositions did reveal to us another characteristics of the millisecond environment. Much high frequency activity is not only episodic, but is also strikingly abrupt in commencement and completion.

They all share the same features:

- a sudden onset of intense activity of submissions and cancellations of limit orders that stops abruptly after a short period of time,
- lack the change in the pattern of executions before, during, or after these high-frequency episodes.

[7] Kosinski (2010)

These suggests that the term high frequency trading that is used to describe some low-latency activity is generally a misnomer: there is indeed high frequency activity, but it does not lead necessarily to intense trading. It simply manifests in intense submissions and cancelations of orders[8].

The millisecond environment therefore consists of activity by some traders who respond to market events and other who seem to cycle in wall-clock time. This is activity could give rise to intense episodes of submissions and cancelations of limit orders that start and stop abruptly, but these episodes need not be accompanied by intensified trading in the stocks. Before we proceed to measure low-latency trading and investigate its impact on market quality, it is useful to discuss the types of market participants whose activities shape the millisecond environment.

2.7 The players: Proprietary Algorithms and Agency Algorithms

Much trading and message activity in U.S. equity markets is commonly attributed to trading algorithms. Not all algorithms serve the same purpose and therefore the patterns they induce in market data and the impact they have on market quality could depend on their specific objectives. Broadly speaking, however, we can categorize algorithmic activity as agency or proprietary. Agency algorithms are used by buy-side institutions to minimize the cost of executing trades in the process of implementing changes in their investment portfolios. Proprietary algorithms are used by electronic market makers, hedge funds, proprietary trading desks of large financial firms, and independent statistical arbitrage firms, and are meant to profit from the trading environment itself (as opposed to investing in stocks).

Agency Algorithms (AA) are used by buy-side institutions and the brokers who serve them to buy and sell shares. They have been in existence for about two decades, but the last ten years have witnessed a dramatic increase in their appeal due to deimalization (in 2010) and increased fregmentation in U.S. equity markets (following Reg ATS in 1998 and Reg NMS in 2005). These algorithms break up large orders into pieces that are then sent over time to multiple trading venues. The algorithms determine the size, timing and venue for each piece depending on order-specific parameters (e.g., the desired horizon for the execution), algorithm-specific parameters that are estimated from historical data, real-time market data, and feedback about the executions of earlier pieces.

The key characteristic of AA is that the choice of which stock to trade and how much of buy or sell is made by a portfolio manager who has an investing (rather than trading) horizon in mind. The algorithms are meant to minimize execution costs relative to a specific benchmark (e.g. volume-weighted average price or market price at the time the order arrives at the trading desk), and they are typically developed by sell-side broker or independent software vendors to serve buy-side clients. Their ultimate goal is to execute a desired position change. Hence they essentially demand liquidity, even though their strategies might utilize nonmarketable limit orders.

[8] See Lauricella and Strasburg (2010)

Proprietary Algorithms (PA) are more diverse, and relative to AA, more difficult to concisely characterize . Nonetheless, these algorithms often belong to the following two bread categories:

- electronic market makers
- statistical arbitrage trading.

Electronic (or automated) market makers are dealers who buy and sell for their own account in a list of securities. These firms use algorithms to generate buy and sell limit orders and dynamically update these orders based on real-time data. Like traditional dealers, they often profit from the small difference between the bid and ask prices and aim at carrying only a small inventory. Another source of profits for such firms is the liquidity rebate offered by many trading venues. These rebates (typically a quarter of a penny per share) are offered to attract liquidity provers and are funded by execution fees paid by liquidity demanders.

Statistical arbitrage trading is carried out by the proprietary trading desks of larger financial firms, hedge funds, and independent specialty firms. They analyze historical data for individual stocks and groups of assets in a search for trading patterns (within assets or across assets) that can be exploited for profit. These profit opportunities might represent temporary deviations from perceived patterns (e.g. pairs trading) or stem from identification of certain trading need in the market (e.g. a large trader that attempts to execute an order and temporarily changes the time-series behavior of prices). Broadly speaking, most of these strategies rely on convergence of prices and the expectations that the market price will revert after temporary imbalances. Some of these traders attempt to profit from identifying the footprints of buy-side algorithms and trading ahead of or against them. Their goal is to profit at the expense of buy-side institutions by employing algorithms that are more sophisticate than typical AA .

Because AA and PA differ in their goals, they differ in the specifications of their algorithms and their technology. AA are based on historical estimates of price impact and execution probabilities across multiple trading venues and over time, and often require much less real-time input except for tracking the pieces of orders they execute. For example, volume-weighted average price algorithms attempt to distribute executions over time in proportion to the aggregate trading and achieve the average price for the stock. While some AA offer functionality such as pegging (e.g. tracking the bid or ask side of the market) or discretion (e.g. converting a nonmarketable limit buy order into a marketable order when the ask price decreases), typically AA do not require millisecond responses to changing market conditions.

Some algorithms simply check the market conditions and execution status every second (or several seconds) and respond to the changes they encounter. Their order reach the market with a lag that depends on the configurations and locations of their computers, generating the sample distributions of remainders. The similarities between the 2007 and 2008 samples suggest phenomena that are pervasive and do not disappear over time or in different market conditions.[9]

[9] One could suggest that even if a significant fraction of market participants were to have their algo-

One might conjecture that these patterns cannot be sustainable because sophisticated algorithms will take advantage of them and eliminate them. While there is no doubt that PA respond to such regularities, these responses only serve to accentuate the clock-time periodicities rather than eliminate them. In other words, as long as someone is sending messages in a period manner, their actions will provoke strategic responses by others who monitor the market continuously (the low-latency traders) and these responses will tend to amplify the periodicity. Since some PA supply liquidity to AA, it is conceivable that clustering of certain times helps AA execute their orders by increasing available liquidity. Furthermore, the clustering of AA means that the provision of liquidity by one investor to another at those times is higher even without elevated PA activity. As such, AA that operate in calendar time would have little incentive to change, making these patterns they identify in the data persist over time.

In contrast to AA, the hallmark of PA is speed: low-latency capabilities. Their need to respond to market events distinguishes them from AA. Therefore, these traders invest in co-location and advanced computing technology to create an edge in strategic interactions. While AA are used in the service of buy-side investing and hence can be justified by the social benefits often attributed to delegated portfolio management (e.g., diversification), the social benefit of PA are more elusive. If we consider electronic market making to be an extension of traditional market making, it provides the service of bridging the intertemporal disaggregation of order flow in continuous markets. Unlike traditional dealers, however, these electronic market making firms have no explicit obligation with respect to market presence or market quality, an issue we further discuss".

The social benefits of other types of low-latency trading are more difficult to ascertain. One could view them as aiding price discovery by eliminating transient price disturbances, but such an argument in a millisecond environment is tenuous. After all, at such speeds and for such short intervals it is difficult to determine the price component that constitutes a real innovation to the true value of the security as opposed to a transitory influence. The social utility in identifying buy-side interest and trading ahead of it is even more problematic.

Furthermore, the race to interact with the market environment faster and faster requires investing in vast resources in technology. PA are at the forefront of such investment, but they are not alone: AA providers respond by creating algorithms that enable clients to implement somewhat more sophisticated strategies that respond to market conditions along pre-defined parameters. Even exchanges such as NASDAQ get into the game by offering clients simple algorithms like pegging or discretionary orders through a platform that is operated by the exchange and connects directly to the execution engine.[10] These al-

rithms cycle in a one-second frequency, the occurrence times would be more smoothly distributed due to randomness in clock synchronizations. They believe that the periodicity can be initiated even by a few, relatively large, market participants.

[10] NASDAQ's RASH (Routing and Special Handling) protocol enables clients to use advanced functionality such as discretion (predetermined criteria for converting standing limit orders to marketable), random reserve (of partially non-displayed limit orders), pegging (to the relevant side of the market or the midquote), and routing to other trading venues.

gorithms collectively constitute ”low-latency trading”, and invite the question of whether they harm or improve the market quality perceived by long-term investors.

2.8 Strategic Runs

The evidence to this point has emphasized message timing. One would ideally like to track low-latency activity in order to decipher its impact on the market. Before turning to the methodology we use to track the algorithms, it is instructive to present two particular message sets that they believe are typical,. It appears that at least some of the activity consists of algorithms that either ”play” with one another or submit and cancel repeatedly in an apparent attempt to trigger an action on the part of another algorithm.

The underlying logic behind each algorithm that generates such strategic runs of messages is difficult to reverse engineer. It could be that some algorithms attempt to trigger an action on the part of other algorithms (e.g. canceling and resubmitting at a more aggressive price) and than interact with them. Whatever the reasoning, it is clear that an algorithm that repeatedly submits orders and cancels them within 10 milliseconds does not intend to interact with human traders (whose respond time would probably take more than 200 milliseconds even if their attention were focused on this particular security). These algorithms operate in their own space: they are intended to trigger the response from (or respond to) other algorithms. Activity in the limit order book is dominated nowadays by this kind of interaction between automated algorithms, in contrast to a decade ago when human traders still ruled. How, then, do these algorithms affect the environment that the human traders observe? How is such activity related to market quality measures computed over minutes rather than milliseconds? In order to answer these questions, we need to create a measure of the activity of these low-latency traders.

They construct a measure by identifying ”strategic runs”, which are linked submissions, cancellations, and executions that are likely to be parts of a dynamic strategy. Since their data do not identify individual traders, the methodology no doubt introduces some noise into the identification of low-latency activity. They nevertheless believe that other attributes of the messages can be used to infer linked sequences. In particular, ”strategic runs” (or simply, in this context, ”runs”) are constructed as follows. Reference numbers supplied with the data unambiguously link an individual limit order with its subsequent cancellation or execution. The point of inference comes in deciding whether a cancellation can be linked to either subsequent submission of a nonmarketable limit order or a subsequent execution that occurs when the same order is resent to the market priced to be marketable. They input such a link when the cancellation is followed within one second by a limit order submission or by an execution in the same direction for the same size. If a limit order is partially executed, and the remainder is cancelled, they look for a subsequent resubmission or execution of the cancelled quantity. In this manner they construct runs forward throughout the day.

The procedure links roughly 60 percent of the cancellation in the 2007 sample, and 55 percent in the 2008 sample. Although they allow up to one second delay from cancellation to resubmission, most resubmissions occur much more promptly. The median resubmission delay in the runs in one millisecond. The length of a run can be measured by the number of linked messages. The simplest run would have three messages, a submission of a nonmarketable limit order, its cancellation, and its resubmission as a marketable limit order that executes immediately (i.e., an "active execution"). The shortest run that does not involve an execution is a limit order that was submitted, cancelled, resubmitted, and cancelled or expired at the end of the day. Their sample periods, however, feature many runs of 10 or more linked messages and the longest run they identify has 93,243 messages. They identify about 57 million runs in the 2007 sample period and 78 million runs in the 2008 sample period.

2.9 Low-Latency Trading and Market Quality

Agents who engage in low-latency trading and interact with the market over millisecond horizons are at one extreme in the continuum of market participants. Most investors either cannot or choose not to engage at the market at these speed[11]. These investor's experience with the market is still best described with the traditional market quality measures in the market microstructure arsenal. Hence, it is natural to ask, how does low-latency activity with its algorithms that interact in milliseconds relate to depth in the market or the range of prices that can be observed over minutes or hours? This question does not have an obvious answer. It seems to resemble the challenge faced by physicists when attempting to relate quantum mechanic's subatomic interactions to our daily life that appears to be governed by Newtonian mechanics. However, if we believe that healthy markets need to attract long-term investors whose beliefs and preferences are essential for the determination of market price, then market quality should be measured using time intervals that are easily observed by these investors.

They therefore seek to characterize the influence of low-latency trading on measures of liquidity and short-term volatility observed over 10-minutes intervals throughout the day. Measures such as the range between high and low prices in these intervals, the effective and quotes spread, and the depth of the exchange's limit order book should give a sense of the market quality. And while would likely not capture every instance of PA in each interval of time, the strategic runs they have identified in the previous section could be used to construct a measure of low-latency activity.

2.10 Empirical Limitations on High Frequency Trading Profitability

Addressing the ongoing controversy over aggressive high-frequency trading trading practices in financial markets, we report the results of an extensive empirical study estimating the maximum possible profitability of such practices, and arrive at figures that are surprisingly modest. Their findings highlight the tension between execution costs and trading horizon confronted by high-frequency traders, and provide a controlled and large-scale empirical perspective on the high-frequency debate that has heretofore been absent. Their

[11] The recent SEC Concept on Equity Market Structure refers in this context to "long-terms investors...who provide capital investment and are willing to accept the risk of ownership in listed companies for an extended period of time" (p.33).

study employs a number of novel empirical methods, including the simulation of an ”omniscient” high-frequency trader who can see the future and act accordingly.

The financial crisis of the past ”two” years has been accompanied by rising alarm - popular media and regulatory - over what is broadly called high-frequency trading (HFT in the sequel). The overarching fear is that quantitative trading groups, armed with modern networking and computing technology and expertise, are in some way victimizing retail (”mom and pop”) trader and other less sophisticated parties. Since modern markets are fundamentally strategic and game-theoretic profitability depends not only on stock fundamentals and macroeconomic conditions, but also on the behavior of other participants - the debate over HFT can be viewed as concern over the practices of one group of ”players” and the resulting costs to other players.

The HFT debate often conflates distinct phenomena - confusing, for instance, dark pools and flash trading, which are new market mechanisms, with HFT, which is a type of trading behavior within both existing and emerging exchanges. The core concern regarding HFT, however, is relatively straightforward: that the ability to electronically execute trades on extraordinarily short time scales[12], combined with the quantitative modeling of massive stores of historical data, permits a variety of practices unavailable to most parties. A broad example would be the discovery of very short-term informational advantages (for instance, by detecting large, slow traders in the market) and profiting from them by trading rapidly and aggressively.

Despite growing controversy over HFT[13], there appears to be no objective, large-scale empirical studies of the potential profitability and impact of HFT. The purpose of this paper is to provide such a study. Main conclusion is a perhaps unexpected one: for at least one broad class of ”aggressive” HFT, the total available market size - that is, the maximum profit that could conceivably be realized using this type of HFT (and hence the maximum cost to other traders) - is surprisingly small: only $21 billion for the *entire universe of U.S. equities in 2008* at the longest holding period, down to $21 million or less for the shortest holding periods. Furthermore, these numbers seem to be vast overestimates of the profits that could actually be achieved in the real world by at least an order of magnitude. These figures should be contrasted with the approximately $50 trillion annual trading volume in the same markets. The findings are of interest in their own right as well as potential relevant to the ongoing debate over HFT.

They make a distinction between *passive* HFT, in which a HFT strategy exclusively places limit orders that are not immediately marketable, and thus act s as a provider of liquidity to the market; and *aggressive* HFT, in which only market orders are used, and thus the HFT must pay the attendant execution costs of crossing the bid-ask spread. In the study they focus on aggressive HFT, and argue that this is the variety of HFT that should be the primary, if not exclusive, focus of any concern, since the presence of passive HFT can only provide price and liquidity improvements to any trading counterparties.

[12] Often measured in milliseconds or less, and aided by the placement of trading servers within very few router hops of the exchanges, a practice known as colocation.

[13] SEC investigation into HFT practices

For aggressive HFT, there is a fundamental tension between two basic quantities: the *horizon* or *holding period*, as measured by the length of time for which a (long or short) position in a stock is held; and the *costs* of trading, as measured by (at least) the bid-ask spread that must be crossed by market orders on entry and liquidation of the position. In order for a trade to be profitable, the position must be held long enough for favorable price movement sufficient to overcome the trading costs. The shorter the holding period, the more extreme (and thus less frequent) the relative price movements must be for profitability. Rational concern over HFT should focus on short holding periods - measured in seconds or less - since at longer holding periods, the advantages of rapid exchange access and low latency are obviously diminished compared to the general trading population. While this tension between horizon and costs is well-understood in quantitative finance, it has not before been empirically studied on a large scale in the context of HFT.

At the core of the experimental study is a novel but simple technique called the *Omniscient Trader Methodology* (or OTM). Given the constraints of aggressive HFT, and armed with two large and rich historical trading data sets, they compute the profit or loss of all possible trades available to the HF trader *in hindsight*, and reach empirical overestimate of profitability by counting *only the profitable trades*. In the way they deliberately remove the greatest difficulty in real quantitative trading - that of predicting which trades will be profitable - and obtain what are certainly gross overestimates of profits realizable in practice to all parties engaging the aggressive HFT. Of course, such a study is interesting only if such overestimates are still surprisingly modest, which is the conclusion they establish.

They begin with the highest-resolution data available for NASDAQ, which permits exact replication of the full historical evolution of the order books for any chosen stock, and use it to simulate the OTM. They employ a two-step process: first they use the OTM to upper bound the profitability of HFT in a small set of the most liquid (and therefore most profitable) NASDAQ stocks, and then use a slightly less detailed data set and regression methods to scale up the estimates to much larger universe of 6,279 US stocks and all major US exchanges.

The background preliminaries demonstrate that traders in modern electronic exchanges have the choice between "passive" and "aggressive" order placement: they can either place limit orders that currently do not instigate any executions and lie in their respective books awaiting possible later execution; or they can place immediately marketable orders that cross the spread, eat into the opposing book, and pay both the spread costs and potentially higher costs for "deeper" shares. With regards to the debate surrounding HFT, *they prove that aggressive orders are the greatest cause for any concern about the negative impacts on trading counterparties.* The reason for this is both simple and well-understood in finance; passive order placement can only *improve* the market for any counterparties, both in prices and volume. By placing passive limit order, a trader can only reduce spreads and provide more shares and available price levels for the market compared to there? absence. Indeed, this is why many exchanges actually give *rebates* for orders that lie in the books but are eventually executed - they are providing liquidity - whereas fees are routinely charged for aggressive orders, which are removing liquidity. Furthermore, if one of the advantages of (and concerns over) HFT is the ability to very rapidly take and liquidate positions to profit on short-term informational advantages, aggressive order placement is necessary; if we have a predictive advantage for 2 second, we can not realize it by waiting for the other side of the market to come to us - we must initiate the entry and exit trades.

For these reasons, in the current study they restrict their attention to aggressive order placement. A concrete example of a type of HFT they are excluding by this choice is market-making, in which a trader tries o perpetually maintain both buy and sell passive limit orders, profiting whenever pairs of such orders are executed without out ever acquiring significant (long or short) inventory. While market making a natural and common type of trading strategy, they again note that it is not generally cited as part of the concern over HFT.

The underlying idea behind the OTM is quite simple: given complete historical data on a given stock, they identify exactly those trades that *would* have been profitable in *hindsight* - that is, given complete knowledge of the trading future of the stock in question. They thus simulate an Omniscient Trader (OT), whose profitability is obviously an upper bound on the profitability of *any* realistic strategy that must make on-line trading decisions based only on the past, and not future data.

The overall methodology is to first apply the OTM to the order book data, which provides us with a very detailed view of profitability and the relationships between a number of fundamental microstructure variables for the limited set of 19 stock on NASDAQ. They then use contemporaneous TAQ data and regression methods on the same 19 stocks in order to construct reliable models for scaling up the estimates to the full universe of U.S. equities and exchanges...

They conclude by offering some perspective on the size of the bounds. The total annual trading volume in the U.S. stock market (measured from TAQ data) is approximately $52 trillion, thus the maximal theoretical profit "unfairly" reaped by HF traders on the US market - i.e. the largest number reported in this work - is less than 0.05% of trading volume. It is also less than half of Goldman Sach's actual gross profit for the year ending September 25th, 2009.
Furthermore, it is important to reiterate that they vastly overestimate, in a number of ways, the profit that could actually be achieved by a real-world trader:

- they assume no trading fees or commissions paid by the HFT,
- they assume the trader is omniscient and can exactly predict the future price movements,
- assume the trader knows not only whether to buy or sell, but also precisely the optimal number of shares to trade at every moment,
- assume that these perfect predictions are made infinitely fast, and that trades execute with zero latency
- assume that the trader's actions do not influence the market or create adverse price movements

- assume that offers taken by the traders remain on the books, allowing the trader to repeatedly profit from a single opportunity as often as 100 times per second.

All of these assumptions fail in practice, and all of these failures reduce the realizable profit. For example, institutional traders might conservatively incur trading fees of about 0.6 cents per share. Since a large fraction of the price fluctuations within short holding periods are very small, these fees can be significant. In order to achieve the $3.4 billion profits reported, the OT would need to trade a total of 195 billion shares; if each such share cost 1.2 cents (since it must be transacted twice to realize profit), the total trading fees would be $2.3 billion, reducing profits by a full two thirds.
The job of predicting short term price fluctuations created by thousands of competing traders, each with a financial incentive to make that task impossible, presumable cannot be preformed with anything like omniscient accuracy. They believe that recognizing even 10% of the profitable opportunities is a phenomenally difficult achievement in the real world. Assuming this bas was not surpassed, the results imply that 2008 HFT profits on the entire US stock market were bounded by $2.1 billion. A real trader will not fail to act on some profitable opportunities, but also mistakenly act in unprofitable cases, causing additional losses. This is especially true when considering trading fees that must be paid regardless of whether a trade was profitable.
Finally, they remark that 2008 was generally believed to be a banner year for HFT profitability due to the volatility of markets during the financial crisis. Preliminary experiments on 2009 data suggest that 2009 HFT profits were indeed about half of the 2008 estimates.
The results demonstrate a surprisingly ...

2.11 Is high-frequency trading inducing changes in market microstructure and dynamics?

They follow up on research previously done in Eisler et. al. (2005); Eisler and Kertesz (2007a,b) and describe the changes in the high-frequency (short period) structure of the equity trading markets that likely have been induced by the spread of HFT firms and strategies. In particular, the paper demonstrates, that since 2005 there has been a marked changed in the correlation structure of stock trading dynamics where amongst many stocks, there has been a measurable departure from the typical H = 0.5 regime and that stronger self-similarity have been steadily increasing over the same period in the time that HFT has become the largest source of market volume...

Given the complex nature of HFT trades and the frequent opacity of firm trading strategies, it is difficult to pinpoint exactly what about HFT causes a higher correlation structure. One answer could be that HFT is the only type of trading that can exhibit trades that are reactive and exhibit feedback on short timescales that traditional trading generates over longer timescales.
Another cause may be the nature of HFT strategies themselves. Most HFT strategies can fall into two buckets (Lehoczky and Schervish, 2009):

- Optimal order execution: trades whose purpose is to break large share size trades into smaller ones for easier execution in the market without affecting market prices and eroding profit. There are two possibilities here. One that the breaking down of large orders to smaller ones approximates a multiplicative cascade which can generate self-similar behavior over time Mandelbrot (1974). Second, the queuing of chunks of larger order under an M/G/∞ queue could also generate correlations in the trade flow. However, it is questionable whether the "service time", or time to sell shares in a limit order, is a distribution with infinite variance as this queuing model requires.

- Statistical arbitrage: trades who use the properties of stock fluctuations and volatility to gain quick profits. Anecdotaly, these are most profitable in times of high market volatility. Perhaps since these algorithms work through measuring market fluctuations and reacting on them, a complex system of feedback used trades could generate self-similarity from a variety of yet unknown process.

Since firm trade strategies are carefully guarded secrets, it is difficult to tell which of these strategies predominate and induces most of the temporal correlations.

They clearly demonstrate that HFT is having an increasingly large impact on the microstructure of equity trading dynamics. They can determine this through several main pieces of evidence. First, the Hurst exponent H of traded value in short time scales (15 minutes or less) is increasing over time from its pervious Gaussian white noise values of 0.5. Second, this increase becomes most market, especially in the NYSE stocks, following the implementation of Reg NMS by the SEC which led to the boom in HFT. Finally, $H > 0.5$ traded value activity is clearly linked with small share trades which are the trades dominated by HFT traffic. In addition, this small share trade activity has grown rapidly as a proportion of all trades. The clear transition to HFT influenced trading noise is more easily seen in the NYSE stocks than with the NASDAQ stocks except NWS. The electronic nature of the NASDAQ market and its earlier adoption of HFT likely has made the higher H values not as recent a development as in the NYSE, but a development nevertheless.

Given the relative burstiness of signals with $H > 0.5$ we can also determine that volatility in trading patterns is no longer due to just adverse events but is becoming an increasingly intrinsic part of trading activity. Like internet traffic Leland et. al. (1994), if HFT trades are self-similar with $H > 0.5$, more participants in the market generate more volatility, not more predictable behavior. The probability of a traded value grater than V in any given time can be given by

$$\mathbb{P}(v \geq V) \sim h(v)v^{-\alpha}$$

where $h(\cdot)$ is a function that slowly varies at infinity and $0 < \alpha < 2$. The Hurst exponent is related to α by

$$H = \frac{3-\alpha}{2}.$$

There are few caveats to be recognized. First, given the limited timescale investigated, it is impossible to determine from these results alone what, if any, long-term effects are incorporating the short-term fluctuations. Second, it is an open question whether the benefits of liquidity offset the increased volatility. Third, this increased volatility due to self-similarity is not necessarily the cause of several high profile crashes in the stock prices such as that of Proctor & Gamble on May 6, 2010. Dramatic events due to traceable causes such as error or a rogue algorithm are not accounted for in the increased volatility though it does not rule out larger events caused by typical trading activities. Finally, this paper does not investigate any induced correlations, or lack thereof, in pricing and returns on short timescales which is another crucial issue.

Traded value, and by extension trading volume, fluctuations are starting to show self-similarity at increasingly shorter timescales. Values which were once only present on the orders of several hours or days are now commonplace in the timescale of seconds or minutes. It is important that the trading algorithms of HFT traders, as well as those who seek to understand, improve, or regulate HFT realize that the overall structure of trading is influenced in a measurable manner by HFT and that Gaussian noise models of short term trading volume fluctuations likely are increasingly inapplicable.

2.12 HFT Wizard

2.13 Supplemental Liquidity Providers & Enhanced Liquidity Providers

Traders Magazine, Flash Point: Equity Industry Clashes over flash and step-up orders

...curious article in the latest edition of Traders Magazine. It is curious mostly because it was allowed to be published, as it definitely peels off the cover of what truly happens at the pantheon of stock exchanges, that *is* dominated by a private group of select high frequency traders, who obtain better and faster pricing than everyone else, and where the group of "select few" is seemingly legally allowed and even encouraged to front-run the "every-one else" *(you, dear reader, are most likely in the latter camp.).* If you ever wondered why HFT generates profits of over $20 billion a year, please read this article.

As for Zero Hedge's intents, we would yet again request feedback from the proper authorities on whether one can derive more than superficial similarities between the method of operation on Direct Edge's Enhanced Liquidity Provider (ELP) program and NYSE's Supplemental Liquidity Provider program (aka. the Goldman (GS) kiss). Amusingly, it is none other than the NYSE's own Larry Leibowitz who raised the most ruckus about the potential abuse of the ELP program.

At an industry conference on market structure in May, a panel on market centers broached the subject of "flash" orders and almost ended in fisticuffs. In one corner was defending champion William O'Brian, CEO of Direct Edge. In the other was Larry Leibowitz, his hot-under-the collar opponent from the Big Board..The head of U.S. execution and global technology at NYSE Euronext assailed Direct Edge's Enhanced Liquidity Prover or ELP program as the "enhanced look" program, ***comparing it to the advance look at orders that NYSE specialists used to get. That practice was seen as giving specialists unfair advantage over the other market participants, and potentially disadvantaging order senders****. O'Brien observed that the NYSE has "more tiers than Yankee Stadium".*

Flash orders are also called "step-up" or "pre-routing display" orders. The rationale *for these order types is simple: Better me than you. They allow a venue to execute marketable order in-house when that market is not at the national best bid or offer, instead of routing those orders to rival markets.* ***They do this by briefly displaying information about the order to the venue's participants and soliciting NNBO-priced responses.*** *If there are no responses, the order can be canceled or routed to the market with the best price.*

All four markets with flash orders treat these orders in a similar way. If they get a marketable buy order, for instance, that would otherwise be routed to a market quoting at the NBBO, they flash the order to some or all of their participants as a bid at the same price as the national best offer. Exactly who sees the flash, how that information is conveyed and the duration of the flash vary by market. ***The maximum allowable time for a flash is 500 milliseconds, or half a second, although most of the markets flash routable orders for under 30 milliseconds.***

NYSE Euronext's anti-flash tirade didn't end with the SIFMA conference. ***The exchange operator, along with market-making firm GETCO and SIFMA, weighed in on the Nasdaq and BATS flash order types with formal letters to the Securities and Exchange Commission****. NYSE and SIFMA urged the SEC to abrogate the Nasdaq rule filing and reject BATS's filing. All three pushed the SEC to study the potential impact of flash orders on the marketplace before deciding whether to give them free rein.*

NYSE and GETCO charged that markets with flash orders were essentially running private markets of quotes for select participants that competed with the public quote stream. With Nasdaq and BATS rolling out new order types to combat Direct Edge, the upshot, in their view, was bad market structure and probably eventual harm to investors.

These firms and SIFMA argued that flash order types call into question some of the basic tenets of the equities market structure. In various combinations, they claimed that the effort to keep flow in-house undermines the concept of a quotation, impairs the meaningfulness of the NBBO, jeopardizes liquidity provision by hurting liquidity providers quoting at the NNBO, and potentially upsets the pursuits of best execution.

So the NYSE is making a mega fuss about a potential market entrant that does what everyone else does - understandable, nobody likes competition, especially not the New York Stock Exchange which has been losing market presence and top line revenue by the boatload recently. Yet the question stands just how much of this ”best kept secret” protocol does the NYSE employ currently to facilitate Supplemental Liquidity Providers, or rather, Provider (singular) - Goldman Sachs. When one firm dominates 50% of principal HFT trading on an exchange and, according to the above logic, can legally front run the other half, what does that mean for the rest of the world?

NYSE Euronext, despite frowning on flash orders, may wind up joining the party. Joe Mecane, executive vice president for U.S. markets at the company, notes that if the SEC allows these flash order types to stand, NYSE Arca would probably convert an existing order type into a flash-type interaction, and would look to more broadly disseminate the information. ”If the SEC is implicitly allowing private access to information, we'll need *to do it to be competitive", he said. NYSE Euronext may decide to offer flash orders on the NYSE as well, Mecane said. Nasdaq, for its part, is implementing a flash-type order this month on Nasdaq OMX BX, its Boston equities market.*

The NYSE is waiting for the deliberations of the same SEC after it did not even care to hear back on whether or not the NYSE's SLP deserves a comment period, objections, and traditional response time, and which waited until the last day to file an extension automatically assuming it would be granted...(And granted of course it was, as the only beneficiary again was Goldman Sachs.)

The primary argument against flash orders is that they create private markets and are therefore a step back for market structure. "These programs are creating a private locked market for a small group of participants, and they are holding up the execution process for that marketable order," Mecane said. He added that the Big Board operator isn't against dark pools, competition or innovative business models. "Our issue is that this creates a tiered market", he said.

Market maker GETCO told the SEC that by creating a two-tiered market, flash orders give professionals receiving the flashes a leg up over other investors. Non-public quotes could also "negatively affect the broader market, including retail investors who rely on the NBBO to ensure that their orders obtain the best, reasonably available price", the firm said. GETCO argued that flash orders, like dark pool liquidity that executes at the NBBO, also leave limit orders that established the best price in the lurch.

One wonders what the response of the SEC will be to this allegation. One wonders less, once it becomes painfully clear that any condemnation of two-tiering and flash orders would potentially automatically preclude Goldman from trading 1 billion PT shares a week for its prop trading accounts.

Ironically, Nasdaq and BATS already may be in enough hot water to really raise the temperature on not only Direct Edge but the NYSE as well:

Direct Edge's O'Brien draws a distinction between how the information his market disseminates is seen and what Nasdaq and BATS are doing. His flashes, he said, are sent out on a different data feed than the ENC's depth-of-book feed, while Nasdaq's and BATS orders are not. As a result, the latter exchange's feeds look like they're locking the market. (Last month exchanges added a flag to flashed orders to identify them for subscribers.)

In Selway's view, this argument clouds the point. The point, he said, is that order messages are being broadcast at prices, that effectively, lock protected quotes. This creates an elite tier of traders with access to better-priced orders than those receiving public quotes through the securities information processors, giving flash recipients an information advantage, he said.

Direct Edge's O'Brien argues that critics of his market's ELP program are twisting a successful innovation into a regulatory concern for purely competitive reasons. He said the ELP program gives participants a choice about how they want their order flow handled, and enables customers to lower their market-impact and transaction costs. He also notes *that critics of the ELP program, which includes dark pools among its participants, are anti-internalization. Internalization refers to the ability of brokers to match customers orders away from public markets. But the ECN's flash orders, on the contrary, O'Brien said, have "democratized access to dark liquidity sources by enabling retail customers to choose to interact with that liquidity to seek larger-size executions and potentially better prices."*

Would one be shocked that the NYSE would be so vocally against Direct Edge when it has the SLP in its back pocket effectively dominating what could be the biggest flash trading in history? Many more questions remain unanswered, but we hope readers now have a much better sense of the continuing fight against the ever more evident extensive informational advantage that Goldman Sachs may probably have thanks to its monopoly of the SLP program.

2.14 Flash Orders - history

Direct Edge was not the first equities market to think up a flash order. The progenitor of the flash among equities markets in the CBSX, which launched in March 2007 as Regulation NMS-compliant market. This is significant because NYSE Euronext argues that flash orders run afoul of Rules 602 and 604 of Regulation NMS. Rule 602, the Quote Rule, requires market centers to publicly disseminate their best bids and offers, and Rule 604 requires customer limit orders to be publicly displayed. The three other markets with flash or step-up orders have all referenced, in public comments or rule filings, CBSX's flash functionality as the basis for their order types.

The CBSX, however, distances itself from the current battle over flash orders. "It's not our fight", said David Harris, CEO of the exchange. His exchange's reason for the flash is different, he said. He also noted that CBSX's flash is an "operational process" approved by the SEC and not an order type.

CBSX developed its flash functionality so its Chicago-based participants could potentially get quicker executions from Chicago-based flash responders and avoid having their order routed 712 miles east for a fill. The exchange's flash process was approved in early 2006 as part of the rule set for the STOC system, the CBOE's small facility to execute securities that were not options. (The initial rule filing for stock trading was submitted to the SEC in 2004.) Rules making STOC NMS-compliant were approved in late September 2006, and the CBXS, which replaced STOC, got the SEC's nod in March 2007.

2.15 NBBO Matches

Another significant issue raised by flash messages is the impact they could have on public limit orders. Ratterman points out that flash orders could affect the price-discovery process by curtailing the executions received by those establishing the NBBO. "A flash order takes advantage of the price-formation risk of someone at the NBBO without rewarding that person with execution", he said.

Economist Harris agrees, but adds a caveat. "At present, executing at the NBBO is subject to the preference of customers", he said. "It seems to be consistent with the present regulatory legal framework that traders on or off the exchange are allowed to match the NBBO on request. If you oppose this system, you are basically saying brokers or exchanges have a responsibility to satisfy clients or other exchanges, and regulators have never said that".

In his view, flash orders could affect the quality of the market, if they come to represent enough liquidity. He describes flash orders as an example of exchanges competing for order flow at the expense of competition for the best price. "The implication is that trader would not post limit orders as aggressively as they otherwise would, and that would not be in the public interest", Harris said. "Theoretically, the bid-ask spread could wide". He added that the SEC "has been willing to protect traders offering liquidity at the NBBO only against trade-throughs in the name of protecting the order submitter [from executing at an inferior price]," but that the regulator may now have to think about the extent of the protection they are willing to offer those establishing the NBBO.

At the SIFMA conference, the SEC's Shillman said the Commission considers competition between limit order critical to robust price discovery. If it appears that an increase in "dark volumes" is undermining the process, he said, the SEC might look at new ways to reward limit orders that establish the best price. One option he highlighted would be to give displayed limit orders at the NBBO protection against executions at the same price. That would require executions at the NBBO to first trade against the price-setting interest available in the displayed market.

Nasdaq's Hyndman doesn't think limit orders are likely to be affected by the flash orders. "We don't think this market structure functionality is bad for limit orders", he said. "We don't know how will it play out, since it's new innovation for exchanges, but we'll see what's changed after a couple months". He added that the SEC is likely to keep its eye on flash orders "in the same vein they're keeping an eye on dark pool growth".

3 Flash Crash

he main theories for the cause of Flash Crash are listed as follows;

1. **Fat Finger Theory**: In the immediate aftermath of the large price drop, several reports[14] indicated that the event may be caused by a *fat finger trade*, an accidental large sell of Proctor&Gambel stock - a trader accidently pressed billion instead of million - which in turn caused a large algorithmic trading orders to sell the stock. This theory was reputed quickly as it was discovered that the decline in Proctor&Gambel trailed, not followed the futures market. The *fat finger theory* is not considered credible.

2. **Impact of High Frequency Trading**: This theory claims that high frequency traders where the cause of the Flash Crash. High frequency traders used *"quote*

[14] CNBC: A Citigroup Trader Made The Big Fat Finger Error

stuffing", sending non-executable orders (orders well outside the current National Best Bid and Offer) in batches. While the true motive of this actions is still unclear, it is claimed that these orders serve as a way to clog the exchanges (causing delays) and outsmart their competitors[15]. The SEC's final report concluded: "...that quote-stuffing – placing and then almost immediately canceling large numbers of rapid-fire orders to buy or sell stocks – was not a "major factor""[16]. Some prominent traders claim that high frequency traders have been a major factor in minimizing and reversing the flash crash[17]. Despite the explanations for different sides*, quote stuffing still is a controversial and unanswered topic[4]

3. **Large directional bets**: There are two large trades that might have caused the Flash Crash; some have suggested that a large purchase of put options on the S&P500 index by the hedge fund Universa Investments may have been the cause of the crash[5]. Joint report by SEC and CFTC has pointed out a single sale of 75,000 eMini S&P500 futures contract by Waddell&Reed as a cause of the Flash Crash.

4. **Delays**: This theory claims that delays in the New York Stock Exchange and Alternative Trading Systems might be the cause of the crash. Technical problems at the NYSE lead to delays in the NYSE quotes reported to the Consolidated Quotation System, while the time stamps of the quotes where indicating that they were current[18]. Market participants that had direct access to the Open Book could see both correct current quotes and the delayed quotes from CQS. Confronted with this issues a lot of market participants decided to pull out of the market, either by stop trading or by sending stub quotes (very low bids and very high offers). What is disturbing, is that this type of delays still occur[7].

3.1 How it started? Economic News

The 6th of May started with unsettling political and economic news from overseas concerning the European debt crisis that lead to growing uncertainty in the financial markets. Increasing uncertainty during the day is confirmed by various markets; high volatility (VIX); flight to quality (rise in the price of gold and the bonds): and an increase in premium for the buying protection against default by the Greek government. This lead to a significant, but not extraordinary, down day in early trading for the securities and the futures markets.

the Greek crisis has culminated during the riots on the streets of Athena, as the Greek parlament has voted laws that would cut the spending and restrict the social rights.

In the course of the day, the S&P500 volatility index, a measure of the expected volatility of the S&P500 Index, increased by 31.7 precent, which was the fourth largest single-day increase in VIX. Prices on gold futures rose 2.5%, while yields on ten-year Treasuries fell

[15] Nanex Llc. "Analysis of the "Flash Crash", Date of Event: 05/06/2010, Part 4, Quote Stuffing"

[16] "Findings Regarding the Market Events on May 6, 2010", Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues

[17] "Jim Simons on Flash Crash: High Frequency Traders Saved the Day", The Wall Street Journal

[4] "Explaining Bizarre Robot Stock Trader Behavior", The Atlantic

[5] "Did a Big Bet Help Trigger 'Black Swan' Stock Swoon?", The Wall Street Journal

[18] "The NBBO is Broken", Nanex Llc.

[7] "Latency On Demand?", Nanex Llc.

nearly 5% as investors engaged in a ”flight to quality”.

Starting at 1:00pm, a overall increase in risk also began to manifest itself in the price volatility of individual equities. The number of volatility pauses, also know as Liquidity Replenishment Points (”LRPs”), triggered on the New York Stock Exchange for individual equities listed and traded on the exchange began to substantially increase above average levels (for how much!!!!??).
By 2:30 pm selling pressure had pushed by Dow Jones Industrial Average (DJIA) down about 2.5%. By this time, buy side liquidity in the E-Mini had fallen from the early-morning level of nearly $6 billion dollars to $2.65 billion (representing a 55% decline). Buy side liquidity in SPY had also fallen from the early-morning level of about $275 million to $220 million (a decline of 20%). Some individual stocks also suffered a decline in both buy-side and sell-side liquidity by this time.

Against the backdrop of negative market sentiment and thinning liquidity, at 2:32pm, a large Fundamental Seller (a mutual fund complex) initiated a program to sell a total of 75,000 E-Mini contracts (valued at approximately $4.1 billion) as a hedge to an existing position. The large Fundamental Seller chose to execute this sell program via an automated execution algorithm that was programmed to feed orders into the June 2010 E-Mini market to target the an execution to 9% of the trading volume calculated over the previous minute, but without regard to price or time.
The execution of this sell program resulted in the largest net change in daily position of any trader in the E-Mini since the beginning of the 2010. Only two single-day sell programs of equal or larger size - one of which was by the same large Fundamental Seller - were executed in the E-Mini in the 12 months prior to May 6th.
High frequency traders and market makers were the likely buyers of the initial batch of orders submitted by the Sell Algorithm and these buyers built up temporary long positions. High frequency traders accumulated a net long position os about 3,300 contracts. Between 2:41pm and 2:44pm, high frequency traders aggressively sold about 2,000 E-Mini contracts in order to reduce their temporary long positions. At this time, high frequency traders stopped providing liquidity and instead began to take liquidity and were competing with the Fundamental Seller for the liquidity.
At the same time, high frequency traders traded nearly 140,00 E-Mini contract or over 33% of the total trading volume. This is consistent with the high frequency traders typical practice of trading a very large number of contracts, but not accumulating an aggregate inventory beyond three or four thousand contracts in either direction.
In a day of very negative market sentiment and high volatility, the combined pressure for the Fundamental Seller, high frequency traders and other traders drove the price of the E-Mini down approximately 3% in just four minutes for the beginning of 2:41pm through the end of 2:44pm.
High frequency traders traded over 1,455,000 contracts on May 6, which comprised almost a third of the total daily trading volume. Yet, net holdings of high frequency traders fluctuated around zero so rapidly that they rarely held more than 3,000 contracts long or short on that day. Moreover, compared to the three days prior to May 6, there was an unusually high level of ”hot potato” trading volume - due to repeated buying and selling of contracts - among the high frequency traders, especially during the period between 2:41pm and 2:45pm. Specifically, between 2:45:13 and 2:45:27, high frequency traders

traded over 27,000 contract, which accounted for about 49 percent of the total trading volume, while buying only about 200 additional contracts net.
At 2:45:28pm, trading on the E-Mini was paused for five seconds when the CME Stop Logic Functionality was triggered in order to prevent a cascade of further price declines. In that short period of time, sell-side pressure in the E-Mini was partly alleviated and buy-side interest increased. When trading resumed at 2:45:33pm, prices stabilized and shortly thereafter, the E-Mini began to recover.

3.2 Nanex LLC. Report

The analysis of the Wadell & Reed e-Mini[19] futures trade led us to an unexpected breakthrough. By process of elimination, and with the SEC report for context, *they* finally have a crystal clear understanding what caused the May 6ht, 2010 flash Crash.

First of all, the Waddell & Reed trades were not the cause, nor the trigger. The algorithm was very well behaved; it was careful not to impact the market by selling at the bid, for example. And when the prices moved down sharply, it would stop completely.

The buyers of those, contracts, however, was *(were?)* not so careful when it came to selling what they had accumulated. Rather than making sure the sale would not impact the market, the did quite the opposite: they slammed the market with 2,000 or more contracts as fast as they could. The scale was furious, it would often clear out the entire 10 levels of depth before the offer price could adjust downward. As time passed, the aggressiveness only increased, with these violent selling events occurring more often, until finally the e-Mini circuit breaker kicked in and paused trading for 5 seconds, ending the market slide.

Because of the arbitration *(arbitrage)*, when the e-Mini changes price with high volume, many ETF's are repriced (quotes updated, traders executed). The component stocks of ETF's are also repriced, along with many indexes. And finally, all the option chains for the ETF's, their components and indexes are also repriced. The entire system simply cannot absorb the impact of the sudden move in the e-Mini on high volume. A sale (or purchase) of 2,000+ contracts which rips through one-side of the depth of book in 50-100 milliseconds, will immediately overload many systems. The impact reverberates for a much longer period of time than the sell (or buy) event itself.

The first large e-Mini sale slammed the market at approximately 14:42:44.075, which caused an explosion of quotes and trades in ETF's, equities, indexes and options - all occurring about 20 milliseconds later (about the time it takes information to travel from Chicago to New York). This surge in activity almost immediately saturated or slowed down every system that processes this information; some more than others. The next sell event came just 4 seconds later at 14:42:48, which was not enough time for many systems to recover from the shock of the first event. This was the beginning of the freak sell-off which became known as the flash crash.

[19] explain

In summary, the buyers of the Waddell & Reed e-Mini contracts, transformed a passive, low impact event, into series of large, intense bursts of market impacting events which overloaded the system. The SEC report uses an analogy of hot-potato. We think it was more like a game of dodge-b all among first-graders, with a few eighth-graders mixed in. When the eighth-graders got the ball, everyone cleared the deck out of panic and fear.

4 Criticism

- Getting a look at orders before someone else does, is commonly called ”cheating”. The National Market System (NMS) was supposed to prevent that; this was the so-called *”innovation”* of Nasdaq, remember? No specialists, no balancing of orders to open a stock, all done by a computer. Equality of access. Up until it became profitable to make some people more equal. The **intent** of a public stock exchange is to insure equality of access to information so that the markets are orderly, not rigged.

- Using flash order information (or anything else) to front-run is **illegal**. In all of its forms, this is an extremely serious matter and it must be stopped.

- To the extent that these HFT systems are in fact using flash (or other) traffic to gent in front of orders and advantage themselves **they are dramatically increasing the violence of market moves**. A stock trading at $20 that has a bid come in with a limit of $20.10 would normally fill (assuming sufficient depth) at $20; this does not materially move the market. But if a HFT system ”sees” the order, steps in front of it and buys up all the shares at $20 and then re-sells them to the customer at $20.04 (one penny better than the next best offer at $20.05) it has **caused** the current ”last” price to move where it otherwise would not. Multiply this by millions of shares an hour and the impact on price moves could be tremendous.

- HFT systems that front-run are able to garner **risk-free** profits. This is in fact the reason such a practice is banned - their ”risk-free” profit **is your guaranteed loss**. The markets are in fact a negative-sum game (due to trading costs) - if there is a ”risk-free” opportunity out there it can only exist because someone else is guaranteed a loss.

Public and fair markets demand transparency. All users must obtain access to order flow at the same time, without exception, and attempts to ”step i front of the line” must be met with both civil and criminal sanctions for market manipulation.

Three relatively-minor changes that would leave those who are using HFT legitimately unharmed but would destroy most of the ability to cheat. These are:

- Eliminate the *”flash order”* entirely. All market participants must get order and flow information at the same time - no exceptions.

- Force all order (e.g. Immediate or Cancel - IOC, etc.) to be valid for a reasonable minimum period that allows **humans** response. 1 second would meet this criteria; it would destroy the ability of the "robots" to use abusive patterns without preventing the legitimate use of "immediate or cancel" orders. The time selected *must be greater than the average human reaction time plus round-trip network transit time within the nation*; visual recognition time for young adults averages a bit over 200 milliseconds (0.2 seconds) exclusive of the response (e.g. a mouse click) and round-trip transit time on high-speed circuits cross-country (corner-to-corner) is approximately 100ms. Thus the minimum acceptable time in the neighborhood of 500ms assuming no intervening computer computational delays (e.g. brokerage servers, etc.); doubling this to provide for a margin (not all people are 20 years old, there are typically multiple computers between exchanges and end user, charting or display software requires time to post the event on screen, etc.) seems reasonable.

- Define as "front running" by law **any** scheme or practice that exposes or discovers orders to any select group of players before the market as a whole, irrespective of how. The unfortunate reality is that there is no mechanism available to prevent computers from exploiting asymmetric information; ergo, you must define the provision or discovery and use of any such asymmetric information **in the public markets** as a criminal offense. Penalties should include treble forfeiture of all profits gained from such abuse and a permanent ban on all access to securities business as well as prison time.

5 High Frequency Trading Strategies

5.1 "Scalping"

In this section we will describe a short term trading strategy called *scalping*. Scalping is a modern version of momentum strategy, a strategy that profits from abrupt, short term directional price changes. In this strategy traders examines the order book and they search for limit orders, either buy or sell limit orders, in the order book with few, relative to other limit orders, shares offered to be bought or sold. We will describe this on a fictions order book; let

$$b_t^k = (p_t^k, s_t^k) \in \mathbb{R}_+^2$$

denote a buy limit order at time t, that has $k = 1, \ldots, n$ buy limit orders at lower price than it, p_t^k denotes the price of the limit order and s_t^k denotes the size of the limit order. For example, a best buy limit order will be denoted by $b_t^0 = (p_t^0, s_t^0)$, the second best buy limit order is denoted by $b_t^1 = (p_t^1, s_t^1)$ etc. Similarly, a sell order will be denoted as

$$a_t^k = (p_t^k, s_t^k) \in \mathbb{R}_+^2,$$

a_t^0 will denote a best sell limit order, the second sell limit order is denoted by a_t^1 etc.
Let $a_t^0 = (100.00, 900)$, $a_t^1 = (100.01, 100)$, $a_t^2 = (100.02, 1000)$ be the resting sell orders in the order book and $b_t^0 = (99.99, 1000)$ the best buying order in the order book.
One can immediately notice that there is an sell order a_t^1 that has few share to sell - only 100 shares - relative to other sell order, the corresponding ratios of shares to be sold, are $\frac{1}{9}, \frac{1}{10}$. The traders strategy is to buy all of the shares up to price 100.01 and post a sell order of 1000 shares at price 100.01. That way his sell order will become the best sell order in the order book. If the trader is able to sell his 1000 shares at price 100.01 (he

bought 900 shares at price 100.00 and 100 shares at price 100.01), he will earn \$0.009 per share, i.e. 0.9 penny's per share.

This simple strategy seems risky since the trader, after acquiring the shares, posts a limit order that does not have to filled. There are several way a traders can hedge[20] in this strategy; the simplest way is to place a stop-loss, that would ensure the trader, in case of an adverse move he would limit his losses. One can hedge himself by buying an put option on that stock or selling an call option. Due to different trading mechanisms in the equity and options markets, e.g. different tick size, time decay of options, these type of hedging is very difficult to implement efficiently. Another possibility is to buy or sell a "highly" correlated stock, again that hedging strategy itself is highly risky, due to microstructural issues - Epps effect and the fact correlation is variable. A recommended hedging strategy is to hedge with the *same* security. For instance, if the trader, like in our example bought 1000 shares he could sell 300 shares to hedge his position. The ratio of bought and sold shares in our example is $\frac{3}{10} = 0.3$.

We will now describe how would a high frequency trader perform a scalping strategy on the order book previously mentioned; first a trader notices which shares he wants to buy and what is his target selling price

1. trader would buy 900 shares at price 100.00 and 100 shares at price 100.01, total cost will be $900 \cdot 100.00 + 100 \cdot 100.01 = 100001.00$,

2. trader wants to sell the 1000 shares he acquired, for the price of 100.01 and conditioned on him selling these shares, he would receive $1000 \cdot 100.01 = 100010.00$,

3. profit is $100010.00 - 100001.00 \to \$9.00$.

Digression: It is a legitimate question to ask, why should a trader buy 100 shares at price 100.01, just to resell them at the same price, but now as a part of a block of 1000 shares! The first reason is the *time priority* of orders in the order book, i.e. if two orders with the same price are in the order book, the one that arrived earlier has the priority in the execution. The second reason is due to market perception; a lot of market activity is triggered by a printing of a price. For instance, someone might have a *stop-loss* at 100.01 and our buying the 100 shares at that price, might trigger someone to be forced to buy at that price - something that would work in our favor; or a buy algorithm might be initiated by the printing of the mentioned price.

Now that the high frequency trader has identified which shares he wants to buy and at which price he wants to sell them, contrary to the "common" way of thinking, he will first post a sell limit order of 1000 shares at price 100.01 - at this point he still hasn't acquired the 1000 shares he just posted to sell! Next, the trader short-sells[21], for example 300 shares at price 99.99 - note that this is his hedge. If the market moves against him i.e. the price moves downward to e.g. 99.95, he would limit his losses - if he has acquired the 1000 shares, or profit - if he didn't acquired the 1000 shares at the time of the adverse price movement.

When the traders has placed his sell order of 1000 shares at price 100.01 and short-sold

[20] the term "hedge" (or "hedging" as a verb) generally means - *"protecting yourself against adverse price movements"*

[21] short selling (also known as shorting or going short) is the practice of selling assets, usually securities, that have been borrowed from a third party (usually a broker) with the intention of buying identical assets back at a later date to return to the lender

300 shares at price 99.99 he can start buying the 900 shares for price 100.00 and 100 shares for price 100.01. He will have 1000 shares, a best selling order of 1000 shares at price 100.01 and a short position of 300 shares.
Notice that, for the hedge to be effective the high frequency trader has to be able to simultaneously have a long and short position in the same security at the same time! In a sense, there have to be two "virtual" traders in order to avoid *position neutralization*, otherwise owning 1000 shares and short-selling 300 shares would result in owning 700 shares.
In case of positive outcome, the trader would sell the 1000 shares for 100.01 earning $9.00, finally he would need to unwind his short position - his hedge - by posting a buy limit order of 300 shares at price 99.99.

5.2 Low Latency Arbitrage

Low latency arbitrage is a widely used term that describes the ability of high frequency traders to get access to the orders and quotes from exchanges before the general public does so. High frequency traders to so in two ways; they *co-locate* their computers as close as they possibly can to the electronic exchanges that execute their trades, and they pay exchanges to give them actual price and order information before the raw data gets conslidated and disseminated to other market participants.

Here is an example[22] of how latency arbitrage works;
Suppose the National Best Bid and Offer for stock WXYZ is \$ 19.99(bid) - \$ 20.00(ask). Assume 1000 shares of WXYZ are offered on INET[23] at \$20.00 and 1000 shares are also offered on ARCA[24] at \$20.00. An investment company wishes to buy 3000 shares of WXYZ. He decides to sent a limit buy order for 1500 shares at \$20.00 to ARCA and 1500 shares at \$20.00 to INET. The orders of an investment company are crossed against the posted orders at INET and ARCA, resulting in trades of 1000 shares at each exchange. The balance of the orders of an investment company - 500 shares at each exchange - should be posted as the new best buying offer at \$20.00 at each exchange. Because the orders of an investment company was the first bid of \$20.00 to arrive at both exchanges, it should be the first order to post, i.e. it should have a *time priority* over all other orders at the same price. Unfortunately, this will typically not happen. Here is the reason:

- UTP[25] Quotation Data Feed (UQDF) - the Securities Information Processor used for Regulation NMS[26] compliance by both exchanges, is slow and will still show the obsolete \$20.00 offer that investments company order removed from both exchanges;

- because UQDF is still showing a \$20.00 offer at ARCA, INET will not allow the investment's company order to post, because it would *lock the market*[27];

[22] example is taken from: "Public Commentary on SEC Market Structure Concept Release", Tradeworx Inc.

[23] INET is an Electronic Communication Network (ECN), an electronic system that attempts to eliminate the role of third party in the execution of orders

[24] NYSE ARCA is a securities exchange on which stocks and options are traded

[25] Unlisted trading privileges

[26] National Market System

[27] Market is said to be *locked* when the best buying order equals best selling order

- because UQDF is still showing a \$20.00 offer at INET, ARCA will not allow the investment's company order to post either, for the same reason as mentioned previously.

In the mean time, any high frequency trader with direct feeds to both exchanges will notice that the offer is gone, but is still displayed on UQDF. Many such high frequency traders will rush to from a new \$20.00 bid, and will circumvent the Order Protection Rule[28] by sending an Inter-market Sweep Order (ISO)[29]. An investment company cannot use an ISO order because it is not a broker-dealer[30]. Most executing brokers do not allow their non-broker/dealer customers to utilize ISO order, because compliance with Regulation NMS can not easily be verified on a pre-trade basis. INET's policy will be to post the investment company's order immediately, but to make it a *hidden order*[31] so that it does not lock the market. ARCA's policy will be to do the same thing, but make the investment's company order visible after the SIP has updated to show the real price. In both cases, the investment company will be **behind** the high frequency trader who sent an ISO orders in priority, even though the investment's company order arrived at both exchanges first!
Here is why this issues matter:

- In a price-time priority market, orders that are at the front of the queue experience the *least* adverse selection, and orders that are at back of the queue receive the *most* adverse selection;
- This is obvious, because if you are the last one to buy on the bid, that means the bid is about to become the new offer. Conversely, if many people are behind you on the bid, that means the bid is likely to hold after you trade;
- Empirically, there is a \$0.017 per share difference in profitability for a posted share that is first in line versus one which is last in line;
- This results in tens of millions of dollars (conservative estimation) of extra trading *costs* for investors and *profits* for the high frequency traders.

The following is a **real-world** example (from Tradeworx Inc.[32] trading activity) of violations in price/time priority caused by the Rule 611 or Regulation NMS. Such violations occur with tremendous regularity throughout the trading day, as an unintended but

[28] Regulation NMS ensures that investors receive an execution price that is equivalent to what is being quoted on any other exchanges where the security is being traded. The order protection rule requires that each exchange establish and enforce policies to ensure consistent price quotation for all NMS stocks, which include those on major stock exchanges as well as many over-the-counter (OTC) stock. The order protection rule is also known as *Rule 611*, or the *trade through rule*

[29] Inter-market sweep order is a limit order designated for automatic execution in a specific market center even when another market center is publishing a better quotation. When send an Inter-market Sweep Order, the sender fulfils Regulation NMS order-protection obligations and NYSE Rules by concurrently sending order to market centers with better prices. These orders are not subject to auto-routing and must be marked with a trader indicator of "F"

[30] Broker-dealer is a company that trades securities for its own account or on behalf of its customers

[31] Hidden orders allow users to hide the limit orders on the order book. They have lower priority than visible orders at the same price level

[32] Tradeworx Inc. is a high frequency trading firm

direct consequence of the ban against *locked markets*.

EXAMPLE: exchange = NASDAQ, date = 02/17/2010, ticker = SPY[33]
Tradeworx Inc. sent a sell order for 1643 shares, trying to hit the buying order at $110.18 and post an selling order there.

<timestamp>	<event>
44740149566	Tradeworx actively fill quote id#151913900, 100 shares, $110.18
44740149566	Tradeworx actively fill quote id#151914097, 200 shares, $110.18
44740149566	Tradeworx actively fill quote id#151915176, 300 shares, $110.18
44740149566	Tradeworx actively fill quote id#151915775, 1000 shares, $110.18
44740149566	Tradeworx post remaining 43 shares as offer - order was repriced to $110.19 to comply with Reg. NMS, quote ID #151918883
44740152636	quote ID #151919503 posts offer at $110.18
44740159434	quote ID #151919503 is filled by incoming buy order at $110.18
44740188406	Tradeworx cancel the order, which was not filled

Let us look at an another example of how latency arbitrageurs make their money. According to the *The Wall Street Journal*, TSF Capital, a $1.1 billion firm that trades for mutual funds and is among those losing out to latency arbitrageurs, decided to conduct a trade to illustrate how it is getting ripped off by them. The *Journal* reports that in March, 2010, a TSF trader sent an order to buy shares of Nordson (NDSN) through an instant message requesting that the order be executed in a specific Dark pool. Dark Pools are "unregulated" alternative electronic stock exchanges, in many cases run by big Wall Street banks, in which buyers and sellers show up anonymously to declare their interest in buying and selling a certain number of shares of stock within some price limit.
The TSF trader asked the broker to execute the order in "broker pool #2", telling the broker not to pay more "than the midpoint between what buyers and sellers where offering, which at the time was $70,49" - most like the *pricing* in that particular Dark Pool was done at midpoint price, according to the *Journal.* But the market price for Nordson shares did not change for a few seconds so the TSF trader "set a trap: He sent a separate order into the broader market to sell Nordson for a price that pushed the midpoint price down to $70,47".
TSF was "almost immediately sold Nordson for $70,49 - the old, higher midpoint - in broker pool No.2, which did not reflect the new sell order", according to the *Journal.* Perhaps what happened was that a latency arbitrageur was able to buy the shares at $70,47 in the broader market and sell them to TSF for two cents more.
Most likely the following happened; high frequency trader having seen that the new best buy order was placed in the broad market - the one that TSF trader placed, should push the midpoint price at the Dark Pool higher, but is still not updated due to slowness of SIP and quotation system at Dark Pool, sent an Immediate or Cancel (IOC) Order[34] to probe if there is any interest in Dark Pool for the older, but not yet changed price of $70,49; if this order is matched he can keep sending IOC orders until there is no interest.

[33] The SPDR© S&P©500 Exchange Traded Fund (Ticker: SPY) is a fund that, before expenses, generally corresponds to the price and yield performance of the S&P500© Index (Ticker: SPTR).

[34]

How the high frequency trader gets rid of the shares he acquired is a real mystery.
(How the latency arbitrageur knew that the higher price is a mystery. But it may have been due to practices of pinging the dark pool. This is the practice of sending a series of small order to the Dark Pool to see if the latency arbitrageur can guess the price at which, say, the seller wants to sell.)
In this case, TSF overpaid by two cents a share for its Nordson stock. All these pennies add up to $3 billion a year that should have gone into the pockets or retail investors but instead, help provide those billion-dollar annual payday for hedge funds.

5.3 Liquidity Rebate Trading

Liquidity rebate trading involves taking advantage of certain "rebates" that some exchanges offer trading firms that are willing to step up and provide shares when needed. Market centers such as the stock exchanges and Electronic Communication Networks (ECN's) offer these rebates in order to attract trading volume[35] Such rebates can range from one quarter to third of a penny[36]. Liquidity rebate traders make a profit by looking for large order flows and then filling a part of a large order, then re-offering , the shares at the same price and collecting the exchange fees for providing liquidity to the market[37]. If the order is filled, the market center pays the broker dealer a rebate and charges a larger amount to the broker dealer who took liquidity away from the market. This has led to trading strategies solely designed to obtain liquidity rebates[38]. Liquidity rebate trading has been scrutinized because it allows rebate traders to basically trader for free by having their commission costs and exchange fees covered by the exchanges and ECN's.
The following is an excerpt from ...."Playing Fair", where Joe Saluzzi and Sal Arnuk provide a real world example of difficulties they econture with rebate trading;
Sal:*...But all the rebate trading just distorts the market. Let me give you a real world example. For one customer of ours, we were buying a stock. We had to buy probably 30,000 or 40,000 shares, which is not very big order, but it is very big order when you consider that the stock trades 5,000-6,000 shares a day. Well, as soon I displayed my first bit of liquidity, I started a chain of events. People stepped in front of me and then someone stepped in front of them. So I cancelled and walked away and said, "Okay, this is not the way to do it. We have to think about this". But while I adjusted the way we were going to play the stock, these two guys - without doing one single trade - and I say "two guys" but I mean the high frequency traders jockeying the quote - changed their quotes 1,600 times in a period of 20 minutes, alternating around the NBBO.*
Joe: *And how many shares traded?*
Sal: *Zero traded.*

James Angel of Georgetown, Lawrence Harris of University of Southern California and Chester Spatt of Carnegie Mellon. According to their paper, "Equity Trading in the 21st Century", the maker/taker model has "...Distorted order routing decisions, aggravated agency problems among brokers and their clients, unleveled the playing filed among

[35] Take a look at Arnuk& Saluzzi
[36] See Hutchinson, Patterson&Rogow
[37] see Hutchinson
[38] see Arnuk&Saluzzi

dealers and exchange trading systems, produced fraudulent trades, and produced quoted spreads that do not represent actual trading costs.”

5.4 Flash Trading

Another concept tied to liquidity trading is know as *flash trading* or issuing flash orders. A flash order is a trade based on access to information for a matter of milliseconds that is not yet public. Flash orders developed due to the competition that exchanges face over the volume of shares posted on their platforms. In order to further encourage trading on their platforms, some exchanges, such as the NASDAQ (up until September, 2009), allow firms to get a 30 millisecond peak at orders before they get sent to other markets. The result is that...

a trading firm can keep its order on a certain exchange for up to half a second without matching an existing buy or sell order on another exchange, a move that puts in a position of poster, rather tan responder. The hope is that another trader who needs to buy or sell quickly steps in on the other side of the trade...[a] dynamic that boosts the chance that flash-orders trader will complete the trade on the exchange and get the rebate.

Traders who benefit from the use of flash orders are shown the buy and sell orders ahead of everyone else in the marketplace in exchange for a fee. With this very small advance notice of market conditions, high frequency traders can use their super-computer to conduct rapid statistical analysis of the changing market state and trade ahead of the public market.

The use of flash orders in automated trading was virtually unheard of until last year when financial blogs started criticizing firms such as Goldman Sachs for gaining unfair profits through the practice. Recently, there has been a great deal of controversy over this practice and the SEC has proposed an outright ban on the use of flash orders.

Keiser isn’t just speculating on this. He claims to have invented one of the most widely used programs for doing the rigging the market. Not that that’s what he meant to invent. His patented program was designed to take the manipulation *out* of the market. It would do this by matching buyers with sellers automatically, eliminating *”front running”* - brokers buying or selling ahead of large orders coming in from their clients. The computer program was intended to remove the conflict of interest that exists when brokers who match buyers with seller are also selling from their own accounts. But the program fell into the wrong hands and became the prototype for automated trading programs that actually *facilitate* front running.

When *”market making”* (matching buyers with sellers) was done strictly by human brokers on the floor of the stock exchange, manipulations and front running were considered an acceptable (if morally dubious) price to pay for continuously *”liquid”* markets. But front running by computer, using complex trading programs, is an entirely different species of fraud. A minor flaw in the system has morphed into a monster. Keiser maintains that computerized front running with HFT has become the principal business of Wall Street and the primary force driving most of the volume on exchanges, contributing not only to a large portion of trading profits but to the manipulation of markets for economic and political ends.

5.5 The *"Virtual Specialist"*: The Prototype for High Frequency Trading

Until recently, most market making was done by brokers called *"specialists"*, those people you see on the floor of the New York Stock Exchange haggling over the price of stocks. The job of the specialist originated over a century ago, when the need was recognized for a system of continuous trading. That mean trading even when there was no *"real"* buyer or seller willing to take the other side of the trade.

The specialist is a broker who deals in a specific stock and remains at one location on the floor holding and inventory of it. He posts the "bid" and the "ask" prices, manages "limit" orders, executes trades and is responsible for managing the uninterrupted flow of orders. If there is a large shift in demand on the "buy" side or the "sell" side, the specialist steps in and sells or buys out of his own inventory to meet demand, until the gap has narrowed.

This gives him an opportunity to trade for himself, using his inside knowledge to book a profit. That practice is frowned on by the Securities Exchange Commission (SEC), but it has never been seriously regulated, because it has been considered necessary to keep markets "liquid".

Keiser's "Virtual Specialist Technology" (VST) was developed for the Hollywood Stock Exchange (HSX), a web-based, multiplayer simulation in which players use virtual money to buy and sell "shares" of actors, directors, upcoming films, and film-related options. The program determines the true market price automatically, by comparing "bids" with "asks" and weighting the proportion of each. Keiser and HSX co-founder Michael Burns applied for a patent for a "computer-implemented securities trading system with a virtual specialist functioN" in 1996, and U.S. patent no. 5960176 was awarded in 1999.

But things went awry after the dotcom crash, when Keiser's company HSX Holdings sold the VST patent to investment firm Cantor FItzgerald, over his objection. Cantor Fitzgeral then put of the program that would have eliminated front-running on ice, just as drug companies buy up competing patents in order to take them off the market. Instead of preventing front-running, the program was altered so that it actually enhanced the fraudulent practice. Keiser notes that this sort of patent abuse is illegal under European Intellectual Property Law.

Meanwhile, the design of VST program remained on display at the patent office, giving other inventors ideas. To get a patent, applicants must list "prior art" and then prove that their patent is an improvement in some way. The listing of Keiser's patent shows that it has been referenced by 132 others involving automated program trading or HFT.

Since then, HFT has quickly come to dominate the exchanges. High frequency trading firms now account for 73% of all U.S. equity trades, although they represent only 2% of the approximately 20.000 firms in operation.

In 1998, the SEC allowed online electronic communication networks, or alternative trading systems, to become full-fledged stock exchanges. Alternative trading systems (ATS) are computer-automated order-matching systems that offer exchange-like trading opportunities at lower costs but are often subject to lower disclosure requirements and different trading rules. Computer systems automatically match buy and sell orders that were themselves submitted through computers. Market making that was once done with a "specialist's book" - something that could be examined and audited - is now done by an unseen, unaudited "black box".

Alternative trading systems allow just about any sizable trader to place orders directly in the market, rather than routing them through investment dealers on the NYSE. They also allow any sizable trader with a sophisticated HFT program to front run trades.

5.6 Flash Trades: How the Game Is Rigged

An integral component of computerized front running is a dubious practice called "flash orders". Flash orders are permitted by a regulatory loophole that allows exchanges to show order to same traders ahead of others for a fee. At one time, the NYSE allowed specialists to benefit from an advance look at incoming orders, but it has now replaced that practice with a "level playing field" policy that gives all investors equal access to all price quotes. Some ATS's, however, which are hotly competing with the established exchanges for business, have adopted the use of flash orders to pull trading business away from the exchanges. An incoming order is revealed (or flashed) to a trader for a fraction of a second before being sent to the national market system. If the trader can match the best bid or offer in the system, he can then pick up that order before the rest of the markets sees it.

The flash peek reveals the trade coming in but not the limit price - the maximum price at which the buyer or seller is willing to trade. This is what the HFT program figures out, and it is what gives the high-frequency trader the same sort of inside information available to the traditional market maker:he now gets to peek at the other player's cards. That means high-frequency traders can do more than just skim hefty profits from other investors. They can actually manipulate the markets.

In fact, high frequency traders may be *removing* liquidity from the market. So argues John Daly in the U.K. *Globe and Mail* citing Thomas Caldwell, CEO of Caldwell Securities Ltd.: *Large institutional investors know that if they start trying to push through a large block of shares at a certain price - even if the block is broken into many small trades on several ATSs and markets - they can trigger a flood of high frequency orders that immediately move the market prices to the institution's disadvantage...That's why institutions have flocked to so-called dark pools operated by ATSs such as Instinet, and individual dealers like Goldman Sachs. The pools allow traders to offer prices without publicly revealing their identities and tipping their hand.*

Because these large, dark pools are opaque to other investors and to regulators, they inhibit the free and fair trade that depends on open and transparent auction markets to work.

A simple question comes out by itself; why doesn't SEC ban flash orders? In order to understand that, must go back to Electronic Communication Networks, rebates, Alternative Trading Systems and Regulation NMS.

Access fees originated in the world of Electronic Communication Networks, or ECN's, which in reality are electronic stock exchanges. Under the Securities Exchange Act of 1934, national securities exchanges are required to be self-regulatory organizations[39] (SRO's), a function that magnifies the cost of operating any stock exchange. The costs of maintaining SRO function was a barrier to entry that would have prevented ECN's arising in late 1980's from competing with the New York and American Stock Exchanges, as well as with NASDAQ, which at the time was regulated as a "securities information processor". The SEC believed that ECN's offered an innovative alternative to the floor-based models of the New York and American Stock Exchanges and to the market maker business model employed by NASDAQ. So, the SEC allowed ECN's to register as broker-dealers, but granted them a host of exceptions because many broker-dealers regulations made no sense when applied to the exchange business. Eventually, ECN's and other variations on the electronic stock exchange them became known as Alternative Trading Systems and the exceptions to broker-dealers regulations were defined and codified as Regulation ATS[40]. Alternative Trading System is a broker-dealer operating under the exceptions governed in Regulation ATS. An ATS that cannot satisfy the conditions of Regulation ATS must register as a national securities exchange.

ECN's originally operated as a closed systems. Only subscribers were permitted to post or take quotations, and subscriber agreements established a transaction-based fee schedule. In that sense, the ECN business model was similar to stock exchanges, which also charged transaction-based fees. ECN's were open to the members of the public, and most importantly, to institutional investors who otherwise were required to access public liquidity through broker-dealer members of various registered exchanges.

Broker-dealers soon discovered that ECN's were a convenient way to display and access institutional liquidity at prices that improved the quotation displayed in exchanges. As a result, something of a two-tier market developed, with institutional and broker-dealers orders being displayed in ECN's at prices that were superior to those displayed in the public markets. Retail order flow generally did not have access to the ECN prices. The SEC views two-tier markets as inconsistent with the objectives of the Exchange Act. Therefore, it eventually required broker-dealers to provide the same quotes in the public markets that they displayed in ECN's.

To avoid losing the broker-dealer business, most ECN's began displaying their quotations in the public markets and provided their broker-dealer clients with useful systems tools that enabled them to provide consistent quotations in ECN's and public markets. However, this also meant that, for the first time, quotations in ECN's could be accessed by broker-dealers that were not subscribers. Broker-dealers, or course, will no pay for something that can be obtained for free. ECN's responded by sending bills to broker-dealers that accessed the ECN quotations in public venues and when some large broker-dealers refused to pay, interesting litigation commenced. Both sides had a point, which is usually the source of litigations. Broker-dealers pointed out that they had never agreed to

[39] write some thing about SRO's

[40] Alternative Trading System

subscribe to ECN's services. Moreover, broker-dealers complained that they were not permitted to charge access fees when persons who were not their customers accessed their public quotes. ECN's argued that, in absence of access fees, no one would pay for use of their services, which would deprive them of any opportunity to make a profit.

Some ECN's hiked up fees for accessing their public quotes and then provided rebates to their subscriber for posting quotes. This practice essentially required non-customers to pay for the service. A business is not required to treat persons who are not its customers fairly. So, something of an arms race developed with access fees increasing periodically as one ECN tried to capture business from others.

This madness was based on conceptual fallacy. ECN's really were not broker-dealers; instead, they were exchanges and should have been regulated accordingly. Exchanges traditionally have charged their members fees for accessing exchange trading services. Persons who are not members can only obtain access to the exchange through the facilities of members. Confusing the broker-dealer and exchange model led to unanticipated consequences.

It is also true that traditional exchanges, protected from competition by regulations that required them to be SRO's, failed to modernize by innovation in trading systems. ECN's introduced a breath of fresh air into an antiquated system, which the SEC found irresistable.

Eventually, this road led to Regulation NMS. To level the playing field, all regulated participants - exchanges, ATS's and broker-dealers - were permitted to charge access fees to non-customers, but the amount that could be charged was capped. Among other things, this forced the transformation of the NYSE into a huge ECN forced to compete on a level playing filed with other exchanges. It also led surviving ECN's to merge with exchanges or to register as national securities exchanges.

It is a nice conceit to imagine that competition spurs innovation, lower prices and increasing social welfare. But innovation and lower prices are not the only ways to compete, and some methods of competition do not increase social welfare. The flash order in an example of competition's dark side.

To attract customers to their systems, some exchanges and a few of the remaining ECN's exploited a loophole in Regulation NMS that permitted favored customers to display their orders to a select group, rather than to the unwashed masses. The loophole recognized that a person owning the security can always decide that she will only sell that security privately to certain people. But if she wants access to the broader array of purchasers in the public markets, the price of that access is that quotes entered in the public markets must be available to all participants.

So, the NYSE complained about flash orders, Congress got involved, and the SEC proposed to close the loophole.

The difficulty is that closing the particular loophole would affect the options markets, which so far have no cap on access fees. In the options market, a "maker-taker" model exists that involves charging access fees and using them to pay rebates to "liquidity providers" in much the same way that was so popular with ECN's prior to Regulation NMS. As a result, the SEC received a lot of comments from options players pointing out that flash orders are used to avoid having orders routed to options exchanges charging higher access fees. The proposal to abolish flash orders is therefore being delayed while the SEC figures out how to deal with access fees at options exchanges.

"...Third, the SEC should ban the use of so-called "flash orders" by high frequency traders. Flash orders allow exchange members who pay a fee to get a first look at share order flows before the general public. By viewing this buy and sell order information for just milliseconds before it goes to the wider market, these "investors" (better said - traders) gain an unfair advantage over the rest.
As the New York Stock Exchange complained to the SEC on May 28, selling flash orders for a fee provides "non-public order information to a select class of market participants at the expense of a free and open market system." To use a baseball metaphor, flash orders allow some batters to pay to see the catcher's signals to the pitcher, while the rest of us don't see them. Market that permit a privileged few to have special access to information cannot maintain their credibility.
Amazingly, it is a loophole in current regulations that allow this unfair practice. This can and should be fixed immediately..."

Charles Schumer, the third-ranking Democrat in the U.S. Senate, asked the Securities and Exchange Commission to ban so-called flash orders for stocks(!!!), saying they give high-speed traders an unfair advantage.
Schumer's letter to SEC Chairman Mary Schapiro yesterday raised the stakes in a debate over the practice offered by NASDAQ OMX Group Inc., Bats Global Markets and Direct Edge Holdings LLC, which handle more than two-thirds of shares traded in the U.S. With flash orders, exchanges wait up to half a second before they publish bids and offers on competing platforms, giving their own customers an opportunity to gauge demand before other traders.
(I have to mention the response by Bats Chairman in Advanced Trading about flash orders!)
"This kind of unfair access seriously compromises the integrity of our markets and creates a two-tiered system, where a privileged group of insiders receives preferential treatment", Schumer wrote in his letter.
Flash orders make up less than 4 percent of U.S. stock trading, according to DirectEdge and Bats. They have drawn criticism from the Securities Industry and Financial Markets Association, which is Wall Street's main lobbying group, and GETCO LLC, one of the biggest firms that uses high-frequency trading strategies to make markets in stocks and options. NYSE Euronext, owner of the world's largest exchange by the value of companies it lists, told the SEC in May that the technique results in investors getting worse prices.
Schumer, a member of the Senate Banking Committee, said he will introduce legislation to ban flash orders if the SEC doesn't act on his request.
NASDAQ and BATS stop flash trading.
"The war on high frequency trading continues apace - and already there have been some high-profile casualties, or at the very least, capitulations. Both Nasdaq and BATS said on ...insert-date... they will stop offering flash trading - mere month after they initiated the service on June 3."
The U.S. Securities and Exchange Commission proposed banning flash orders after lawmakers said the practice may give hedge funds an advantage over other investors.
SEC commissioners unanimously voted today to seek public comment on a rule barring exchanges and trading platforms from giving clients access to information about stock orders a fraction of a second before the market.

"Investors that have access only to information displayed as public quotes may be harmful if market participants are able to flash orders and avoid the need to make the orders publicly available", Chairman Mary Shapiro said.
Democratic Senators Charles Schumer and Ted Kaufman urged the commission to halt the practice, arguing frequent traders use technology to profit from access to information not available to retail investors. Direct Edge Holding LLC has relied on flash orders to take market share from NYSE Euronext.
Nasdaq OMX Group Inc. and Bats Global Market voluntarily dropped the flash orders ...insert time... after the practice drew scrutiny from Congress and SEC.
The SEC's proposed ban requires a second vote at a larger public meeting to become binding.

5.7 Flash Orders

BATS would support a ban on flashed orders based on the rational concerns listed in the July 7th newsletter. If a ban would relieve pressure and give the industry the pause necessary to review and reconsider flashed order functionality, then "let's" collectively go that route. NASDAQ, DirectEdge, CBSX are you open to coordinated approach to withdrawing flashed orders? "We" are.

At the same time, we would like to point out that some of the recently hyped allegations surrounding flashed orders are profound, and allude to seemingly scandalous behavior. Four such allegation regarding flashed orders are discussed below:

Misstatement 1: Flash orders create unique front-running opportunities.

Response: BATS BOLT orders do not create unique front-running opportunities.

Longer explanation: We think it is disingenuous to suggest that orders that are widely published to an exchange's full membership, distributed to several market data vendors, and even displayed in real time on several mainstream financial websites, could create meaningful front-running opportunities. That said, we recognize that illegal front-running can occur in a variety of contexts, and nothing about BOLT orders raises unique issues in this regard. Importantly, exchanges and member firms have responsibility to detect, and take enforcement action against, illegal front-running, regardless of the context in which it occurs.

Misstatement 2: Flash orders are used solely by high speed trading firms.

Response: BATS BOLT orders are used by a wide variety of firms.

Longer explanation: The types of customers using this functionality are broad and encompass all kinds of firms, from retail brokers, to institutional participants, to proprietary trading firms, to automated market markers, to agency brokers, and so on. There is no single profile that describes firms that are using the functionality.

Misstatement 3: Flashed orders disadvantage retail investors.

Response: BATS BOLT orders benefit traders and investors, including retail investors.

Explanation: Flashed orders typically provide price improvement to investors to execute against them. While orders are being flashed, they are effectively compressing the spread between the bid and the ask down to zero. For example, stock XYZ might be quoted in the national market as $23.10 bid x $23.11 offered. A BOLT enabled order first executes against all available bids (or offers) at BATS and then is disseminated to all BATS members as a quote one penny better than was available before the order came in. In the previous example, an order to buy at $23.11 would execute against all available orders on BATS to sell at $23.11, and when the book has been totally cleared, the remainder of that order is published as a (better priced) bid at $23.11. Market participants, retail investors included, are then able to benefit from executions against these flashed orders and receive the price improvement they offer.

Misstatement 4: Exchanges intended to flash quotes solely to their own member.

Response: BATS actually prefers to disseminate BOLT orders publicly.

Explanation: A primary driver behind BOLT was to publish better prices quotes as widely as possible given the current regulatory limitations. BATS would prefer that BOLT quotes be publicly disseminated through the consolidated tape. They represent better priced orders and we believe they need to be published widely to the entire industry. Doing so, however, would create temporary conditions where bids are equal to offers and offers are equal to bids (i.e. locker markets), and this is not allowed under the current regulatory environment. All BATS orders, including BATS orders, are published through our market data feeds (which are provided free of charge). BATS quotes, including BOLT quotes, are even published in real time on the Internet through several of the most popular financial portals. The only place where you can't find BOLT orders is on the consolidated tape, and they would post them there too if the regulatory structure allowed it. Simply stated, the BATS implementation made flashed quotes available to as wide an audience as currently allowed by securities law.

Provocative suggestion: If the existing regulatory structure was modified to make locked quotations acceptable, and all quotes (flashed or otherwise) were then reflected in the consolidated tape, then the potential issues associated with flashed orders might find acceptable resolutions. Maybe flashed orders are really just innovative implementation designed to work around the limitation of a regulatory framework that currently seeks to avoid locked markets. It's likely to be yet another hot topic, but one that we propose be revisited in light of how the market has evolved over the last couple of years.

It needs to be said again - there are valid reasons to review, debate, and even ban flashed orders. I listed six of those reasons in our last newsletter. It's equally important to make clear, however, that there are accusations being circulated recently that are distorting the perception how these systems actually work. The difference between the real concerns and the hype may be difficult for people outside of our industry to discern, so we felt it was important to shed some factual light on this topic.

There is grave danger in allowing misstatements and myths to perpetuate, and I hope that we have been able to clear up some of the confusions here. The escalation of misconceptions currently being witnessed is alarming. We need to take a deep breath, look around, and realize that things in the secondary trading markets may not be as broken as they are currently made out to be. Evolving regulatory rule-making and oversight is good (even vital), and we should let the regulatory process work as it was intended. Otherwise we could find ourselves dealing with the unintended consequences of snap reactions, and possibly make things worse rather than better.

...Our position on this debate is clear: the SEC currently deems the BOLT and Flash functionality legal and compliant with the Act. BATS stands ready, however, participate in an industry review of potential issues associated with this functionality. From our perspective, potential issues common to BOLT, Flash and ELP worth further investigation are as follows:

Issue 1: *Public market venues circulating quotes to an exclusive and private network of users.*
There is a possibility of creating a ”two-tiered” market, where the best quotations from specific markets are made available to a limited number of market participants.

Issue 2: *”Price forming” resting orders at other markets being traded around.*
Customers who take the risk of displaying orders in a public market are helping to establish reference prices as a vital part of price discovery process. Under Regulation NMS, such orders that form a market’s Protected Quote are protected from being traded through their limit prices but are not protected from another market trading at the same price. While that protection continues to exist with the implementation of BOLT/Flash/ELP, it is likely that BOLT/Flash/ELP create a greater frequency of instances in which such Protected Quotes are denied an execution they might have otherwise received, thereby creating a disincentive to post aggressive limit order and harming the price discovery process.

Issue 3: *Locked markets in a regulatory structure designed to avoid them.*
In the current Regulation NMS environment, both the spirit and the letter of the law speak in the avoidance of locked markets. While the debate around whether locked markets are good or bad is interesting, it’s irrelevant in an environment where the regulations prohibit it. In all three cases (BOLT/Flash/ELP), depending on you vantage point and definition of ”immediate”, it’s possible to see how a virtual locked market situation is created. During the period in which marketable orders are being exposed to a private network of users, those same orders are effectively priced at a locking price to other displayed markets quotes.

Issue 4: *Disconnected consolidated tape stream that doesn’t reflect the markets’ best prices.*
The consolidated tape has long been the industry reference for the market’s top of the book quotes. It can be used as a benchmark for best execution, and it can be used as

a basis for determining the NBOO (National Best Bid and Best Offer). In BOLT, Flash and ELP, "exposed orders" are not reflected in the consolidated tape, which might create a potentially harmful disconnect in the public quote stream over time.

Issue 5: *Confusion between accessing Protected Quotes versus achieving best execution.*
The general obligation to seek access to "reasonable available" prices that are better than the NBBO is a well established tenet of the best execution obligation. BOLT, Flash and ELP raise questions about whether the better quotations temporarily show to a private network of users are truly "reasonably available" to many brokers handling agency orders. To be compliant with their best execution obligations, are all such brokers required to go to the expense of taking in direct feeds and seeking access to these fleeting better-priced quotations? Or, in doing so are they running the risk that not only will they fail to access better priced quotations but will also miss the published markets for their customers? It may be that firms' experience in seeking to access these quotations will provide the only concrete guide to determining their relevance to satisfaction of best execution obligations, but it seems that this remains open for discussion.

Issue 6: *Increased requirement to take and digest all direct feeds.*
As the previous issues are being considered, a common central theme also emerges. In order to resolve these issues, brokers may discover that they have a new requirement to take and digest all available direct feeds. Without taking each individual direct feed, a broker may not be aware of all the best priced orders in the market. Whereas the consolidated tape may have been adequate previously, a broker may be forced to assume the hassle and expense of subscribing to all direct feeds individually.

As stated above, BATS recognizes the potential benefit of further discussion and debate surrounding the BOLT/Flash/ELP functionality and the associated potential concerns. While it may spark an even stronger and wider debate, we would welcome a discussion around whether locked markets are in fact good for the industry. If locker markets were accepted by the regulatory framework, and **all** quotes were then reflected in the consolidated tape, many, if not all of the issues above might find acceptable resolution.

In conjunction with a regulatory review of the BOLT/Flash/ELP functionality in the equities market, BATS would suggest that the "step-up auction" functionality found in the options market also be reviewed. The common theme in both cases is the definition of "immediate" as a critical point in determining how to best handle marketable orders.

5.8 Automated Market Making

Automated market maker (AMM) firms run trading programs that ostensibly provide liquidity to the NYSE, NASDAQ and Electronic Communication Networks. AMM's are supposed to function like computerized specialists or market makers, stepping in to provide inside buy and sell, to make it easier for retail and institutional investor to trade. AMM's, however, often work counter to real investors. AMM's have a ability to "ping" stock to identify reserve book orders. In pinging, an AMM issues an order ultra fast, and if nothing happens, it cancels it (IOC). But if it is successful, the AMM learns a tremendous amount of hidden information that it can use to its advantage.

To show how this works, this time our institutional trader has input a discretion into the algorithm to buy shares up to $20.03, but nobody in the outside world knows that. First, the AMM spots the institution as an algo order. Next, the AMM starts to ping the algo. The AMM offers 100 shares at $20.05. Nothing happens, and it immediately cancels. It offers $20.04. Nothing happens, and it immediately cancels.

Then it offers $20.03 - and the institutional algo buys. Now, the AMM know it has found a reserve book buyer willing to pay up to $20.03. The AMM quickly goes back to a penny above the institution's original $20.00 bid, buys more shares at $20.01 before the institutional algo can, and then sell those shares to the institution at $20.03.

Automated market makers co-locate their servers in the NASDAQ or the NYSE building, right next to exchanges' servers. AMM's already have faster servers than most institutional and retail investors. But because they are co-located, their servers can react even faster. That's how AMMs can issue IOC order - immediate or cancel - something known as "cancel and replace". They issue the order immediately, and if nothing is there, it is canceled. And that's how AMMs get trades faster than any other investor, even though AMMs are offering the same price. AMMs pay large fees to exchanges to co-locate, but it obviously has a decent return on investment. According to *Traders Magazine*, the number of the firms that co-locate at NASDAQ has doubled over the last year.

5.9 Quote Stuffing

In the wake of the Flash Crash, buy-side traders have expressed concern about high frequency traders flooding the exchanges with orders and then rapidly canceling 95 percent of them. There even is talk of predatory strategies and targeted *denial-of-service (DoS)* attacks, intended to increase high frequency trader's speed advantage over other market participants, being launched against exchanges. While this has largely been refuted by the U.S. exchanges - both NYSE and Direct Edge executives were quoted recently in a Bloomberg News report as saying they did not see a problem here - some sources remain suspicious.

"The SEC is concerned about quote stuffing. Apparently someone has been doing denial-of-service attacks on the exchanges," says Larry Harris, professor of finance and business economics at USC's Marshall School of Business, who was interviewed before the SEC released its final Flash Crash report (related article, page 11). "If you send too many requests for service, you deny other people the opportunity get there in time. TO get 1,500 quotes to buy and have all 1,500 canceled in a second, that really means someone is trying to overwhelm the exchange."

But Tim Mahoney, CEO of BIDS Trading, maintains that quote stuffing did not occur, adding that there is a lot of misinformation in the marketplace. "Suggesting that someone has such good insight into the exchanges - you have to know which symbols are on which servers. You have to be able to target the symbols that are clumpled on the same servers to have the same impact each time and slow them down," he says in explaining what it would take to launch a targeted DoS attact that would not interfere with an HFT firm's own orders. "I have a hard time believing that's the case."

Real or not, FNY's Schenk suggests a remedy to DoS attacks: HFT orders, he says, should be required to stand for a minimum period of time, and HFT firms could be charged a fraction of a penny for each message, which would potentially discourage them from sending to many. According to UCLA's Harris and others, however, these remedies would force high frequency traders to alter their strategies and maintain quotes for longer than they want to, which could constrain the growth of high frequency trading, an area of the markets that is growing and expanding overseas to Europe and Latin America.

Then, of course, there are the market participants who are worried about the unintended consequences of tinkering with the equity structure and who argue that the equity markets are working well. Yes, we have had some blips, they admit - if you consider the Flash Crash, when $862 billion in value was erased in 20 minutes, a "blip" - but over all, the market work well. All this talk of imposing market maker obligations on high frequency traders or placing a tax on HFT shops that blast high volumes of canceled messages isn't necessary, they insist. The fact that the market recovered on May 6th within a few minutes, after prices hit rock bottom, could be a sign the HFT is self-healing, they suggest.

With so may questions about and diverging opinions on the Flash Crash, there is obviously more work to be done on explaining the inner-workings of HFT firms and why they submit and cancel so many messages. In the meantime, buy-side traders have to live with the fact that a Flash Crash may happen again.

When one buy-sider at Advanced Trading's Summit in October asked what the institutional trader should do if a similar event occurred, a speaker suggested that the trader should stop trading and pull its orders from the market until things calmed down. But that is exactly what caused the liquidity crisis in the first place.

5.10 NYSE Disregards Fairness for Retail Investors

One might have hoped that big exchanges like the New York Stock Exchange (NYSE) would try to keep a level playing field in trading for all market participants. But according to Arnuk, that is not the case. The NYSE is opening a $500 million, 400.000 square foot co-location facility in Mahwah, N.J.

In pitching this facility to HFT's, the NYSE makes a ig deal about how it will connect every participant's computer by 1.000 feet of cable to those of the NYSE. Therefore, if Renasissance's computers are 100 feet from those of the NYSE, its trade will take them the same amount of time to reach the NYSE computers as Citadel's computer, which is 800 feet away. Arnuk thinking's it's too bad that the NYSE is not more concerned about the fairness of this for the retail investor.

5.11 Supplemental Liquidity Provider

A Supplemental Liquidity Provider (”SLP”) is a member organization that electronically enters proprietary orders from off the Floor of the Exchange into the systems and facilities of the Exchange and is obligated to maintain a bid or an offer at the National Best Bid (”NBB”) or the National Best Offer (”NBO”) in each assigned security in round lots averaging at least 10% of the trading day and for all assigned SLP securities, adds liquidity of an average daily volume (”ADV”) of more than 10 million shares on a monthly basis.

5.12 Light Pools

Help is finally on the way for institutional investors in their losing battle for profits against high frequency traders, and it’s not coming from regulators.

In March Credit Suisse is planning to launch Light Pool, a new stock-trading market that will block orders from high frequency traders when they cross certain level, Barron’s report.

Since high frequency traders are usually co-located at exchanges, they’re able to outmaneuver slower moving investors, forcing them to pay more for buys and sell stocks for less than they could have. But Credit Suisse’s new venue will also slow down executions, leveling the playing filed.

From Barron’s:
Credit Suisse in March will launch what it calls the Light Pool, a trading venue for mutual funds and institutional investors that purposely puts high-frequency traders at a disadvantage. This is revolutionary.

High frequency traders are courted by the 13 major stock exchanges because they deliver trading volume and pay big bucks for concierge services, like the direct data feeds from the exchanges that give them crucial information head start of several microseconds.

Dan Mathisson, managing director of Credit Suisse’s advanced-execution services, says the trading firms will have to route trades to the Light Pool through an outside stock exchange. ”That extra hop could add 100-to-200 milliseconds to a trade, enough time to be very discouraging to high frequency traders”, he says.

The latest data show the first major weekly inflow of retail investment money into domestic equity funds since the flash crash this past May. These investors plunked down $3.8 billion into the equity funds the week ending Jan. 12, according to the Investment Company Institute. During 2010 they withdrew an estimated $82 billion, in part because they were spooked by the flash crash, when the Dow plunged more than 700 points in 10 minutes and then climbed 300 points in the next 10. That thrill ride was aided and abetted by high-frequency traders using over-clocked computers to front-run panicked retail investors.

These traders program their computers to buy and sell millions of shares of stock every minute, based on short-term trends, not the underlying fundamentals of the companies.

Risk-averse to an extreme, their goal is to make a penny or so on each trade. It's easier for a machine to predict correctly if it is looking ahead only by a second or two. If the traders execute the same trades simultaneously, they can trigger dramatic market swings. High-frequency trading firms love to buy from and sell to "dumb" individuals and institutional investors. Individuals tend to place market orders rather than using limit orders at or below the bid price. Thus, they pay the maximum. As for mutual fonds and other institutional investors, they are easily front-run by the new trading operations, which have faster access to market data as well as faster trading computers. If the funds are buying a particular stock, the traders' computers can detect this activity, buy up shares ahead of the fund and sell it back to the fund for a profit of a cent or two. This runs up the costs for mutual fund investors.

The SEC has been mulling some curbs on high-frequency trading to shield long-term investors. But the plodding agency likely will take a year or two to enact any changes, and by then the math whizzes at the trading firms will have figured out another way to make chops out of the retail and institutional lambs.
Fortunately, there is a promising free-market response. Credit Suisse in March will launch what it calls Light Pool, a trading venue for mutual funds and institutional investors that purposely puts high-frequency traders at a disadvantage. This is revolutionary. High-frequency traders are courted by the 13 major exchanges because they deliver trading volume and pay big bucks for concierge services, like the direct data feeds from the exchanges that give them a crucial informational head start of several milliseconds. Dan Mathisson, managing director of Credit Suisse's advanced-execution services, says the trading firm will have to route traders to the Light Pool through an outside stock exchange. "That extra hop could add 100-to-200 milliseconds to a trade, enough time to be very discouraging to high-frequency traders", he says.
High-frequency traders claim they bring benefits to the market, such as liquidity, and that critics exaggerate their alleged abuses. Yet Light Pool is getting strong indications of interest from institutional investors. Sal Arnuk of Themis Trading inn Chatham, N.J. compares the new venue to the "tipping of a hat" to criticism of the new traders that he and colleague Joe Saluzzi raised in 2008.
Too bad there's no Light Pool for individuals yet. Out among the wolves, they're apt to get eaten up again and again.

6 Legal&Security

6.1 Code theft at Goldman Sachs

In July 2009, Goldman Sachs's proprietary algorithmic trading code was allegedly stolen by a Russian immigrant named Sergey Aleynikov. The platform that Aleynikov tried to steal was the proprietary trading system that Goldman uses in its algorithmic trading of stocks and commodities, a high frequency trading platform that Aleynikov himself supposedly helped create. Federal authorities claimed the platform contained Goldman's top secret mathematical formulas and algorithms that the company utilizes to generate massive profits.

Because this theft also coincided with the current U.S. recession, speculation became rampant over this ”new” technology. Some in the national media have speculated that such theft could collapse the economy. Some in the financial press have portrayed Aleynikov sa a mass criminal who sought to use this technology to unhinge the fabric of our society; others have speculated that such computer programs could derail on entire bank with the push of a button.

High frequency trading had finally gone public, and it wasn't pretty. In the months that followed, dozens of other news stories discussed the practices of Goldman Sachs, specialized HFT firms, and high frequency trading in general. Since then, various lawmakers, government agencies, and financial executives have battled over the merits and drawbacks of HFT.