The Five-Second Pause

Donald MacKenzie

What goes on in stockmarkets appears quite different when viewed on different timescales. Look at a day's trading, and market participants can usually tell you a plausible story about how the arrival of news has changed traders' perceptions of the prospects for a company or the entire economy and pushed share prices up or down. Look at activity on a scale of milliseconds (thousandths of a second), however, and things seem quite different.

When two American financial economists, Joel Hasbrouck and Gideon Saar, did this a couple of years ago, they found strange periodicities and spasms. The main periodicity was large peaks of activity separated by almost exactly a thousand milliseconds: they occur ten to thirty milliseconds after the 'tick' of each second. The spasms, in contrast, seemed to be governed not directly by clock time but by an event: the execution of a buy or sell order, the cancellation of an order, or the arrival of a new order. Average activity levels in the first millisecond after such an event are around 300 times normal. There are extended periods (lengthy, at least, when time is measured in milliseconds) in which little or nothing happens, punctuated by spasms of thousands of orders for a corporation's shares and cancellations of orders. These spasms seem to begin abruptly, last a minute or two, then end just as abruptly.

Little of this has to do directly with human action. None of us can react to an event in a millisecond: the fastest we can achieve is around 140 milliseconds, and that's only for the simplest stimulus, a sudden sound. The periodicities and spasms found by Hasbrouck and Saar are the traces of an epochal shift. Twenty years ago, the heart of most financial markets was still a trading floor on which human beings did deals with each other face-to-face. An 'open outcry' trading pit in Chicago, for example, was often a melee of hundreds of sweating, shouting, gesticulating bodies. Now, the heart of many markets (at least in standard products such as shares) is an air-conditioned warehouse full of computers supervised by only a handful of maintenance staff.

The deals that used to be struck on trading floors now take place via 'matching engines', computer systems that process buy and sell orders and execute a trade if they find a buy order and a sell order that match. The matching engines of the New York Stock Exchange, for example, aren't in the exchange's century-old Broad Street headquarters with its Corinthian columns and sculptures, but in a giant new 400,000 square foot plain-brick data centre in Mahwah, New Jersey, 30 miles from downtown Manhattan. Nobody minds you taking photos of the Broad Street building's striking neoclassical façade, but try photographing the Mahwah data centre and you'll find the police taking an interest: it's classed as part of the critical infrastructure of the U.S.

Human beings can, and still do, themselves send orders from their computers to the matching engine, but this now accounts for under half of all US share trading. The remainder is algorithmic: it results from share-trading computer programs. Some of these programs are used by big institutions such as mutual funds, pension funds and insurance companies, or by brokers acting on their behalf. The drawback of being big comes when you try to buy or sell a large block of shares. Such an order typically can't be executed straightaway (if it's a large order to buy, for example, it will usually exceed the number of sell orders in the matching engine that are close to the current market price), and if traders spot a large order that has been executed only partially they will change their own orders and their price quotes so as to exploit the knowledge. The result is what market participants call ‘slippage’: prices rise as you try to buy, and fall as you try to sell.

So big institutions often use ‘execution algorithms’, which take large orders, break them up into smaller slices, and try to choose the size of those slices and the times at which they send them to the market in such a way as to minimise slippage. For example, ‘volume participation’ algorithms calculate the number of a company’s shares bought and sold in a given period such as the previous minute, and then send in a slice of the institution’s overall order whose size is proportional to that number, the rationale being that there will be less slippage when markets are busy than when they are quiet. The most common execution algorithm, known as a volume-weighted average price or VWAP algorithm (it’s pronounced Veewap), does its slicing in a slightly different way, using statistical data on the volumes of shares that have traded in the equivalent time periods on previous days. The clock-time periodicities found by Hasbrouck and Saar almost certainly result from the way VWAPs and other execution algorithms chop up time into intervals of fixed length.

The goal of execution algorithms is to avoid losing money while trading. The other major classes of algorithm are designed to *make* money *by* trading, and it is their operation that gives rise to the spasms found by Hasbrouck and Saar. ‘Electronic market-making’ algorithms replicate what human market makers have always tried to do (continuously post a price at which they will sell a corporation’s shares and a lower price at which they will buy them, in the hope of earning the ‘spread’ between the two prices), but revise those prices as market conditions change far faster than any human being can. Their doing so is almost certainly the main component of the flood of orders and cancellations that follows even minor changes in supply and demand.

‘Statistical arbitrage’ algorithms search for transient disturbances in price patterns from which to profit. For example, the price of a corporation’s shares often seems to fluctuate around a relatively slowly-moving average. A big order to buy will cause a short-term increase in price, and a sell order will lead to a temporary fall. Some statistical arbitrage algorithms simply calculate a moving-average price, buy if prices are more than a certain amount below it and sell if they are above it, thus betting on prices reverting to the average. More complicated algorithms search for disturbances of price patterns involving more than one company’s shares. An example of such a pattern that was explained to me by one former statistical arbitrageur involved the shares of Southwest Airlines, Delta, and ExxonMobil. A rise in the price of oil would benefit Exxon’s shares and hurt Delta’s, while having little effect on Southwest’s (because market participants knew that, unlike Delta, Southwest entered into hedging trades that offset its exposure to changes in the price of oil). In consequence, there was normally what was in effect a rough equation amongst relative changes in the three corporations’ stock prices:

$$\text{Delta} + \text{ExxonMobil} = \text{Southwest Airlines}$$

If that equation temporarily broke down, statistical arbitrageurs would dive in and bet (usually successfully) on it reasserting itself.

No-one in the markets contests the legitimacy of electronic market making or statistical arbitrage. Far more controversial are algorithms that in a sense prey on other algorithms. If, for example, an algorithm can successfully detect the electronic signature of a big VWAP (doing so is called 'algo sniffing') then it can earn its owner substantial sums: if the VWAP is programmed to buy a particular corporation's shares, the algo sniffing program will buy those shares faster than the VWAP, and then sell them to it at a profit. Algo sniffing often makes users of VWAPs and other execution algorithms furious: they condemn it as unfair, and there is a growing business in adding 'anti-gaming' features to execution algorithms to make it harder to detect and exploit them. However, a New York broker I spoke to last October defended algo sniffing:

> I don't look at it as in any way evil … I don't think the guy who's trying to hide the supply-demand imbalance [by using an execution algorithm], why is he any better of a human being than the person trying to discover the true supply-demand? I don't know why … someone who runs an algo sniffing strategy is bad … he's trying to discover the guy who has a million shares [to sell] and the price then should readjust to the fact that there's a million shares to buy.

Whatever view one takes on its ethics, algo sniffing is clearly legal. More dubious in that respect is a set of strategies that seek deliberately to fool other algorithms. An example is 'layering' or 'spoofing'. A spoofer might, for instance, buy a block of shares and then issue a large number of buy orders for the same shares at prices just fractions below the current market price. Other algorithms and human traders would then see far more orders to buy the shares in question than orders to sell them, and are likely to conclude that their price was going to rise. They then might themselves buy the shares, causing the price to rise. When it did so, the spoofer would cancel its buy orders and sell the shares it held at a profit. It's very hard to determine just how much of this kind of thing goes on, but it certainly happens. In October 2008, for example, the London Stock Exchange imposed a $35,000 penalty on a firm (its name has not been disclosed) for spoofing.

Some, but not all, automated trading strategies require ultra-fast 'high-frequency trading'. Electronic market making is the clearest example. The 'spread' between the price at which a market-making program will buy shares and the price at which it will sell them is now often as little as one cent, so market-making algorithms need to change the quotes they post very quickly as prices and the pattern of orders shift. An algo sniffer or statistical arbitrageur may have a little more time: I've been told, for example, that statistical arbitrage programs may hold a position for as long as a day (and in some cases even longer) before liquidating it, hopefully at a profit. Even in those cases, however, an opportunity will vanish very quickly indeed if another algorithm spots it first.

Speeds are rising all the time. In Hasbrouck and Saar's data, which come from 2007 and 2008, the relevant unit of trading time was still the millisecond, but that's now beginning to seem almost leisurely: time is often now measured in microseconds (millionths of a second). The London Stock Exchange, for example, says that its Turquoise trading platform can now process an order in as little as 124 microseconds.

Some market participants are already talking in terms of nanoseconds (billionths of a second), though that's currently still mainly marketing hype, not technological reality.

Because the timescales of trading have changed, the significance of space has also altered. A few years ago, it was common to proclaim the 'end of geography' in financial markets, and it's certainly true that if one is thinking in terms of hour-by-hour or even minute-by-minute market movements, it doesn't really matter whether a trader is based in London, New York, Tokyo, Singapore or São Paolo. However, that's not the case in high-frequency trading. Imagine, for example, that your office is in Chicago, the second largest financial centre in the U.S., and you want to trade on the New York Stock Exchange. You are around 800 miles away from the matching engines in Mahwah, and sending a message that distance, using the fastest fibre-optic route between Chicago and New Jersey that I know of, takes around 16 milliseconds. That's a huge delay: you might as well be on the moon. Technical improvements in the amplifiers needed to boost signal strength and in other aspects of fibre-optic transmission will reduce the delay somewhat, as would straightening the route (fibre-optic cables still tend to follow railway lines because it's easy to negotiate rights of way there, but railways don't usually run in straight lines for hundreds of miles, instead going via centres of population). Ultimately, however, the finite speed of light is an insuperable barrier. If Einstein is right, no message is ever going to get from Chicago to Mahwaw in less than four milliseconds.

The solution is what's called 'co-location': placing the computer systems on which your algorithms run next to the matching engines in data centres such as Mahwah. That's the reason the Mahwah facility is so big: four-fifths of the space in it is available for rent. Co-location isn't cheap − a single rack on which to place your server can cost you $10,000 per month, and it has become a big earner for exchanges and other electronic trading venues − but it's utterly essential to high-frequency trading. Even the precise whereabouts of your computers within data centres is a matter of some sensitivity: you hear tales (possibly apocryphal) of traders gaining entry to centres and trying to have holes drilled in walls so that the route from their server to the matching engine is shorter. The New York Stock Exchange has put quite a lot of effort into ensuring that no one spot within the Mahwah facility is better than any other in terms of speed of access to the matching engines.

Tales of computers out of control are an established fictional theme, so it's important to emphasise that it is not at all clear that automated trading is any more dangerous than the human trading it is replacing. One way in which increased danger would manifest itself is in higher volatility of the prices of shares traded algorithmically. The evidence on that is not conclusive − like-for-like comparison is obviously hard, and the academic literature on automated trading is still small − but most such work has if anything suggested that automated trading reduces volatility. For example, statistical arbitrage algorithms that buy when prices fall and sell when they rise can be expected normally to reduce volatility.

The bulk of the research also suggests that automated trading makes the buying and selling of shares cheaper and usually easier. Renting rack space in a data centre may be expensive, but not nearly as expensive as employing dozens of well-paid human traders. Twenty years ago the 'spread' between the price at which a human market maker would buy and would sell a share was sometimes as much as 25 cents; as noted above, it is now often as little as one cent. That translates into substantial savings for mutual funds, pension funds and other large institutions, almost certainly outweighing by far their losses to algo sniffers. When assessed on criteria such as the cost of trading, the effects of automation are probably beneficial nearly all of the time.

What needs weighed against this is most dramatically one strange and disturbing episode that occupied a mere 20 minutes of the afternoon of 6 May 2010, beginning around 2:40 p.m. The overall prices of U.S. shares, and of the index futures contracts that are bets on those prices, fell by 5% in around five minutes, a fall of almost unprecedented rapidity (it's typical for broad market indices to change by a maximum of only 1-2% in an entire day). Overall prices then recovered almost as quickly, but gigantic price fluctuations took place in some individual shares. For example, the shares of the global consultancy Accenture had been trading at around $40.50, but dropped to a single cent. Sotheby's, which had been trading at around $33, suddenly jumped to $99,999.99. The market was already nervous that day because of the Eurozone debt crisis (in particular the dire situation of Greece), but no 'new news' arrived during the critical twenty minutes that could account for the huge sudden drop and recovery, and nothing had been learned about Accenture to explain its shares losing almost all their value.

Sotheby's price of $99,999.99 is, of course, the giveaway. What happened between 2:40 p.m and 3 p.m. (the 'flash crash' as it's called) was primarily an 'internal' crisis of the financial markets, rather than a response to external events. For five months, large teams of the staff of the Securities and Exchange Commission (SEC) and Commodity Futures Trading Commission (CFTC) researched in great detail what had gone wrong, ploughing through terabytes of data. While some market participants disagree with specific aspects of the analysis they published in September, most seem to feel that it is most likely broadly correct.

The trigger was indeed an algorithm, but not one of the sophisticated ultra-fast high-frequency trading programs (a fact greeted with clear relief by those who deploy these programs). It was a simple ‘volume participation’ algorithm, and while the official investigation does not name the firm that deployed it, market participants seem convinced that it was the Kansas City investment managers, Waddell & Reed. The firm’s goal was to protect the value of a large position in the stockmarket against further declines, and it did this by programming the algorithm to sell 75,000 index future contracts. (These contracts track the S&P 500 stockmarket index, and each contract was equivalent to shares totalling around $55,000. The seller of index futures makes money if the underlying index falls; the buyer gains if it rises.) The volume participation algorithm calculated the number of index futures contracts that had been traded over the previous minute, sold 9% of that volume, and kept going until the full 75,000 had been sold. The total sell order, worth around $4.1 billion, was unusually large, though not unprecedented – the SEC/CFTC investigators found two efforts in the previous year to sell the same or larger amounts of futures in a single day − but the pace of the sales on 6 May was very fast.

Both those previous times, the market had been able to absorb the sales without crashing. In the first few minutes after the volume participation algorithm was launched at 2:32 p.m. on 6 May it looked as if the market would again be able to do so. Electronic market-making algorithms bought the futures that the volume participation algorithm was selling, as did index-arbitrage algorithms. (These programs exploit discrepancies between the price of index futures and the price of the underlying shares. A large sell order in the index-futures market will often create just such a discrepancy, which can be profited from by buying index futures and selling the underlying shares.) Algorithmic trading was still in the benign zone that it occupies most of the time: electronic market makers and arbitrageurs were 'providing liquidity', as market participants put it, making it possible for the volume participation algorithm to do its intended large-scale selling.

However, high-frequency traders usually program their algorithms to be 'market neutral', in other words to insulate their trading positions from fluctuations in overall market levels. From around 2:41 p.m., therefore, those algorithms started to sell index futures to counterbalance their purchases, and the electronic index-futures market entered a spasm of the kind identified by Hasbrouck and Saar. One algorithm would sell futures to another algorithm, which in its turn would try to sell them again, in a pattern that the SEC/CFTC investigators call 'hot potato' trading. For example, in the 14 seconds following 2:45:13 p.m., more than 27,000 futures contracts were bought and sold by high-frequency algorithms, but their aggregate net purchases amounted to only around 200 contracts. By 2:45:27 p.m., the price of index futures had declined by over 5% from its level four and a half minutes previously. The market had entered a potentially catastrophic self-feeding downward spiral.

Fortunately, though, the electronic trading platform on which these index futures were being bought and sold (the Chicago Mercantile Exchange's Globex system) is programmed to detect just such a spiral. Its 'Stop Logic Functionality' is designed to interrupt self-feeding crashes and upward price spikes. A ‘stop’ is an order that is triggered automatically when prices reach a pre-set adverse level. For example, buyers of index futures will sometimes try to protect themselves from catastrophic losses by placing stop orders that will sell those futures if their prices fall below a given level. These sales, however, can potentially begin a cascade, causing further price falls that in their turn trigger further stop orders. The goal of the Stop Logic Functionality is to halt this process by giving human traders time to assess what is happening, step in and pick up bargains.

At 2:45:28 p.m., the price falls triggered Globex’s Stop Loss Functionality, and it imposed a five-second pause in trading. It worked. As Alison Crosthwait of Instinet (one of the oldest electronic trading venues) told the readers of an internet discussion forum hosted by the TABB Group, a capital markets research and advisory company, the five-second pause ‘provided ample time for market participants to consider their positions and return to the market or not depending on the conclusions they reached. [It] allowed market participants to regain confidence.’ Their purchases stopped the downward spiral of the price of index futures when trading restarted at 2:45:33 p.m.

The crisis, however, was not yet over. Index arbitrage and other mechanisms tie the index-futures market intimately to the underlying stockmarket, and by 2:45 p.m. the latter was becoming largely paralysed. High-frequency trading systems are often programmed to cease operating if unusually large price movements occur, and other systems are monitored by human beings who have what is in effect a large red stop button on their screens. Throughout the United States the automated systems stopped and the red buttons were pushed. Some market participants told the SEC/CFTC investigators that they were scared that the price falls meant that some catastrophe had occurred, but that somehow they had not heard. Others seem simply to have worried that there was a technical fault such as corruption of the incoming data feeds that carry price information. Orders were cancelled on a massive scale and no replacements posted. In the case of some corporations' shares, the market effectively ceased to exist.

The world of human trading that algorithmic trading has largely replaced had at its heart a subtle compromise. To be a market maker on the steps of Chicago's raucous open-outcry trading pits or in the marble-walled main trading room of the New York Stock Exchange brought with it privileges. For example, unlike other market participants a New York Stock Exchange 'specialist' (as the exchange's official market makers were called) could see the 'book' of buy and sell orders that had not yet been executed. In return for this considerable advantage, specialists were required to keep trading going, even in the event of a considerable imbalance between buy and sell orders, by using their own capital to fill the gap, all the while adjusting prices until the imbalance disappeared. Market makers in both Chicago and New York did sometimes overstep the line, opportunistically exploiting their privileged positions. However, as the sociologist Mitchel Abolafia documented in his 1996 book, *Making Markets*, in general such opportunism was held in check, not just by the formal rules but by the presence of informal norms amongst people who interacted with each other face-to-face day in, day out, year after year.

These delicate social ecosystems have not survived the transition to fully electronic trading. Market makers' privileges have largely vanished (for instance, you don't now need to be a market maker to get fast access to the New York Stock Exchange's 'book', but need only to pay for what is called 'level two' access and to rent a few racks at Mahwah to ensure that the data arrive with minimal delay), and their obligations have been reduced commensurately. However, some traces of those obligations still linger. For example, if you are an official market maker you still have always to quote a price at which you would buy and a price at which you would sell the shares in question.

Pushing the red button on an official market maker's system, therefore, did not remove all the bids to buy and offers to sell, but reduced those bids to the lowest possible price that could be entered into electronic trading systems (a cent), and increased the offers to sell to the maximum possible price ($99,999.99). These 'stub quotes' – as market participants call them – allow market makers to fulfil their formal obligations, while being so hopelessly unattractive that under normal circumstances no-one would ever want to take a market maker up on them. In the case of several stocks, however, the evaporation of the market by around 2:45 p.m. was so complete that stub quotes were the only ones left. In consequence, 'market orders' (orders simply to buy or to sell at the best available price) were executed against stub quotes, hence Accenture's price of a cent and Sotheby's $99,999.99.

The recovery was gradual, although largely complete by 3 p.m. It seems to have been led by futures prices on Globex stabilising and then rebounding after the five-second pause. Human traders began to spot what appeared to be extraordinary opportunities, although they were later often to be disappointed when exchanges cancelled sales at a cent and purchases at $99,999.99 on the grounds that they were 'obviously erroneous'. The fact that what largely had happened was simply that trading had stopped meant that the sums of money lost (and made) were only modest, and in consequence wider financial damage was limited. Even Waddell & Reed (if the volume participation algorithm was indeed the firm's) probably did not lose overwhelming amounts. The algorithm simply kept going through the turmoil − algorithms, after all, don't panic − finally completing its 75,000 sales by 2:51 p.m. By then, futures prices were already well on the way back up, thus limiting the losses caused by the algorithm selling at temporarily very low price levels.

Despite the modesty of the losses incurred, many market participants and regulators found the flash crash deeply unnerving, and I think they were right to do so. What troubles me most about the whole episode is not something that happened, nor even something that was said, but something that was not said. Ms Crosthwait's posting elicited only five comments from other TABB Forum members, and none disagreed with her judgement that five seconds was 'ample time for market participants to consider their positions'. She was certainly right to identify the triggering of the Stop Logic Functionality as the turning point, and the stabilisation of futures prices after the five-second pause shows that she was correct: five seconds *was* enough time. Given, however, that she was talking about human beings coming to decisions and not computer systems recalibrating themselves (since we don't ordinarily talk of computers 'considering' things and 'regaining confidence'), it points to a situation that in the terminology of the organisational sociologist Charles Perrow is one of 'tight coupling': there is very little 'slack', 'give' or 'buffer', and decisions need taken in what is, on any ordinary human scale, a very limited period of time. It takes me five seconds to blow my nose.

Simultaneously, too, stock trading (especially in the US) has become a *system*, and it is one of at least moderate complexity. True, there is nothing too dreadfully complicated about trading on any one exchange. The programs controlling Globex were, in consequence, perfectly well able to detect a dangerous condition and pause trading accordingly. However, while the Chicago Mercantile Exchange retains a dominant position in the trading of 'derivatives' such as index futures, the traditional stock exchanges such as the New York Stock Exchange and London Stock Exchange have been losing business rapidly to other electronic trading venues. There are now some fifty such venues on which U.S. shares are traded, and they don't operate in isolation. They are tied together by algorithms exploiting discrepancies in prices amongst them, and also by rules imposed by the Securities and Exchange Commission, which for decades has been trying to fuse the diverse exchanges of the U.S. into a 'National Market System'. For example, the SEC requires that brokers don't simply execute their customers' orders on their preferred venue, but achieve the most favourable prices (the 'national best bid and offer', as those prices are called). As Steve Wunsch, one of the pioneers of electronic exchanges, puts it (also on TABB Forum), U.S. share trading 'is now so complex as a system that no one can predict what will happen when something new is added to it, no matter how much vetting is done'. If Mr Wunsch is correct, there is a risk that attempts to make the system safer, for example by trying to find mechanisms that would prevent a repetition of last May's dramatic events, may actually have unforeseen and unintended consequences.

In his 1984 book *Normal Accidents*, Perrow argues that systems that are both tightly coupled and highly complex are inherently dangerous. To put his argument at its crudest, high complexity in a system means that if something goes wrong it takes time to work out what has happened and to act appropriately, while tight coupling means that one doesn't have that time. More profoundly, he suggests, a tightly coupled system needs centralised management, while a highly complex system can't effectively be managed in a centralised way because we simply don't understand it well enough, and therefore it requires organisational decentralisation. Systems that combine tight coupling and high complexity thus embody an organisational contradiction, argues Perrow: they are 'a kind of Pushmepullyou out of the Doctor Dolittle stories (a beast with heads at both ends that wanted to go in both directions at once)'.

Perrow's theory is just that, a theory. It has never been tested very systematically, and certainly never proved conclusively, but it points us in a necessary direction. When thinking about automated trading, it's easy to focus too narrowly, either pointing complacently at its undoubted benefits or exaggeratedly invoking our culture's fear of out-of-control computers. Instead, we have to think about financial systems as a whole, desperately hard though that kind of thinking often is. One such system is the credit system that failed so spectacularly in 2007-08, which is slowly recovering but without governments having solved the systemic flaws that led to the crisis, such as the combination of banks that are too big to be allowed to fail and 'shadow banks' (institutions that perform bank-like functions but aren't banks) that are regulated only more weakly. Share trading is another such system: it is less tightly interconnected in Europe than in the United States, but it is drifting in that direction here as well. There has been no full-blown stockmarket crisis since October 1987: last May’s events were not on that scale.[*] But that is an absence that we cannot guarantee will continue.

19 April

[*] Donald MacKenzie wrote about the 1987 crash in the *LRB* of 4 August 2005.