FINDING – AND FIXING – FLAWS IN FINANCIAL MARKET MICROSTRUCTURE

Brian F. Mannix[*]

[5/31/'16 Draft Submission to JLEP. NumWords = 9,171.]

Abstract: The automation of financial trading has dramatically reduced the cost of transactions, but at the same time has raised persistent questions about the effect of automation on market fairness, stability, and economic efficiency. This paper argues that there are indeed flaws in market microstructure, but they are not the sort that are easily addressed by regulation. Instead, technological innovation – especially the introduction of temporally buffered trading – is likely to provide a satisfactory resolution of existing problems. Temporal buffering gives market participants the option of trading more slowly, while limiting their exposure to predation by higher-speed traders. Three varieties are considered: short random delays (as used by ParFX), short fixed delays (as used by IEX), and short batched auctions (as proposed by Budish, et al).

Contrary to a common misunderstanding, an "efficient market" cannot mean the fastest possible market, because speed incurs real resource costs. Temporal buffering allows market participants to choose their preferred speed, and improves market efficiency in two ways: it avoids wasteful expenditures on high-speed "racing," and it reduces the transient information asymmetries that otherwise tend to be ubiquitous in high-speed markets. Regulators' priorities should be to: (1) avoid creating barriers to constructive innovations, (2) provide a regulatory framework that allows markets operating at different speeds to co-exist, and (3) rely on competition to sort out which innovations are useful and which are not.

INTRODUCTION

In the 21st century, automated algorithmic trading by computers has become the dominant method of exchanging securities, commodities, derivatives, and currencies in major markets around the world. Many more trades take place, at dramatically lower costs per trade, than in the days when human traders stood on a trading floor – or even when human traders sat at computer terminals, and controlled them in real time. There is little

[*] Research Professor, George Washington University Regulatory Studies Center.

doubt that automated trading has brought some substantial improvements to the efficiency of financial markets.

At the same time, many participants, regulators, and observers of financial markets have a sense that something has gone seriously awry: that the explosive growth of high-frequency trading (HFT) is somehow excessive, costly, unfair, and/or destabilizing. There are at least two persuasive indications that HFT entails some loss of efficiency. The first is the amount of real resources being invested in the arms race for zero latency. Tens of billions of dollars are spent to achieve miniscule temporal advantage in trading. Ships repeatedly cross the oceans laying fiber optic cables, each time stretching them a little bit tighter in order to render the previous cables obsolete. Where possible, traders will erect microwave towers, despite their relative inefficiency, to beat the traders who are using fiber optics, in which the speed of light is slightly slower.[1]

The second indicator is the amount of effort being made on the defensive side of the arms race. Large banks, mutual fund operators, and other sophisticated institutional traders try various methods to insulate their own transactions from the high frequency traders. If the latter were merely providing a useful service to the broader market, one would not expect large investors to go to such great lengths to avoid being serviced.

Several ideas for changing the rules have been discussed. Without a coherent explanation of exactly what is wrong, however, it can be very difficult to develop a promising remedy.

The object of this paper is to offer one such explanation: that the digitization of the trading infrastructure, in combination with ubiquitous but fleeting information asymmetries, has stimulated a dramatic expansion of racing. By racing I mean the wasteful expenditure of resources in a contest to trade ahead of other market participants; that is, racing – like its cousin, queuing – is an example of a directly unproductive profit-seeking (DUP) activity whose costs erode the gains from trade that otherwise would be available to participants in the market.

The paper also offers a specific remedy: the optional use of randomizing temporal buffers in the order flow. By slightly slowing the

[1] The best empirical paper documenting this arms race is Budish, Eric B. and Cramton, Peter and Shim, John J., "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response" (February 17, 2015). Chicago Booth Research Paper No. 14-03. Available at SSRN: http://ssrn.com/abstract=2388265

pace of trading, such buffers will allow market-data dissemination processes to saturate (i.e., will allow information asymmetries to dissipate) a little bit faster than order execution processes, so that price discovery and trading can operate more efficiently in an environment with more symmetrical information. By decoupling order flow from market-data flow, this remedy should also help reduce the likelihood of chaotic feedback instabilities in automated trading markets.

Racing and its associated costs have received a good deal of attention in other contexts, particularly the race-to-fish in certain fisheries.[2] Most analyses of financial markets appear to overlook the inefficiency of racing, however, in part due to a widespread misunderstanding of the efficient market hypothesis (EMH). Because the EMH emphasizes the speed with which information is incorporated into prices, many people tend to confuse speed with economic efficiency, thinking that faster must always be better. This is nonsense, of course. Real-world markets can always be made to operate a little faster, for a cost; but they can never be instantaneous. As the speed of trading approaches instantaneity, the cost will approach infinity.

It follows that the optimum speed of trading – the efficient speed, in the ordinary economic sense of efficiency – must be finite. In order to have a complete understanding of what an economically efficient market looks like, therefore, we need to be able to explain what it means for a market to be trading too fast, as well as too slow. And we need to know what conditions might cause a market to operate at the wrong speed, and how such conditions might be corrected so that the market can find its optimum speed.

I. RECOGNIZING RACING AND RETHINKING EFFICIENCY

One way or another, markets clear. Ideally, they clear at low cost by discovering a price acceptable to the buyer and the seller, with the price determining how the gains from trade will be divided between them. When, for whatever reason, the price mechanism is not functioning ideally, other mechanisms will assert themselves to close the gap between buyer and seller. Price controls on gasoline produced some spectacular *queues* in the United States in the 1970s. Economic regulation of airlines produced extra

[2] For a dramatic example see the first season of Discovery Channel's "Deadliest Catch." Later seasons feature an ITQ type of fishery management, and racing ceased to be such an important factor.

legroom, extra elbow room (i.e., empty seats), flying piano bars, and other forms of extravagant *non-price competition*. Trade barriers have fostered bribery, even to the point of measurably degrading GDP in some nations; a vast literature on *rent-seeking*[3] contains many more examples of Directly Unproductive Profit-seeking (DUP) activities[4] that waste real economic resources even as they appear to be privately profitable. *Racing* is one of those DUP activities, and it is commonplace. We see it in currency runs, in land and mineral rushes, in patent races, in fisheries with short and frantic seasons, and in a variety of other situations where temporal priority is rewarded.

Both racing and queuing dissipate economic rents by wasting resources, but in racing the waste can be more difficult to spot. When we see people waiting hours in line to buy gasoline, the real-resource losses are obvious. When commuters arrive at work early just to get a parking space, it is not immediately obvious, but is nonetheless true, that mispriced parking is causing a net welfare loss. It is all too easy to mistake racing for productive effort. In still other contexts, racing may be described as a "panic," but that label is misleading. Rational people will still trample each other to flee an inferno, or a collapsing currency.

Commercial fisheries provide some of the most instructive examples of racing. At the level of biologically and economically sustainable yields, the market price for fish is often much higher than the cost incurred in catching them. The difference represents an economic rent on the resource; but capturing that rent, without destroying it, is a challenge. In the absence of property rights in free-swimming fish, unrestricted competition will cause a fishery to collapse. Short fishing seasons is one common mechanism for preventing a collapse, but the response tends to be a more rapid expenditure of fishing effort – larger and faster boats, larger nets, etc. – in a race against the clock until a frantic equilibrium is achieved.[5]

The overcapitalization of a fishery – excess investment in fast boats and

---

[3] Beginning with Gordon Tullock, "The Welfare Costs of Tariffs, Monopolies, and Theft," Western Economic Journal 5 (3) (1967): pp. 224–232; and Anne O. Krueger, "The Political Economy of the Rent-Seeking Society" American Economic Review, 64 (1974): pp. 291-303.

[4] Jagdish N. Bhagwati, *Directly Unproductive Profit-seeking (DUP) Activity*, JPE 1982 p. 988 vol 90 no. 51 U. Chic.

[5] The Environmental Defense Fund, among others, has documented the dynamics of fisheries collapsing under traditional management regimes, and the advantages of using property rights instead. https://www.edf.org/oceans/how-turn-around-overfishing-crisis.

other capital that may be used only a couple of weeks out of the year – is so obviously wasteful that fishery managers may impose "gear restrictions" and other regulatory impediments in an attempt to reduce the waste. But when one factor of production is constrained, extra effort is channeled into another factor; the race continues on whatever margin is available until it is no longer worth it, the rents are exhausted and the market clears. Note that competition in the race-to-fish will drive profits to zero, but that emphatically does *not* mean that it will drive costs to zero. The deadweight loss is real: the waste is not that someone is making a profit, but that no one is.

But if racing is wasteful, then it should not exist in a ideally functioning market; there must be an underlying market failure that causes the misallocation of resources. Often that market failure is an absence of well-defined property rights, as in a common property resource. Indeed, the classical "tragedy of the commons" can be seen as an example of racing: the tragedy is not that there are too many sheep on the town commons, but that the sheep are turned out too early, eating the grass shoots before they have a chance to grow.[6] Overgrazing and overfishing are both symptoms of the same underlying problem, and solving that problem is the key to avoiding the loss. The enclosure movement in Great Britain, and barbed wire in the U.S., solved overgrazing; Individual Tradable Quota (ITQ) management plans, by creating property-like shares in a fishery, are well on their way to solving overfishing.

In fisheries that succumb to racing, we don't fret about whether faster boats have an "unfair advantage," nor do we complain that the fishery is "rigged." Some people may violate the rules, and we take pains to enforce them; but no one is under the illusion that better enforcement of rules will solve the underlying problem. Whether it is fair or unfair, lawful or unlawful, racing is economically disastrous because it destroys wealth, for everyone involved – those who win the race, as well as those who lose it.

A. Racing the News

Racing in financial markets bears a superficial resemblance to racing in fisheries. Indeed, the reported investments in high-speed data centers, fiber-optic linkages, and other accoutrements of high-frequency trading bear an uncanny resemblance to the overcapitalization that one sees in poorly

[6] Garrett Hardin (1968). "The Tragedy of the Commons." Science 162 (3859): 1243–1248.

regulated fisheries. They are costs incurred in the pursuit of profit; but, to the extent that they are unproductive, they erode the economic rents (i.e., the returns on investment) that would otherwise be available in the market. Here the remedy must be different, however, because the underlying market failure is different. The cause of racing in financial markets is not a failure of property rights, but rather an asymmetrical distribution of market-relevant information.

Information asymmetry is a well-understood market failure[7] albeit one that, in the context of financial trading, has a history of some controversy. This arises, in part, from the tension between two views of information as an economic good. One view is that information asymmetries, whatever their origin, cause unfairness and inefficiency; much of our regulatory system is designed to ensure that public information is available to everyone at the same time. The other view is that those who trade on information are improving price discovery and thereby helping make the market more efficient; their profit is simply the reward they receive for the service they are providing. From this latter perspective, the majority of market participants appear to be free-riding on those few who make the needed investment to produce accurate information and, through trading, to share it.

Over several decades this argument has not been settled, most likely because there is merit in both points of view. Information is valuable, but once produced can be copied for free; and it cannot be characterized neatly as a pure public good nor as a pure private good. Our legal institutions that deal with the ownership of information (e.g., the patent system, copyright and fair-use doctrine, etc.) tend to strike a balance between these two extreme views of information as an economic good. Financial markets have their own complicated set of contractual and legal institutions for handling information.[8]

In all of these fields, the digital revolution has upset the pre-existing balance between the private-good and public-good models of information and has forced a reexamination of institutions that govern the use of information. Thus we should not be surprised that the digitization of trading has dramatically altered the way that information is processed and rewarded in financial markets.

---

[7] George Akerlof (1970), "The Market for Lemons: Quality Uncertainty and the Market Mechanism," Quarterly Journal of Economics (The MIT Press) 84 (3): 488–500.

[8] For an early description of how information markets and security markets are intertwined see Henry Manne, *Insider Trading and the Stock Market* (New York: The Free Press, 1966).

B. Finding Inefficiency in an EMH-Efficient Market

The speed of automated trading certainly appears to be a good thing, in that it brings us closer to the ideal of a market that almost instantaneously reflects all of the available information. So how can we possibly reconcile the Efficient Market Hypothesis (EMH)[9] with the claim made here that racing is a manifestation of inefficiency? The simple answer is that these are two different uses of the same word.

The phrase "efficient market" as used in the EMH typically has a static meaning. The EMH states that markets quickly reach an equilibrium, but people forget that it is the equilibrium that is efficient – not necessarily the quickness of reaching it. We tend to take it for granted that faster information incorporation translates into superior resource allocation, and that the profits made by news traders therefore represent compensation earned for a productive activity. But it is not necessarily so. The speed at which a market's prices incorporate new information is, in part, the product of competition among traders to profit by trading early on breaking news. Real resources are expended in that competition; and, to the extent that they are devoted to unproductive racing, they represent a real loss.

The typical statement of the EMH glosses over this point, implicitly treating instantaneity as if it were an optimum. From Eugene Fama: "[W]e should note that what we have called *the* efficient markets model . . . is the hypothesis that security prices at any point in time 'fully reflect' *all* available information." [Emphasis in original.] From Burton Malkiel: "The logic of the random walk idea is that if the flow of information is unimpeded . . . prices fully reflect all known information."[10]

But, of course, prices do not instantaneously. To see where economic inefficiency may be hiding in an otherwise EMH-efficient market, consider an alternative informal paraphrasing of the hypothesis:

> "If ***t*** is the last moment in which a particular bit of information has no trading value because no one knows it yet, and ***t+1*** is the earliest moment in which it has no trading value because now everyone effectively knows it, then ***t*** and ***t+1*** are very close

[9] Eugene F. Fama, "Efficient Capital Markets: A Review of Theory and Empirical Work," The Journal of Finance, Vol. 25, No. 2, May 1970, pp. 383-417.

[10] Burton G. Malkiel, "The Efficient Market Hypothesis and Its Critics," Working Paper, April 2003. https://www.princeton.edu/ceps/workingpapers/91malkiel.pdf.

together and getting closer all the time."

This restatement captures the essence of the EMH, for which there is extensive empirical confirmation in the literature, but makes it also makes it clear that the EMH says nothing about what happens in between time ***t*** and ***t+1***. However brief that interval may be, there is (at least today) a great deal of trading that happens within it. And, because information during that interval is not symmetrically distributed and prices are not in equilibrium, we should not expect trading during that interval to be efficient in the usual economic sense. Nor should we expect empirical tests of the market's static efficiency to be able to identify a dynamic inefficiency of the sort that racing represents.

Today ***t*** and ***t+1*** may be only microseconds apart, but by one important measure – the latency/jitter ratio – they are farther apart than ever. We will come back to that concept later in the paper. For now, suffice it to say that high-frequency trading thrives, and exacts its toll, within this ephemeral realm. Markets that are EMH-efficient are nonetheless bleeding billions of dollars of value through the temporal interstices that are opened up by the digitization of trading.

The information asymmetries that drive this inefficiency arise because news does not break instantaneously. Those who learn it first may profit by placing orders to buy or sell securities, later unwinding their position after prices have adjusted. News traders may expend real resources in an attempt to surf the leading edge of any bit of breaking news. Nice traders – those whose have some exogenous reason to trade, rather than any particular news – will widen bid-ask spreads, withdraw temporarily from a turbulent market, or otherwise take defensive action in response to the heightened risk of being on the wrong end of a trade.[11] This is the lemon effect: the classic description of a market impaired by information asymmetries.

At the very short time scales in which computer programmed high-frequency trading takes place, another complication arises. Some high-frequency trading programs may examine the flow of the trading data itself and trade on the news it contains – essentially racing the tape. This is feasible because the dissemination of market news and the processing of market orders use the same digital technology. Both processes have the

[11] This terminology comes from Fischer Black. Initially he distinguished "news traders" from "noise traders" (unfinished working paper, personal communication, 1994), and then changed this to "news traders" vs. "nice traders" in his "Equilibrium Exchanges," Financial Analysts Journal 51 No. 3 (1995), published posthumously.

same "relaxation time," and are therefore strongly coupled. The net effect can be destabilizing as trading programs attempt to outrun each other in the direction of any perceived trend, or else defensively withdraw causing liquidity to evaporate. The "flash crash" of May 6, 2010, did not appear to be a panic, nor (because it so quickly rebounded) was it simply a rapid adjustment to a new equilibrium; it may in part have been a manifestation of market instability associated with high-frequency racing of market data.

Of course, it remains true that a market could not function without news traders. But those who spend real resources to learn in a microsecond what everyone will know, for free, in a millisecond are not performing a service. Those resources are directed not at creating real value, but at redistributing value. The distinction, above, between trades that takes place at equilibrium prices and those that take place "between the ticks" is an artificial one; in reality there is a continuum that is not so easily parsed. Even so, at very short time scales, we can infer that the benefits of price discovery become vanishingly small while the risks of costly and destabilizing racing become large. For this reason trading strategies that depend upon very high speed are more likely to be associated with inefficient racing than those that occur at lower speed.

Before looking more closely at the high-frequency trading, however, it will be helpful to go through an example that illustrates (because so many doubt it) exactly how a news trade can be presumably profitable and yet unambiguously inefficient.

C. The Helicopter & the Drilling Rig

The following example is an actual trade, but not one that took place at high speed. Indeed, the advantage of this trade is that it unfolded over weeks, so that it is easy to see all the moving parts, to examine the motivations of the participants, and to make some judgments about the consequences. The trade took place in 1972 in the stock of Amax Exploration, Inc., which at the time was listed on the Vancouver Stock Exchange.[12]

Among Amax's assets was a speculative mineral claim in the Yukon

[12] I learned of the details of this transaction from the helicopter pilot, personal communication, 1973. Note that after 1972 the Vancouver Stock Exchange thoroughly reformed its trading systems – several times, in fact – so that no implication should be drawn from this discussion regarding the quality of execution today on that particular exchange. The lessons of this story apply to any continuously trading platform.

Territory thought to contain recoverable quantities of zinc, copper, and associated minerals. Like many such remote deposits, this staked claim would remain idle until someone determined that it was worthwhile to make the investment in an access road. In the spring of 1972 Amax decided to test the ore deposit, and sent in a crew with a bulldozer that towed a drilling rig.

Learning of this, an equity trader contracted with the helicopter pilot to shadow the drilling crew. Because of the distances involved (satellite phones had not yet been invented), the trader built a radio repeater tower, powered by a generator, in the intervening wilderness. Through the tower the pilot would be able to reach the trader in Whitehorse, where there was a landline connection to Vancouver. The trader instructed the pilot to hover over the rig and watch the emerging drill core; a high-quality zinc ore would have a characteristic flat-black appearance. On cue, the pilot reported the buy signal: "It looks black to me."

It is not obvious which side of this this transaction one would want to be on. The helicopter was expensive; it likely cost more per hour to keep it hovering in the air than it cost to keep the drill bit turning in the ground.[13] We can only assume that the resulting trade was marginally profitable, after taking into account that the trader would have incurred the same expense hovering over a dry hole (and might then have made some money taking a short position). But the resources expended on the radio link and the helicopter were nonetheless pure waste.

It is true that some information about the ore deposit was incorporated into Amax's stock price a few days earlier than it otherwise might have been. But that information was vastly inferior to what the drilling crew possessed, since they could test the core chemically, measure the thickness of the ore deposit and its overburden, etc. Moreover, having access to that information sooner could not possibly increase the real returns from the mine. Amax could not begin to build a road until the following summer, and could not begin mining until the summer after that. Ultimately the net returns to Amax stockholders from developing that site would be diminished not only by the cost of the drilling rig but also by the cost of the helicopter. If the mine had been financed privately there would have been no helicopter; it would have served no purpose. The cost of the helicopter was pure waste, and it was incurred because the expedition was financed on

[13] This was a test hole in a shallow sedimentary deposit – far easier than drilling through hard rock for oil or gas.

a continuously trading public market that created the opportunity and the incentive to engage in racing.

Note that competition would be expected to drive excess profits to zero, even among helicopter traders; perhaps it already had. But competition would not drive costs to zero. The fact that traders were not making an excess profit from racing strategies did not mean that there was no problem. The helicopter was still there, the real resource losses were being incurred, and, through the market, the costs were being distributed among those traders who hired helicopters and those who did not. Everyone's combined returns were lower than the returns from an identical venture financed privately or by some racing-proof mechanism.[14]

In many respects, the helicopter is a more modern example of Rothschild's pigeon. When Wellington defeated Napoleon at Waterloo in June of 1815, that news briefly had trading value across the Channel on the London Bourse, where the sovereign bonds of all the European powers had been in play ever since Napoleon's escape from Elba 100 days earlier. Baron Nathan Rothschild allegedly received the news in London first, via carrier pigeon from a confederate traveling with Wellington, and he proceeded to make a profit in the market.[15]

Today, news with trading value crosses the English Channel through fiber optic connections. These may soon be obsolete, however, now that an HFT firm has undertaken to construct a slightly faster pair of microwave towers – tall enough to compensate for the curvature of the earth as they reach across the Channel.[16] The race goes on.

---

[14] Note the striking similarities between this trade and the case brought by the SEC against the Texas Gulf Sulfur Company, described in Manne (1966) p 51ff. In both cases the "insider" information consisted of a drill core from a Canadian zinc/copper deposit. Since Amax was traded on a Canadian exchange, however, it was not subject to SEC jurisdiction.

[15] While the story of Rothschild's pigeon has appeared in many sources, its accuracy has recently been disputed. See Brian Cathcart, *The Rothschild Libel: Why has it taken 200 years for an anti-Semitic slur that emerged from the Battle of Waterloo to be dismissed?* Independent, Sunday May 3, 2015. Available at:
http://www.independent.co.uk/news/uk/home-news/the-rothschild-libel-why-has-it-taken-200-years-for-an-anti-semitic-slur-that-emerged-from-the-10216101.html.
See also:
https://www.rothschildarchive.org/contact/faqs/nathan_mayer_rothschild_and_waterloo.

[16] Tim Cave and James Rundle. *High-Speed Trader DRW Proposes Thousand-Foot-Plus Tower in Rural England*, Wall Street Journal, Jan. 4, 2016. *available at*:
http://www.wsj.com/articles/high-speed-trader-drw-proposes-thousand-foot-plus-u-k-tower-1451937343.

II. WATSON'S THUMB AND THE GENESIS OF RUNAWAY RACING

A. The Digitization of Jeopardy!

The previous examples suggest that racing on information asymmetries takes place at slow speeds as well as fast, and that it has been going on for as long as we have had continuous financial trading. If information asymmetries are perhaps a mixed blessing, and in any event are ubiquitous and largely unavoidable, and if racing on breaking news has been a feature of financial trading for centuries, then what has changed? What is new and different about automated trading, other than the things – like cost, speed, and accuracy – that seem to be unambiguous technological improvements?

The answer to that question is subtle, and it will help to illustrate it with a recent experiment – one that pitted a computer against two humans. In 2011 an IBM computer, nicknamed Watson, appeared in the TV game show Jeopardy!, along with two human Jeopardy! champions – Ken Jennings and Brad Rutter. Watson was actually a very large custom-built computer in the back room, with vast databases of information to consult, but no connection to the internet. What IBM and Jeopardy! thought they were testing was the ability of the computer to understand questions posed in ordinary English, and to extract answers from the mostly unstructured database. (Actually, because this was Jeopardy!, the questions were answers and vice versa . . . but that matters not. We will refer to them as clue and response.)

In the event, Watson performed very well. But it struck many observers that his strongest performance was in pressing the signaling device that gave him the opportunity to respond to a clue. While Jeopardy! host Alex Trebek is reading a clue, the contestants' signaling devices (handheld buttons) are inactive. They become active as soon as the host finishes reading, and the Jeopardy! board lights up to signal to the players that their devices have been activated. The first contestant to press his or her button is given a five-second opportunity to provide a single response. If a contestant pushes the button too soon, however, his button is deactivated for one-quarter of a second, or 250 milliseconds.

So the first margin on which Jeopardy! contestants compete is the speed with which they press a button. And here is where Watson had a distinct edge. The average male college student, pushing a button in response to a

visual stimulus, has a response time of 190 milliseconds. Watson pushed his button using a solenoid that had a response time, or latency, of just 8 milliseconds.

Human contestants have other strategies available to them. Instead of waiting for the light that indicates buzzer activation, they can instead listen to the cadence of the host's voice. Switching to an auditory cue is, by itself, enough to lower the human response time to 160 milliseconds. More importantly, by listening to the host read the clue, humans can anticipate when he will finish. This strategy will fail when they buzz-in too soon; but it will enable them, some of the time, to beat Watson to the buzzer.

Moreover, it is a strategy that Watson cannot effectively imitate. Listening to the clue, rather than reading it, would be a challenge by itself for a computer. But even if Watson were able to do it well, it would not confer any latency advantage. There is another human in the loop – call him buzzerman – who sits off-camera listening to the host read the clue, and then presses his own button to activate the contestants' devices. His performance is necessarily variable, and there is no reason to think that a computer could mimic him with any greater success than another human could. So Watson's best strategy is to wait for the activation light and then use the raw speed of his solenoid to leave a very small window for his human opponents to shoot for. And his success rate with this strategy was high.

Let us pause here to note that we are not going to be saying anything about the fairness of this Jeopardy! contest. First of all, both IBM and Jeopardy! made it very clear that this was not a real contest but a demonstration, and the reward structure had been changed accordingly. The human contestants understood all of this in advance. Watson's winnings went to charity. Second, keep in mind that the Jeopardy! format had been selected for this demonstration specifically because it presented numerous seemingly insurmountable obstacles for the computer. Watson acquitted himself remarkably well in overcoming these. While he had an advantage in this one aspect of the game, there isn't space here to list all of the ways in which Jeopardy! favored human contestants.

B. Watson, Wharton, & Wilson

So the point of this discussion is not about fairness; indeed, it is not about computers vs. humans at all. We now need to extend the demonstration a little further by doing a thought experiment. What if Brad

Rutter were replaced with a second computer – call her Wharton. Suppose that Wharton is not quite as smart as Watson, but she is equipped with a solenoid with a latency of 6 milliseconds. By buzzing in consistently ahead of Watson, Wharton should prevail. Now let's introduce Wilson, a computer who gets a little over half the questions right. But Wilson, with a 4-millisecond solenoid, should be able to shut out both Wharton and Watson.

It is not hard to imagine that this would fundamentally change the character of the contest. Jeopardy! would become much less fun to watch – and not merely because it lacked a "human interest" element. What was once a game of wits would become a game of thumbs.

But why exactly is that? It is because computers are consistent, in a way that humans are not. When humans play Jeopardy!, their individual response time is initially an important competitive edge. But, with a little practice, everyone achieves an adequate level of competence with the signaling device. Differences in thumb speed do not disappear altogether, but they do tend to fade into the noise, while differences in knowledge, and in the speed of retrieving it, come to the fore.

"Fade into the noise" is the key phrase here. Human performance is variable, and the variability *between* humans is not much greater than the variability in performance of a single human in repeated trials. If I am 5 percent faster than you on average, I will not win every race. I will likely win a majority of races between us, but it might only be 60 out of 100. Some days I will not do my best, or you will. In contrast, if my computer is 5 percent faster than yours, it will beat you every time. Such is the consistency of digital systems: absent some external source of variability, they will produce the same result repeatedly. If computers play Jeopardy! under the same rules that work perfectly well for humans, the result will be a very different, and rather boring, game. Only one of them will ever get the initial opportunity to answer questions, and it will be the one with the fastest solenoid. Innovation and investment will focus on reducing latency; over time, competition will produce ever faster solenoids, but not smarter contestants.

To be clear: the problem is not that computers are too fast. Other things being equal, speed is a good thing. Nor is the problem that humans find themselves at a disadvantage. The problem is that the pre-existing rules of competition, which work well for humans, work very poorly for computers. They place far too great a premium on speed, at the expense of

intelligence. Computer systems are characterized not only by a low latency, but also by a very low jitter – the variability of latency. That predictability, when combined with Jeopardy's rules that favor temporal priority, will reward competitors who invest resources in gaining a speed advantage.

From time to time we change the rules of sports to make a game more interesting, and we could expect Jeopardy! to do the same – to change the rules so as to allow computers to compete on the basis of their ability to answer questions rather than push buttons. What might that change look like? After reading each clue, the responder could be chosen by lot from all those who pushed the buzzer within the first 250 milliseconds. Or, somewhat equivalently, a random delay could be added to the response time of the signaling device. This would introduce a synthetic variability in latency, removing some of the returns to speed, and shifting the competition to other margins.

Automated financial trading seems to be degenerating in much the same way we would expect an automated game of Jeopardy! to degenerate. Much of the digital infrastructure associated with high-frequency trading may be useful, but some of it is simply Watson's thumb, grotesquely overgrown.

III. TEMPORALLY BUFFERED TRADING

The problem with using digital computers to play Jeopardy! is similar to the problem of using automated digital systems in financial trading: in both cases, the competitive energy is channeled into an unproductive latency race. Investments in speed are disproportionately rewarded. Below I describe a proposed remedy in two different ways: once as a continuous lottery for priority, and then as an injection of temporal noise into the order flow. These are essentially the same remedy, but it is helpful to look at it from these different perspectives.[17]

How can a lottery operate in a continuous trading environment? Suppose arriving orders are not exposed to the market right away, but instead are place in a buffer, or queue. But this queue is not a first-in/first-out queue; instead, orders would be drawn out at random. In this sense it is more of a pool than a queue – call it a pooled queue. The average waiting

[17] The author has a U.S. patent pending on the use of a randomizing temporal buffer in financial trading: "System, Method, and Computer-Readable Medium for Improving the Efficiency and Stability of Financial Markets," U.S. PTO Non-Provisional Appl. No. 13/828,398 (Publ. No. US-2013-0297478-A1, 11/7/2013).

time may be very brief, but some orders will be kept waiting longer than others. In effect, when the timing of access to the trading floor is precious, it is allocated by lottery.

In order for the pooled queue mechanism to function properly, all orders must be subject to the same delay mechanism – including cancellation orders. A "buy" order, for example, can be cancelled by entering an offsetting "sell" order, but the party placing the two orders should have no control over when, exactly, each order is processed, or which one will be processed first.

By imposing random delays on incoming orders, the pooled queue mechanism renders racing at short time scales impractical. These random delays can be very short – less than one second – and still have the effect of diminishing the opportunity and incentive to race. A brief delay will be of little consequence to nice traders and to most news traders. It will, however, discourage traders who are seeking to profit from "news-with-a-fuse" – information whose trading value is expected to vanish almost immediately because it will be widely available almost immediately. In particular, it will discourage racing the tape.

Although a random delay sounds like something traders would want to avoid, it is not. The pooled queue lottery forces all market participants to bear some short-term timing risk, but this is beneficial because that risk is unavoidable anyway. Trading a security in a buffered market should produce higher returns than trading an otherwise identical security in an unbuffered, "real-time" market. Order buffering produces higher returns by avoiding the costs and risks associated with the very short-term transient information asymmetries that exist in the real-time market. Short-term racing is a negative-sum game, and most traders will be happy to avoid playing it. The pooled queue buffering mechanism allows market makers, nice traders, and most news traders to trade with each other, and to separate themselves from news-with-a-fuse traders.

One useful feature of temporal buffering is that it can be adjusted to accommodate varying market conditions as they develop, while maintaining continuous and orderly trading. For example, the average delay could be set at a very small number, even zero, for normal market conditions. The average delay (size of the buffer) could be increased quickly – up to some predetermined limit – in response to sudden price movements, unusual trading volume, unusually one-sided order flow, unusually low liquidity, or other indicators of a turbulent market. This promises to be more effective

and less disruptive than circuit breakers, which, instead of discouraging racing, can create new opportunities to engage in it.

Note that it is not necessary to create a physical buffer to implement the pooled queue mechanism; it suffices to impose randomly distributed short delays to the incoming order flow. In effect, the pooled queue mechanism suppresses racing by introducing a synthetic jitter – a random variability in the timing of a trade. In other contexts this is called dithering, and it has an interesting history.

Bomber crews during World War II noticed that the mechanical computers used in navigation and bomb sights appeared to operate more reliably during flight than they did on the ground. The reason was mechanical vibration – it acted as a lubricant and kept the gears from sticking, and torque from accumulating in the mechanical parts. Engineers soon began to attach small vibrating motors to earthbound computers in order to achieve the same result.

With the advent of digital computing, dithering did not disappear, but took on a new form. The digital processing of analog (continuous) data tends to introduce distracting artifacts at the higher frequencies; by adding high-frequency noise (often called "blue" noise, because blue is at the high-frequency end of the visible spectrum), these artifacts can be, if not removed, rendered invisible.

If you are reading this paper on a computer screen, chances are good that the computer's audio circuit uses sonic dithering with blue (here, meaning high-pitched) noise to remove audible artifacts from digitized music. The video adapter likely uses spatial dithering with blue (here, pixel-scale) noise to remove digital artifacts from displayed photographs and movies. If it is a high-end system designed for gaming, it may also use temporal dithering with blue (here, brief delays) noise to provide a fluidity of movement that digital rendering may otherwise find difficult to achieve.

What the pooled queue mechanism provides to continuously trading financial markets is temporal dithering, or high-frequency timing noise. Just as it does with movies and video games, this noise supplies a fluidity of movement. Indeed, the very concept of continuity in a digital system is something of a challenge. This is not a problem as long as the digital processes are much faster than the processes they are controlling – megahertz and now gigahertz computers have no trouble providing the illusion of continuity to music we listen to on a kilohertz scale. Similarly,

computers have no trouble suppressing vibration in machine tools. However, when a continuous process being controlled by a computer has patterns that resonate in the same frequency range in which the computer operates, digital artifacts and instabilities may appear. Temporal noise erases those.

One of the lessons of fishery regulation is that it is all too easy to suppress one rent-dissipating mechanism only to have another one pop up elsewhere. Even if the random delay mechanism succeeds in suppressing HFT racing, how can we be sure that we are not just shifting the inefficiency somewhere else?

To answer this question, we need to think in terms of a competition for "market share" among different market-clearing mechanisms. Prices, races, queues, and lotteries all may compete simultaneously to clear a market. When the prizes get unusually large, for example, people will often get up early (racing) to get a good place in line (queuing) to buy (pricing) lottery tickets (lottery). Similarly, rush-hour traffic on a congested toll road may be simultaneously governed by a dynamic combination of prices, races, queues, and lotteries.

The random delay mechanism allows an essentially costless lottery to occupy the high-frequency space in a financial exchange – the space where racing ordinarily would occur. It effectively blocks access to that space where information asymmetries are prevalent (or, more accurately, can be bought), and where trading is thereby inefficient. By shifting trading to lower frequencies, it allows the price mechanism to operate on a time scale where public information is more evenly distributed. The result is not just a symptomatic treatment; the random delay mechanism is designed to mitigate the underlying market failure and thereby make trading more efficient.

Experiments with random delays and with other forms of temporal buffering are already taking place. We will briefly comment on three.

A. ParFX and the Random Delay

In London, a coalition of banks has built a random delay mechanism into a new currency trading platform called ParFX,[18] using an average trading delay of 80 milliseconds. Since it first began trading in April, 2013,

[18] www.ParFx.com.

the company reports that the buffers are working as intended, and some competing foreign exchange platforms have begun to adopt a similar technology. The random delay seems to be especially popular with banks exchanging Australian dollars with other currencies. Australia is a major commodity exporter, leading to a large demand for currency exchange. Because of its location, there is inevitably a substantial latency when trading on the major exchanges in London – and lots of incentive to engage in latency racing. The ParFX random delay mechanism makes it feasible to trade without having to make the investment needed to engage in racing, or to defend against it.

B. IEX and the Deterministic Delay

IEX is an equity trading platform in New York, whose story has been well told in the Michael Lewis best-seller, *Flash Boys*.[19] Since it began trading in October, 2013, IEX has gained market share; within two years it accounted for 10 percent of all equity trading on alternative platforms. Now IEX has applied to the SEC to become a full-fledged exchange, prompting competitors to raise a number of questions about its trading system.

Instead of a random delay, IEX disrupts HFT strategies by imposing a 350 microsecond delay on all incoming orders. This deterministic delay constitutes a synthetic latency – in contrast to the synthetic jitter (variability of latency) imposed by a random delay. But the intent is similar: the delay provides assurance to customers trading on IEX that they are trading with other customers who also are willing to tolerate a brief delay. The fixed delay, because it is predictable, may be more susceptible to gaming. On the other hand, according to *Flash Boys*, some of the IEX team believe that a random delay would be more easily gamed.

In October 2012 I had proposed to Brad Katsuyama and the IEX management team that they consider incorporating randomizing temporal buffers into their new exchange; we met in January 2013 to discuss it. I did not meet with the IEX technical staff, dubbed the "Puzzle Masters;" but they apparently also read my proposal. Here is how Michael Lewis describes their reaction:

> [O]ne professor suggested a "randomized delay." . . . The Puzzle Masters instantly spotted the problem: Any decent HFT firm would simply buy huge numbers of lottery tickets – to

[19] Michael Lewis, *Flash Boys: A Wall Street Revolt*, W. W. Norton, March 2014.

> increase its chances of being the 100-share sell order that collided with the massive buy order. "Someone will just flood the market with orders," said Francis. "You end up massively increasing the quote traffic for every move."[20]

The "Puzzle Masters" were wrong about how a randomizing temporal buffer would work. First, "massive" orders would not be monolithic; typically they would be broken into smaller pieces, each with its own random delay. Second, the system would not allow orders to be cancelled without also imposing a random delay on the cancellation, so that anyone "flooding the market" with exploratory sell orders would find those orders being crossed – i.e., being matched with the component parts of any buy orders with which they "collided."[21]

Crossing orders is exactly what a financial exchange is supposed to do. Could an HFT firm nonetheless use this flood-the-market strategy to uncover information about the existence of a large unfilled supply or demand? Sure, if the HFT firm was willing to accept the resulting trades. But the information it thereby gained about the state of the market would be partial, and would emerge at a pace that provided little advantage to the most extreme speed-based trading algorithms. The randomizing buffer system is not intended to hide information indefinitely, nor to prevent *any* market movement in response to large orders; it is simply intended to dampen the bleeding-edge latency arbitrage that depends for its success on high-cost high-speed strategies.

In our conversations Katsuyama was unsure whether the Puzzle Masters had uncovered a real vulnerability, but he gave other reasons why IEX decided not to use what I had proposed. Probably the most compelling of these was that a random delay might be viewed by the SED as a violation of Regulation NMS. His own uniform "fixed delay" solution had the advantage in that he could implement it in the form of the famous shoebox – a 60 kilometer coil of optical fiber through which all incoming orders were received. The SEC will have difficulty finding fault with this: the fiber doesn't discriminate; all exchanges use fiber for access; there are no rules governing the length or routing of the fiber, and there are precedents for the use of coils. While I think random delays are a theoretically more elegant solution, I have to acknowledge that Katsuyama's architecture

---

[20] Michael Lewis. *Flash Boys*, p. 174.

[21] Budish, et al (2015) make the same error. They dismiss random delays as ineffective because "each of the fast trading firms sends infinitely many messages." (p. 61). Clearly that is not a practical strategy if the trading messages cannot be canceled.

nicely finesses the more problematic aspects of existing SEC regulations.

C. Frequent Batch Auctions

In addition to randomizing and fixed temporal buffers, batched call auctions are another option that has been discussed as a solution to the excesses of HFT racing. Budish, et al, make a persuasive case that batching of orders will mitigate many of the difficulties inherent in trying to maintain "continuous" trading. They do not specify the size of a batch, but argue that it can be less than a second, and still be effective. Some have objected that batched trading will involve thousands of predictable opening and closing events each trading day, creating lots of small opportunities for HFT strategies to arbitrage. On the plus side, the call auction mechanism that is used for price discovery in batched trading has some important advantages, including a tendency to erase the distinction between makers and takers of liquidity.

IV. COMPETITION ACROSS TRADING PLATFORMS

Budish, et al, propose that batched auctions be required for all trading. They acknowledge that there is another possible approach: "A second area for future research is the nature of competition among exchanges. Suppose that one or more exchanges adopt frequent batch auctions while other exchanges continue to use continuous trading: what is the equilibrium? Can an entrant exchange that adopts frequent batch auctions attract market share?"[22] Not only are those the right questions to be asking, they are questions that should be addressed first, before even considering the possibility of a regulatory mandate.

Similarly, former SEC Commissioner Larry Harris (2012) has proposed mandatory random delays: "Regulatory authorities could require that all exchanges delay the processing of every posting, canceling, and taking instruction they receive by a random period of between 0 and 10 milliseconds." Even if you believe, as I do, that random delays will create a more efficient trading platform, that is no reason to mandate them.

As they are developed, temporally buffered trading mechanisms, running alongside real-time markets, will give market participants a choice of how fast they want to trade. The racing hypothesis implies that slightly slower trading will appeal to many investors, and will produce superior

[22] Budish, et al, p. 51.

returns. But it will be far safer for regulatory agencies to loosen regulations in order to allow these competitive experiments, than to tighten regulations and impose a uniform remedy. Buffered financial markets can exist side-by-side with continuous real-time markets without difficulty. Arbitrage between these markets will keep them synchronized, with the caveat that arbitrageurs must follow the rules in each market they trade in. We have plenty of experience with different markets operating at different speeds, such as the retail market for mutual funds, trading once per day, and the market for Exchange Traded Funds (ETFs), trading continuously, or the venerable London gold fix, even while gold is traded continuously and sometimes frantically elsewhere.

Ironically, regulators are likely to make much more rapid progress by allowing innovations, than they will by mandating them. One reason is the heterogeneity of market participants. Even is a temporally buffered market is more efficient for most traders, it may be intolerable for an important subset. Mandating its use would create difficulties for firms that are attempting to keep an ETF aligned with its underlying market basket, for example. Mandating any such reform is likely to ban trading strategies that, for some participants, are essential and perfectly legitimate. There will be strong resistance to imposing such restrictive mandates.

This problem is aggravated because innovative trading platforms may need to impose some very specific restrictions, such as the order types that they will process. An exchange using a random delay, for example, will need to put restrictions on how orders may be cancelled. It is neither necessary nor desirable to impose these restrictions on the entire market; they are only needed for orders that are processed on that particular exchange.

Regulators of all types of financial trading, in the U.S. and around the world, will be challenged to provide a regulatory framework that allows different trading platforms to experiment with a variety of market structures, and that encourages them to interoperate, to compete, and to evolve in response to customer demand. One essential ingredient of such a regulatory framework will be a more sophisticated understanding of time, as it is measured across a spatially distributed trading system. With that in mind, we turn to a final topic: the physics of space-time.

A. On the Special Relevance of Special Relativity

> The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are

> radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality.[23]

The pace of financial trading is running into the physical limits set by the speed of light, and this has implications for how we think about market microstructure. The theory of special relativity[24] helps us understand the nature of the constraints that traders face. For example, some commenters have proposed that all markets should be synchronized to a master clock, failing to appreciate that – at the speed of today's markets – there is no such thing as a master clock. Space-time is structured in a way that makes absolute time impossible. Critics of the IEX application to become an exchange have objected to the speed bump that may cause transactions to take place at "stale" prices, but that claim needs to be evaluated in a context where, at some level, all prices are somewhat stale.

> So we see that we cannot attach any absolute signification to the concept of simultaneity, but that two events which, viewed from a system of co-ordinates, are simultaneous, can no longer be looked upon as simultaneous events when envisaged from a system which is in motion relatively to that system.[25]

In the theory of relativity an "event" has a precise meaning; it is a specific set of coordinates – a four-dimensional point – in space and time. From any such point, one can imagine a burst of light traveling in all spatial directions. The set of all points that can be reached by that burst of light is the event's "future light cone" (so called because of its appearance when time is graphed on the y-axis, as in figure x), and it contains all events that are unambiguously subsequent to the event at the origin. There is a second light cone that contains all past events. In addition, there is a set of points that lie outside either the past or future light cones – these points are "causally disconnected" from event at the origin. Thus the envelope of a event's light cone is sometimes called the "causal boundary," because if two events lie outside each other's light cones, there can be no information flow, and thus no causal connection,[26] between them.

---

[23] Hermann Minkowski's address to the 80th Assembly of German Natural Scientists and Physicians (Sept. 21, 1908), published later as: Minkowski, Hermann (1909), *Raum und Zeit* ("Space and Time"), Physikalische Zeitschrift 10: 75–88. Minkowski was Einstein's physics teacher.

[24] Albert Einstein, Zur Elektrodynamik bewegter Körper ("On the Electrodynamics of Moving Bodies"), Annalen der Physik, Bern 1905. pp. 891 - 921. English translation available at: http://www.fourmilab.ch/.

[25] Einstein (1905), op.cit.

[26] Since this is a law journal, it is important to clarify that the phrase "causally connected," in the present context, does not imply actual causality. It merely indicates that

High frequency traders attempt to gain advantage in the sequence of market events – the acquisition of information and the execution of trades – by skating ever closer to edge of the light cones that connect those events. But no technology can operate outside the limits of the causal boundary. And that provides one method of trying to avoid disclosing information to the HFTs. A smart order router can break a large order into multiple components, and direct them to multiple exchanges. If it takes into account the latency of delivering those orders to their destinations, and controls the timing so that the arrival events lie outside each other's light cones, each component order will be able to execute at its destination, without being influenced by the simultaneous existence of the others.

So, for example, IEX mentions this technique in a patent application:[27] "Ensuring Simultaneous Information Delivery to Geographically Distinct Trading Systems" (para. 130): "[M]any trading systems may target information delivery on a temporal plane." Note that IEX's "temporal plane" corresponds to the plane labeled "hypersurface of the present" in Figure 1.

Here we have to take note of a peculiar wrinkle in time: simultaneity is a relative concept – it depends on the frame of reference of the observer.[28] So, while we can draw a hypersurface of simultaneity, or "the present," it is in fact arbitrary. Observers traveling at different velocities will always be able to agree about the ordinal sequence of events that are causally connected, but they will not be able to agree about the sequence of causally disconnected events. Two disconnected events will appear simultaneous to one observer, while other observers will put them in different order. Regulatory agencies are anxious to improve audit trails and to include precise time-stamps, so that after any market dislocation they will be able to reconstruct the sequence of events across multiple linked markets. For a spatially separated set of transactions, however, there is an irreducible ambiguity to the sequence of events, which can never be resolved more precisely than the laws of physics allow.

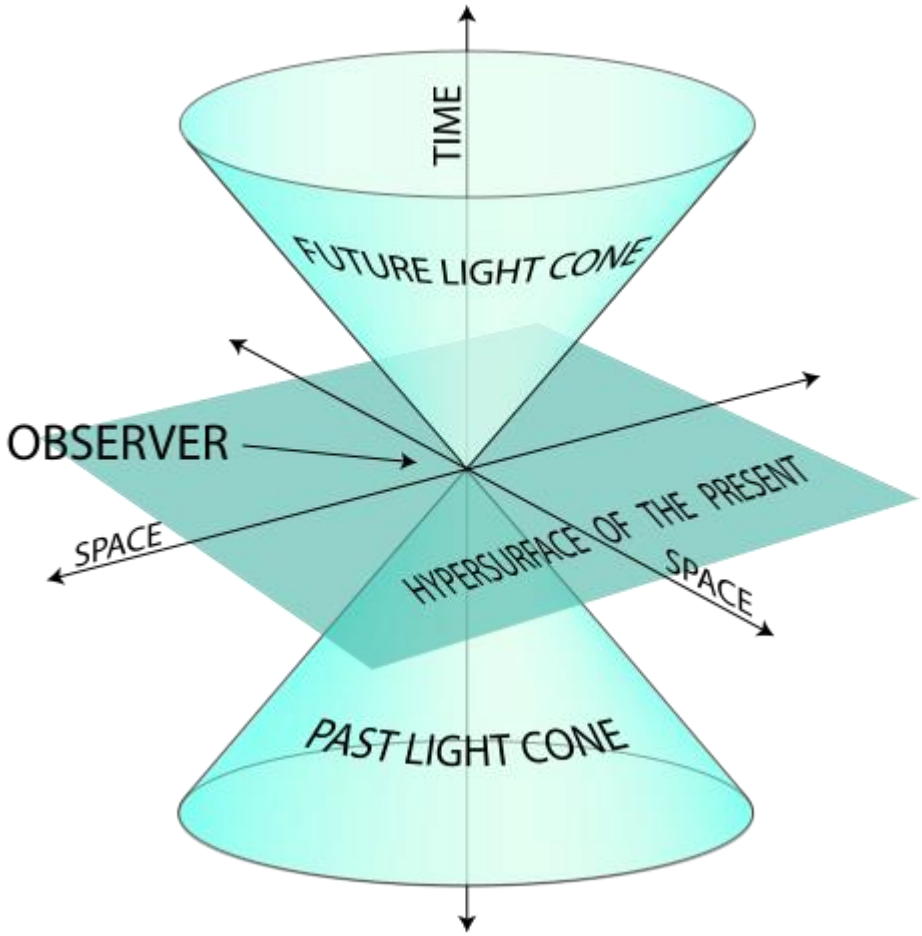
---

events can be traversed at less than the speed of light – along a "timelike curve," or worldline, in the parlance of special relativity.

[27] Katsuyama, *et al* (assignee: IEX Group). *Transmission Latency Leveling Apparatuses, Methods, and Systems*. U.S. Patent Application, Publication No. US 2015/0073967 A1. Filed July 3, 2104. Pub. Date: Mar 12, 2015.

[28] Different observors traveling at different velocities.

Figure 1



"World line" by K. Aainsqatsi at en.wikipedia. Licensed under CC BY-SA 3.0 via Commons – https://commons.wikimedia.org/wiki/File:World_line.svg#/media/File:World_line.svg.

This has implications when thinking about the meaning of such terms as "stale" prices. First, we need to recognize that, in any spatially distributed system of trading centers, it will not be possible to avoid some degree of price staleness. Indeed, it will not be possible to make an unambiguous definition of staleness. Nonetheless, it is true that temporal buffering will increase staleness in the sense that transactions will take place that could have been processed sooner at a different location. Is that a problem? Remember, an efficient market cannot be the fastest possible market –

speed has a cost, and infinite speed has an infinite cost. The prices on the temporally buffered exchange may be preferable for two reasons. One, they can be accessed without having to go the expense of trading at high speed. Second, many traders will prefer to accept a slower pace, as long as they are sure that they are trading with others who are similarly patient. The buffering encourages a self-selection process: Those who need fast execution can obtain it on real-time markets; those who can tolerate a brief delay will choose buffered markets. For them, staleness may be a virtue.

CONCLUSION

The primary challenge for regulators of financial trading is neither to decide which practices to ban, nor which to mandate. Instead, it is to build a framework in which different trading mechanisms can compete, innovation is encouraged, and more stable and efficient markets are permitted to evolve.

* * *