THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Regulatory Studies Center

Working Paper
January 2013

Races, Rushes, and Runs: Taming the Turbulence in Financial Trading

Brian F. Mannix[1]

[1] Brian F. Mannix is a Visiting Scholar at the George Washington University Regulatory Studies Center (www.regulatorystudies.gwu.edu), 805 - 21st St. NW, Suite 609, Washington DC 20052. This views in this working paper are those of the author, who can be reached at Brian@Mannix.com, and not those of the Regulatory Studies Center or GW University. Comments welcome.

Races, Rushes, and Runs:
Taming the Turbulence in Financial Trading

Brian F. Mannix
Working Paper, January 2013

Introduction

Many participants, regulators, and observers of commodity and security markets have a sense that something in recent years has gone awry: that the explosive growth of high-frequency digital trading is somehow excessive, costly, unfair, and/or destabilizing. 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 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.

[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.

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.

A few notes at the outset:

- First, this paper does not rely for explanatory value on panics, manias, nor any other psychological or behavioral phenomena. Perfectly rational *homo economicus* will engage in racing as described herein, as will computers unburdened with any emotional baggage. "Cooling-off periods" and other psychological remedies will not cure what is broken.
- Neither does the racing hypothesis depend on allegations of cheating, conspiracies, underhandedness, unlawful or unfair or "toxic" behavior, nor anything that needs to be stamped out. This is not to deny that there may be multiple abusive practices at work, but they are not a necessary part of the explanation of why racing is inefficient and why it is a growing problem. This is important because it also means that even successful efforts to eliminate abuses will not solve the underlying problem.
- This paper does not blame racing on any defect in regulation or market monitoring, and it contains no prescription for a regulatory fix. Rather, it alleges that an existing market imperfection has recently been severely aggravated by a technological change, the digitization of trading systems, and that the adverse effects can largely be mitigated by another technological change, the voluntary use of temporally buffered trading systems in parallel with real-time trading systems. Market participants should find it in their interest to adopt this remedial technology; if they do not, there is no compelling reason for regulators to impose it on them. Competition will be the best test of theory.
- Finally, this is not an empirical paper; it describes and illustrates the theory of racing and its application to financial markets. Additional theoretical, empirical, and experimental work will be needed to quantify the effects of racing and to test the proffered remedy. In the sections that follow, I hope at least to start that conversation.

I. Recognizing Racing and Rethinking Efficiency

"I don't have to outrun the bear, I just have to outrun you!"

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 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 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.

[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.

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

What's 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 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 an asymmetrical distribution of market-relevant information.

Information asymmetry is a well-understood market failure[5] 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

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

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.

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 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.

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)[6] 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 always 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*

[6] 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.

available information." [Emphasis in original.] From Wikipedia: "The semi-strong-form EMH claims both that prices reflect all publicly available information and that prices instantly change to reflect new public information."[7]

But, of course, prices do not "instantly change." 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 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.

Between t and t+1, falls the shadow.

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. Noise 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. This is the lemon effect: the classic description of a market impaired by information asymmetries.

[7] Wikipedia, "Efficient-market hypothesis," http://en.wikipedia.org/wiki/Efficient-market_hypothesis, accessed 5/7/12.

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 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 tape racing.

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, I want to give an example that illustrates (because so many doubt it) exactly how a news trade can be presumably profitable and yet unambiguously inefficient.

II. Diary of a DUP Transaction: the Helicopter & the Drilling Rig

The example I want to give 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, and I learned of it from the an American helicopter pilot who had recently returned from Vietnam. He was not yet ready to return to the United States and was working as a contract pilot in northern Canada. Apropos of nothing, and while standing on a high ridge along the Yukon/Northwest Territory border where we had landed, he made the comment: "Unless you really know what you are doing, do not invest your money in the Vancouver Stock Exchange."[8] He then told the following story.

[8] 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 that particular exchange. The lessons I draw from his story apply to any continuously trading platform.

Amax Exploration, Inc., at the time was listed on the Vancouver Stock Exchange. Among its assets was a speculative mineral claim in the Yukon Territory thought to contain recoverable quantities of zinc 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.

And what did you see? I asked. The pilot shrugged: "It looked black to me."

The pilot was right, I did not want to be trading with someone who was using such methods to get an informational edge over naïve traders. But I also did not want to *be* that trader. The helicopter was expensive: at one point I calculated that it cost more per hour to keep it hovering in the air than it cost to keep the drill bit turning in the ground.[9] I 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 the 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 a continuously trading public market that created the opportunity and the incentive to engage in racing.

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

Note that competition would be expected to drive excess profits to zero, even among helicopter traders; perhaps it already had. But it would not drive costs to zero. The fact that traders were not profiting 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.

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 Rothschild received the news in London first, via carrier pigeon from a confederate traveling with Wellington, and he proceeded to make a killing on the market.

Some people hear that story and think it illustrates the efficiency of markets in capturing information; others think it shows inefficiency or unfairness. Since I had already heard the helicopter story, my own first thought was absurd: "Isn't that a waste of a perfectly good pigeon?" Of course it isn't. Rothschild's profit was simply a return on his own resourcefulness and cleverness, for which I don't begrudge him. One would like to think, however, that a truly efficient market would encourage clever people to employ their talent in a productive manner, and not merely a profitable one.

III. Watson's Thumb and the Genesis of Runaway Racing

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 he 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.

Human Jeopardy! champion Ken Jennings eyes his real competition – not the computer in the back room, but Watson's incredible thumb, the buzzer-pressing solenoid on the desk.

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 (OK, college male) 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 seemed to have an advantage in this one aspect of the game, there isn't space here to list all of the ways in which Jeopardy! favored humans.

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 is as dumb as a soccer ball – OK, he 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, combined with the characteristics of digital technology, place far too great a premium on speed at the expense of intelligence. Specifically, computerized trading 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, rewards 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. We now turn to possible remedies.

IV. Buffered Trading: the PoolQ and the Art of Noise

The PoolQ

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.

How can a lottery operate in a continuous trading environment? Suppose arriving orders[10] 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, or PoolQ for short. The average waiting 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 PoolQ 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.[11]

By imposing random delays on incoming orders, the PoolQ 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 noise 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

[10] This mechanism could be applied to all orders arriving at an exchange, to a separate pool, or to a particular class of financial contracts.

[11] With experience, it might be possible to adopt a more relaxed version of this constraint. For example, a cancellation order, after being held in the PoolQ, might then execute immediately if the original order is still uncrossed. This creates an asymmetry that might be exploited, however, so the safest initial procedure is to treat cancellations no differently from any other order – i.e., they have no particular connection to the original.

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 PoolQ 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 PoolQ buffering mechanism allows market makers, noise traders, and most news traders to trade with each other, and to separate themselves from news-with-a-fuse traders.

Because this solution is advantageous to most traders, there is no need to impose it by regulation. Buffered financial markets can exist side-by-side with 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, including the retail market for mutual funds, trading once per day, and the market for Exchange Traded Funds (ETFs), trading continuously. For an investor averse to racing, the limited ability to trade mutual funds should be considered a feature, and not a bug. But it does not create a problem to have a real-time ETF market running simultaneously, for those with a taste for a faster game.

There are several ways the PoolQ mechanism can be applied, including:

- *A distinct buffered security.* A financial exchange could create a separate contract that could be traded only through a PoolQ. For example, a futures exchange could create a "b-mini" equity index contract that was identical to an existing e-mini contract, except that the b-mini could only be traded through a buffer.

- *An off-exchange trading pool.* A trading pool, or so-called "Dark Pool," could offer order buffering as a service to its customers. All orders arriving at the pool would be subject to a brief delay, order matching would be done internally, and only net positions would be traded externally. Note that it is a common practice among dark pools to withhold or delay the availability of information about their order flow and trading; this is one mechanism by which racing losses can be mitigated. A distinctive feature of the PoolQ mechanism is that it delays the exposure of the order itself to the market, as opposed to merely delaying the availability of information about the order.

- *An entire exchange.* A financial exchange could adopt the PoolQ mechanism broadly to stabilize trading.

One useful feature of the PoolQ mechanism is that it can be adjusted to accommodate varying market conditions as they develop, while maintaining continuous and orderly trading. For example, the average PoolQ 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 PoolQ mechanism; it suffices to impose randomly distributed short delays to the incoming order flow. In effect, the PoolQ mechanism suppresses racing by introducing a synthetic jitter – a random variability in the timing of a trade. In other contexts this is called dithering.

In Defense of Dithering

Sometimes it is necessary to state the obvious, just to have it on record. "Dithering" would not have been my first choice of words to describe the use of temporal noise to smooth financial trading. Client: "I'm worried that our returns are being eroded by the HFTs; we just can't keep up with those guys." Adviser: "Perhaps you should try a bit more dithering." No, it doesn't sound right at all. Unfortunately, the word is by now too well established and we won't be able to avoid it. So here is a brief history of dithering.

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 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 PoolQ 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.

Noise traders have no particular preferences about which part of a second their order executes in; therefore they should have no trouble tolerating high-frequency temporal noise. Many news traders will not be troubled by it, either; because the information they possess has a durability measured in seconds or longer. But there is a subset of news traders who will find it intolerable. Those are the news-with-a-fuse (NWAF) traders, and the fact that they find it intolerable is the very reason that the noise traders will find it attractive. Temporal noise will allow the noise traders to separate themselves from the NWAF traders.

The Black and Blue Handshake

When contemplating the value of dithering, it is essential to keep in mind that applying it to my own trade has no value whatsoever, by itself. The advantage comes from the access it gives me – the opportunity to trade with others who are also dithering. And they will not trade with me unless they are quite sure that I am doing it too. And, since all of us would be vulnerable if information about our orders were to leak out to the real-time market, we need to be quite sure that the PoolQ is dark – that no one can trade on information about the orders it contains.

Fortunately, it should be possible to provide this assurance. While an order is delayed in a temporal buffer, it is not exposed to the market. There is no need for anyone other than the order originator to know what it contains – not even the operator of the exchange or pool. In effect, until it is exposed, it is not an order but an order commitment. Its precise contents can be encrypted until the point when it is exposed. In this sense the PoolQ can not only be dark, it can be absolutely black, with no visibility inside or out.

But that raises another problem: How can the pool operator be sure that a trader is not changing an order commitment at the time it is decrypted? Even a difference of a single bit could change a buy order into a sell, depending on what key is used to decrypt it. For this reason, the encryption mechanism needs to allow the pool operator to confirm that the order exposed to the market is the same order that was committed to a few moments earlier.

The process will look like this: one message conveys an encrypted order commitment to the PoolQ, which assigns a random delay, after which a second message conveys a key that "opens" the order commitment and verifies its uniqueness. This two-step handshake process allows the customer to be sure that the order commitment was kept confidential (encrypted blackness), and the pool operator to be sure that the mandatory delay was not compromised (encrypted blueness).

Just as the handshake between knights in the middle ages was intended to signify an empty weapon hand, the black and blue handshake signals that a trader is not in the possession of news with a fuse. It is a signaling mechanism that creates the confidence for investors to trade with each other. NWAF traders may still prefer to trade in the real time market, and there is no reason to stop them from doing so. Over time, however it will become a less attractive place to trade.

Competition Among Market-Clearing Mechanisms

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 PoolQ 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 PoolQ 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 PoolQ mechanism is designed to cure the underlying market failure and thereby make trading more efficient.

V. Conclusion

As they are developed, temporally buffered trading mechanisms,[12] 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 returns. To the extent that wasteful racing is suppressed, confidence in financial markets should be restored. Temporal noise will cloak the higher frequencies and give buffered markets a more workable approximation of fluidity[13] and continuity. The average investor should once again be able to take a random walk down Wall Street, without fear of stepping on the cracks.

[12] Patents pending.

[13] The relationship between fluidity and liquidity in a buffered market is a subject for future research.