“Will the Market Fix the Market?”

Eric Budish

AEA / AFA Joint Luncheon Talk

Prepared Remarks

SLIDE 1-2: EMH

I’ve followed the stock market since I was a little kid, as I imagine is the case for many of us here in this room. I distinctly remember reading the newspaper with my father looking at stock quotes the morning after Black Monday in 1987 and talking with my father about whether that day, Tuesday, the market would go up or down. As an optimistic 9 year old I predicted the market would go up. I’m pretty sure I even told him to “buy IBM” – I checked, and it went up 11 points that day, and another 7¾ points the day after!

As I’m sure is also the case with many of you here in this room today, the first serious academic idea I encountered about the stock market was the efficient markets hypothesis. In my case, first via Burton Malkiel’s classic A Random Walk Down Wall Street, and more rigorously in John Campbell’s asset pricing course in graduate school.

Gene Fama famously distinguished between three forms of the efficient markets hypothesis: weak, semi-strong, and strong. He wrote that the EMH is “obviously an extreme null hypothesis” that we do not expect to be literally true. So it would be useful to “pinpoint the level of information at which the hypothesis breaks down.”

The weak form says you can’t beat the market – make risk-adjusted excess returns – using just past prices as your information set. The semi-strong form says you can’t beat the market using both past prices and other forms of public information, like company news and earnings releases. The strong form says you can’t beat the market even using private information. Fama concluded that there was evidence against this strong form, but that the weak and semi-strong forms held up pretty well. The plain-English translation is that to beat the market you have to know something that the rest of the market doesn’t know. My market call as a 9 year old was luck, not skill!

Our modern understanding of market efficiency, as summarized by the 2013 Nobel Committee, is that stock prices are “next to impossible” to predict in the short run, but that there is meaningful predictability in the longer run, for instance using the market’s price-earnings ratio, or by favoring value stocks over growth stocks.

The main debate is over how to interpret this long-run predictability: to what extent does it reflect variation in risk, across time and across stocks, that we may not fully understand but that is perfectly consistent with market efficiency, and to what extent does it reflect behavioral biases which are not consistent with market efficiency. That the market is hard to beat in the short run, on the other hand, is relatively uncontroversial – the Nobel committee even writes that “such a situation would reflect a rather basic malfunctioning of the market mechanism”

SPREAD NETWORKS

In 2010, a company called Spread Networks spent $300 million to dig a high-speed fiber optic cable between New York and Chicago.

The important feature of this cable is that it was dug in a straight line, whereas previous cables zigzagged along railroad tracks, around mountains, etc.

The straightness of this cable shaved round-trip data transmission time between New York and Chicago by 3 milliseconds. That’s 3 thousandths of a second, or roughly 1/100th of the time it takes to blink your eye.

Industry observers described 3ms as an “eternity”, and joked that the next innovation would be to dig thru the earth, to get an even straighter line by “avoiding the planet’s pesky curvature”.

And this joke quickly became true! Nobody dug a tunnel, but microwaves are a different way to get a straighter line from Chicago to New York, because light travels with less refraction through air than through glass, and lo and behold the first microwave connection between Chicago and New York went live about a year later.

When I first read about this cable --- and I still remember this viscerally --- I really didn’t understand it.

It was pretty obvious that it was some kind of arms race for speed, but what wasn’t at all obvious to me – and if this makes me sound unsophisticated, so be it – was what the speed was for.

3ms seemed too short to be about fundamentals – again, this is 1/100th of the blink of an eye. To put this in perspective, companies issue their earnings releases once per quarter, which is 8 billion milliseconds – and this is usually after the market is closed anyways.

And my academic, efficient market instinct was to be deeply skeptical of trading strategies that are short-run and purely technical. When academics hear phrases like 200-day moving averages, or head-and-shoulders patterns, or “support points” or “resistance”, we kind of roll our eyes. As Burton Malkiel wrote, “Technical strategies are usually amusing, often comforting, but of no real value”.

MARKET DESIGN PERSPECTIVE, I

My collaborators Peter Cramton, John Shim and I tried to make sense of the HFT arms race, approaching the topic from the perspective of market design

- Market design research assumes that participants in a market act rationally in their best interests given market rules – if HFTs are investing large sums of money on tiny speed advantages there must be some incentive to do so
- But, market design research takes seriously the possibility that the rules themselves are flawed
- And. Market design research takes seemingly tiny institutional details extremely seriously. All too often, such seemingly unimportant small details of a market end up having a dramatic effect on market performance. Al Roth calls this "economic engineering"
- Even Milton Friedman himself emphasized the importance of getting the right "rules of the game" in place, before letting market competition do the rest.

Indeed, our research identified a simple structural flaw in the design of modern financial exchanges – essentially, a glitch introduced in the transition from humans to computers.

The flaw is that exchange computers treat time as a continuous variable, and process requests to trade serially – that is, one-at-a-time in order of receipt.

Essentially, when markets transitioned from being run by humans to being run by computers – which on the whole has been a good thing – we made a mistake and forgot to put time into units.

MARKET DESIGN PERSPECTIVE, II

This flaw causes a violation of the EMH, built directly into the market design.

In an efficient market, to make money, you either have to (i) take risk, or (ii) have private information

What I'll show is that "built in" to the current market design are profitable arbitrage opportunities that are (i) riskless, and (ii) use only symmetric public information – information revealed to the whole market at exactly the same time, with economically obvious implications.

Let me underscore this: the details of the market design cause a violation of the EMH, built right in.

These arbitrage opportunities aren't supposed to exist in a well-designed market, and they harm the market in two ways

First, they make markets less liquid – as you'll see, they are like a tax on liquidity provision

Second, they induce a never-ending arms race for speed

Our paper then proposes an alternative market design that directly addresses the problem. The basic idea is simple: put time into units. Then, if multiple trade requests arrive at "the same time", award the trade to whoever offers the best price – that is, run an auction.

This market design, called frequent batch auctions,

Fixes the failure of the EMH, and in doing so

Improves liquidity

And stops the arms race

PLAN FOR TALK

In the first part of this talk, I want to tell you about this research with Cramton and Shim in some detail.

- I'll give you new empirical facts about how the current market design behaves at high-frequency time horizons
- I'll then go over the simple theory model that shows what's wrong with the current market design
- Last, I'll show how moving from continuous+serial -> discrete+batch directly fixes the problem

I'll then organize the second part of the talk around the main question I've gotten about this research, across academia, industry, and regulators: <u>will the market fix the market.</u>

- That is, will market forces on their own fix the problem identified in BCS, or is a regulatory intervention necessary
- In this part of the talk I'll discuss some new research on the IO of stock exchange competition that is joint with Robin Lee and John Shim, and which provides a framework for analyzing the private vs. social incentives for market design innovation
- I'll also make a few brief remarks that reflect on the lessons I've taken from what, to be frank, has been an unusual few years – as a young, untenured academic, I published a paper and found myself in the middle of a high-stakes, highly charged, policy debate.

DESCRIPTION OF THE CLOB

Just to make sure we're on the same page, let me spend one slide describing the current market design, often called the "Continuous Limit Order Book"

- The basic building block is a limit order, which specifies a price, quantity and direction
- Traders may submit limit orders to the market at any time during the trading day
- They also may cancel their outstanding limit orders at any time
- These orders and cancelations are processed by the exchange one-at-a-time in order of receipt, that is, as a *continuous-time, serial process*
- Trade occurs whenever a new order is submitted that is either an offer to buy at a price higher than some outstanding offer to sell, or vice versa. For example, an order to buy MSFT at 62.36, or to sell MSFT at 62.35, would initiate a trade in the order book shown on the right

Our data is, roughly, the "play by play" of the continuous limit order book.

CONTINUOUS MARKETS DON'T WORK

I want to begin with some empirical facts that show that continuous-time markets don't work as expected in continuous time

This is a plot of the price paths over the course of a trading day for the two most liquid instruments that track the S&P 500 Index: The E-mini future and the SPY exchange traded fund.

This is millisecond level direct-feed data, the same data that HFT firms subscribe to

Over the course of a trading day, these two instruments are very highly correlated, as we'd expect, since they track the same index

Here's an hour of data: [very highly correlated]

A minute of data: [very highly correlated]

And this is what the market starts to look like when we zoom into high-frequency time scales: this is a 250 millisecond slice of the day

When you zoom into high-frequency, the correlation between assets falls apart. The correlation is basically 1 at a minute, hour, day, but the correlation is basically 0 (<0.01) at a millisecond. [It doesn't matter whether or how you take relativity into account.]

The reason this phenomenon matters is that it creates <u>**obvious mechanical arbitrage**</u> opportunities. For example, when ES jumps and SPY hasn't reacted yet, the arb is to buy cheap in New York and sell expensive in Chicago, or vice versa.

The **usual story with obvious arbitrages is that**, **once discovered, the arb gets competed away**. I now want to show you that this is **not what happens under our current market design**

DURATIONS

We do find that the **durations** of arbitrage opportunities have come way down over time, from in excess of 100ms at the start of our data to about 7ms by the end of our data

[The RHS shows the distribution by year, which is a more sophisticated way to show that the market is getting faster and faster]

PROFITS

The **PROFITS** per arb opportunity, however, have stayed flat over time.

The arbs are not being competed away over time.

FREQUENCY

The FREQUENCY of arb opportunities DOES fluctuate over time, but we find that this frequency is driven mostly by market volatility, as opposed to market forces competing away the arbitrage

CORRELATION OVER TIME

This figure is a complementary way of showing the same phenomena

On the RHS of this graph, we see that in each year of our data the market is getting faster – this shows up as the 100ms correlation getting higher each year

However, on the LHS of the graph, we see that in all years, at high-enough frequency, the correlation is always approximately zero.

CONSTANT

So, to summarize, competition does raise the bar for how fast you have to be to capture obvious arbs

But competition does not eliminate the arbs or even reduce its total size

This suggests that we should think of latency arbitrage and the resulting arms race as a "constant" of the market design – I will come back to this point in the theory.

SIZE OF THE PRIZE

In our data, we compute the size of the ES-SPY arbitrage to be $75 million per year. We suspect this is an underestimate for reasons we detail in the paper, both the way we treat transactions costs and some details of the CME data feed at the time of our study.

But more important to emphasize is that ES-SPY is really just the tip of the iceberg in the race for speed.

Our analysis comes from ES-SPY arbitrage because that's where we could get the data.

But, conceptually, all continuous electronic markets around the world are similar. Lots of highly correlated instruments, and nothing in the market architecture to enable prices to move at exactly the same time.

As one example, here are correlated pairs of US Treasury Bond instruments,

Here are other equity indices

Here are some currency pairs, some precious metals, Oil & Gas, and

In honor of Al, here's coffee

… and here's simply a long list of obviously correlated pairs that John and I wrote down without too much effort.

SIZE OF THE PRIZE

In US equity markets, there are arbitrage trades that are even simpler than ES-SPY, because each stock trades on as many as 13 different exchanges and 50+ dark pools. These within-symbol HFT opportunities were at the heart of the Michael Lewis book Flash Boys.

There is a race to respond to public news, such as company news, government data, or other macro indicators such as the consumer confidence numbers. This race to respond to public news was at the heart of the New York Attorney General's initial interest in high-frequency trading, what he termed Insider Trading 2.0.

Last, there is a race to the top of the book, that is an artifact of the minimum price tick.

I still don't think it's possible to put a precise estimate on the total prize at stake in the speed race, given the data available to academic researchers. But common sense extrapolation from our ES-SPY estimates and the other numbers academic researchers have been able to put out for various aspects of the speed race suggests that the sums are substantial – easily single-digit billions per year in US equities alone. Michael Lewis reports on a figure from an industry study of as much as $20bn per year in US equities – a bit more than 1 penny per share traded – though this is higher than can be extrapolated to based on what's known to academics.

If you extrapolate from US equities to all financial instruments around the world that trade on continuous limit order books, it's easy to get to double-digit billions per year.

If you take an NPV of these annual amounts – and since latency arbitrage is like a tax on trading, the size of the prize grows with market capitalization, that is, the denominator r-g in the Gordon growth model is relatively low – it's easy to get to in excess of $100bn NPV.

I don't think this is the #1 policy problem in finance, but it's big, and worth trying to solve.

THEORY

I now want to move on to the theory model

The model is really quite simple, and I think it's pretty teachable, so I want to go over the setup and main ideas in some detail

The model will show, in I think a quite transparent way, that the mechanical arbs we saw in the data are "built in" to the current market design, that these arbs harm liquidity, and that they induce a never-ending arms race for speed.

The model is a descendant of the famous Glosten Milgrom (1985) model. If anything, it's simpler.

There is a security, let's call it *x*, that trades on a continuous limit order book market

There is a public signal, let's call it *y*, of the value of security x

And I want to make a purposefully strong assumption, which is that x is worth exactly y. The fundamental value of x is *perfectly* correlated to the public signal y, and can always be costlessly liquidated at this value.

The goal of this assumption is to create a best case scenario for price discovery and liquidity provision on the continuous limit order book, assuming away issues like asymmetric information and inventory costs, to isolate the course flaw

This public signal y evolves as a compound Poisson jump process, symmetric with mean zero, with arrival rate lambda jump, and "jump size" distribution J. [In words, with probably lambda jump the signal changes, and what we care about economically is the distribution of the absolute value of these changes, which is J.]

PLAYERS

There are two kinds of players in the model, Investors and Trading Firms

Investors we can think of as representing end users of financial markets

We are going to model them very simply. They arrive to the market needing to either buy or sell 1 unit of x, with arrival rate lambda_invest.

When they show up, they trade immediately, buying at the ask or selling at the bid

Trading Firms don't have an intrinsic demand to buy or sell x, their goal is simply to buy low and sell high.

Initially, I'm going to assume there are N g.t. 2 Trading Firms exogenously present in the market. Later, I'll endogenize entry via investment in speed.

LATENCY

Initially, I want to assume away all latency. When y changes, everybody sees this with zero time delay, for free. There is no latency in sending messages to the exchange or receiving information from the exchange.

Again, I'm trying to create a best-case scenario for the continuous market, to try to isolate the problem.

SNIPING

Given the setup of the model, with no asymmetric information, no inventory costs, no latency, etc. – you might conjecture that competition among trading firms will lead to effectively infinite liquidity for investors.

That is you, trading firms should offer to buy or sell x at price y, in unlimited quantity, at zero bid-ask spread.

Again, we've turned off all of the usual sources of costly liquidity provision in financial markets.

But that is not what happens, due to an issue with the continuous limit order book market design, that we call "sniping".

SNIPING

Suppose that y jumps, say from y1 to y2. This is the moment at which the correlation between y and x temporarily breaks down, like we saw in the empirics.

Trading firms providing liquidity in the market for x send a message to the CLOB, to withdraw their old quotes, and replace them with new quotes

[use hands to illustrate]

<u>However</u>, at the exact same time, *other* trading firms send a message to the market, attempting to trade at the stale quotes before they are cancelled

Since the market design processes messages <u>serially</u> – that is, one at a time in order of receipt – it is possible that a message to snipe a stale quote will get processed before the message to cancel the stale quote

In fact, it's not only possible, but <u>probable</u>, because for every 1 TF trying to cancel, <u>all other</u> N-1 TFs will try to snipe.

SNIPING GRAPHICS

So I've been explaining this sniping phenomenon with my hands for several years now, but given that this is a big room I figured we should make some graphics, so here is that argument, quickly, one more time.

[show sniping graphics]

SNIPING

Hence, in the continuous market, symmetric public information creates arbitrage rents. Obvious mechanical arbs like ES-SPY are simply “built in” to the market design

I want to underscore, again, that this really isn’t supposed to happen in an efficient market. Symmetric public information is supposed to get into prices for free, and asset prices are supposed to be hard to predict in the short run.

In equilibrium, these arbitrage rents are ultimately paid by investors

EQUILIBRIUM EFFECT ON LIQUIDITY

There are some details of the equilibrium analysis that I’ll skip, but the basics are pretty simple.

In equilibrium, trading firms are indifferent between liquidity provision and stale quote sniping.

Liquidity provision earns revenues, as investors show up and pay the bid-ask spread, and has costs, when jumps occur and the liquidity provider gets sniped.

Sniping has benefits when jumps occur. In these expressions, notice that the cost to every one liquidity provider, is exactly the same as the benefits to the N-1 trading firms who will try to snipe him.

Making TFs indifferent between these two activities yields an equation that characterizes the equilibrium cost of liquidity for investors. What it says is that the profits TFs earn from sniping come out of the pockets of investors via the cost of liquidity.

ENDOGENOUS ENTRY

Now, let’s endogenize entry by having a simple speed technology.

TFs can either see jumps in y with a small time delay, delta-slow, for free.

Or they can pay a cost, c-speed, to see jumps in y with less delay, delta-fast.

Think of delta-slow as the old cable and delta-fast as the new cable, or delta-fast is the microwave, etc.

Equilibrium is very similar to above. It is now characterized by two equations: the first one is the indifference condition from above, between liquidity provision and stale-quote sniping, and the second is a free entry condition, that characterizes how many trading firms invest in speed.

The new equation that comes out of the model says, in words, that all of the revenues trading firms earn from sniping, they dissipate by investing in speed, so that the prize in the arms race, expenditure on speed, and the cost to investors are all the same in equilibrium.

WHATS THE MARKET FAILURE

So I'm a Chicago guy, and I have to ask the Chicago question: what exactly is the market failure here? Isn't the arms race for speed just healthy competition?

The model tells a tail of two market failures. The first market failure is sniping – arbitrage rents from public information, in violation of efficient markets theory. And the second market failure is that these sniping rents then induce an arms race for speed, which mathematically boils down to a prisoners' dilemma. If all trading firms could commit not to spend money on the latest speed technology, they'd be better off, but that's not an equilibrium because each individual trading firm has an incentive to try to be slightly faster than the rest.

REMARKS

HFTS

I want to make two quick remarks about the model.

First, the model does /not/ say that HFT is bad for markets. Rather, it says that HFTs endogenously do two things – they provide liquidity, and they snipe stale quotes – the first of which is useful for investors, and the second of which is harmful to investors.

Don't be misled by the fact that sniping looks like zero-sum HFT on HFT combat. This misses the economics, which is that sniping is like a tax on liquidity provision, which in turn harms investors.

CONSTANT

The second remark is, if you'll notice, the size of the sniping problem in our model had nothing to do with the speed technology per se – it didn't matter whether the faster guys were faster by seconds, milliseconds, microseconds, nanoseconds, etc.

It also doesn't matter whether the speed technology is cheap or expensive – if it's cheap, there will be lots of entry, if it's expensive there will be less entry.

This tells us that the arms race is an <u>equilibrium</u> feature of the market design – it will keep going as long as we have this market design.

And, to illustrate, here are some highlights from the HFT arms race since we first started this research, in Oct 2010

ARMS RACE SLIDES

Here’s the first microwave connection between Chicago and NYC, launched in 2011. It wasn’t very straight, but since light travels about 50% faster thru air than glass, it was straight enough to be faster than the Spread Networks cable.

This first connection actually starts around the corner from here, at the corner of Randolph and Columbus

Here then is all the progress since that first connection between Chicago and NYC since 2011. You’ll notice that there are connections to DC too. That’s because government numbers – perhaps the purest form of symmetric public information there is -- are disseminated from DC. The location is literally on K street.

Here is a microwave path currently in construction, with data as of Dec 2016, that appears to be aimed at sending info from Chicago to Seattle by microwave, and then the rest of the way to asia by underwater cable.

And here’s a poster I saw at a recent industry conference, for a hardware device for HFTs, with the tagline “Nanoseconds matter”. A nanosecond is a billionth of a second. Every time you blink your eyes, 400,000,000 nanoseconds go by.

FBA

I now want to explain why moving from the current continuous market design, to our proposed discrete-time market design, directly solves the problem

HIGH LEVEL WHAT IS FBA

At a high level, our proposed market design, frequent batch auctions, modifies the current market design in just two ways.

First, we put time into units. Time is treated as a discrete variable, not continuous.

Second, we process orders in batch, using an auction, not serially.

FBA: DEFINITION

Now let’s define frequent batch auctions more completely.

During the batch interval --- to fix ideas for the second part of the talk let's say 1 millisecond, but it could be longer like 100 milliseconds --- traders submit bids and asks

> These are exactly like standard limit orders
>
> They can be freely modified, canceled, etc.
>
> If an order isn't executed in one batch interval, it remains outstanding for the next, and the next, etc., until it is either executed or canceled

At the end of each batch interval, the exchange batches together all outstanding orders – both new orders that arrived this interval and outstanding orders from previous intervals – and computes market-level supply and demand curves

If supply and demand don't cross, no trade happens, and all orders just carry forward to the next auction

If supply and demand do cross, transactions occur at the market-clearing price, that is this is a uniform-price auction [just like in the treasury primary market]

Priority is still price then time, but with time discrete. This means that if my order has been resting in the book for several intervals, and yours is new this interval, I have priority over you, but if we both entered in the same interval we have the same priority. [If necessary to break ties there is rationing]

The information policy is that the same information is disseminated as in the continuous market – trades, outstanding orders, cancels, etc. – but with the information disseminated in discrete time, after each batch interval. This is an important detail to underscore, it's economically important that information is disseminated in discrete time, not continuously throughout the batch interval.

SLIDE: 3 CASES

To explain how and why frequent batch auctions work, I want to go over 3 cases

CASE 1

The first case is that nothing really happens. This is a very common case, because most instruments, in most seconds there is zero trade, let alone in most milliseconds. Even the most active symbols have activity in only about 5% of milliseconds, and have trades in only about 1% of milliseconds. For a stock like Google, there is activity in less than about 0.5% of milliseconds, and trade in less than about 0.1% of milliseconds.

In this case, all outstanding orders simply carry forward to the next time interval, and the state of the book is displayed publicly, flashing on your screen or to your algo in discrete time.

This is just like displayed liquidity in a limit order book market. The bottom of the supply curve is the ask, the top of the demand curve is the bid

CASE 2

A second case is a small amount of trade happens, for instance an investor shows up and wants to buy a small amount at market

This case, too, is just like current practice. The investor can “buy at the ask” or “sell at the bid” just like in the limit order book market.

CASE 3

The third case is there is a burst of activity during the interval, for example in response to public news (a jump in y).

This case is where discrete time and continuous time are importantly different.

REASON 1

The first, and more obvious, reason is that discrete time reduces the economic relevance of tiny speed advantages.

As long as the batch interval is long relative to the difference between fast and slow traders – and I don’t want you to think of fast and slow as HFT versus grandma, but more like cutting-edge HFT versus other sophisticated market participants – then most information arrives at a time during the batch interval when all traders see it equally.

It’s only a small sliver of the interval, of proportion delta/tau, where if public information arrives during that sliver the speed advantage is relevant

REASON 2

The second, and more subtle, reason is that the auction changes the nature of competition. Instead of competing on speed, trading firms compete on price

The easiest way to see this is to suppose that public information actually /does/ arrive during the critical window, and there are some slow traders with stale quotes in the book.

In the continuous market, this will lead to a race by the fast traders to snipe the stale quotes.

Whereas in the discrete auction market, the fast traders compete on price. So if the information is truly public and obvious, the auction will compete away the arbitrage profits.

EQUILIBRIUM ANALYSIS

In the interest of time I'm going to skip some of the details of the equilibrium analysis and instead just give you the main takeaways.

If we treat the # of HFTs in the market as exogenous – which is probably the right case for thinking about an initial entry of a frequent batch auction exchange, or a pilot test – then FBA eliminates sniping for *any* batch interval tau. I interpret this "any" as long enough, given computational and communications technology, to enable genuine batch processing of HFTs acting on essentially the same signal at essentially the same time. This is probably on the order of 1ms or less.

If we treat the # of HFTs as endogenous – this is the relevant case for thinking about a market-wide reform aimed at stopping the speed race – then the equilibrium analysis says that the interval has to be long relative to the speed advantages in play. A rough calibration, given the scale of the modern speed race, points to intervals that are a tenth of a second or less being sufficiently long.

COMPUTATIONAL BENEFITS

In addition to stopping sniping, enhancing liquidity, and stopping the arms race, discrete time also has significant computational advantages over continuous time.

The basic conceptual point is that computers and communications technology are not infinitely fast, whereas a continuous-time market implicitly assumes that they are.

To highlight just one example of the complexity this causes, consider the market's paper trail – for regulators surveilling markets, for academics trying to analyze markets, for investors trying to assess best execution. In a continuous-time market, this paper trail has to be adjusted for geographical latency, exchange latency, clock synchronization failures – you have to know about relativity to know how to read the play-by-play of the market. Whereas in discrete time, all this complexity goes away.

One other example is the intrinsic tradeoff, in continuous markets, between error-checking and speed. Every additional line of code reduces speed.

ALTERNATIVE RESPONSES

There have been numerous other responses to the HFT arms race discussed in recent years, which we discuss in some detail in the paper. The only thing I want to mention here is that the Bans seem to misunderstand cause and effect, and that my thoughts on IEX's market design are fairly mixed and nuanced, and if you’re interested I’d encourage you to look at my SEC comment letter on their exchange application.

SUMMARY

So, to summarize Budish, Cramton and Shim

We look at the HFT arms race from the perspective of market design

The root problem isn’t “evil HFTs”, it’s a market design glitch – continuous-time, serial process trading.

This glitch causes a built-in failure of the EMH, and it’s simple to fix, by moving to a discrete-time, batch process market design. This market design eliminates sniping, enhances liquidity, stops the arms race, and simplifies the market computationally.

TRANSITION FROM BCS TO BLS

Paper released publicly in July 2013, pretty quickly took on a life of its own. [I guess not that surprising in retrospect, given the subject matter, but it was a surprise at the time]

Attention came not just from within academia, but from many different kinds of stakeholders, including exchanges, HFTs, the large investment banks and broker-dealers, institutional investors, trade groups, and many different regulatory bodies in the US and Europe.

My approach to the attention that the paper got – and I have no idea if this was the best approach, personally or professionally was:

First and foremost simply to invest significant time talking not just at academic seminars and conferences, but in lots of private meetings with stakeholders and lots of panel discussions and the like at industry conferences, where sometimes it felt a bit like being in the lion’s den. I probably visited or at least spoke to half a dozen each of the largest HFT firms, exchanges, broker-dealers, and institutional investors, and numerous regulatory bodies around the US, UK and Europe. Essentially, if a credible stakeholder expressed sincere interest in the work, I made time and often got on a plane. My colleague Austan Goolsbee called this “shoe leather” costs.

I learned an enormous amount from these interactions, first about institutional details, and second about how to communicate the work in a language that would translate and resonate. I actually think a lot of the improvements to the paper itself, between the first working paper version and the final published version, were shaped by these discussions.

I guess the other main element of my approach is that I’ve avoided any financial ties. This is a high stakes, highly charged debate, and I’ve tried my best to be an independent, objective voice.

The work of course had a range of reactions.

Some of the response was quite positive

Highest profile support came from the NY AG, Eric Schneiderman

Who among other things noted that as a U of C economist, I am, quote, "not an enemy of free markets". I of course agree with this.

As well as Bloomberg and Goldman Sachs, where, full-disclosure, I worked for a few years before graduate school as an extremely low-level investment banker

Support came as well from within academia, including, to my delight, across the ideological spectrum of my colleagues

And perhaps my favorite compliment for the work came from the quant hedge fund manager Cliff Asness.

What those who liked the work liked was

- First of all, the overall market design approach, trying to identify and then fix the underlying structural issue, as opposed to attacking symptoms or moralizing
- Second, I think it helped that the theory and data were both quite simple and transparent
- Economic idea that most resonated was that of transforming competition, from competition on speed to competition on price.

Some of the response was of course quite caustic

I actually got called "communist" A LOT. Which I found kind of odd for a paper proposing AUCTIONS, but that's kind of besides the point.

One prominent exchange executive said at a conference that if US markets adopted discrete-time, that he would take continuous markets and the associated technology to North Korea, and, quote, "show them how they could become the center of finance in about three weeks"

MOST COMMON QUESTION

But I think by far the modal response, and this spans academia, industry, and regulators, was some version of the question that is the title for today's talk. The modal sentiment was something like,

- You're probably right … but how do we get from here to there?
  - Is a regulatory mandate required? (Hence, "communist")
  - Or, can market forces alone fix the problem? [That is, will the market fix the market?]

- So what I want to do in this part of the talk is try to answer this question, can the market fix the market? I will tell you about some new research, still very much in progress and not yet released, with Robin Lee and John Shim on the economics of stock exchanges in the modern era: how they compete, and what are their incentives to innovate on market design.

TITLE

SEC CHAIR WHITE

To frame the issue, I want to start with a quote from the Chair of the SEC, Mary Jo White. This quote is an excerpt from a high-profile speech the Chair gave in June 2014, in the wake of the release of Flash Boys and all the surrounding controversy

She first acknowledges the possibility of an arms race, which is notable in and of itself

She then says she's wary of a market design mandate, but is receptive to exchanges innovating on market design, including FBA.

Last, she says that the SEC's job, as the market regulator, is to ensure that it doesn't inadvertently stand in the way of innovation.

PRIVATE VS. SOCIAL

The implicit presumption in Chair White's remarks is that market forces correct inefficiency, so long as regulators don't get in the way.

This is a natural instinct, and it's surely the standard case in economics, but as we all know it's not the only case.

PRIVATE vs SOCIAL II

The goal of this paper is to build a model of stock exchange competition so that we can understand the private and social returns to innovation, and ultimately try to answer the question "will the market fix the market?"

The first part of the paper builds a theory of status quo competition – competition amongst the continuous markets. The second part of the paper uses a variety of data to, I'll use the word "validate" the model empirically – it's a very simple and parsimonius model of a complex industry, so I want to show you that it's sensible.

Last, we use the model to study the incentives to innovate. The bulk of the scientific contribution of the paper is in the first two parts, but for the purpose of today's talk the third part is most pertinent, so I'll go thru the details of the first two parts at relatively high frequency.

EXCHANGE COMPETITION GAME

The starting point for our model is the BCS model of continuous trading, but instead of having a single exchange that is passive, we have multiple exchanges and they are active strategic players in the game.

One technicality I should mention to avoid confusion is that we also will assume shares are perfectly divisible – investors still arrive to market wanting to buy "1" share, but they can split this across multiple exchanges however they like. Think of "1" as a metaphor for a modest share purchase by an institutional investor

Initially, all exchanges use the continuous market design, and have two prices that they can set strategically.

First are trading fees, denoted f. For simplicity, we assume this fee is paid equally by both sides of the trade, avoiding some of the complexity of modern fee schedules. As you'll see empirically, this simple convention does a good job of capturing the economics.

Second are exchange-specific speed technology fees, F. I didn't emphasize this in my discussion of the previous paper, but many of the key technologies in the speed race are sold by the exchanges themselves:

Specifically,

1. co-location of servers, which means the right to put your computer near the exchange computers, with fast connectivity; and
2. proprietary fast data feeds

To be fastest at any given exchange, trading firms need to buy both general purpose speed technology, like microwave connections and fast code, as well as this exchange-specific speed technology, F

KEY INSTITUTIONAL DETAILS

There are two key institutional details that shape how we model stock exchange competition

The first is UTP, which basically says that stocks are fungible across exchanges. You can buy a share of stock on any one exchange, sell it on any other exchange, even if all the while it is technically listed on some third exchange. Where a stock is listed isn't as economically important as many people seem to think.

The second is Reg NMS, which is a long and complicated piece of regulation but economically has two key aspects. First, it requires that traders, on an order-by-order basis, send their trade to whatever exchange or exchanges have the best displayed price. This obligation can be fulfilled by the trader, their broker, or by an exchange acting on behalf of the trader. Second, it requires that exchanges make data about their quotes easily electronically accessible. We are going to model Reg NMS by assuming that investors and Trading Firms can frictionlessly search all exchanges any time they wish to act – that is, search is zero cost.

Note parenthetically that the precise language of Reg NMS did not anticipate the importance of latency in modern markets; that's a story for a different day, for the purpose of today's talk we model Reg NMS as frictionless search.

We are also going to assume that Frequent Batch Auctions are allowed under Reg NMS – a June 2016 SEC ruling suggests this is the case, for auctions conducted every millisecond or faster, but there is still some regulatory ambiguity that needs to be resolved.

UTP + REG NMS

This combination of Fungibility and Frictionless Search means that, even though stock exchanges look like a platform market, that should be studied with the IO tools of platform competition -- associated with Rochet and Tirole, and Mark Armstrong, etc. – in fact we should think about stock exchanges as what we will call a "virtual single platform".

Investors and Trading Firms use fungibility and frictionless search to "stitch together" a synthetic single market out of what look like many disparate markets. Once we have single market, we can start to think about the economics in terms of traditional supply and demand – in this case, the supply of liquidity and the demand for liquidity – and exchange fees can be thought of as like a tax wedge between supply and demand. Frictionless search then drives this fee towards marginal cost, which in this case is basically zero.

EQUILIBRIUM

The characterization of equilibrium is a bit involved notationally, so I am going to omit some of the math, but the economics are fairly straightforward. There are four key features of equilibrium.

First, stock exchanges are a virtual single platform, on which exchange market shares coordinate behavior. If an exchange has 20% share, this means that TFs supply 20% of liquidity on this exchange, and Investors route 20% of their desired trades to this exchange. The marginal unit of liquidity is indifferent across all exchanges, because of this coordination between liquidity demand and liquidity supply. Bid-ask spreads are equivalent across all exchanges.

Second, trading fees are perfectly competitive, as mentioned.

Third, where exchanges do have market power is in the sale of exchange-specific speed technology. If an exchange has 20% share, this means that 20% of sniping opportunities happen on this exchange, and the exchange, uniquely, can sell exchange-specific fast access to these sniping opportunities.

Last, there is a money pump constraint. An exchange with just 10% share would like to lower its prices even lower than the competitive level to gain share, and hence be able to earn more from selling speed technology. But, this bumps up against a zero lower bound – if an exchange charges a negative price to trade, this creates a money pump.

All four features of this equilibrium are borne out in the data. I'll go over this quite quickly in the interest of time.

EMPIRICS: VIRTUAL SINGLE PLATFORM

First, re the virtual single platform theory, we see the volume-depth relationship holds up robustly in the data, and that at least for the most actively traded stocks, the modal number of exchanges at the best price at any given time is, roughly speaking, "all of them".

EMPIRICS: TRADING FEES

Second, re competitive trading fees, while fee schedules are notoriously complicated, when you cut through the complexity and try to ask what is the average cost to trade a share of stock during regular trading hours, it is about 0.01 pennies on each of the three major exchange families. 0.01 pennies isn't quite zero, but it's pretty small. [Added up across all shares traded in the US, it's about the same annual revenue as my employer, Chicago Booth.]

EMPIRICS: COLO / DATA I

Third, re market power for Exchange-Specific Speed Technology, exchanges in the US do indeed make a large fraction of their revenue from Colocation, Connectivity, and Data. On the LHS is a revenue pie chart for the BATS family of exchanges, which has the cleanest data among all US stock exchange families, whereas on the RHS is the same revenue pie but for the Chicago Mercantile Exchange, a large futures exchange. Whereas for the CME, Data and Access is about 15% of total revenue, for BATS it's nearly 70%.

EMPIRICS: COLO/DATA II

In fact, for BATS' US Equities business, trading fees alone are below operating costs, which is consistent with trading fees not only being competitive, but even being modestly subsidized, though as a caveat this is just 10-K data which isn't as granular as one might like. NASDAQ and NYSE's financial data looks similar, but the data are less clean because they have more ancillary businesses. Across NYSE, NASDAQ and BATS, US equities colo and data revenue looks like it is around $1bn per year.

EMPIRICS: COLO/DATA III

Nasdaq currently sells four different tiers of co-location and connectivity, the latency differences among which are measured in microseconds, or millionths of seconds

EMPIRICS: MONEY PUMP

Last, you can see the money-pump constraint bind in the in trading fees data. If you look to the right column, which we call a "Max User" – basically, the fee charged to a high-volume Trading Firm or Broker – fees on 8 of the 10 exchanges controlled by the 3 largest exchange families actually get slightly negative, as low as negative 1/100th of a penny per share per side. This is right around where the money pump constraint binds in actuality – it's a bit less than zero, because of per-share taxes that must be paid to the SEC and FINRA.

INCENTIVES FOR MARKET DESIGN INNOVATION

So the data suggests that the simple model is sensible, and I now want to turn to the main question at hand, which is the Incentives for Market Design Innovation.

Let's start with the good news …

**

Suppose a startup exchange enters and adopts Frequent Batch Auctions – which I'll call Discrete – and charges a fee of zero. The other exchanges all still use the Continuous market design, and all still charge a fee of zero to trade, they charge positive fees for speed technology.

A reasonable prior – our prior before we started the study – is that there's basically the usual coordination problem, in which there are multiple equilibria. [[If I have an idea for a ride-sharing platform that is 10% more efficient than Uber, and launch Budish-Rides, one equilibrium possibility is for both passengers and drivers to ignore it.]]

But in fact, in our model there's a unique equilibrium, which is that the Discrete market design gets 100% share.

The key reason is the frictionless search. Trading Firms strictly prefer to offer liquidity on Discrete than to offer liquidity on Continuous, to avoid the cost of getting sniped, <u>IF</u> investors notice it. If they could be sure investors notice the liquidity, they'd be willing to offer a slightly narrower spread on Discrete than Continuous.

The key thing is, because search is frictionless, Investors notice. In a sense, Reg NMS mandates that they notice.

So, the unique equilibrium is 100% Discrete. Essentially, if two otherwise identical markets operate in parallel, and one has a tax but the other doesn't, the one without the tax will get all the trades if there are otherwise zero frictions.

This same argument works if Discrete charges a positive fee, provided the fee is smaller than the per-share savings from sniping. The economic interpretation is that the market design innovator is getting paid for eliminating sniping.

BUT…BERTRAND TRAP

OK, so that's the very good news. The bad news is that, if an Entrant innovates, Incumbents have incentive to copy.

Suppose an initial market design innovator adopts Discrete and charges a positive fee … then an incumbent will want to switch to Discrete and charge a slightly lower fee … and this process continues until fees get competed down towards the perfectly competitive level.

This is great for the market – we get a better market design, at competitive fees -- but bad for the entrant, whose profits get competed down to zero.

Conceptually, this is the classic Bertrand trap. Innovators are unable to capture the social value of innovations, if other firms can easily copy.

AND…INCUMBENT RENTS

A second source of tension between private and social incentives to innovate is seen by considering the incentives of an Incumbent to adopt.

If an Incumbent adopts the Discrete market design, the analysis is similar to before – the unique equilibrium is that the market tips to the market design that eliminates the sniping tax, and Bertrand competition competes fees down to zero.

However, there is a key difference, which is that the incumbent's profits /used/ to be positive – they used to have rents, under the old equilibrium, from exchange-specific speed technology.

Formally, you can model the game among incumbents as a repeated prisoner's dilemma, in which the status quo is an equilibrium.

NASDAQ QUOTE

These theoretical ideas -- The Bertrand Trap, and the idea that incumbents are in a Prisoners' Dilemma where it would be privately optimal to adopt Discrete, if you could be the only one to do so -- might seem a bit fanciful, but here is a quote from a senior executive at Nasdaq, at an academic event a few years ago. He said:

Technologically, we could adopt FBA.

But it would cost time and effort to get the SEC to approve it, and if it got approved, it would be immediately copied, so there'd be no first-mover advantage. And hence, no incentive.

SUMMARY OF ANALYSIS

So, to summarize the analysis:

If an FBA enters, the new market design wins significant share. In the stylized model, 100%, though of course that's an abstraction.

However, the fierce competition on fees that ensues drives the entrant's profits down towards zero.

And an incumbent's incentive to adopt is even worse, because it foregoes the rents from speed technology.

So, things don't look so good for the market fixing the market!

*

I want to first put a more optimistic spin on the analysis, and then talk about some potential regulatory responses.

The optimistic spin is that the model suggests where a market solution is more versus less likely to come from. The thing about profits of zero, in a theoretical model, is that zero is pretty close to being positive. So, while it's hard to squint at the model and see entering as an FBA being an entrepreneurial home run, you can certainly squint and make the case that there is positive incentive. For example, if the Bertrand trap doesn't mind literally, or immediately, and the adoption costs are relatively low.

This could lead to a private solution coming either from a new entrant, like an IEX, or from an incumbent with low market share and hence little to lose from the status quo, like the Chicago Stock Exchange, or, as of this fall, IEX, which entered with a continuous market design, and currently has share on the order of about 1.5-2%.

A private sector solution could also come from outside of US equities markets – for example, in futures markets, where the Bertrand trap is not an issue because futures contracts aren't fungible across exchanges. Futures markets may therefore be better able to capture the value created by eliminating sniping, either through trading fees or market share.

Least likely would be one of the large incumbent stock exchanges, with large revenues from the speed race.

So, I don't rule out hope for a private sector solution – what I conclude robustly is that the private incentives to innovate on market design are dramatically lower than the social incentives to innovate, and, for the largest incumbents, these private incentives are likely negative.

REGULATORY RESPONSE

Since the private incentives to innovate on market design are dramatically lower than the social incentives, and the private incentives may even be negative, there may be a role for a regulatory intervention.

I first want to describe what I think is a quite modest regulatory response, that our model suggests would facilitate the market fixing the market. It would have three elements

First, reduce adoption costs for the first entrant. One simple way to do this would be to proactively clarify that frequent batch auctions are allowed under Reg NMS. [What market designs are and are not allowed under Reg NMS is currently ambiguous.]

Second, reduce tick-size constraints. I didn't emphasize tick sizes in today's presentation of the theory, but a fat tick size can be a constraint against the market tipping towards a more efficient design. [The current tick size of one penny may not sound like very much, but, especially for lower priced stocks, it has a large effect on behavior, including driving trade off-exchange to dark pools, and to exchanges with non-standard fee structures. I encourage you to look at the research of Mao Ye on the many perverse effects of tick size constraints.]

Third, a light-touch regulatory response might have to find some way to insulate an initial entrant from the Bertrand trap – for instance, a modest exclusivity period, so that, reframing the words of the Nasdaq executive, it is not "immediately copied" and so there is a modest "first-mover advantage".

**

That said, the most direct regulatory response to the research I've presented today would be to simply put time into discrete units.

Such a policy, properly designed, would fix latency arbitrage, stop the speed race, and would also dramatically simplify the market computationally. While it's hard with the data that's available to date to express 100% confidence, I personally would support such a policy.

POLITICAL ECONOMY OF REGULATION

As is often done in economics talks, I spent a long time trying to convince you that there is a market failure, and then somewhat quickly described a potential regulatory response. If there's anything that I've learned at the Chicago lunch table, though, it's to think about the incentives of regulators as well. And, in this instance, the political economy of solving the problem is tricky.

Most centrally, addressing sniping and the speed race is a classic example of dispersed benefits versus concentrated harm: harm to HFTs with a comparative advantage at speed, to exchanges who sell speed, to speed technology providers, and so forth

Compounding the concentrated/dispersed problem is that the subject matter is both technical and nuanced, both of which make it more difficult to organize dispersed interests.

The regulatory risk-reward of fixing the problem also is not great. By this I mean that the benefits of fixing the market are subtle – an improvement in liquidity, computational simplicity – but the potential cost is that if regulators touch market rules, and there is a flash crash or other extreme event, then even if one has nothing to do with the other regulators will get blamed.

Last, but certainly not least, there's a bit of a chicken-and-egg problem, where each country's regulators would kind of like some *other* country to try it first. Even a pilot test would be a major undertaking. My one bit of progress to date on overcoming the chicken-and-egg problem is with the Financial Conduct Authority in the UK, which is gathering new kinds of data from exchanges to enable more direct measurement of latency arbitrage.

Bottom line, the political economy for regulators to address the problem isn't great, and, to be honest, I'm somewhat sympathetic.

SO WHERE ARE WE

- We've got a well identified market failure, that strikes at the core of what we mean by an efficient market [at the core of how efficient markets are supposed to work]
- It's causing an industrial arms race, harms investors, and makes markets unnecessarily complex.
- It's probably >$100bn NPV. Not the biggest problem in finance, but it's worth fixing, and the solution is pretty simple.
- The case for market forces for fixing this market failure – I don't want to say it will never happen, but the economics aren't great.
  - It's not like our paper came out and exchange CEOs started calling and saying this is great, let's do your thing
  - And this kind of makes sense

  - If you take our model at face value, the private incentives for an Entrant to fix the problem are zero, and the private incentives for an Incumbent to fix the problem are negative.
- And the political economy for the regulator to fix the market also isn't great. Again, I don't want to say it will never happen, but the political economy is tricky, because of the concentrated interests, how technical the subject matter is, etc.
- So, something I've been thinking about a lot is what to do about this.
- And I want to emphasize, that while this problem is important, and I'm of course quite proud of the work, there are much bigger ideas from economics with the same structure, where you have a well-defined market failure, a nice solution from economics, and the political economy for fixing the problem is lousy.
  - Carbon taxes probably the biggest example

And i have two thoughts I'd like to end with

MILTON FRIEDMAN

- My first thought is just be patient and keep doing the work. There's plenty left to do. More theory, more empirics, more time with stakeholders.
- This is the sentiment of a famous quote of Milton Friedman, from Capitalism and Freedom, where he talked about the "inertia of the status quo" and how it "takes a crisis", but my favorite part of that quote is how he described our job as economists, as keeping good ideas alive, keep pushing, until an opportunity comes along, so that if and when there's an opening politicians or industry leaders reach for one of our good ideas

ROTH AND ZINGALES

My second thought is more of a question, and is whether there is more we can do as a profession – whether there are institutions we can build, or professional norms and incentives we can shape – to bring our good ideas from research into the world

- Al Roth, in The Economist as Engineer – and this is without a doubt the single paper that's had the biggest influence on my professional life -- wrote that we need to foster a new kind of literature in economics, akin to the relationship between engineering and physics, that's considerably more applied, and more tailored to institutional details, than we're used to / may be comfortable
- My colleague Luigi Zingales argued in his AFA presidential address that we should be more accepting of and encouraging of engagement in policy debates, as long as the engagement is disciplined by theory + data

- Both Roth and Zingales are thinking about the gap between what academics naturally do in research, and what academics might be able to do, and need to do, to get good research ideas implemented in the world

CONCLUDING SLIDE

Reflecting on the unusual experience I've had over the past few years

- as a relatively young researcher,
- being thrust into a high stakes, highly charged, fast-moving debate,
- and at times, frankly, feeling outgunned, no matter how clever were my talking points, by the interests on the other side

I've come to the I guess obvious conclusion that the changes Roth and Zingales are encouraging are /especially/ important for ideas that are high social value, but where concentrated interests are opposed – that is, when our ideas benefit the dispersed

When SV and PV are aligned, then natural economic forces help build the bridges from research to practice.

I'm not saying they happen automatically or without serious effort, but at least the winds of change are with us. I especially have in mind some famous examples in finance, like

- Index funds
- Derivatives
- Sophisticated portfolio management

These are all ideas that are extremely socially valuable, and there were private incentives to bring the idea out into the world, and the process kind of worked

But the other case, where SV is positive but Private interests are opposed – [and what I've talked about today is one example, but there are numerous other considerably more important examples] – then not only is there not a natural magnet pulling the good idea from research to practice, but there's active opposition.

In the end I'm an optimist, and believe that, eventually, good ideas win – and I'll wager that we see discrete-time trading /eventually/, if we measure time in decades

But I'm left wondering if we can speed up.

- If high-frequency traders are investing billions of dollars on improving the speed of information transmission between markets by literally billionths of seconds

Is there more WE can do to speed up the transmission of our ideas, from Decades to something a bit more high frequency.