September 9, 2019

Vanessa Countryman, Secretary
Securities and Exchange Commission
100 F Street NE
Washington DC

Re: Release 34-86168; File Number SR-CboeEDGA-2019-012; Cboe EDGA Exchange, Inc.; Notice of Filing of a Proposed Rule Change to Introduce a Liquidity Provider Protection on EDGA ("Filing")

Dear Ms. Countryman:

This is in response to Cboe's recent comment on the Filing,[1] where Cboe has included a few graphs.[2] So far as I can tell the graphs are the only empirical evidence Cboe has provided for the Filing.

The graphs

Cboe's description of the graphs is that they display:

> ...markouts for liquidity providers on EDGA in SPY during the month of July 2019 based on whether or not the transaction involved a "missed cancel" - i.e. where the liquidity provider attempted and failed to cancel or replace their quotation within four milliseconds after an execution. These statistics illustrate the difference between the execution price and the midpoint price at the time of the trade and in the milliseconds following an execution. *The yellow line shows markouts in situations involving a missed cancel.* [Emphasis added.] As illustrated, the midpoint price moves dramatically in the milliseconds immediately following transactions in this category, which often involve a handful of faster firms that are routinely able to predict and profit from prices that are about to change. That is, prices immediately move against the resting order in the milliseconds following the trade, indicating that the trade was likely to have been executed at a stale price. *By contrast, the purple line represents the same markouts for all other transactions.* [Emphasis added.] These markouts, which represent the majority of trading activity conducted on the Exchange, show relatively stable prices following an execution.[3]

Cboe concludes:

> The LP^2 delay mechanism would reduce the effectiveness of latency arbitrage strategies by offering a four millisecond period for liquidity providers to update their posted quotations before trading at a stale price.[4]

[1] Letter from Adrian Griffiths, Assistant General Counsel, Cboe, to Vanessa Countryman, Secretary, Securities and Exchange Commission, August 22, 2019. ("Cboe 1")

[2] Cboe 1, page 4 and pages 18-20.

[3] Cboe 1, pages 3-4.

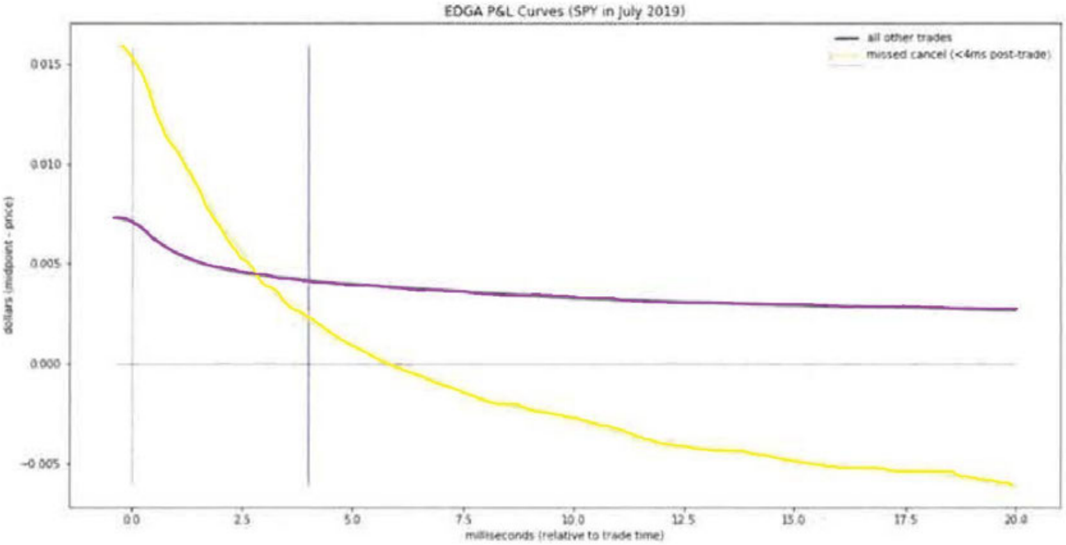
[4] Cboe 1, page 4.

Cboe includes the following details on its calculations:

> For a given time t_1 relative to a trade that occurred at time t_0, the markout for a buyer (seller) is calculated as m_1 - p (p - m_1), where m_1 is the midpoint price at time t_1 and p is the execution price at time t_0.[5]

The following discussion is involved. Briefly, I believe the graphs don't show anything like what Cboe believes they show, and they might flatly contradict the Filing's narratives. Even taking them as they are, however, the "harm" Cboe thinks is done is trivial while the subsidy the Filing grants to Cboe EDGA's market makers is enormous.

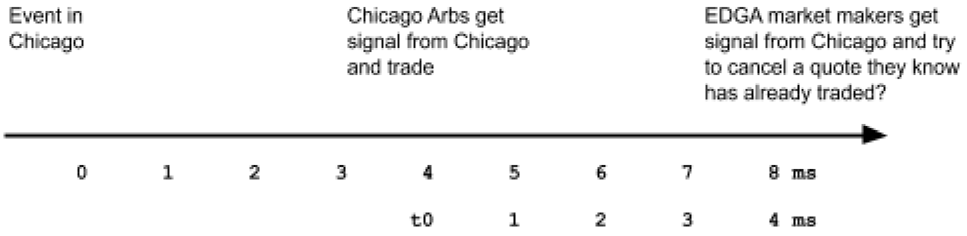
A paradox



Let's start by noting an apparent paradox with this graph. Since the yellow line[6] sample selection ("Yellow Sample") conditions on a failed market maker cancel, if Cboe EDGA's narrative in the Filing is correct, we can assume at t_0 in the graph (trade time) we are four milliseconds along from the relevant event in Chicago. According to the Filing, that's because there is a group of superfast latency arbitrageurs using high-speed microwave connections to get to Cboe EDGA's data centers in New Jersey from the futures markets in Chicago in as little as four milliseconds ("Chicago Arbs").

On the other hand, as Cboe EDGA described it in the Filing, its market makers get their signals from Chicago in about eight milliseconds because they use slow fiber. Since its market makers rely on those slower signals from Chicago before trying to cancel, it will take at least another four milliseconds after t_0 before market makers will even know they should cancel their quotes.[7] And then their cancels fail because the Chicago Arbs have already traded with their quotes. Now we're left with the paradox of why the market maker is trying to cancel a quote it surely *already knows* was traded out milliseconds ago by the Chicago Arbs, by the firms that got to Cboe EDGA so much faster because they use microwave connections. Does it take Cboe EDGA four milliseconds or more to report a trade?

[5] Cboe 1, page 4.
[6] All images have been enhanced to make the yellow and purple lines more legible.
[7] See the discussion of microwave and fiber and their associated latencies in the Filing at page 6.


Event in Chicago
Chicago Arbs get signal from Chicago and trade
EDGA market makers get signal from Chicago and try to cancel a quote they know has already traded?
0
1
2
3
4
5
6
7
8 ms
t0
1
2
3
4 ms

Unless it takes Cboe EDGA four milliseconds to report a trade, I don't see how the Yellow Sample shows Chicago Arbs picking off stale quotes. If anything, the purple line - trades where market makers didn't attempt and fail to cancel their quotes ("Purple Sample") - includes Chicago Arb activity. Cboe describes the Purple Sample as showing "relatively stable prices following an execution."

All this suggests that for the Yellow Sample the SPY signal to cancel is coming from somewhere closer than Chicago, or that contrary to the Filing some or all of the Cboe EDGA market makers use something faster than fiber. It might also suggest this graph and the others like it are nonsense. We can't resolve this paradox here, though Cboe EDGA might by telling us how many of its market makers use microwave.

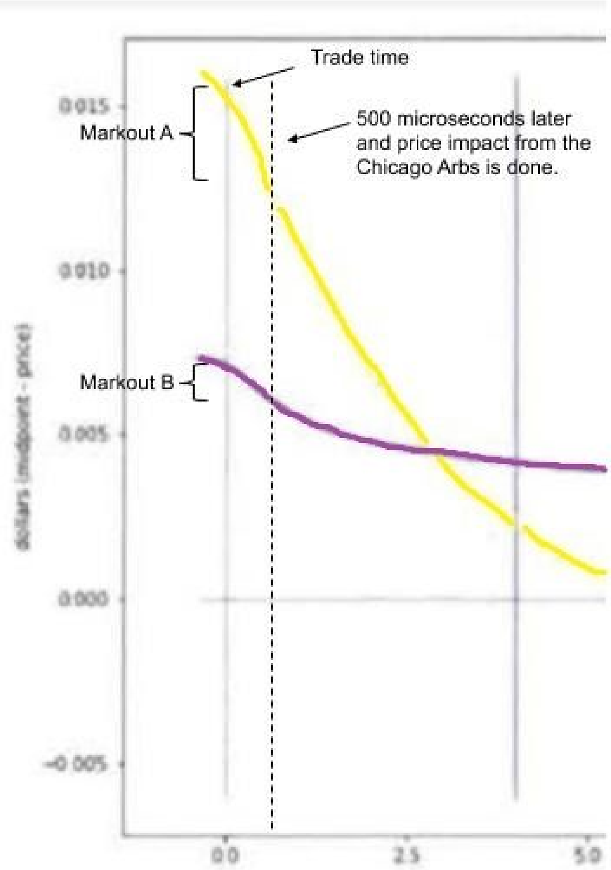
Gift card

Moving ahead anyway and stipulating to Cboe's presentation, let's deconstruct the graph. The main area of interest is on the left of the graph, in the first few moments after trade time. If the Chicago Arbs get from Chicago to New Jersey in about four milliseconds, at trade time - that is, t_0 in its markout calculations or at 0.0 milliseconds in the graph (the graph origin) - for the Yellow Sample, and assuming all those trades were initiated by Chicago Arbs, as noted above we're already four or more milliseconds along from the relevant event in Chicago. In that time the Chicago Arbs have shipped their data to New Jersey and are picking off stale quotes wherever they can (hence the trade itself in the Yellow Sample, of course).

The relevant analysis period to understand the price impact of the Chicago Arbs, then, is in the first few hundred microseconds after the trade at t_0 and not any further along. That's because by the trade time at t_0 the Chicago Arbs as a group have received their data from Chicago, processed it, and executed their trades. All that's left is to generate a new NBBO.[8] We'll generously add 500 microseconds past t_0 for Cboe EDGA and other exchanges to report trades and new quotes to the SIP and for the SIP to process them and update the NBBO (the NBBO determines the price impact shown in the graph).[9] After those 500 microseconds, *all price impact from the Chicago Arbs is long since finished regardless of when market makers try to cancel their quotes.*

[8] If the Yellow Sample includes all trades where a market maker tried but failed to cancel its quote before a trade was executed against it, following Cboe's narrative all successful arbitrage behavior from the Chicago Arbs as a group must then be included in the graph at t_0.

[9] Famously, IEX's universal speed bump is 350 microseconds. IEX arrived at that figure because of the geography of New Jersey equities market data centers. In other words, a 500 microsecond window allows time enough for Cboe EDGA and other exchanges to send trade reports and quote updates to the SIPs and for the SIPs to process that data.



The markout attributable to the Chicago Arbs - and that's assuming there aren't any coincidences and all the activity in the Yellow Sample is due to arbitrage[10] - is about .25 cents ("Markout A" in the graph). The control sample, the Purple Sample, has a markout about half as large, or about .125 cents ("Markout B" in the graph). In other words, taking the graph as presented and under the most favorable assumptions to the Filing, the "harm" done on Cboe EDGA by the Chicago Arbs is (Markout A - Markout B), or about .125 cents per share.

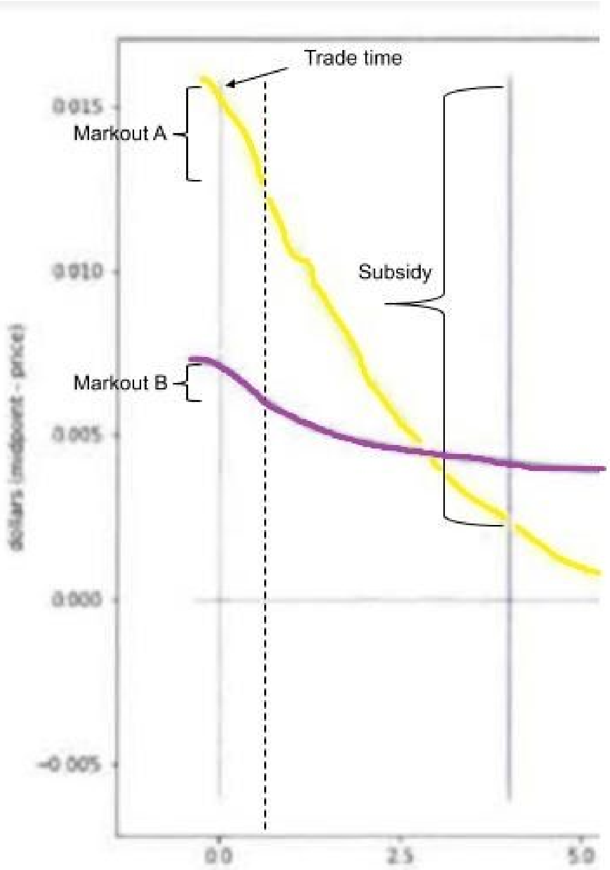
What's that work out to? We'll have to estimate that because Cboe hasn't provided any data. Looking at 2019 trades on EDGA in SPY, it could be as high as $90 a day.[11] Cboe EDGA wants to make one of the most significant changes to U.S. market structure in decades because the Chicago Arbs are costing its market makers less than $100 a day. We could all chip in for a nice Starbucks gift card instead.

[10] Without supporting evidence Cboe suggests that all the trades included in the Yellow Sample are done by Chicago Arbs and all the trades in the Purple Sample are done by "investors." If I understand Cboe, in the entire month of July 2019 there aren't any coincidental failed cancels in the Yellow Sample, there aren't any investors taking liquidity in the Yellow Sample, and there aren't any Chicago Arbs taking liquidity in the Purple Sample.

[11] Based on back-of-the-envelope calculations for Q2 2019 SPY trade data on EDGA. Assumptions here include that the Yellow Sample is 5% of all EDGA trade volume and a .125 cents per share markout attributable to Chicago Arbs, as above. If Cboe doesn't like this estimate, it can provide its own.

The subsidy

If the speed bump is implemented, I've argued that Cboe EDGA's market makers will receive a valuable regulatory subsidy.[12] We can use the graph to estimate what that investor-funded subsidy amounts to for trading in SPY.



With a four millisecond speed bump that investor-funded subsidy could be $900 a day or more in SPY.[13] If we extrapolate that to all stocks, as we'll see in the next section, the total daily subsidy on Cboe EDGA is $75,000.[14] Extrapolating that to the market at large, conservatively, it works out to about $1 billion a year.[15]

[12] Letter from R. T. Leuchtkafer to Vanessa Countryman, Secretary, SEC, July 12, 2019 ("Leuchtkafer 1").

[13] Assumptions are the same as in note 11, above. The markout from t_0 to the four millisecond speed bump boundary is about 1.25 cents per share. If Cboe doesn't like this estimate, it can provide its own.

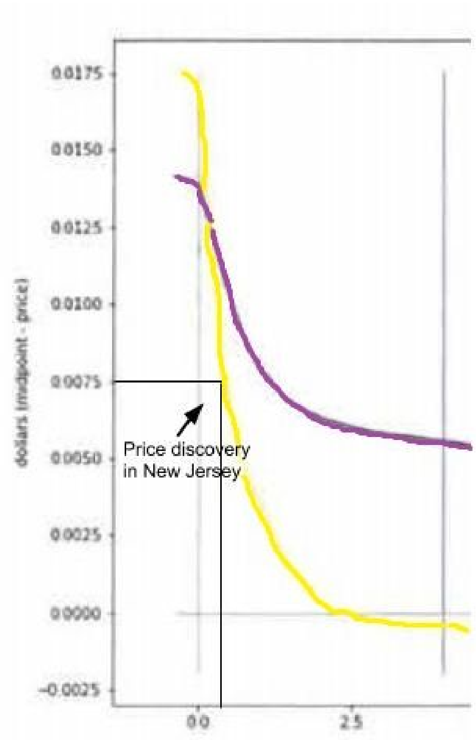
[14] Based on these simple variables: 150 million shares/day traded, 5% sweep/arb fade rate, a penny per share price impact avoided. If Cboe doesn't like this estimate, it can provide its own.

[15] Taking $75,000 a day for Cboe EDGA and then estimating for the market based on Cboe EDGA's approximate 2% market share of NMS stocks. If Cboe doesn't like this estimate, it can provide its own.

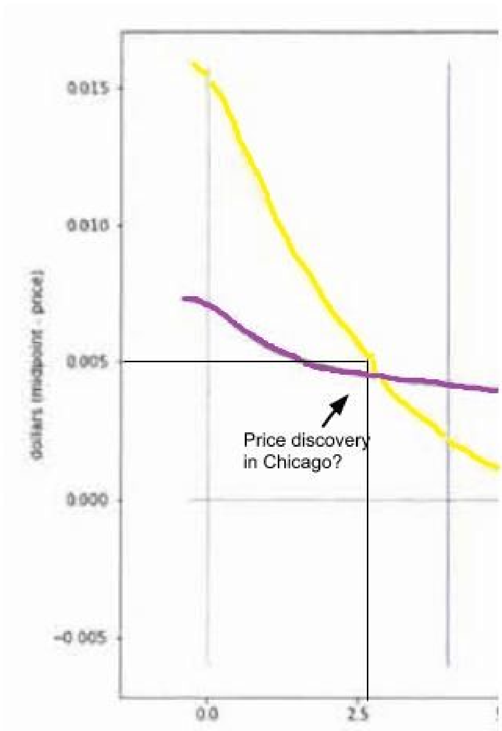
The real action

Though Cboe EDGA justifies its proposal because of the Chicago Arbs and a handful of ETPs like SPY, I believe the real purpose of the Filing is found nowhere near Chicago. The real action is in the data centers in New Jersey and with every security traded in those data centers.

There's a good sense of that in the graphs Cboe provides in the appendix to its letter.[16] The graph for UTX is instructive. In that example the price in the Yellow Sample drops like a stone in the first 200 microseconds or so, realizing approximately half of its overall price movement. In contrast, it takes more than ten times longer for the SPY Yellow Sample to realize half of its overall price movement (2.5 milliseconds).



UTX



SPY

Are the Chicago Arbs arbitraging UTX against the S&P futures contracts, trading against the equities markets in UTX much more rapidly than they do the markets in SPY? Of course not. The UTX graph is likely showing the effect of investor equities market sweeps. It has little or nothing to do with the Chicago futures markets. Since all the major equities market data centers are clustered in northern New Jersey, and since price discovery for a corporate stock is in the equities markets themselves, the price impact of an equities market sweep will unfold very quickly after t_0. Here we see that half the price impact in UTX is realized within about 200 microseconds and nearly all of it is realized within 2 milliseconds. In SPY, it takes about 2.5 milliseconds to realize half the price impact and more than 15 milliseconds to realize most of it. (And of all that in SPY, the Chicago Arbs might explain less than 10% of the total price impact, limited to the first few hundred microseconds.) The Chicago Arb story doesn't scale for corporate stocks, but Cboe EDGA's proposed speed bump will be the same in UTX as it is in SPY.

[16] Cboe 1, pages 18-20.

Research

Commenters have debated what research on TSX Alpha, a Canadian exchange, shows or doesn't show about discriminatory speed bumps at an inverted fee exchange. An academic study is pretty gloomy about it.[17] An IIROC/Bank of Canada study[18] is less gloomy, though hardly enthusiastic, and that's the study Cboe highlights.[19] Cboe offers a quotation from an IIROC notice about the study, that TSX Alpha " 'did not adversely affect the quality of Canadian markets,' "[20] but that's pretty weak beer if Cboe is looking for any kind of endorsement. Contrary to Cboe EDGA's pitch that its own speed bump "will improve market quality for investors"[21] and "make better markets"[22] and "promote liquidity provision"[23] and "improve displayed prices"[24] and "benefit all market participants,"[25] the IIROC/Bank of Canada study found TSX Alpha "did not impact market-wide liquidity" and could not "identify any significant impacts on effective spreads, price impact or quoted depth."[26] For certain participants, though, it found negative effects. Buy-side investors experienced "higher price impacts and effective spreads..."[27] The Ontario Securities Commission ("OSC") staff notice about the study echoes IIROC/Bank of Canada and goes further.[28] OSC reported that its own "market quality measures examined did not materially change..."[29] and that in OSC's survey of market participants, TSX Alpha "added complexity into routing decisions" and "in certain situations, fill rates on Alpha have decreased, often for orders that are expected to go through multiple price levels or need to be split and sent to multiple marketplaces simultaneously (e.g. institutional orders). Some dealers reported initial fill rates to be much lower on Alpha in these circumstances..."[30] None of which seems like any kind of an endorsement at all, and to my ear sounds like quote fading ahead of institutional investor equity market sweeps, making displayed prices less accessible.

A zero sum game

Taking what we've derived from the graphs, a rough estimate for the investor-funded regulatory subsidy a four millisecond speed bump grants to market makers on Cboe EDGA can run to at least $75,000 a day and $1 billion a year if extended to the whole market. It's a zero sum game. That subsidy will come from investors attempting but failing to trade with displayed prices and paying worse prices as a result.

Critics were scolded for years that ever speedier markets impounded new information into prices more efficiently, with spectacular benefits for all investors. But when investors caught on to the game, suddenly the story became speed kills and the SEC was approached for relief. No one would say "Thanks to *Flash Boys* many institutional investors now use fast smart routers, and we want a discriminatory speed bump so our high frequency traders can pull quotes in front of an investor's sweep in the dispersed price/time equities marketplace we designed, built, or bought."[31] They won't say so because the SEC's explicit

[17] Chen, Foley, Goldstein, and Ruf, "The Value of a Millisecond: Harnessing Information in Fast, Fragmented Markets," (2017).
[18] Anderson, Andrews, Devani, Mueller, Walton, "Speed Segmentation on Exchanges: Competition for Slow Flow," Bank of Canada Staff Working Paper 2018-3, January, 2018. ("IIROC/BC")
[19] Cboe 1, pages 10-11.
[20] Cboe 1, pages 10-11.
[21] Cboe 1, page 17.
[22] Filing, page 2.
[23] Filing, page 4.
[24] Filing, page 33.
[25] Filing, page 33.
[26] IIROC/BC, page 16.
[27] IIROC/BC, page 16.
[28] Ontario Securities Commission Staff Notice 21-712 ("OSC Notice"), available at https://www.osc.gov.on.ca/documents/en/Securities-Category2/20180202_21-712_sn-alpha-impact.pdf.
[29] OSC Notice, page 3.
[30] OSC Notice, page 4.
[31] Though Cboe inadvertently comes close to knocking on that door when it writes that "a very significant amount of institutional order flow is managed through broker-dealer algorithms that could respond to market information in less than this timeframe." (Cboe 1, page 10.)

policy goal for nearly a half-century has been to make displayed quotes *more* accessible to investors. But what else do the graphs show? The story about fat money arbitrageurs is just bait for regulatory rent-seeking.

That story collapses at a glance. The SEC is being asked to make displayed liquidity *less* accessible, to make it more difficult for investors to trade with displayed prices. That's what a discriminatory speed bump does, by design. Now that everyone understands how things work,[32] if market makers can't figure out how to trade in the marketplace they helped create, they should make room for firms that can.

Sincerely,

R. T. Leuchtkafer

[32] In 2018, Chris Concannon, President and COO of Cboe Global Markets, commented that *Flash Boys* had been a significant catalyst of direct feed sales because routing brokers (among others) realized they needed to have them. Not long after that the *Wall Street Journal* reported Cboe EDGA was considering a speed bump.