MANIPULATING ALGORITHMIC MARKETS*

Pedro Tremacoldi-Rossi
COLUMBIA UNIVERSITY

(Job Market Paper)

Click here for most recent version

Abstract

This paper develops a new methodology for causal price impact in high-frequency financial markets to study a widespread form of market manipulation and its consequences. I identify directly from data when a trader takes both sides of the same transaction but instead of letting orders cross uses a compliance tool to prevent legal exposure. This functionality is offered by every major exchange and in US futures markets its default use option allows the tool to be exploited strategically. This form of self-trading can effectively signal demand at artificial prices and result in disproportionate liquidity removal from markets. I introduce a source of variation that generates systematic differences in information exposure to traders. This leverages an institutional feature of electronic limit order books where as-good-as random delays between when a trade happens and the market learns about it can be used to assign treatment. By comparing trades occurring almost at the same time facing an identical information set, except for the news about a reference trade, I implement an empirical approach that estimates dynamic responses robust to microstructure noise and confounders. My findings show that self-trading successfully moves prices in the direction that benefits the trader, both by making liquidity providers revise quotes and enticing others to trade. I then use these estimates to quantify the role of self-trading in flash events: brief moments of substantial price increases or declines. Using a causal attribution framework, I separate information shocks — price adjustments based on news — from manipulative price impact to be able to assess the role of each factor individually and in combination. I find that almost 10% of flash events in US futures markets are driven by attracting others to trade in the direction consistent with profitable self-trading.

*I thank Dan Bernhardt, Matthieu Gomez, John Griffin, Harrison Hong, Scott H. Irwin, Andrei Kirilenko, Tatiana Mocanu, Conner Naughton, Alexei Orlov, Noémie Pinardon-Touati, Michel Robe, José Scheinkman, Simon Schmickler, and participants at the Commodity & Energy Markets Association Annual Meeting and the Financial Economics Colloquium at Columbia University for helpful comments and suggestions. I also thank the CME Global Command Center for discussions on CME's Globex infrastructure and for providing data. No trading strategy or conduct is linked to identifiable market participants in this study. The Office of Futures and Options Research (OFOR) at UIUC generously provided funding.
Correspondence: pt2614@columbia.edu.

"*It is not illegal to be smarter than your counterparties*" — Circuit Judge Richard Sullivan, in dismissed CFTC manipulation suit

1 Introduction

Trader could in theory simply delete order But exchange latency for a trade is almost double than exchange latency for cancel! That is, market participants take longer to detect self-trade

Manipulation and fraud have been part of trading since trading has existed. Stockjobbers (akin to exchange floor specialists) at London's Exchange Alley, the 17th century precursor of the London Stock Exchange, commonly made their brokers spread false negative stories while selling stocks to attract unaware sellers. As the selling pressure built, the stockjobber would have brokers promptly waiting to buy on the market at discounted prices. During the US railroad expansion in the early 1900s, several prominent cases of wash trades — when the same party takes both sides of a fictitious transaction — led to large volatility episodes, as when a Rock Island Company official ordered dozens of their brokers to buy shares of the company to lure in other buyers and inflate the company's share prices (Ripley (1911)).

In pre-electronic markets, bids and offers were matched by traders at the exchange floor. Floor brokers relied on hand signals, shouting, and paper tickets to make their quotes public, find and negotiate with opposite-side traders. Electronic limit-order books centralize this sequence of bilateral engagements by adopting a set of algorithmic rules that dictate how, when, and which orders are matched. Exchanges' matching engines are disciplined by an order priority rule and a host of checks intended to sustain market efficiency and stability. At the same time, high-frequency traders developed automated strategies exploiting the infrastructure of algorithmic markets that either actively employ or give the appearance of manipulative behavior.

Terms like quote stuffing, spoofing, and layering became the focus of regulatory policymaking and financial enforcement, especially after the triggering event of the 2010 Flash Crash was attributed to a British trader's spoofing. While most of these practices are banned and offer legal exposure or at minimum hefty fines by regulatory agencies, detecting defined manipulative behavior and meeting the legal burden of intent are challenging empirical tasks. Confounding behavior can arise in equilibrium (Baruch and Glosten (2013), Hasbrouck (2018), Williams and Skrzypacz (2021)). Seemingly abnormal market activity may be due to information or liquidity needs, and ultimately market conditions influence and are influenced by manipulation. Indeed, the U.S. Securities and Exchange Commission (SEC) never prosecuted more than 40 manipulation cases a year, whereas the U.S. Commodity Futures Trading Commission (CFTC) prosecutes only about 9 cases annually since its expanded mandate by Dodd-Frank.

This paper uses a new methodology and design features of algorithmic financial markets to provide a comprehensive study of market manipulation and its consequences. Our empirical approach identifies

directly from trading data a widespread form of market manipulation — self-trading. Because parallel management of multiple executing algorithms often puts a trader on both sides of the market, unintended order crossing is much more likely in modern financial markets than when humans traded at exchange floors. Every major exchange in the world offers participants self-trading prevention (STP) functionalities to stop orders from the same trader from fully crossing and potentially characterize wash trades, which are illegal.

When passive and marketable orders from the same trader cross, the STP functionality must decide which of the orders to cancel. Many exchanges leave it up to the trader to choose the default option, but at the Chicago Mercantile Exchange (CME) — the world's largest derivatives market — only passive orders are canceled by the STP in the event of a self-trade. The marketable order remains alive in the book, promptly matching with other traders' limit orders. My data contain direct signatures of order cancellations at the CME that can only be triggered by the STP functionality, effectively revealing all *prevented* self-trades at the exchange, the lifetime of the passive orders canceled and of the marketable order that remains publicly displayed.

Our paper proceeds in three main parts. In the first part, I focus on documentation. Using data from several of the largest futures markets in the world, this step underscores the pervasiveness and algorithmic nature of self-trading. At least 0.5% of posted liquidity self-trades without ever being filled by other traders' orders. In corn futures alone, this amounts to $3.2 billion/year of front-contract liquidity provision removed due to self-trading. As much as 7% of trades include at least one self-match, implying that a non-trivial share of marketable orders removes more liquidity from the limit order book than their actual entry size — both by normal execution against standing orders and by triggering STP cancellations of the self-trader's limit orders.

A quarter of self-trades happen very fast — within 100 milliseconds, with over 10% executing under 1 millisecond and about 1% under 10 microseconds. This implies that not only a considerable share of self-trading is an algorithmic activity, but it is performed by a subset of market participants with access to latency edge, as execution times approaching sub-microsecond require cutting edge hardware, e.g., field-programmable gate arrays. These findings are robust intraday and during long time series, over a variety of assets and underlying market regimes. Although at this step of our analysis I ascribe no intent to these aggregate statistics, the volume of self-trading I document matches a 2013 ominous remark by CFTC's then commissioner, Bart Chilton: "*If this were 0.4% of trading I wouldn't be giving speeches about it. These are whole percentages... high-frequency traders engage in wash trades in voluminous instances*".[1]

The second part of the paper focuses on quantifying the market impacts of manipulation. While self-trading may be accidental, there are material advantages to exploit the practice. In our framework, self-trading involves two sequential types of market impact. From the moment the self-trader's limit

[1]This also tracks a 2015 Joint Staff Report by the U.S. Department of the Treasury, the Board of Governors of the Federal Reserve System, the Federal Reserve Bank of New York, SEC, and CFTC studying a liquidity dry-up episode on October 15, 2015 in the treasuries market which identified high levels of self-trades (up to 5% of total volume).

order enters the book until it is cancelled by the STP functionality, it acts much like spoofing if the trader never intended to provide actual liquidity. Up to 40% of the limit orders that eventually self-trade improve quoted prices when entered in the book. They are also larger than regular limit orders. Both price and size signals increase the probability of imbalanced order flow build-up in the short-term by other trades.

Above and beyond spoofing, however, self-trading involves a trade execution, triggered by the trader's marketable order on the opposite side of the order flow build-up. Imagine that the manipulator wants the price of an asset to jump. The trader enters a new higher bid order and after other traders join the same side of the order book, she executes a marketable sell order. This increases the trade price of the asset and triggers the STP tool to delete the resting buy order. Traders receive two conflicting directional signals, making the net trading pressure impact ambiguous.

Detection algorithms from high-frequency traders scanning trade imbalance (e.g., momentum strategies) identify a seller-initiated trade with potentially significant liquidity taken off the book (actual fills and STP-cancellations). This is a sell signal and could trigger a marketable sell order in response. Very short-term directional execution is a common response by algorithms that would interpret the seller-initiated order by the self-trader as informed price pressure. By attempting to trend-chase, they amplify their signal reading. Other detection algorithms incorporating information on cumulative return and order flow imbalance receive the opposite signal — relative to the self-trader's bid entry, trade price jumps because of the seller-initiated execution at a higher price. This fact combined with the net buying order flow built-up during the spoofing-like stage sends a buy signal and could trigger buy orders. As long as, even if temporarily, aggregate net buying dominates selling, the self-trader is able to successfully make the asset's price jump.

Naturally, observing price movements, order flow, or trade imbalance following the entry of the self-trader's limit order is insufficient to tease out the market impact of manipulation. To empirically connect book updates to a self-trader's actions, I develop a methodology that segments order flow that could be reacting to information introduced by a particular order. I exploit the fact that most exchanges send out public updates about changes to the limit order book slightly *after* those changes are processed in the exchange's matching engine because of system latency. This allows us to (1) disconnect orders arriving before traders can learn about a market update, and (2) measure the probability that when the self-trader begins her strategy, other traders were already trying to trend-anticipate or ignite momentum.

Our approach suggests large directional order flow movements during the spoofing component of self-trading. Buy (sell) limit orders from self-traders attract large volumes of other buy (sell) orders, even when using placebo bids (offers) from other traders entered almost at the same time, but that the market learns with delay. Because limit orders cancelled by the STP functionality are larger than the average contemporaneous order, flow processing trading algorithms likely assign a higher weight to the signal provided by the self-trader's order. Indeed, after controlling for size and price, order flow build-up

following the self-trader's order is identical to the average order *unconditional* on entry time. This is consistent with the self-trader not trend-chasing and instead contributing to momentum ignition.

I then document trading pressure around the moment when orders from the same trader cross and trigger the STP functionality. I find a two-part market response. Within one second following the self-match, net trading pressure with the same direction of the self-trader's marketable order builds. After that, trading pressure reverses and follows the same direction as the order flow build-up. Larger existing order flow imbalances during the spoofing stage lead to more long-lasting directional trading pressure. Taken together, these findings suggest that, if the average self-trade intended to affect the price of an asset, the market response associated with it is consistent with the strategy being plausibly profitable.

In the third and final part of the paper, I turn to causally linking market design to incentives and returns to manipulation. The use of STP shields traders from the legal possibility of executing a wash sale in case of crossing orders. Whereas self-trading has a similar effect on markets as wash trading, abuse of the STP functionality is considered by exchanges only as "disruptive practice". This amounts to little more than a slap on the wrist of violators compared to severe penalties to wash sales in place since the Commodity Exchange Act of 1936. But the interplay between microstructure design and algorithmic manipulation goes well beyond the strategic use of compliance tools to potentially manipulate prices.

Leveraging two separate natural experiments that unexpectedly modified the protocols used by the exchange to match orders in the limit order book, I show how matching algorithms that either reward price improvement, speed, or size of liquidity provision generate additional gains to self-trading. The first experiment changed the quantity traders that improve quoted prices have reserved when matched with liquidity takers. In 2018, the CME increased by ten-fold this allocation parameter in corn futures markets, while leaving other grain contracts operating under the same matching rules unchanged. The second experiment affected interest rates markets in 2020, when the exchange gave complete matching priority to the fastest traders quoting on a price level in the 3-year treasuries market, making its matching algorithm the same used in 5 and 10 year T-notes. Prior to the change, 3-year rates followed the same allocation rules used in 2-year treasuries, which remained the same.

To connect the shocks introduced by these rule changes to order flow activity, I identify three sophisticated strategies that exploit self-trading to decrease the inventory cost — either in terms of execution risk or bid-ask spread — necessary to gain allocation priority. These strategies involve the sequential execution of several order types on opposite sides of the market, making it highly unlikely for them to happen by chance. Indeed, extensive and intensive margin responses induced by the rule changes indicate that traders exploiting the STP functionality avoid paying the full cost of price impact. Finally, I complement our high-frequency market impact results by instrumenting for the intensity of self-trading with the reforms.

Related work. I contribute to a growing literature on market manipulation detection and its effects. Studying manipulative behavior empirically is challenging because multiple phenomena can generate

similar observable patterns, as with quote stuffing which may potentially be intentional (Ye et al. (2013)) or arise in equilibrium from high-frequency trading competition (Hasbrouck (2018)). Most of the literature has exploited publicly available filings of successfully charged cases of misconduct (e.g., Aggarwal and Wu (2006), Akey et al. (2020)), since forensic-type analysis as done by Lee et al. (2013) may also suffer from many confounding factors. I provide a direct assessment of manipulative behavior by studying a sequence of events linked to an order lifespan that identify self-matches from market data. This allows us to cleanly measure market effects from manipulative behavior.

Studying self-trading in electronic markets has two advantages over analyzing other types of market misconduct. First, self-matched orders are indirectly flagged by the exchange data feed, which objectively defines the event space and does not depend on subjective criteria to define a certain practice as predatory (e.g., how many cancellations one needs to observe to infer manipulative behavior). Second, similar to quote snipping (Aquilina et al. (2021)) and strategic runs (Hasbrouck and Saar (2013)), non-accidental self-matches are accompanied by a characteristic path of quoting behavior.

2 Data and Setting

I begin by noting a technical difference between a **match** and a **trade** events. A match occurs when a marketable order on one side of the book matches with one or more resting orders on the other side of the LOB. Under normal circumstances, a match results in a trade, where the exchange sends out a trade message confirming execution to the parties involved in the match. This is followed by inventory changes in each side's brokerage account during settlement and netting. The distinction between match and trade in our context is important because a match may fail to execute a trade, for example, when the only orders matching are from the same trader and the self-trading prevention functionality is active. In this case, orders cross and match, but the tool steps in to prevent a trade between these orders.

Because exchanges and regulators refer to the practice that triggers the STP functionality as self-trading, even though the only activity involves self-matching, I use the two terms interchangeably throughout the paper, only making a distinction when relevant.

2.1 Regulatory Background

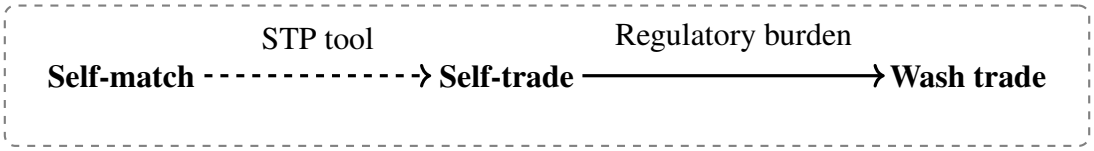
Wash trades have been illegal in the US since the Commodity Exchange Act of 1936. The Law specifies that any person entering into or confirming the execution (i.e., brokerage firms) of a wash trade (also referred in its text as accommodation, fictitious, or cross trade) engages in unlawful trading behavior. In compliance with these rules, exchange regulations explicitly forbid wash trading. For example, CME's trading rulebook

> RULE 534 (WASH TRADES PROHIBITED) — *No person shall place or accept buy and sell orders in the same product and expiration month, and, for a put or call option, the same*

> *strike price, where the person knows or reasonably should know that the purpose of the orders is to avoid taking a bona fide market position exposed to market risk (transactions commonly known or referred to as wash sales). Buy and sell orders for different accounts with common beneficial ownership that are entered with the intent to negate market risk or price competition shall also be deemed to violate the prohibition on wash trades. Additionally, no person shall knowingly execute or accommodate the execution of such orders by direct or indirect means.*

clearly states that both individuals engaging in self-trading and brokerage firms enabling the conduct are subject to regulatory consequences.

Sidestepping how to define the legal threshold necessary to determine whether a market participant "*knows or reasonably should know*" that orders placed would cross-trade, the necessary condition for a wash trade is that a trade event occurs. Therefore, even if orders from the same trader were to cross — a match — as long as a trade event is not triggered, that action would not qualify even as a potential wash trade. Building off of this principle, exchanges began offering in the early 2010s order management tools that would ensure a trade event would not occur whenever two orders from the same account crossed, essentially insulating market participants from technically registering a trade after a self-match event.



2.2 Self-Trade Prevention Functionalities

In June 2013, CME introduced an optional functionality in its electronic trading platform, Globex, to enable brokerage firms to avoid self-trading. CME's Self-Match Prevention (SMP) tool automatically cancels a trader's resting order in case of a match with an aggressing order from the same trader. The functionality allows users to change the default option, canceling the incoming (newest) order instead of the resting order (oldest). The functionality is offered for free and adds no latency penalty as it filters out the aggressing order in the trading engine, not in the outer exchange gateway.

Defining "same trader". Brokerages choose at which operational level their orders, if self-matched, would imply a potential wash-trade risk under CME's definition of common beneficial ownership. The common interpretation in the industry is to consider orders submitted for the same account or accounts that have common ownership (e.g., different desks at a proprietary high-frequency shop operating in a common market). In practice, a trading firm requests an SMP ID tag with the CME, which tags all orders submitted by the firm within or across clearing centers as belonging to the "same trader". Because the proper use of the SMP tool is to the advantage of the brokerage, firms have incentives to narrowly define common ownership.

Self-trading prevention in other asset classes. Virtually all centralized exchanges offer self-trade prevention functionalities. These functionalities largely follow the same engine as CME's SMP. The Intercontinental Exchange (ICE) has self-trading protection services since 2013 and makes the use of the tool mandatory for proprietary traders using algorithmic trading applications. In April 2021, Eurex also made the use of SMP mandatory to proprietary algorithmic firms. Self-trading protection tools at Nasdaq and NYSE are optional. Across these exchanges, traders have the choice of canceling the resting or aggressing order, or both, in case of a self-match.

2.3 Enforcement

Despite being a classic form of market misconduct, successful prosecution of wash trading is relatively rare. Enforcement ability stems from ideally establishing intent ("*knows or reasonably should know*") or at minimum quantifiable prejudice (harm) to other traders. These are empirically difficult to show beyond reasonable doubt. Self-trading prevention tools are designed to provide "insurance" to trading firms against wash sales risk by preventing self-matches from turning into executed trades. By deleting only resting orders when a self-match happens, however, these functionalities are not market neutral. This implies that they can be exploited strategically. While the use of STP functionalities always insulates traders from potential wash trading, manipulation of the compliance tool is still a potential violation of trading rules — a much less severe offense and enforced only by exchanges.

The CME RULE 575 (DISRUPTIVE PRACTICES PROHIBITED) bans orders not entered "*in good faith for legitimate purposes*". The rule's text directly addresses the possibility of manipulative use of the self-trading prevention tool: "*The use of self-match prevention functionality in a manner that causes a disruption to the market may constitute a violation of Rule 575. Further, if the resting order that was cancelled was non-bona fide ab initio* [to begin with], *it would be considered to have been entered in violation of Rule 575*".[2]

To the extent that enforcement cases of wash sales are rare, cases involving disruptive trading practices are even more scant. Since 2013, only a couple of instances of abusive use of the STP functionality were punished. In 2015, the high-frequency trading firm Allston Trading was investigated by the CFTC following the complaint of another trading firm, HTG Capital, that the firm used CME's STP tool to spoof treasury futures. In 2018, CME issued a notice of disciplinary action against a trader that violated RULE 575.

The notice describes that the trader "*entered orders on one side of the market at the best bid/offer without the intent to trade. After other market participants joined his resting orders,* [the trader] *entered an aggressive order on the other side of the market at the same book level.* [The trader] *used self-match prevention software, which caused his resting orders to be cancelled within the same millisecond of entry of the aggressive order. The aggressive order then immediately traded opposite the market participant*

[2]See it here: https://www.cmegroup.com/rulebook/files/cme-group-Rule-575.pdf. RULE 575 broadly targets spoofing practices.

who joined or bettered [the traders] *resting bid/offer and turned the market*." The trader had to pay about $90,000 in fines for a $10,000 disgorgement.

2.4 Trading Message Data

This paper uses high-frequency trading data from multiple futures markets from the Chicago Mercantile Exchange (CME). The CME represents an ideal setting to study the impact of self-trading in modern financial markets because of how its matching engine handles self-trades. With sufficient institutional knowledge, it is possible to identify directly from the data orders that cross from the same trader.

CME commercializes a version of feed data packaged as Market-by-Order (MBO), which provides all information necessary to identify self-trades, as well as native and synthetic iceberg orders, every update during an order's lifetime, and detailed information on trade events, including partial fills.[3] This level of detail on market messages combined with the default option of CME's self-trading prevention functionality — which deletes the resting order involved in a self-trade — allow me to construct a simple algorithm that completely recovers the subset of self-trades.[4]

Data sample. I obtain MBO message data from the CME spanning several months and markets. The core analysis uses outright futures contracts of E-mini S&P 500, 2-year and 10-year t-notes, gold, and oil from October 2019 to March 2020. These markets are among not only the most liquid futures in the world, but the most widely traded products across all asset classes. To give a sense of the magnitude of the message data, there are over 60 billion updates to the limit order book across these markets in the 6-month period.

Institutional details of trading infrastructure. The data contain message packets disseminated by the CME in FIX/FAST protocols. Messages are timestamped in nanoseconds (with microsecond precision guaranteed), where I observe two timestamp types. Transaction timestamp — generated when an incoming message leaves CME's gateway and is processed by the matching engine — and sending timestamp — generated when an outbound message to the data feed leaves the limit order book software system. This is when the information about what happened at the transaction timestamp becomes public.

[3] Because CME's messages are disseminated under a FIX protocol instead of ITCH or similar systems, the exchange usually offers to the public data that only shows partial book depth (either top of the book, first five or ten best price levels), and tracks markets only by level (or price) updates. These data have no comprehensive order IDs or order-level status changes, which confounds several different market phenomena. For example, trade summaries — message packets sent out to market participants detailing all matches involved in a trade — from market depth data make it impossible to properly track iceberg or implicit orders, which may result in type I or II errors when attempting to identify self-trades. In contrast, with MBO data one can track orders and trade summaries to identify orders involved in self-matches.

[4] Deletion of the aggressing order instead of the resting order results in no direct market implications, except for using the exchange's data feed and potentially queuing other incoming orders by imposing latency externalities.

The gateway is the outerwall of CME's electronic market Globex and clocks the arrival time of orders as they are routed and processed to the matching engine (see Figure 1). Transaction timestamping is crucial to track self-trading because a trade execution message and a delete message with same transaction timestamp were hardware-clocked at the same instant. Sending timestamps may be in theory identical because of transmission latency from the exchange between when a change is recorded in the limit order book and sent out to market participants, not because those events were recorded at the same time. In the data, however, sending timestamps almost never coincide.

Trading at CME happens almost without interruption, except for an afternoon period from 1:20 PM to 7:00 PM and a brief pause in the morning during 7:45–8:30 AM. Commodity markets operate without designated market makers, while treasury futures have a number of designated firms, including Goldman Sachs, Morgan Stanley, Nomura, and Allston Trading. Contract specifications vary widely across products — including maximum lot and tick sizes — as well as which algorithm the exchange uses to match orders. These differences in the plumbing may shift incentives to self-trading across products.

3 Anatomy of Self-Trading in Futures Markets

3.1 Detection Algorithm

The first step is to identify self-trades in the market-by-order data. These are strictly flagged as self-traded and handled by the STP functionality. Though the data do not have a "flag" for an order that self-matches, a simple detection algorithm guarantees that I can observe every time traders used the default option in CME's STP tool. In Figure 2, the best ask price of $100 has six orders of varying lot sizes. As usual in US trading data, market participant IDs are completely masked, so that the traders behind these six separate orders are unknown. An incoming buy order at $100 of size 700 then hits the market. Panel B in the figure shows the trade summary message for the event, which tracks the history of each order's fulfillment against the aggressing order. Note that even though the best ask level was 635-lot deep, only 160 out of the 700-depth of the market buy was filled.

Inspecting the trade summary records more closely reveals the reason for the sub-fulfillment. At the same exact timestamp, 3 out of the 6 ask orders are canceled with a filled quantity of zero, while the other 3 orders are completely filled. The only possibility for the three no-fill deleted orders is that they were removed from the book by the self-trading tool because they were sent by the same account as the crossing buy order.

This simple detection algorithm exploits the fact that CME's STP feature prevents traders from aggressing into their own resting orders by canceling the trader's passive quotes. Crucially, the marketable order remains alive in the order book. These audit-type records allow me to sidestep forensic or indirect approaches to *infer* self-trades. This is an important feature: it is very rare for researchers or

regulators to *know* when questionable market conduct happens. Some signals happen so often that are almost uninformative — a trader canceling too many orders — or so sparsely that separating them out from "noise" is costly and usually requires targeted financial incentives, which despite being potentially large, tend to be paid only if a successful prosecution or settlement are reached. In short, one faces significant measurement error or sample selection issues already from the start of the analysis.

3.2 General Patterns

Self-trading happens often. I begin by establishing broad patterns of self-trading activity in futures markets. Table 1 shows that on average 0.65% of posted liquidity is removed from the limit order book via self-trading. This statistic tracks every unique order across all first-nearby futures markets in our sample (buy and sell sides) from its entry time until it eventually leaves the book. About 380,000 limit orders were entered and canceled by the STP functionality during our sample period. This amounts to an annualized notional value of over $46 billion in posted liquidity that effectively is never supplied.

When tracking every trade event — where a marketable order can match with multiple resting orders — almost 4% of trades in futures markets involve at least one STP-cancellation. Marketable orders in self-trades are about twice the average size of a regular order crossing the spread — 10 to 5 contracts — executing an annualized value in excess of $100 billion. Remember that because self-trades necessarily remove posted liquidity by triggering the STP tool, the total liquidity take-up from these marketable orders is even greater than twice that of regular trades (in aggregate, an annualized notional liquidity taking impact in excess of $150 billion).

Interestingly, self-trades are less likely to fully consume liquidity at the best price level on the other side of their entry. Sweeping self-trades are about 4% while regular trades amount to around 5%. Because marketable orders that self-trade are larger and also consume more posted liquidity, a lower sweep rate is consistent with self-trades executing against deeper top-of-book levels on average.

Self-trades consume more liquidity than regular trades. Table 2 shows the magnitude of additional liquidity taken from the limit order book due to automated cancellations by the STP functionality. On average, over 37% of liquidity taken when a trade involves at least one self-match is a STP-triggered cancellation. When smaller aggressing orders are executed, this fraction is as high as 48%. Most self-trades are triggered by small marketable orders, between 1 and 5 contracts. As the size of marketable orders in self-trades increases, more liquidity is taken via fulfilment. Note that this result is not necessarily mechanical — if self-traders jointly scaled resting and marketable orders, quantity slippage could remain relatively constant across aggressing size. The negative relationship shown in the table is consistent with considerations regarding inventory control or price impact.

The last column of the table computes the notional value in aggregate of fleeting liquidity. These are limit buy or sell orders that are deleted by the trader within 500 milliseconds. While this time window

is much longer than the average high frequency trading activity, it is short enough to filter out human corrections to "fat-finger" errors (e.g., a trader manually deleting an order initially entered by mistake). Fleeting liquidity induced by quote flickering (e.g., Hasbrouck and Saar (2009), Baruch and Glosten (2013)) is an interesting benchmark in this context because even if all of this liquidity was bona fide at entry time (that is, no layering, wire warming or bandwidth congestion), ultimately cancellations occur to evade execution. Similar to STP-triggered cancellations, unless other traders are faster, in practice fleeting liquidity is not fully executable liquidity.

3.3 The Economics of Self-Trading

Self-trading involves two sequential types of market impact. From the moment the self-trader's limit order enters the book until it is cancelled by the STP functionality, self-trading acts much like spoofing. This sends the first market signal, generating order flow impact. Above and beyond spoofing, self-trading also involves trade execution, triggered by the trader's marketable order on the opposite side of her limit order. This is the second market signal, which generates trading pressure impact. Whether either of these pressure sources affect other traders is an empirical question I investigate next.

Step 1: Order flow buildup. The first market impact component provides order flow signal to other traders across two potential dimensions: price and size. Traders only observe demand and supply signals at existing quotes. New quotes that narrow the bid-ask spread send credible price signals since the entering trader faces execution risk alone. Size also matters — larger orders increase depth at newer or existing price levels more than smaller orders. Price and size signals combined increase the probability of imbalanced order flow build-up in the short-term.[5]

When a self-trader betters the outstanding top-of-book bid or offer quotes by adding liquidity, other traders receive an updated demand price signal. Even if the self-trader only adds liquidity to an existing price level, large enough orders still contribute to perceived imbalance of order flow. When successful, bona fide liquidity from other traders joins the self-trader's price level — or further narrows the spread. In the spoofing equilibrium of Williams and Skrzypacz (2021), bona fide order flow imbalance arises as a consequence of the market maker being unable to pinpoint whether an order cancellation *before* execution reflects a sincere change in liquidity needs or spoofing. In our context, a strategic trader has no need to manually cancel a posted quote — execution risk is at minimum partially offset by the ability to self-trade, which triggers the STP functionality to cancel an order during execution. Up until a self-match, other traders have no reason to consider whether order flow from a self-trader is any different from a sincere trader's flow.

Figure 3 illustrates an example of this dynamics when the self-trader enters a passive buy order without setting a new price, i.e., affecting the limit order book only through size. Following the order

[5]Traders never fully observe true supply and demand since most inventory needs are managed through block trades in sequential, coarse execution, which are latent at a given point in time (Donier et al. (2015)).

entry, other traders' orders join the buy-side of the market, generating considerable net buying order flow until the self-trader executes a marketable sell order.

Step 2: Trading pressure impact. Self-trading not only potentially generates order flow impact, but also induces trading pressure impact. To see how, imagine that a self-trader wants the price of an asset to jump. She enters a limit buy order and after order flow buildup, the trader enters a marketable sell as in Figure 3. This increases the trade price of the asset and triggers the STP tool to delete the resting offer. What signal does this execution event send to other traders?

Traders receive two conflicting directional signals, making the net trading pressure impact ambiguous. Detection algorithms from high-frequency traders scanning trade imbalance (e.g., momentum strategies) identify a seller-initiated order with potentially significant liquidity taken off the book (actual fills and STP-cancellations). This is a sell signal and could trigger a marketable sell order in response. Very short-term directional execution is a common response by algorithms that would interpret the seller-initiated order by the self-trader as informed price pressure. By attempting to trend-chase, they amplify their signal reading.

Other detection algorithms incorporating information on cumulative return and order flow imbalance receive the opposite signal — relative to the self-trader's offer entry, trade price jumps because of the seller-initiated execution at a higher price. This fact combined with the net buying order flow built-up during the spoofing-like stage sends a buy signal and could trigger buy orders. As long as, even if temporarily, aggregate net buying dominates selling, the self-trader is able to influence the price upward.

Empirical considerations. The picture I depict above is a general characterization of market manipulation in what the literature names "trade-based" manipulation (e.g., Allen and Gale (1992), Putniņš (2012)), which broadly boils down to pump-and-dump dynamics and bear raids. Attributing order flow build-up or trading pressure to self-trading faces the key identification challenge in empirical market microstructure — a trader's action affects the limit order book and is affected by it. Spelling out the relevant counterfactual in our case is simple: what would have happened to order flow and trading pressure impact without the self-trader's action? To answer this, some additional considerations regarding the conditions in which self-trading I first need to determine whether self-trading activity consists primarily of trend-chasing or trend-setting.

Trend-chasing. Identifying informed order flow is crucial for high-frequency traders to predict returns through anticipatory trading, back running and other trigger-style strategies (Yang and Zhu (2019), Baldauf and Mollner (2020)). While theory usually links large orders to informed flow (Glosten (1994), Easley et al. (1997), Biais et al. (2000)), in practice the use of meta-orders and randomized execution algorithms makes trend signals much more dispersed across order sizes. Yet, order flow remains highly serially-correlated (Gabaix et al. (2003), Gabaix et al. (2006), Tóth et al. (2015)), which suggests both

that trend signals are predictive and that some subset of traders pick up on these signals and amplify order flow persistence.

Trend-setting. Momentum ignition strategies involve setting or exacerbating trends directionally. The standard approach to limit losses with the strategy is to execute small marketable orders — 20 lots or less in the E-mini S&P 500 futures market according to Clark-Joseph (2013) — and profit from price impact when supply is sufficiently inelastic. Self-trading enables a trader to provide trend signals at lower expected inventory cost than simply executing against the market.

3.3.1 Order Flow Buildp and Market Signals

In Figure 1, I illustrated the stylized routing system of CME's trading infrastructure. Our data clocks two timestamps — a transaction and a sending time. In short, the transaction timestamp records the moment a certain event that will alter the state of the limit order book *happens*. For example, when a marketable order executes against liquidity. However, traders not participating in the event, i.e., the wide market, can only *learn* about the event when the public data feed is refreshed and disseminates a market update message about the event. Because of latency within this infrastructure, there is a delay between the moment an event happens and the market knows about it. Our historical data feed contain both timestamps and therefore I can organize order flow temporally depending on when a change to the limit order book occurs and when it become public information.

The first column in Table 3 shows the frequency of self-trades that are executed within 10 microseconds, 100 microseconds, 1 millisecond, 100 milliseconds, 500 milliseconds, 1 second, 10 seconds, 30 seconds, 1 minute, and after 1 minute from book entry. These are regular event times to make cross-market and time-series comparisons easier. A quarter of self-trades happen very fast — within 100 milliseconds, with a non-trivial share of 10% executing under 1 millisecond and about 1% under 10 microseconds.

Execution times approaching sub-microsecond are impressive even for high-frequency trading firms. Cutting edge hardware including field-programmable gate arrays (FPGA) are necessary to provide this level of latency so that a self-trader can respond with a marketable order after sending a limit order. This implies that not only a considerable share of self-trading is an algorithmic activity, but it is performed by a subset of market participants with access to latency edge. These are likely relatively few firms (Boehmer et al. (2018)).

Limit orders entered by self-traders further provide price signal — on average 9% of orders in oil markets improve bid or offer quotes (on average by 8 basis points), and as much as 40% of self-trades executed within 1 millisecond in treasuries narrowed the spread (on average by 12 basis points). Other traders react to the entry of these limit orders — and fast. On average, 2 unique orders join the price level within 1 millisecond in corn futures, with an average of 623 contracts added up to one minute. The average notional value of a corn contract in the period is about $18,500, which implies that over

$11.5 million in liquidity is supplied to the same price level of the self-trader's limit order in just one minute. Until the execution of the self-trade, on average another $15.5 million in liquidity is supplied. The average self-trading strategy execution time in corn futures is just over 2 minutes.

Responses in the wheat and soybean markets are broadly similar to corn futures, only smaller in absolute magnitude as these order books are less deep. Nonetheless, response time of other traders is similarly fast. A non-zero average quantity of orders respond to the self-trader's limit order entry in less than 10 microseconds, and I identify several instances in the data of sub-microsecond reaction from other traders. For reference, the human blink lasts about 100 milliseconds (100,000 microseconds). In wheat markets, 10% of the total added liquidity following the entry of the self-trading limit order happens within 100 milliseconds, which is certainly provided by algorithms.

These results speak to the ability of the spoofing component of a self-trading strategy in being followed by order flow impact.

4 A Framework for Estimating Causal Price Impact

Establishing causality has remained a long-standing challenge in empirical work with high-frequency financial data. Greater time granularity helps to separate events in time which could look contemporaneous in lower frequency. But time granularity increases the role of noise relative to economic signals, making strategies to shut off endogeneity sources difficult to implement. To a large extent, causally answering "what would have happened to prices without the trader's action?" remains a key challenge in the empirical study of trading.

This paper makes progress in this direction by combining data on an institutional feature of modern algorithmic markets and a new identification strategy...

The two main ingredients of my approach generate variation in the following way. This framework can be applied in any context where data available contains a public and internal timestamp,

The duration of this latency cannot be jointly explained by a wide range of market activity indicators, ,

including persistence in past latency (, message traffic (demanded the exchange's system factors including,

and market and minute fixed effects to ensure to try to capture whether

persistence in past latency,

explain empirics

My results show that self-trading moves prices

the prominent role of "noise" () becomes particularly important.

While noise in trading demand has motivated much of the theoretical models in market microstructure, its oversized role in

$$\mathbb{E}\left[p_{t+m} - p_{t-1}\right] - \mathbb{E}\left[p_{t^{\text{cont}}} - p_{t-1}\right]$$

Note that this rolling measure captures the peak momentum (or the steepest part of a flash event). To illustrate, consider a flash bull definition tracking minute-returns for a 3-minute period. If there are two top flash events with time overlap, I keep the largest one and drop other. Sequential and disjoint flash events are included in the sample

always taker: (patriot) someone who would've served regardless of the value of instrument: the treatment here is observing the self-trader's trade. So I measure the impact on the trade of the first "revealed" trade. They see the same book information as the self-trader (and the control group) did. What do I need to rule out? That this trade was driven by private information. Because this would've been a trade that would've always happened. This would've been an always taker.

$\mathbb{E}\left[Y|G=1\right]$: % large trades around self-trades (both for flash bull/bear or not) (expect flat)

$E[Y(T,1)|G=1]E[Y(0,1)|G=1]$

Information shock shouldn't be predicted by signed self-trader's trade (either way). In other words, the timing could be endogenous, but on average the sign x_{t+m} should be orthogonal to x_t controlling for pre-existing trend (which I are).

$p_{\tau_0-1} - p_{\tau_0-1}$: instantaneous price impact

$$ATT = \mathbb{E}\left[p_{s_0<t<s_1} - p_{t_0-1}\right] - \mathbb{E}\left[p_{t_0<t<s_0} - p_{t_0-1}\right] \tag{1}$$

The first term captures the transaction price of the first trade after s_0 but before any s timestamp of the contemporaneous orders. Trades in this interval may be reacting to the same information set as the self-trader and to the self-trader's order, which is already public information. They cannot however be reacting to contemporaneous trades because these were not yet made public to the market.

Another way to think about exchange latency is to see it as generating as-if-random variation in exposure to each trader's to publicly displayed information in the limit order book.

In short, beyond the instantaneous impact of the trader on the limit order book (which combines liquidity removed and potentially walking the book, an effect usually described as "transitory price impact"), any observed market response is subject to endogeneity.

Time t is tracked by two different time-stamps: τ is the transaction time and s the sending time.[6]

The causal parameter being estimated is the average treatment effect on the treated $\mathbb{E}\left[Y(1) - Y(0)|G=1\right]$

If there are no always-takers (), the causal estimand can be decomposed into $ATT \times P(X)$.

[6] trader$\longrightarrow$exchange (BGS) $\longrightarrow$matching engine$\longrightarrow$ {private feed, sending} $\longrightarrow$trader

4.1 A New Source of Exogenous Variation in High-Frequency Data

where δ_k measure the importance of the aggregate quantity of inbound messages of type k (trade, new, update, cancel) in that minute. I allow for potential persistence in latency, captured through ω_τ. Our goal with this specification is to understand how much of the variation in within-minute latency from the exchange's market data feed remains unexplained by a wide range of market activity indicators. Even if latency is as-if random with respect to market activity, external factors may affect its timing. For example, the exchange could perform routine checks of distribution servers in the morning or partially "cool off" some of its market data processing power during overnight sessions, when trading demand is more sparse. Minute fixed effects μ_q in the above specification (e.g., min 1, min 2, ..., min 5451, ...)

4.2 Empirical Setup

We are interested in estimating the causal price impact of a reference trade that occurs in $t = 0$ over $t + h$ subsequent events. Let x_0^* denote the reference self-trade with a transaction time τ_0^* and sending time s_0^*. The time interval $s_0^* - \tau_0^*$ measures exchange latency. Because the duration of this latency is as-good-as random (as shown in Table 3), it can be used to assign treatment.

Treatment assignment. $z_{n,t}$ corresponds to whether a trade (or order) was exposed to information about a reference event, with the change in treatment status between $t - 1$ and t being denoted by $\Delta z_{n,t}$. Treatment is assigned based on variation generated by exchange latency, which quasi-randomly exposes trades occurring almost at the same time to the same information set, except the "news" about the reference trade. Though in this setting self-trades are the reference events, more generally any group of orders in the limit order book (e.g. institutional trades, designated market maker quotes) can play the role of interventions.

With a treatment assignment plan, I now define the two building blocks of the identification strategy: treated and control trades. Trades (or quote updates) timestamped $\tau_0 \in (\tau_0^*, s_0^*)$ cannot observe x_0^*, and because no other update to the limit order book has become public since τ_0^*, I assume they had public information contemporaneous to the reference trade. I denote this set of trades by $\{x_0^{C_n}\}_n$, and they comprise our pool of control units (or clean comparison group).

Consistent with the standard market efficiency assumption in market microstructure models, I assume that all publicly available information up to τ_0 is incorporated in the limit order book, so that residual information (even if correlated across a subset of market participants) is treated as innovations. As I discuss below, such idiosyncratic shocks (e.g. private information, liquidity needs) are allowed to determine $x_0^{C_n}$, as long as the timing of execution is not completely orthogonal to the history of the

displayed limit order book. This further allows traders to differ in how they extract and use signals based on public information, including by endogenously determining order entry.[7]

Exchange latency around the reference trade directly assigns treatment to a subset of potentially impacted trades $\{x_0^{I_n}\}_n$ for which $s_0^* < \tau_0^{I_n} < \min_n\{s_0^{C_n}\}$. These orders are exposed to the same public information as contemporaneous trades, except for the information update from the reference order. By limiting their transaction time to occur prior to the public communication of the earliest contemporaneous order, I also ensure they are not affected by control units. Treated trades are allowed to be placed strategically (e.g. à la Kyle (1985)) and due to any "primary" reason (including manipulative intent), as long as what determines their submission timing is not completely uncorrelated with the state of the order book. Because this relevance-type assumption is critical to interpret $z_{n,t}$ as a sharp treatment assignment function rather than "fuzzy", or an intent-to-treat, I develop a strategy to assess its plausibility later in the paper.

With impacted trades serving as treated observations and contemporaneous trades as the control group, I have the basic ingredients to estimate causal effects of an intervention in a potential outcomes framework. The framework outlined below can achieve that with two sets identifying assumptions. To recover the immediate treatment effect in $h = 0$, the empirical strategy implements linear projections whose requirements are identical to a standard difference-in-differences: no anticipation and parallel trends in outcomes. These are sufficient to recover the average treatment effect on the treated as the causal estimand. For $h > 0$ dynamic treatment effects, an additional assumption is required to prevent changes in future control trades induced by earlier reference trades. I call this assumption homogeneity in indirect treatment effects, while direct effects are allowed to be heterogeneous and vary over time.

4.2.1 Immediate Price Impact

I estimate the immediate price impact of a self-trade in $t = 0$ using the following linear projection regression

$$p_{n,t} - p_{n,t-1} = \delta + \beta_0 \Delta z_{n,t} + \varepsilon_{n,t}$$

for the estimation sample restricted by the assignment of $z_{n,t}$:

$$\begin{cases} \text{impacted trades } (\Delta z_{n,t} = 1): & \{x_t^{I_n}\}_n \\ \text{or contemporaneous trades } (\Delta z_{n,t} = 0): & \{x_t^{C_n}\}_n \end{cases} \tag{2}$$

where $p_{n,t-1}$ is the last publicly displayed price up to the reference self-trade, impacted trades (i.e. with $s_t^* < \tau_t^{I_n} < \min_n\{s_t^{C_n}\}$) transition from $z_{n,t-1} = 0$ to $z_{n,t} = 1$, and contemporaneous trades (those with $\tau_t^* < \tau_t^{C_n} < s_t^*$) remain untreated with $z_{n,t-1} = z_{n,t} = 0$.

[7] To recover dynamic price effect responses following an immediate price impact ($h > 0$), I will later restrict how control trades may respond to the original self-trade relative to trades that get treated later.

Identification. The specification above is identical to a difference-in-differences model with binary treatment, two groups, and two periods. As a consequence, β_0 recovers the average treatment effect on the treated (ATT):

$$ATT = \mathbb{E}\left[\Delta p_{it} | \Delta z_{it} = 1\right] - \mathbb{E}\left[\Delta p_{it} | \Delta z_{it} = 0\right]$$

under the same standard identification assumptions as the difference-in-differences: *(i)* parallel trends and *(ii)* no anticipation. I discuss those in more detail when I write out the full dynamic empirical model below. Equation (2) therefore estimates the causal effect of self-trading on immediate returns.

No transitory price impact. Note that what I call immediate price impact is by construction different from the usual first-lag response estimated in vector autoregressive specifications commonly used in market microstructure (e.g. Hasbrouck (1991), Brogaard et al. (2019)). The first-period response estimated in these papers also captures the trade's own effect on the order book, a price response known as transitory price impact (when a trade walks the book by executing at different prices). In contrast, model (2) considers price changes only using trades following the self-trade.

4.2.2 Dynamic Price Impact

The impact of changes in the order book may not be fully absorbed immediately. This suggests a dynamic version of the one-event linear projection regression in (2), where self-trading may have a price impact up to $t + h$ lags. Specifically, I set $h = 5$ throughout the paper. While in electronic markets this still represents a very small time-frame (on average 0.2 second in the data), estimating increasingly distant lags require an extremely large cross-section to offset data losses in the time series of each event. Because treatment assignment effectively depends on traders reacting as fast as exchange latency, longer differences can be estimated very imprecisely, as I discuss below.

To estimate the dynamic price impact of a reference self-trade, I use a similar full linear projection specification as in Dube et al. (2023):

$$p_{n,t+h} - p_{n,t-1} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$$

for the estimation sample restricted by the assignment of $z_{n,t}$:

$$\begin{cases} \text{impacted trades } (\Delta z_{n,t+h} = 1): & \{x_t^{I_n}\}_n \\ \text{or contemporaneous trades } (\Delta z_{n,t+h} = 0): & \{x_t^{C_n}\}_n \end{cases} \tag{3}$$

using as control units trades that have not experienced a change in treatment status at $t + h$.

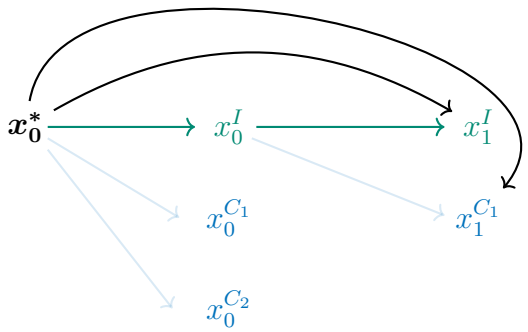
Identification. To fix ideas, consider Figure 4 which illustrates the estimation of treatment effects up to the first lag, $h = 0, 1$. There are two trades during the exchange latency around the reference

self-trade x_0^*. Because the information about these trades becomes public after the self-trade, they are used as contemporaneous trades for $h = 0$: $x_0^{C_1}, x_0^{C_2}$. There is one trade after the self-trade becomes public and before the earliest contemporaneous trade is displayed: x_0^I therefore comprises the set of impacted trades. In this case, β_0 is estimated off of comparisons between x_0^I and $x_0^{C_1}, x_0^{C_2}$.

To estimate β_1, I compare the trades impacted by the first set of impacted trades (x_1^I) to the trades contemporaneous to that trade ($x_1^{C_1}$). This targets the effect of the self-trade at $h = 1$ through its immediate impact at $h = 0$. For example, suppose the self-trade causes a change in price of 0.2% of x_0^I relative to the contemporaneous trades. If I estimate that x_0^I causes the price change of x_1^I to be 0.4% higher than the trades contemporaneous to x_0^I, this last effect is attributed to the self-trade. This causality chain can be seen as a contagion mechanism, or a directed spillover in time. Every set of impacted trades is only exposed to additional information coming from the most recently impacted trades, netting out the differential effect of any other changes to the limit order book since the reference self-trade.

Even though I focus on the chain $x_0^* \longrightarrow x_0^I \longrightarrow x_1^I$, more generally there are other potential treatment effects that I must take into account to cleanly estimate (3). As emphasized by the growing literature on the shortcomings of two-way fixed effects to recover treatment effects with heterogeneous and dynamic responses (de Chaisemartin and D'Haultfuille (2020), Goodman-Bacon (2021)), the use of "forbidden" observations as controls can result in bias so severe that the sign of estimates may be flipped. These unclean comparisons involve inadvertently using previously treated units as controls for units that receive treatment in the future. Since the treated controls may still be experiencing lagged effects from their own intervention, they are not a suitable counterfactual.

The linear projections specification developed by Dube et al. (2023) avoids these comparisons by specifying a group of clean controls (equivalent to the conditions for $z_{n,t}$ in equations (2) and (3)). In our setting, this means that a particular potential channel of treatment from the original self-trade may contaminate estimates of β_h for $h > 0$.



The diagram above shows causal relationships of interest between the sets of trades in the Figure 4 example. For simplicity, I omit factors external to the limit order book that could also determine trades, as I discussed before. The problematic arrow to estimate β_1 is $x_0^* \longrightarrow x_1^{C_1}$. Since that contemporaneous

trade can already observe the self-trade (as opposed to $x_0^{C_n}$), it may be experiencing a delayed response to it. This would invalidate the trade as a counterfactual for x_1^I. A (strong) way to rule this out is of course to assume that the information content of the reference self-trade has already been completely incorporated into prices and the only channel through which lagged responses operates is the main causality chain. While this may seem more acceptable for longer lags, in most high-frequency data settings like liquid futures markets, time responses between even multiple events is very small. For example, the average time difference between the self-trade and the contemporaneous orders in $h = 5$ is 221 milliseconds, even though there have been several dozen trades within this time.

Indeed, as I show in **Appendix X**, unless one imposes assumptions on how traders update on the set of public information, the probability that an order responds at τ to *any* extent to a self-trade in τ_0 is always 50%. One of the advantages of using double differences in this setting is that I can avoid making this arguably strong assumption. Instead, I exploit the potential direct effect of the self-trade on the impacted trade at $h = 1$, $x_0^* \longrightarrow x_1^I$. Recall that the fact that the trades at $h = 1$ are assigned to treated or control groups is by chance: they execute at the exchange almost at the same time. This means that even if the information effects of the self-trade move non-linearly in time (that is, over different h lags), as long as this effect is the same for impacted and contemporaneous trades within the same h, time fixed effects in model (3) will absorb them. This is stated more formally by the assumption below.

ASSUMPTION 1 (Homogeneity in Direct Effects with Same Delay): $\mathbb{E}\left[x_{t+h}^I | x_0^* = 1\right] - \mathbb{E}\left[x_{t+h}^I | x_0^* = 0\right] = \mathbb{E}\left[x_{t+h}^C | x_0^* = 1\right] - \mathbb{E}\left[x_{t+h}^C | x_0^* = 0\right]$, for all $h > 0$.

Crucially, note that the assumption above to recover dynamic treatment price impacts does not impose homogeneous responses to the reference event (through the main causality chain) nor that the

Details on the construction of $\Delta z_{n,t}$. At a very basic level, if there are no eligible contemporaneous trades at any h, there is no control group. Similarly, without a set of impacted trades — perhaps because information about all contemporaneous trades becomes public before any new trade — there is no treated group. The h-th moment any of these two happen in the data, it is no longer possible to estimate dynamic effects for that self-trade.

The common practice of displaying how much variation quasi-experimental sources generate for causal identification therefore takes the role of

TABLE showing number of impacted, contemporanous etc

4.3 Estimates

Figure 5 reports the first set of price impact estimates. Panels **(a)** and **(b)** show price responses for $h = 0, ..., 5$ periods for a buy and sell self-trade. The immediate price impact following a buy is positive,

quickly flipping into negative returns. The exact opposite pattern manifests for a self-trade sale: the initial price impact is negative, becoming positive up to the fifth event following the immediate impact. $h = 5$ corresponds to an average time since the self-trade of about 0.2 second. Magnitudes of estimated effects in both panels are reported in percentage points. For example, self-trade sales increase returns by up to 0.05 percentage points (5 bps).

There are two main important patterns from this first set of results.

First, relative to trades executed almost at the same time, buy and sell orders whose public information set incorporates the self-trade, execute at a different average price. Because I condition treatment effects on the same pre-shock price, estimates of β_h in practice measure incremental responses of impacted trades relative to contemporaneous orders. While the possibility of a pre-existing common trend may affect external validity considerations on the magnitudes estimated because of selection — something I address below — this effectively nets out order flow persistence effects in our causal estimand. This is an important advantage of the linear projection approach specified with exchange latency when compared to for example vector autoregressive models.

Second, the return path is consistent with self-trading being profitable: a self-trader who wishes to lower prices submits an offer and then matches against that offer with a buy trade (panel **(a)**); conversely, to raise prices she self-trades with a sell order (panel **(b)**). While it is possible to trace out price estimates for longer lags, a drawback of this empirical framework is that it demands relatively liquid markets as it mechanically throws away trades or shocks that do not qualify for treatment assignment status. For example, if there are no observations within the exchange latency interval, there are no candidate contemporaneous trades. If a contemporaneous trade becomes public before the self-trade, it also gets incorporated into the information set of impacted trades. This is why I only consider impacted trades as those happening after the public display of the self-trade but before any other trade is published in the limit order book. As a consequence, further-out lags have increasingly less data — there are almost 4 thousand times more observations at $h = 0$ compared to $h = 10$. Because of that, confidence intervals become too wide, which is why I stop the analysis at shorter lags.

Unpacking price impact estimates. *How* do prices respond to self-trading? One possibility is that impacted trades walk the price ladder — because they're too large relative to liquidity posted at the top of the book, they execute at a higher average price by increasingly matching against worse-priced limit orders. A second possibility is that contemporaneous trades end up front-running impacted trades. Because contemporaneous trades have a transaction time slightly earlier than that from impacted trades, it is possible that they instead consume all liquidity posted at the outstanding best quote worsening execution prices for impacted trades. A third possibility is that liquidity providers update what they consider stale quotes as soon as they learn about the self-trade, but before impacted trades happen. This also represents a type of implementation shortfall, but in this case the effect is attributed to $\Delta z_{n,t}$ instead of the control trades.

Panels **(a)** and **(b)** in Figure 6 show dynamic price effects for buy and sell self-trades conditioning the control group only to trades that walk the book, i.e., establish a new best quote after executed. This exercise tests whether differences between impacted and contemporaneous trades are primarily attributed to front-running by control observations. Effects for buy self-trades first rover around, and then become indistinguishable, from zero. This contrasts with strong negative price responses in the baseline specification. Patterns for sell self-trades are more similar to the ones in Figure 5, but magnitudes are much more muted.

Panels **(c)** and **(d)** show estimated price responses comparing impacted trades that do not establish a new price to contemporaneous trades that also do not walk the price ladder. In this subsample, price impact can only be attributed to changes in quotes happening between contemporaneous and impacted trades, but after the exchange latency window. Estimates are similar to the baseline model, with two notable differences. First, once I turn off temporary price impact from trades, estimates are actually generally larger and more precise. Second, $\widehat{\beta}_0$ no longer has the opposite sign of lagged treatment effects. Across all self-trade events, the first price response was opposite to what the self-trader would want: buy orders had positive impact in $h = 0$ and sell orders decreased prices. Conditioning returns on outstanding quote changes pits liquidity providers against liquidity takers in an arms-race (Aquilina et al. (2021)) won by the first group. By executing against worse quotes, immediately impacted trades may well reinforce the self-trader's desired signal, explaining the growing effects over events.

Taken together, these finds show: *(i)* self-trading triggers immediate and dynamic price responses; *(ii)* those responses are consistent with the return direction a profit-seeking self-trader would want, and *(iii)* treatment effects are primarily driven by liquidity providers also responding to the self-trader *after* the trade — just like they respond during the spoofing-like component showed in **Section 3.**

4.4 Specification Tests and Robustness

I now conduct a battery of tests to assess the empirical plausibility of some of the assumptions underlying the empirical strategy, as well as the robustness of my baseline findings. I conduct three main tests: one to assess the plausibility of treatment assignment (similar to a relevance-style test), one the check for violations of the Stable Unit Treatment Value Assumption (SUTVA), and the last test to investigate potential selection on the slope of the price trend self-traders decide to trade.

4.4.1 Orthogonal trades

My identification strategy cleanly estimates treatment effects even if impacted trades are motivated by private information or other sources of microstructure noise. This includes trading driven by omitted variables (like a price increase in another market with cross-asset intermediaries, leading to inventory control or arbitrage), purely information-driven (insiders or short-lived private information e.g., Akey et al. (2022)), or any other idiosyncratic source that can determine the decision to trade and partly when

to trade. This means that I need only to assume that the timing of these trades — the fact that a trade is observed in $h = 2$ and not during $h = 5$ — is not completely orthogonal to the information in the limit order book. This assumption can be seen as a relevance-type condition.

While this requirement cannot be perfectly tested, I can draw on insights provided by standard microstructure models to assess its plausibility empirically. A natural benchmark is the sequential trade framework pioneered by Glosten and Milgrom (1985)) where traders arrive randomly. This is more useful to my setting than strategic order placement models after Kyle (1985) as a main feature in these models is exactly endogenous responses to the limit order book.

I consider a very stylized trading determination model that gives predictions for expected treatment effects if the dominant type of trades labeled as impacted and contemporaneous were to be orthogonal to the limit order book. That is, by rejecting these predictions I argue that the data is inconsistent with traders ignoring publicly displayed information, which lends credibility to my identification strategy.

Microstructure model. Assume that traders arrive randomly, with informed traders trading with probability α, $0 \leq \alpha \leq 1$, and noise traders with probability $1 - \alpha$. Informed traders exhibit positive autocorrelation in their order flow, meaning their trades are clustered and directional: buys follow buys and sells follow sells. Noise traders, in contrast, trade randomly, with equal likelihood of buying or selling, and their order flow is assumed to have no autocorrelation. Note that in our framework this captures exactly the reason for trading I want to rule out: signed trades completely orthogonal to market data.

The observed order flow (in the market data) x_t at time t is drawn from either informed or noise traders, based on their arrival probabilities:

$$x_t = \begin{cases} x_t^{\text{informed}}, & \text{with probability } \alpha, \\ x_t^{\text{noise}}, & \text{with probability } 1 - \alpha, \end{cases}$$

where x_t^{informed} and x_t^{noise} represent contributions from informed and noise traders, respectively. As orders from different traders arrive, they queue for immediate execution and get sorted into trade market data.

In this setting, exchange latency z_{t+h} effectively picks the trade flow observations x_t and randomly places them into h consecutive non-overlapping subgroups. Each subgroup will have some observations assigned to control $z_{t+h} = 0$, which correspond to our contemporaneous trades, and the others to treated $z_{t+h} = 1$ (our impacted trades). Just like empirically, the first observations in each h are the control, and the latter the treated units.

Without loss of generality, regress a buy indicator $\mathbf{1}\{x_t > 0\}$ on the treatment dummy z_{t+h}: $\mathbf{1}\{x_{n,t+h} > 0\} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$. This specification captures differences in buy probabilities between those assigned to impacted relative to those assigned to contemporaneous trades, with the re-

gression coefficient for each subsample β_h given by:

$$\beta_h = P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 1) - P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 0).$$

The value of β_h depends on the contributions of informed and noise traders to order flow. For informed traders, positive autocorrelation implies that the probability of a buy in the "latter" positions within each h ($z_{t+h} = 1$) exceeds that in the "earlier" ($z_{t+h} = 0$):

$$P(\mathbf{1}\{x_t^{\text{informed}} > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{informed}}) = P(\mathbf{1}\{x_t^{\text{informed}} > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{informed}}) + \rho(x_t^{\text{informed}}),$$

where $\rho(x_t^{\text{informed}})$ captures the autocorrelation in informed trader flows. In contrast, for noise traders, the probability of a buy is independent of z_{t+h}:

$$P(\mathbf{1}\{x_t^{\text{noise}} > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{noise}}) = P(\mathbf{1}\{x_t^{\text{noise}} > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{noise}}) = 0.5.$$

I can then rewrite the expression for β_h as:

$$\begin{aligned} \beta_h = \quad & \alpha P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{informed}}) + (1-\alpha) P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{noise}}) \\ - \quad & \alpha P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{informed}}) - (1-\alpha) P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{noise}}) \\ = \quad & \alpha \rho \left(x_t^{\text{informed}}\right) \end{aligned} \tag{4}$$

Thus, we expect that magnitude of estimated coefficients of signed trade flow to increase with the proportion of informed traders and to vary based on the autocorrelation in the flows of informed trades. Alternatively, $\widehat{\beta}_h = 0$ implies that either the proportion of informed traders is very low, that informed flow is uncorrelated (which would imply extremely fast offsetting good and bad news), or that this is not a good microstructure model of the data.

Results. I can take the expression for β_h in (4) directly to the data. A first result in Figure 7 that is inconsistent with orthogonal trades comes from estimated effects at $h = 0$. By running conditional buys on a self-trade buy and separately sells on a self-trade sell, we would not expect orthogonal traders to trade against the direction of x_t^*. The negative and precisely estimated effects contradict this.

This first-pass result around $h = 0$ could be misleading if there is positive autocorrelation in the trade flow, particularly because the self-trader trades against her desired price movement. That is, a self-trade buy intends to drive prices down: if prices are already declining, a sell in $t - 1$ would predict a sell in $h = 0$. If by chance impacted trades happen to select more informed traders than those labeled as contemporaneous, I could "spuriously" obtain $\widehat{\beta}_h < 0$.

To rule this out, dynamic effects are useful. A second result in Figure 7 comes from how the difference in signed trades between "treated" and "control" evolve over time. Strong negative immediate

responses quickly revert, becoming either weakly positive or zero until $h = 2$. The average trade response for $h > 2$ is either weakly negative or zero. This variation in the sign of estimated effects around a self-trade is not consistent with "random" trading. If self-trades select into auto-correlated order flow (which could be seen as trend-amplifying), expression (4) shows that unless positive news happen to turn to bad news around the same instant as x_t^*, subsequent values of β_h should be positive.

A third and final result is the large presence of null differences between impacted and contemporaneous trades. Zero treatment effects are only consistent with a dominance of noise trading or lack of serial correlation in informed flow. With the latter ruled out by assumption, it may well be that at such fine temporal resolution and with enough data, I obtain zero effects based on too much noise. However, to obtain relatively stable changes in the sign of effects as those in Figure 7, the share of noise traders in the total trade flow would need to vary strongly from e.g., $h = 0$ to $h = 1$. Since z_{t+h} selects trades as-good-as randomly, it is unlikely that changes in how often noise traders arrive relative to informed traders would not be reacting to information in the limit order book.

This analysis strongly suggests that pure-noise order flow does not dominate trading activity in the futures markets I study, at least around self-trades. As a consequence, self-trade events are likely relevant to other traders and subsequent price dynamics.

4.4.2 Contemporaneous trades with short-lived private information

The institutional setting of futures markets in the US enables me to credibly stipulate the earliest moment someone can learn about an order from public data. Even the fastest trader, with access to the lowest latency data feed, cannot learn about a trade earlier than the sending time, which I exploit to assign information-based treatment to high-frequency events.

There exists however one possibility where contemporaneous trades could have known about the event trade during the exchange latency window. This stems from a technical feature related to how CME transmits information about a trade to the publicly displayed order book and to the accounts participating in that trade. Aggressing and passive orders involved in a transaction receive a trade confirmation in their private gateway used to submit orders. Because of the way this system is set up within the exchange's infrastructure, it tends to run slightly faster than the sending time. In practical terms, the concern would be that some of the trades I assume are informationally unaffected when in reality they learned about the event and reacted during the exchange latency window. This would constitute a violation of the Stable Unit Treatment Value Assumption (SUTVA).

While this may pose a threat to my identification strategy, this "liquidity-provision" learning mechanism (Aït-Sahalia and Salam (2024)) is different from snipping motivated by a correlated shock outside the order book as studied by Budish et al. (2015). While in both cases traders are attempting to remove — snipe — stale quotes as they learn about changes in the asset value, in the exchange latency period only one type of trade is profitable for snipers. Specifically, they need to trade the same direction as

the private signal they learned. This is true even if the sniper does not form expectations on the price response to the self-trade, i.e., she does not need to conjecture about the path of $\widehat{\beta}_{t+h}$.[8]

This suggests a simple test where one modifies the control group using only trades that rationally would not be reacting to the information in the event trade. Figure 8 replicates our baseline estimates conditioning contemporaneous trades on the direction of the self-trade at $h = 0$ and subsequently on the direction of each impacted trade. While this approach throws away data unnecessarily — contemporaneous trades unaware of the self-trade but that traded in the same direction anyway — it can provide an upper bound on the bias imposed by using a potentially contaminated control group. Dynamic price effects are of very similar magnitude and overall pattern as in the baseline model, with some longer lagged effects becoming more imprecisely estimated. Overall, it seems the potential contamination of contemporaneous trades that could have short-lived private information on the event trade is negligible.

5 Is Self-Trading Responsible for Flash Events?

In the previous section, I developed a framework to estimate average treatment effect responses to self-trading. While the estimates I obtained are robust and non-negligible (up to 5 bps within half a second), at such high frequency one may wonder whether these even add up to meaningful aggregate effects. I showed in **Section 2** that self-trading is common, but these average treatment effects are likely to be small relative to other one-time shocks like news or aggregate liquidity dry-ups. However, if self-trading activity tends to cluster, or intensify during periods with large market movements, their contribution to price trends may add up and exacerbate fundamentals-driven responses.

I study this type of setting by focusing on flash events, defined as the largest positive (bull) and negative (bear) 10-minute return events across trading days. Flash events may be caused by a host of reasons — some "right" like reacting to news — and perhaps some bad, like malfunctioning of execution algorithms or manipulation. A skeptic stance on manipulative-like behavior requires accepting that they do not happen at random — self-trades are more likely to happen when self-traders think the potential benefits of an action that could be perceived as market misconduct are anticipated to be larger. Secondly, there is a magnitude attribution problem: even if self-trading triggers price responses causally — which they do — if they select into events with strong price trends, their contribution to that event may not be quantitatively very important.

This type of causal question — would a flash crash have occurred without the presence of self-trading? — can be tackled with a framework known as causal attribution.

first proposed by Pearl (1999) and Rosenbaum (2001) and to the best of my knowledge, . For this analysis, I draw closely from the potential outcomes framework in Ganong and Noel (2022) which in-

[8]Suppose we select a buyer-initiated trade at time t. Up to this point, the asset was trading at \$100. Say that the best prevailing bid and ask immediately before the trade is recorded are \$100 and \$101 and that $p_t = \$101$. That is, the buyer crossed the spread, traded at the best offer, and the value of the asset jumped. A sniper can only profit if she replicates the crossing trade immediately after t: buying at the now stale \$101 before quotes and prices adjust upward.

cludes two potential causes for a binary event (the occurrence of a flash event in my setting) and recovers their separate and combined causal contribution. The natural benchmark for large price movements in financial markets is the role of informed trading, or news being incorporated into prices more broadly. The challenge with this is that private information is by definition *private*, so market participants and economists alike can only infer from observed quotes and trades which ones are likely to be informed. Ganong and Noel (2022) show how to leverage reverse-regression — a technique where a flawed proxy for a latent variable is put on the left-hand side of the regression instead of the right, and then regressed on the outcome of interest.

— [] explain, give example, say will base largely on the potential outcomes framework in QJE.

but I consider two main sources: information and self-trading.

5.1 Causal Attribution Framework

We work with a potential-outcomes framework where ... .

The potential outcome Y represents a flash event, defined as the largest positive (bull) and negative (bear) 10-minute return events across trading days. Flash events may be caused by a host of reasons, but I consider two main sources: information and self-trading. An information-driven flash crash meets regulatory scrutiny of "fair price"...

Let T^* be a dummy representing a private-information trade, G a dummy representing a self-trade, the outcome Y

represents a flash bull or bear, and T is the cumulative number of large trades, a noisy measure for informed trading T^*. The potential outcome function is $Y(T^*, G)$, so that each flash bull can .

ASSUMPTION 1: $Y(0,0) = 0$. A flash event needs information-driven trades or self-trades.

The assumption effectively rules out purely noise-driven, sustained price movements. While the qualitative distinction matters, from an econometric perspective whether informed trades were truly informed or not isn't relevant.[9]

ASSUMPTION 2: $Y(1,1) \geq Y(0,1), Y(1,0)$. Flash events are more likely with informed trades and self-trades.

i.e., a flash bull requires a self-trader or an information shock. Sustained price movements need to be maintained by information impounding or by the belief that there's information happening.[10]

[9] This means that "other reasons" correlated with large orders

[10] More broadly, order-flow imbalance (i.e. momentum) may persist in thinly traded markets or during pockets of liquidity dry-ups. Unlikely to be the case.

ASSUMPTION 3: The potential outcome $Y(T^*, G)$ is orthogonal to an informed trade T^* conditional on self-trading.

This conditional exogeneity assumption applies naturally in this context where informed trades come from innovation $\varepsilon_{2,t}$, orthogonal to e.g. trades and quotes. Allows for heterogeneity.

When informed trading happens, flash events with and without self-trades have the sa

This is similar to allow for non-perfect compliance in treated units.

n a market with positive momentumthe arrival rate of buy market orders is higher than the arrival rate of market sell orders

This means that a flash bull without a self-trader must have an information shock. There are different ways to see that this is reasonable. In classic microstructure models, trading is either information or noise-driven.

Note that this assumption requires only that larger orders make it more likely to have information.

income with life event is different than income without life event

share of large trades increases on average with informed trades.

underwater = with self-trade (G=1) overwater = without self-trade (G=0)

default = flash bull/bear T = large trade T* = informed trade

PRIVATE TRADE = LIFE EVENT (CASH FLOW + DOUBLE TRIGGER) SELF-TRADER = NEGATIVE EQUITY (STRATEGIC + DOUBLE TRIGGER)

Using trade size as proxy for informed trading. Assigning larger trades more information content than smaller ones has a long tradition in microstructure models. Larger trades have more. Another set of microstructure models sees larger trades as capturing liquidity needs. A trader may receive a negative shock that requires immediate execution, for example to , for reasons unrelated to information about the asset. While this idiosyncratic motive is interesting in theory and can lead to specific market implications (e.g. XXX), from an identification perspective, it's identical to a private information shock ε_t^j, for trader j. (TRUE THOUGH YOU CAN THINK THAT PRIVATE INFORMATION SHOCK HITS MORE THAN ONE TRADER, WHILE THIS LIQUIDITY NEED MAY BE MORE IDIOSYNCRATIC. BUT IT CAN ALSO BE DUE TO CORRELATED SHOCKS ORTHOGONAL TO THE ASSET, LIKE A SPIKE IN VIX).

Identifying assumptions. As it is standard in difference-in-differences designs, this rules out time-varying unobservables. A special case may be of interest: if causal trader is trading because of a private information shock.

This means that each event around a self-trader can be seen as a separate one —

[11] But this would mean that the

[11] Note that observationally a successful self-trader through manipulation and one that "anticipated" a price trend look the same.

$$\alpha_{\text{informed}} = \frac{\mathbb{E}\left[Y\left(T^*,1\right)|G=1\right] - \mathbb{E}\left[Y\left(0,1\right)|G=1\right]}{\mathbb{E}\left[Y\left(T^*,1\right)|G=1\right]} = \frac{\mathbb{E}\left[T|Y=1,G=1\right] - \mathbb{E}\left[T|G=1\right]}{\mathbb{E}\left[T|Y=1,G=0\right] - \mathbb{E}\left[T|G=1\right]}$$

where each group corresponds to:

$\mathbb{E}\left[T|G=1\right]$: change in % of large trades around self-trades (both for flash bull/bear or not) (expect flat)

$\mathbb{E}\left[T|Y=1,G=0\right]$: change in % large trades in flash bull/bears without self-traders

$\mathbb{E}\left[T|Y=1,G=1\right]$: change in % large trades in flash bull/bears with self-traders

These two should have similar trends — that is, % of large trades in flash events should be the same with or without self-traders

they look at the outcome "up to" the default (-2, -1, 0); final three minutes of flash event "end"

$$\underbrace{\frac{TradeSize_t}{TradeSize_{pre}}}_{\text{Noisy proxy for informed trading}} = a + \kappa \mathbf{1}\{STP\} + \gamma \mathbf{1}\{t \in \tau^{flash}\} + \beta \mathbf{1}\{t \in \tau^{flash}\} \times \mathbf{1}\{STP\} + \varepsilon_t$$

$$\frac{\%LargeTrades_t}{\%LargeTrades_{pre}} = a + \kappa \mathbf{1}\{STP\} + \gamma \mathbf{1}\{t \in \tau^{flash}\} + \beta \mathbf{1}\{t \in \tau^{flash}\} \times \mathbf{1}\{STP\} + \varepsilon_t$$

Interpretation of γ: % of large trades increases on average by $\widehat{\gamma}$% as a share of the baseline from the beginning to the end of a flash event

$$\alpha_{\text{informed}} = \frac{\overbrace{\mathbb{E}\left[T|Y=1,G=1\right]}^{\text{change in \%large trades in flash events with self-traders}} - \mathbb{E}\left[T|G=1\right]}{\underbrace{\mathbb{E}\left[T|Y=1,G=0\right]}_{\text{change in \%large trades in flash events without self-traders}} - \underbrace{\mathbb{E}\left[T|G=1\right]}_{\text{change in \%large trades around self-traders}}} = \frac{\left(\widehat{\gamma}+\widehat{\beta}\right) - \widehat{\phi}}{\widehat{\gamma} - \widehat{\phi}} \tag{5}$$

Regress large trade change 8, 9, 10 minutes after self-trade vs t-1, for all self-trade

The increase in the average trade size from informed trading is the same in flash events with and without self-trading. Note that this allows for systematic differences in uninformed trades between each case — for example, with more frenzy and amplification when self-traders trade. This does rule out the possibility that informed traders split their orders more or less aggressively when there is self-trading, for example, relative to without self-trading. This would be one violation of this assumption. Alternative measures of T ameliorate this concern.

Assumption

that when a life event does occur, above- and underwater borrowers have the same average decline in income

Given a self-trade, probability of flash event

When a large trade does occur, flash events with and without self-trades have the same trade sizes

Use trade size in flash events without self-traders (who always have $T^* = 1$, or informed trading), to learn about $P(T^*)$ for flash events with self-traders.

The researcher infers the probability of treatment from the share of underwater borrowers with an observed income decline, which is 50.5 percent (P(T) = 0.505)

What fraction of flash events with self-traders would be eliminated without them

underwater default probability => flash event probability with self-trades

5.2 Alternative Measures of Trading Information

5.3 Alternative Definitions of Flash Bulls and Bears

5.4 Other Flash Events: Volatility Spikes

6 Sufficient Statistic for Manipulative Intent

where β measures the .

We can decompose the causal effect on returns from self-trading as follows:

More broadly, I can decompose the probability that a trade

$$P\,() \tag{6}$$

Note this is essentially a difference-in-differences, where the treatment is the self-trade, the treated group is the impacted trades, the control group is the contemporaneous trades, and the pre-treatment outcome is the same for both groups: p_{t-1}. Treatment assignment is as-if-random because exchange latency — which determined when s_0 becomes publicly known — is basically exogenous.

Note that in principle, this design appears to violate what is known as "initial condition" in the difference-in-differences literature. This means that no group starts the analysis as being already treated. A related assumption commonly held in the literature is that past treatments do not affect oucomes after a new treatment, essentially because when estimating the causal effect of the second treatment, the ATT bundles the prior response (de chaistain 2023). Because of the same pre-treatment price reference, both treated and control will be "affected" by the previous treatment exactly the same way.

Another way to interpret β is that it measures the incremental causal impact of the self-trader on the asset price. But this means that large portions of the price run-up, where traders are not part of the subpopulation for the did estimation, are not captured by the analysis.

The regression I would run in the attribution analysis is the probability of a flash event based on a self-trade?

$$\alpha_{\text{self-trading}} = 1 - \frac{\mathbb{E}\left[Y\left(T^*, 0\right)|G=1\right]}{\mathbb{E}\left[Y\left(T^*, 1\right)|G=1\right]} = \frac{ATT \times P(T^*)}{\mathbb{E}\left[Y\right]} =$$

$$\alpha_{\text{self-trader}} = \frac{\mathbb{E}\left[Y\left(T^*, 1\right)|G=1\right] - \mathbb{E}\left[Y\left(T^*, 0\right)|G=1\right]}{\mathbb{E}\left[Y\left(T^*, 1\right)|G=1\right]}$$

$$\alpha_{\text{self-trader}} = \frac{\overbrace{\mathbb{E}\left[Y\left(T^*, 1\right) - Y\left(T^*, 0\right)|G=1\right]}^{\text{negative equity causally increases default probabilities by about 30 percentage points}}}{\underbrace{\mathbb{E}\left[Y\left(T^*, 1\right)|G=1\right]}_{\text{underwater default probability (60\%)}}}$$

$$\alpha_{\text{self-trader}} = \frac{\overbrace{\mathbb{E}\left[Y\left(T^*, 1\right) - Y\left(T^*, 0\right)|G=1\right]}^{\text{self-trading causally increases flash event probability by about 30 percentage points}}}{\underbrace{\mathbb{E}\left[Y\left(T^*, 1\right)|G=1\right]}_{\text{self-trades part of flash event (60\%)}}}$$

Probability of the outcome conditional on treatment/group status. So probability of a self-trade given a self-trade.

60% of self-trades happen in flash events

What is flash trade?

Price increase of x% in the next 10 minutes

Causal did will give something life: self-trade causes price to increase by 1%.

Flash event:

$P(Y) = P(Y|T=1)P(T=1) + P(Y|T=0)P(T=0)$

Self-trader increases the price by 2%.

Specifically, I regress the probability that the event is a flash event

$$\mathbf{1}\{Flash(m=10)\}_i = \sum_{m=2}^{10} \delta_m \mathbf{1}\{STP\}_{i,m} + \mu_i + \mu_m + \varepsilon_{i,m} \tag{7}$$

on the presence of self-trading in each minute of that event, for a sample that includes successful and failed flash events. Specifically, a comparison between δ_9 and δ_2 informs whether self-trading activity is more intense closer to the peak of a flash event relative to the beginning, compared to the same difference in a 10-minute period that did not sustain the same price trend strength.

Because given a price trend self-trading is most profitable at the moment of maximum momentum, one way to interpret $\delta_9 < \delta_2$ is that self-trading activity stopped "too short"; another, that it failed to correlate with a flash event increasingly poorly. This interpretation is possible since the sample used in (8) includes flash 1, 2, ..., 10-minute events. For instance, a 2-minute flash event no longer builds

momentum by minute 3, so larger values of self-trading activity later capture either increasingly failed attempts to trigger price responses or of timing to exit with potential profits.

Use "wrong" sign self-trades as placebo

$$\mathbf{1}\{\mid p_t > p_{t-1} \mid\} = \sum_{t=2}^{10} \delta_m \mathbf{1}\{STP\}_{i,m} + \mu_i + \mu_m + \varepsilon_{i,m} \tag{8}$$

Thus, estimated coefficients bundle enter and exit of self-trading responses in

The probability of a flash event can thus be seen as the probability that prices jumped by 5% at the 10-th minute.

6.0.1 Causal effect on sniping

Geographic spillovers to control groups close to treated ones. There are many appli- cations where the effect of the treatment might spill over onto untreated groups geographically close to a treated group. For example, if a county increases its minimum wage, individuals from contiguous counties may decide to go work there to benefit from the higher minimum wage. Similarly, if a coal-fired power plant adopts an emissions-control technology, this will reduce air pollution in the county where the plant is located, but as air pollution travels this can also reduce air pollution in neighboring counties downwind of the treated one. This leads to a violation of the SUTVA assumption in (2.1). Then, a treated unit may experience both a direct treatment effect, arising from its own treatment, but it may also experience an indirect effect, arising from the treatment of treated units located close to it. The sum of such direct and indirect effects is sometimes referred to as the treatments total effect. If a parallel-trends assumption holds but SUTVA fails, Clarke (2017) and Butts (2021b) show that a TWFE regression that ignores such violations will not estimate the average total effect of the treatment across all treated groups. Instead, this regression will estimate the average total effect, minus the proportion of untreated groups that are affected by treated groups treatment, times the average indirect effect of the treatment across those affected untreated groups. Assuming that both effects are of the same sign but that the indirect effect is closer to zero than the total effect, the TWFE regression suf- fers from an attenuation bias. If one is ready to assume that a subset of untreated groups is not affected by treatment groups treatment, for instance because they are geographically far to all treated groups, then the population can be partitioned into treated groups, affected untreated groups, and unaffected untreated groups. Under a parallel-trends assumption, Butts (2021b) shows that a TWFE regression restricting the sample to treated and to unaffected untreated groups estimates the average total effect of the treatment across all treated groups. This ratio- nalizes the common approach in applied work of excluding groups located close to the treatment area. Then, Butts (2021b) also shows that a TWFE regression restricting the sample to affected untreated groups and to unaffected untreated groups estimates the average indirect effect of the treatment across all affected untreated groups.

7 Conclusion

References

Aggarwal, RajeshăK. and Guojun Wu (2006) "Stock Market Manipulations," *The Journal of Business*, Vol. 79, No. 4, pp. 1915–1953.

Akey, Pat, Vincent Gregoire, and Charles Martineau (2020) "Price revelation from insider trading: evidence from hacked earnings news," *Available at SSRN 3365024*.

Akey, Pat, Vincent Grégoire, and Charles Martineau (2022) "Price revelation from insider trading: Evidence from hacked earnings news," *Journal of Financial Economics*, Vol. 143, No. 3, pp. 1162–1184.

Allen, Franklin and Douglas Gale (1992) "Stock-Price Manipulation," *The Review of Financial Studies*, Vol. 5, No. 3, pp. 503–529.

Aquilina, Matteo, Eric Budish, and Peter ONeill (2021) "Quantifying the High-Frequency Trading "Arms Race"'," *The Quarterly Journal of Economics*, Vol. 137, No. 1, pp. 493–564.

Aït-Sahalia, Yacine and Mehmet Salam (2024) "High frequency market making: The role of speed," *Journal of Econometrics*, Vol. 239, No. 2, p. 105421.

Baldauf, Markus and Joshua Mollner (2020) "High-Frequency Trading and Market Performance," *The Journal of Finance*, Vol. 75, No. 3, pp. 1495–1526.

Baruch, Shmuel and Lawrence R Glosten (2013) "Fleeting orders," *Working Paper*, No. 13-43.

Biais, Bruno, David Martimort, and Jean-Charles Rochet (2000) "Competing Mechanisms in a Common Value Environment," *Econometrica*, Vol. 68, No. 4, pp. 799–837.

Boehmer, Ekkehart, Dan Li, and Gideon Saar (2018) "The Competitive Landscape of High-Frequency Trading Firms," *The Review of Financial Studies*, Vol. 31, No. 6, pp. 2227–2276.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2019) "Price Discovery without Trading: Evidence from Limit Orders," *The Journal of Finance*, Vol. 74, No. 4, pp. 1621–1658.

Budish, Eric, Peter Cramton, and John Shim (2015) " The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, Vol. 130, No. 4, pp. 1547–1621.

de Chaisemartin, Clément and Xavier D'Haultfuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, Vol. 110, No. 9, p. 296496.

Clark-Joseph, Adam (2013) "Exploratory trading," *Unpublished job market paper. Harvard University, Cambridge, MA*.

Donier, J., J. Bonart, I. Mastromatteo, and J.-P. Bouchaud (2015) "A fully consistent, minimal model for non-linear market impact," *Quantitative Finance*, Vol. 15, No. 7, pp. 1109–1121.

Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M Taylor (2023) "A Local Projections Approach to Difference-in-Differences," Working Paper 31184, National Bureau of Economic Research.

Easley, David, Nicholas M. Kiefer, and Maureen O'Hara (1997) "The information content of the trading process," *Journal of Empirical Finance*, Vol. 4, No. 2, pp. 159–186, High Frequency Data in Finance, Part 1.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H Eugene Stanley (2003) "A theory of power-law distributions in financial market fluctuations," *Nature*, Vol. 423, No. 6937, pp. 267–270.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley (2006) "Institutional Investors and Stock Market Volatility," *The Quarterly Journal of Economics*, Vol. 121, No. 2, pp. 461–504.

Ganong, Peter and Pascal Noel (2022) "Why do Borrowers Default on Mortgages?" *The Quarterly Journal of Economics*, Vol. 138, No. 2, pp. 1001–1065.

Glosten, Lawrence R. (1994) "Is the Electronic Open Limit Order Book Inevitable?" *The Journal of Finance*, Vol. 49, No. 4, pp. 1127–1161.

Glosten, Lawrence R. and Paul R. Milgrom (1985) "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders," *Journal of Financial Economics*, Vol. 14, No. 1, pp. 71–100.

Goodman-Bacon, Andrew (2021) "Difference-in-differences with variation in treatment timing," *Journal of Econometrics*, Vol. 225, No. 2, pp. 254–277, Themed Issue: Treatment Effect 1.

Hasbrouck, Joel (1991) "Measuring the Information Content of Stock Trades," *The Journal of Finance*, Vol. 46, No. 1, pp. 179–207.

——— (2018) "High-Frequency Quoting: Short-Term Volatility in Bids and Offers," *Journal of Financial and Quantitative Analysis*, Vol. 53, No. 2, p. 613641.

Hasbrouck, Joel and Gideon Saar (2009) "Technology and liquidity provision: The blurring of traditional definitions," *Journal of Financial Markets*, Vol. 12, No. 2, pp. 143–172.

——— (2013) "Low-latency trading," *Journal of Financial Markets*, Vol. 16, No. 4, pp. 646–679, High-Frequency Trading.

Kyle, Albert S. (1985) "Continuous Auctions and Insider Trading," *Econometrica*, Vol. 53, No. 6, pp. 1315–1335.

Lee, Eun Jung, Kyong Shik Eom, and Kyung Suh Park (2013) "Microstructure-based manipulation: Strategic behavior and performance of spoofing traders," *Journal of Financial Markets*, Vol. 16, No. 2, pp. 227–252.

Pearl, Judea (1999) "Probabilities Of Causation: Three Counterfactual Interpretations And Their Identification," *Synthese*, Vol. 121, No. 1, pp. 93–149.

Putniņš, Tālis J. (2012) "MARKET MANIPULATION: A SURVEY," *Journal of Economic Surveys*, Vol. 26, No. 5, pp. 952–967.

Ripley, W. Z. (1911) "Railway Speculation," *The Quarterly Journal of Economics*, Vol. 25, No. 2, pp. 185–215.

Rosenbaum, Paul R. (2001) "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot," *Biometrika*, Vol. 88, No. 1, pp. 219–231.

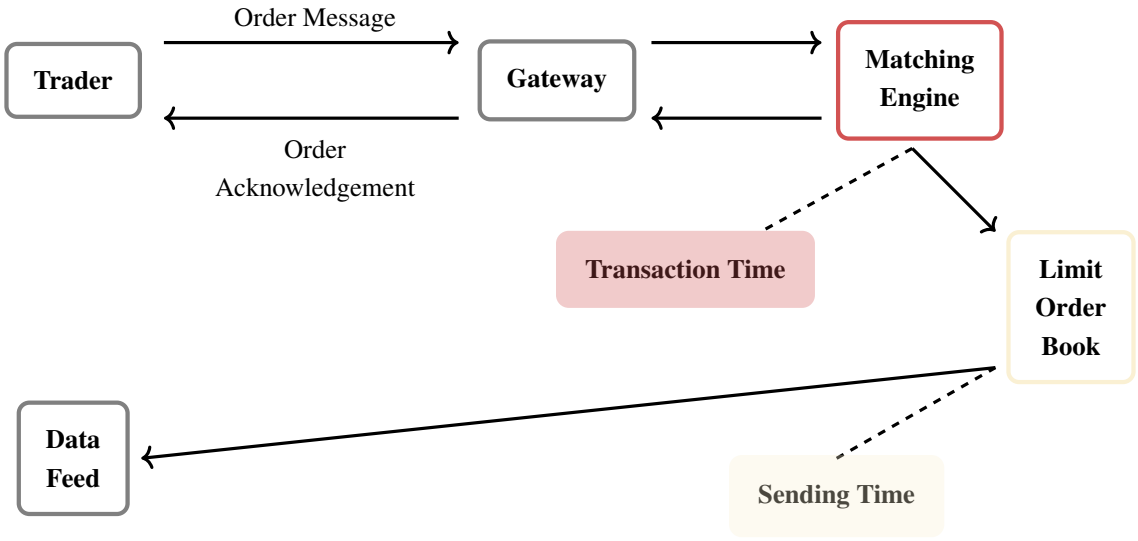
Tóth, Bence, Imon Palit, Fabrizio Lillo, and J. Doyne Farmer (2015) "Why is equity order flow so persistent?" *Journal of Economic Dynamics and Control*, Vol. 51, pp. 218–239.

Williams, Basil and Andrzej Skrzypacz (2021) "Spoofing in Equilibrium."

Yang, Liyan and Haoxiang Zhu (2019) "Back-Running: Seeking and Hiding Fundamental Information in Order Flows," *The Review of Financial Studies*, Vol. 33, No. 4, pp. 1484–1533.

Ye, Mao, Chen Yao, and Jiading Gai (2013) "The externalities of high frequency trading," *WBS Finance Group Research Paper*, No. 180.

Figures and Tables



NOTES: This figure shows a stylized version of the routing system used by the electronic market Globex at the Chicago Mercantile Exchange (CME). When a trader submits an order message to the exchange, CME's gateway records the order arrival and sends a message back to the trader acknowledging the order was received. Then, the order is routed to the exchange's matching engine where it interacts with orders from other traders in the limit order book following allocation rules prescribed by a matching algorithm. This changes the state of the public limit order book. To inform that such a change occurred, the exchange sends a message update to its data feed, which all traders and market participants can access. The transaction timestamp is recorded when the exchange's matching engine allocates a trader's order. The sending timestamp is recorded when the exchange sends the outbound message to the public data feed communicating a change in the limit order book.

FIGURE 1: STYLIZED ROUTING SYSTEM AND MEASUREMENT POINTS OF TIMESTAMPS

(a) Resting Orders at Best Ask and Offer

Top of Book Level	$100	Sell orders	Quantity	**100**	**270**	55	80	25	**105**
			Order ID	**1**	**2**	3	4	5	**6**
	$98	Buy orders	Quantity	75	150	65	150	50	235
			Order ID	7	8	9	10	11	12

(b) Trade Summary After Aggressing Order

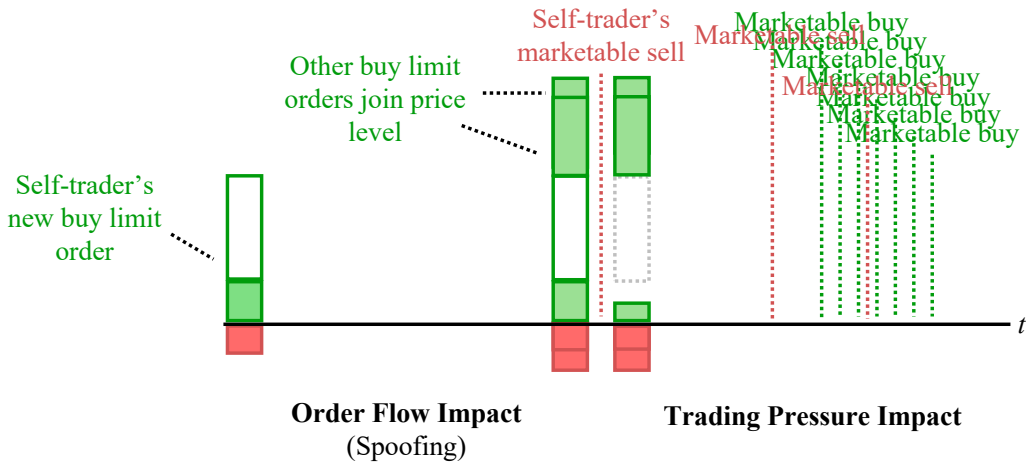
Activity	Quantity	Orders Matched Against	Price	Filled Quantity	Order ID	Timestamp (microsecond)	
Trade	700	3	$100	160	13	8:45:23.587663	*Crossing Trade*
Delete	**0**	**1**	**$100**	**0**	**1**	**8:45:23.587663**	*Self-Trade*
Delete	**0**	**1**	**$100**	**0**	**2**	**8:45:23.587663**	*Self-Trade*
Delete	0	1	$100	55	3	8:45:23.587663	*Filled*
Delete	0	1	$100	80	4	8:45:23.587663	*Filled*
Delete	0	1	$100	25	5	8:45:23.587663	*Filled*
Delete	0	1	$100	0	6	8:45:23.587663	*Self-Trade*

(c) Resting Orders at Best Ask and Offer After Trade

Top of Book Level	$100	Sell orders	Quantity						
			Order ID						
	$100	Buy orders	Quantity	540					
			Order ID	13					
	$98	Buy orders	Quantity	75	150	65	150	50	235
			Order ID	7	8	9	10	11	12

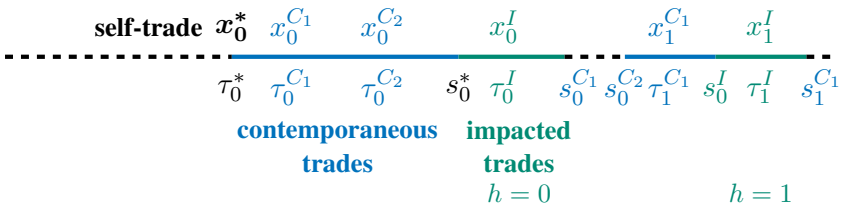
NOTES: xpiring month.

FIGURE 2: IDENTIFYING SELF-TRADES IN MESSAGE DATA



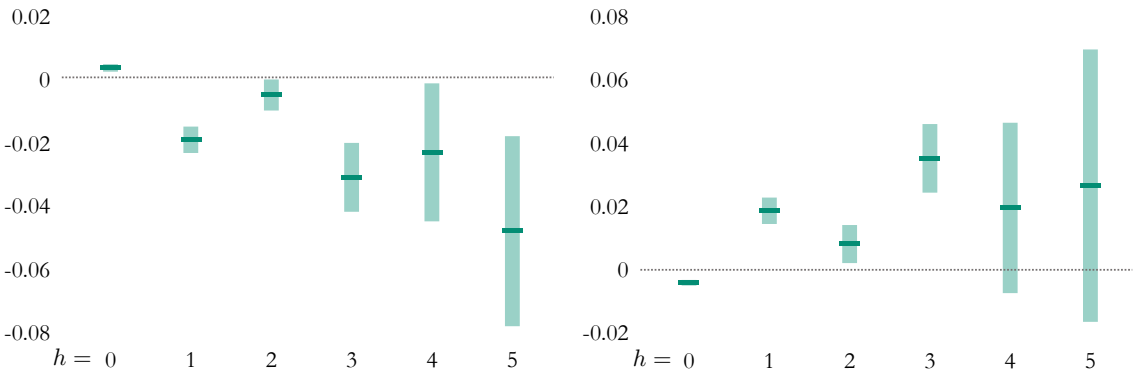
NOTES: This figure illustrates the two sequential types of market impact associated with self-trading activity. From the moment the trader enters a limit order (buy, in this case) until forces a self-match with a marketable order (sell),

FIGURE 3: DECOMPOSING THE IMPACT OF SELF-TRADING



NOTES: This figure

FIGURE 4: IDENTIFICATION STRATEGY FOR CAUSAL PRICE IMPACT



(a) Price change for $Buy_{\text{self-trade}}$ **(b)** Price change for $Sell_{\text{self-trade}}$

NOTES: This figure reports dynamic price effect estimates $\widehat{\beta}_h$ of the linear projection model

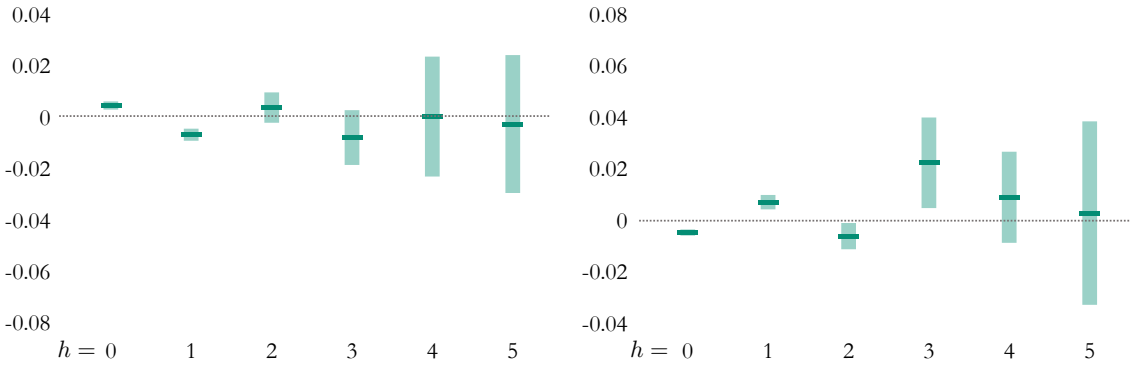
$$y_{n,t+h} - y_{n,t-1} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$$

for the estimation sample restricted by the assignment of $z_{n,t}$:

$$\begin{cases} \text{impacted trades } (\Delta z_{n,t+h} = 1): & \{x_t^{I_n}\}_n \\ \text{or contemporaneous trades } (z_{n,t+h} = 0): & \{x_t^{C_n}\}_n \end{cases}$$

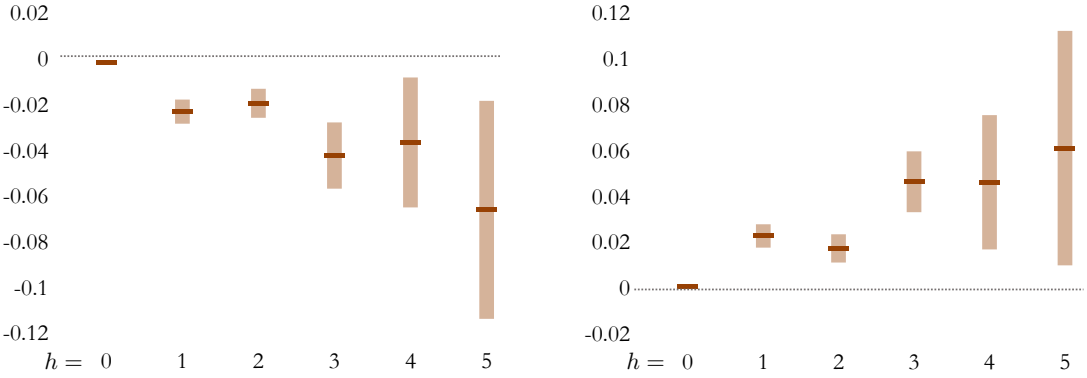
for $h = 0, ..., 5$ events following a self-trade. This specification uses as control units trades that have not experienced a change in treatment status in each $t + h$ period, assigned by exchange latency. Vertical bars are confidence intervals calculated with standard errors clustered at the day level. Data corresponds to futures on 2-year t-notes, 10-year t-notes, and E-mini S&P 500 from October 2019 to March 2020.

FIGURE 5: PRICE IMPACT ESTIMATES FOR SELF-TRADING



(a) Price change for $Buy_{\text{self-trade}}$

(b) Price change for $Sell_{\text{self-trade}}$



(c) Price change for $Buy_{\text{self-trade}}$

(d) Price change for $Sell_{\text{self-trade}}$

NOTES: This figure reports dynamic price effect estimates $\widehat{\beta}_h$ of the linear projection model

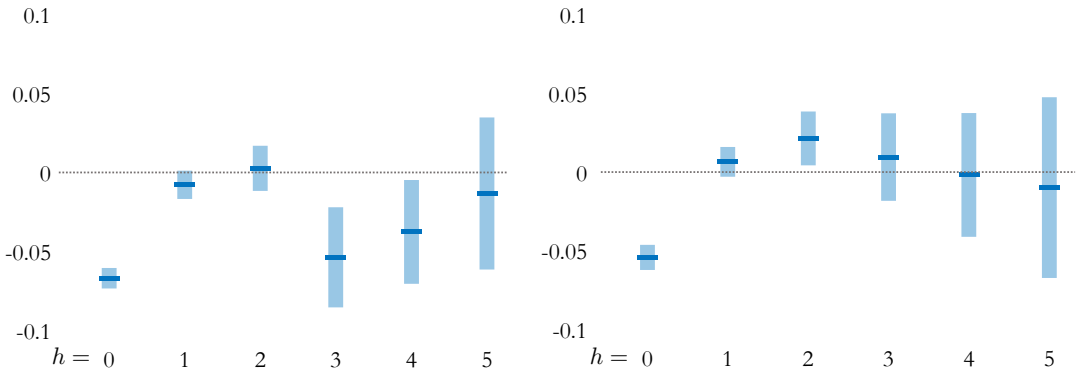
$$y_{n,t+h} - y_{n,t-1} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$$

for the estimation sample restricted by the assignment of $z_{n,t}$:

$$\begin{cases} \text{impacted trades } (\Delta z_{n,t+h} = 1): & \{x_t^{I_n}\}_n \\ \text{or contemporaneous trades } (z_{n,t+h} = 0): & \{x_t^{C_n}\}_n \end{cases}$$

for $h = 0, ..., 5$ events following a self-trade. This specification uses as control units trades that have not experienced a change in treatment status in each $t + h$ period, assigned by exchange latency. Vertical bars are confidence intervals calculated with standard errors clustered at the day level. Data corresponds to futures on 2-year t-notes, 10-year t-notes, and E-mini S&P 500 from October 2019 to March 2020.

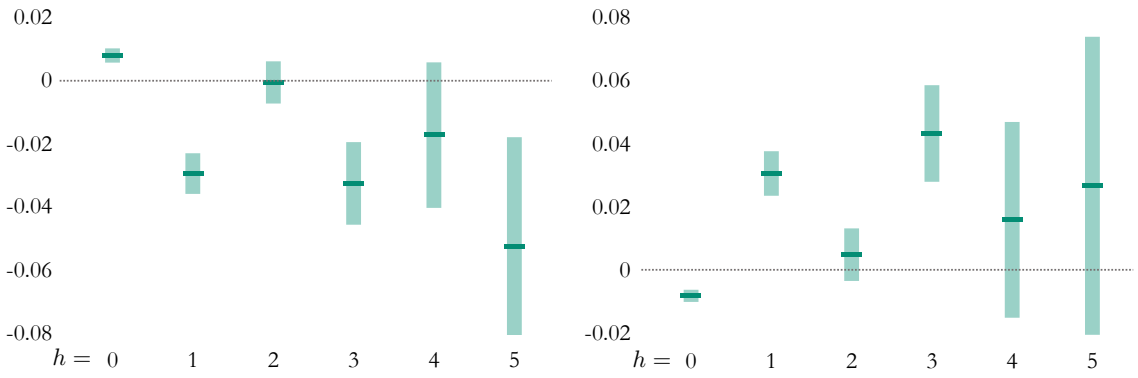
FIGURE 6: PRICE IMPACT ESTIMATES FOR SELF-TRADING: SWEEP



(a) $P(Buy|Buy_{\text{self-trade}})$ **(b)** $P(Sell|Sell_{\text{self-trade}})$

NOTES: This figure reports dynamic estimates $\widehat{\beta}_h$ of the model outcome$_{n,t+h} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$ for the estimation sample restricted by the assignment of exchange latency $z_{n,t}$, where $\Delta z_{n,t+h} = 1$ for impacted trades and $z_{n,t+h} = 0$ for contemporaneous trades. Panel **(a)** has as outcome $\mathbf{1}\{x_{n,t+h} > 0\}$ and is conditioned on observations following a self-trade buy. Panel **(b)** has as outcome $\mathbf{1}\{x_{n,t+h} < 0\}$ and is conditioned on a self-trade sell. Vertical bars are confidence intervals calculated with standard errors clustered at the day level. Data corresponds to futures on 2-year t-notes, 10-year t-notes, E-mini S&P 500, gold, and oil from October 2019 to March 2020.

FIGURE 7: PRICE IMPACT ESTIMATES FOR SELF-TRADING: PROB BUY BUY OR SELL SELL



(a) Price change for $Buy_{\text{self-trade}}$ **(b)** Price change for $Sell_{\text{self-trade}}$

NOTES: This figure reports dynamic price effect estimates $\widehat{\beta}_h$ of the linear projection model

$$y_{n,t+h} - y_{n,t-1} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$$

for the estimation sample restricted by the assignment of $z_{n,t}$:

$$\begin{cases} \text{impacted trades } (\Delta z_{n,t+h} = 1): & \{x_t^{I_n}\}_n \\ \text{or contemporaneous trades } (z_{n,t+h} = 0): & \{x_t^{C_n}\}_n \end{cases}$$

for $h = 0, ..., 5$ events following a self-trade. This specification uses as control units trades that have not experienced a change in treatment status in each $t + h$ period, assigned by exchange latency. Vertical bars are confidence intervals calculated with standard errors clustered at the day level. Data corresponds to futures on 2-year t-notes, 10-year t-notes, and E-mini S&P 500 from October 2019 to March 2020.

FIGURE 8: PRICE IMPACT ESTIMATES FOR SELF-TRADING:NO SNIPING

TABLE 1: SELF-TRADING IN FUTURES MARKETS — BIG PICTURE

	Period Average		Period Aggregate	
	Frequency	**Lots**	**#**	**Notional Value**
Resting Orders ⟶ Self-Trades	0.65%		379,018	$48.50 billion
Trades Including Self-Trades	3.82%		357,632	
Volume Executed (All Trades)		5.21		$9.88 trillion
Volume Executed (Self-Trades)		10.41		$104.06 billion
Sweeping Trades[a]	4.83%			
Sweeping Self-Trades	3.73%			

NOTES: This table shows average statistics computed out of all limit orders that are involved in self-trade events and self-trades of reported quantities for futures markets. Futures contracts are: 2-year and 10-year t-notes, gold, oil, corn, soybeans, and wheat, during the period 01/2019 to 06/2020. I use first-nearby contracts (nearest expiration) only for computations. Notional values are calculated as lot size (contracts)×standardized contract quantity×posted price of each observation, then annualized. [a]**Sweeping trades**: marketable orders that sweep the entire best price level on the opposite side of the market (both by filling orders and triggering STP-cancellations).

TABLE 2: LIQUIDITY TAKEN BY STP-TRIGGERED CANCELLATIONS

	Period Average		Period Aggregate	
	%Liquidity Taken[a]	**Frequency**	**Notional Value Taken**	**Notional Fleeting Liquidity**[b]
Lot Size				
1−5	48.13%	57.69%	$4.75 billion	$513.18 billion
5−10	30.42%	14.93%	$4.37 billion	$225.36 billion
10−20	22.79%	11.11%	$6.62 billion	$275.20 billion
20−50	16.72%	10.10%	$12.33 billion	$344.53 billion
50−100	12.36%	4.06%	$9.91 billion	$314.71 billion
100+	8.94%	2.12%	$10.52 billion	$593.53 billion
All	37.22%	100%	$48.5 billion	$2.27 trillion

NOTES: This table shows average statistics computed out of all self-trade events, broken down by size of the marketable order entered by the self-trader in each event, and aggregate notional values of reported quantities for futures markets. Futures contracts are: 2-year and 10-year t-notes, gold, oil, corn, soybeans, and wheat, during the period 01/2019 to 06/2020. I use first-nearby contracts (nearest expiration) only for computations. Notional values are calculated as lot size (contracts)×standardized contract quantity×posted price of each observation. [a]**Liquidity taken**: defined as the ratio of resting volume by the STP functionality in a self-trade event to the total quantity matched (STP deleted/(STP deleted + filled)). [b]**Notional fleeting liquidity**: defined as new limit orders entered out-of-touch (outside the bid-ask spread) and cancelled within 500 milliseconds. I include flash cancellations only outside the top of the book as many trading strategies at the touch are pegged and dynamically adjust their quote as the market midprice moves. Depending on the execution pipeline of the trader, this high-frequency revision may involve order replacement (cancel stale resting limit order and enter limit order with new price) rather than parameter modifications of the same order. Lot sizes for the notional fleeting liquidity column correspond to the size bucket of the resting rather than marketable reference order. Tables X-X replicate these results using other futures maturities, different cancellation windows and orders posted at all price levels to define fleeting liquidity.

TABLE 3: ORDER FLOW IMPACT

	Period Average			
	Execution Frequency	**# Orders Joining Price Level**	**Lots Joining Price Level**	**% Orders with Price Improvement**
Time From Order Entry				
10 microseconds	1.01%	0.24	3.82	11.11%
100 microseconds	2.73%	0.78	11.18	12.48%
1 millisecond	10.94%	2.12	29.54	14.16%
100 milliseconds	25.59%	6.15	81.91	12.14%
500 milliseconds	29.79%	7.81	118.75	11.67%
1 second	32.23%	9.11	144.58	11.30%
10 seconds	50.19%	18.37	331.57	9.40%
30 seconds	65.12%	26.67	496.70	8.32%
1 minute	76.94%	32.01	623.41	7.95%
$>$ 1 minute	23.06%	25.15	843.20	6.32%

NOTES: Monthly

TABLE 4: EXCHANGE LATENCY

	Exchange Latency$_t$		
	(1)	(2)	(3)
Latency Persistence			
Latency$_{t-1}$	61.540***	53.935***	50.968***
	(4.164)	(4.536)	(3.800)
Latency$_{t-2}$	40.660***	36.080***	34.417***
	(3.153)	(3.262)	(2.487)
Latency$_{t-3}$	26.043***	23.108***	21.551***
	(2.319)	(2.386)	(1.983)
Latency$_{t-4}$	28.605***	24.698***	23.480***
	(2.461)	(2.434)	(2.196)
Latency$_{t-5}$	34.473***	30.846***	26.891***
	(2.971)	(2.911)	(2.821)
Message Traffic			
# Trades		0.004***	0.006***
		(0.001)	(0.001)
# Cancellations		0.002**	0.003***
		(0.001)	(0.001)
# Updates		−0.001	−0.001
		(0.001)	(0.001)
# Entries		−0.003***	−0.002*
		(0.001)	(0.001)
Fixed effects			
Minute			✓
Market			✓
***R*-squared**	**0.01**	**0.04**	**0.08**

NOTES: The table reports estimates from the regression

$$\text{Latency}_{m,t} = \delta_0 + \sum_k \delta_k \text{Messages}_{m,t} + \sum_{\tau=1}^{5} \omega_\tau \text{Latency}_{m,t-\tau} + \mu_q + \mu_m + u_{m,t}$$

where δ_k measure the importance of the aggregate quantity of inbound messages of type k (trade, new, update, cancel) in minute q, ω_τ capture persistence in latency, μ_m are market dummies, and μ_q minute dummies. Standard errors are clustered at the market and minute.

Technical Appendix

7.1 Parsing Order Flow

7.1.1 Derivation

For a reference order entered at $t = 0$ and another at $t = \tau > 0$, the number of all combinations of order-flow during $[0, \tau]$ is its non-empty cardinality, $\mathcal{C} = 2^{\#orders_{t\in[0,\tau]}} - 1$. To choose all subsets of $\mathcal{C}$ that contain the reference order at $t = 0$, one must simply count over $\#orders - 1$ elements, which yields the parsing formula: $\phi = \dfrac{2^{\#orders-1}}{2^{\#orders} - 1}$.

Note that ϕ is a well-defined probability because it is non-negative for any $\#orders$ (which is always at least equal to the order entered at $t = 0$) and $\sum_{k=1}^{\#orders} \dfrac{2^{\#orders-k}}{2^{\#orders} - 1} = 1$.

Example 1. A sell limit order enters the trading book at time $t = \tau$. There are 3 other sell others entered since $t = 0$, when s_0 become publicly displayed. The order s_τ can be reacting to

$$
\begin{aligned}
& s_0 \text{ or } s_1 \text{ or } s_2 \text{ or } s_3 && (25\%)\\
\text{or } \{s_0, s_1\} \text{ or } \{s_0, s_2\} \text{ or } \{s_0, s_3\} \text{ or } \{s_1, s_2\} \text{ or } \{s_1, s_3\} \text{ or } \{s_2, s_3\} && (50\%)\\
\text{or } \{s_0, s_1, s_2\} \text{ or } \{s_0, s_1, s_3\} \text{ or } \{s_0, s_2, s_3\} \text{ or } \{s_1, s_2, s_3\} && (75\%)\\
\text{or } \{s_0, s_1, s_2, s_3\} && (100\%)
\end{aligned}
$$

which is $2^{4-1}/2^4 - 1 = 8/15$.

Now suppose I want to track more orders during $(0, \tau)$. That is, the I want to assign the probability that a reference order at $t = \tau$ responds to another order entered at $t = 0$ or a second order entered during $0 < t < \tau$. The total number of combinations given $\#orders$ is still $\mathcal{C}$. To choose from them, one must consider the order-flow combinations containing one or two reference orders, adjusted for common counts in each case.

Example 1 (cont). A sell limit order enters the trading book at time $t = \tau$. There are 3 other sell others entered since $t = 0$, when s_0 become publicly displayed. I also want to track s_2. The order s_τ can be reacting to

$$\begin{aligned}
&s_0 \text{ or } s_1 \text{ or } s_2 \text{ or } s_3 && (25\%)\\
\text{or } \{s_0, s_1\} \text{ or } \{s_0, s_2\} \text{ or } \{s_0, s_3\} \text{ or } \{s_1, s_2\} \text{ or } \{s_1, s_3\} \text{ or } \{s_2, s_3\} && (50\%)\\
\text{or } \{s_0, s_1, s_2\} \text{ or } \{s_0, s_1, s_3\} \text{ or } \{s_0, s_2, s_3\} \text{ or } \{s_1, s_2, s_3\} && (75\%)\\
\text{or } \{s_0, s_1, s_2, s_3\} && (100\%)
\end{aligned}$$

$$\begin{aligned}
&\text{or } s_0 \text{ or } s_1 \text{ or } s_2 \text{ or } s_3 && (25\%)\\
\text{or } \{s_0, s_1\} \text{ or } \{s_0, s_2\} \text{ or } \{s_0, s_3\} \text{ or } \{s_1, s_2\} \text{ or } \{s_1, s_3\} \text{ or } \{s_2, s_3\} && (50\%)\\
\text{or } \{s_0, s_1, s_2\} \text{ or } \{s_0, s_1, s_3\} \text{ or } \{s_0, s_2, s_3\} \text{ or } \{s_1, s_2, s_3\} && (75\%)\\
\text{or } \{s_0, s_1, s_2, s_3\} && (100\%)
\end{aligned}$$

$$\begin{aligned}
&\text{or } s_0 \text{ or } s_1 \text{ or } s_2 \text{ or } s_3 && (0\%)\\
\text{or } \{s_0, s_1\} \text{ or } \{s_0, s_2\} \text{ or } \{s_0, s_3\} \text{ or } \{s_1, s_2\} \text{ or } \{s_1, s_3\} \text{ or } \{s_2, s_3\} && (25\%)\\
\text{or } \{s_0, s_1, s_2\} \text{ or } \{s_0, s_1, s_3\} \text{ or } \{s_0, s_2, s_3\} \text{ or } \{s_1, s_2, s_3\} && (50\%)\\
\text{or } \{s_0, s_1, s_2, s_3\} && (100\%)
\end{aligned}$$

which is $(20 - 8)/15$ picks containing s_0, s_2, or both.

To generalize the above, note that each step is equivalent to calculating $\binom{\#orders}{k-1}$ and then discarding $\#reference\ orders - k + 2$ repeated counts, for $k = 2, ..., \#orders$. This will however also discard one occurrence of the full subset $\#orders$, which needs to be added back. Putting it all together yields:

$$\phi = \frac{\sum_{k=2}^{\#orders} \left[\binom{\#orders}{k-1} - (\#reference\ orders - k + 2) \right] + 1}{2^{\#orders} - 1}$$

7.1.2 Extensions

Orthogonal liquidity. When deriving ϕ, I assume that the order entered at $t = \tau$ is responding to the state of the limit order book. In reality, it is possible that the order not only does not respond to other reference orders, but to the limit order book *at all*. Although this is certainly empirically more plausible in the case of marketable orders, e.g. due to fire sales or book-invariant execution algorithms, such as TWAP strategies — I can modify ϕ to incorporate the possibility. This will correspond to the counting the empty set in the cardinality $\mathcal{C}$, so that the denominator in the simple and generalized versions of ϕ changes to $2^{\#orders}$.

7.1.3 Implementation

We can then study every $\tau > 0$ following the self-trader's enter as an event. However, often times other self-traders' orders enter the limit order book during τ. That is, self-trader events overlap.

The problem reduces to counting

TABLE 5: 2020 RULE CHANGE — INTEREST MARKETS

	Before Change				**After Change**			
	2-year	**3-year**	5-year	10-year	2-year	**3-year**	5-year	10-year
Parameter								
%FIFO	40%	**40%**	100%	100%	40%	**100%**	100%	100%
%Pro-rata	60%	**60%**	0%	0%	60%	**0%**	0%	0%
Tick size	0.00390625	0.0078125	0.0078125	0.015625	0.00390625	0.0078125	0.0078125	0.015625
Max lot size	19,999	19,999	19,999	19,999	19,999	19,999	19,999	19,999

NOTES: Monthly

8 Simulation Framework

To validate the theoretical predictions, I construct a simulation framework mimicking the behavior of informed and noise traders within a simplified trading environment. Our setup captures the interplay between these two trader types and their influence on observed order flow. This section outlines the simulation parameters, results, and key takeaways.

8.1 Simulation Design

8.1.1 Trader Behavior

Informed Traders: Their order flow exhibits positive autocorrelation, with buys tending to follow buys. Specifically, informed traders are simulated using an AR(1) process:

$$Y_t^{\text{informed}} = \rho_{\text{informed}} Y_{t-1}^{\text{informed}} + \varepsilon_t,$$

where $\rho_{\text{informed}} = 0.9$, $Y_t^{\text{informed}} \in \{-1, 1\}$ (Sell or Buy), and ε_t represents random noise. This ensures persistence in their trading behavior.

Noise Traders: Their trades are generated independently as random $\{-1, 1\}$ values, with equal probabilities of buying or selling. Noise traders actions introduce randomness into the combined flow.

8.1.2 Simulation Parameters

- **Combined Flow:** The combined flow consists of 20 trades per unit, sampled randomly from informed and noise flows based on the arrival rate α ($0 \leq \alpha \leq 1$).
- **Trader Proportions:** α determines the proportion of informed trades, with $1 - \alpha$ representing the share of noise trades.
- **Pair Construction:** Each units combined flow is divided into 10 pairs, with consecutive trades forming a pair (e.g., positions 12, 34, etc.).
- **Regression Model:** For each pair, I regress a dummy variable Y (indicating a buy) on another dummy Z (indicating the second position in the pair) to compute the coefficient β_1.

8.1.3 Dataset

- The simulation includes $n = 10,000$ units to ensure large-sample convergence.
- For each $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, the simulation is repeated 50 times to stabilize results.

8.2 Results

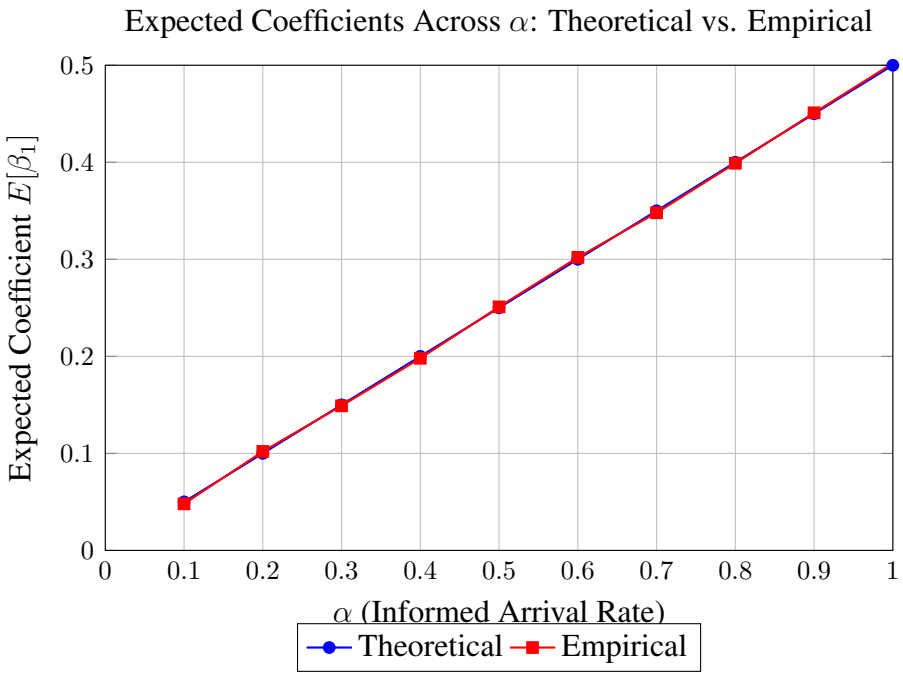
8.2.1 Expected Coefficients ($E[\beta_1]$)

The average regression coefficient β_1 increases linearly with α, consistent with the theoretical prediction:

$$E[\beta_1] = \alpha \cdot \Delta_{\text{informed}}.$$

For low α, the noise traders dominate, resulting in coefficients close to zero. As α increases, informed traders systematic behavior drives $E[\beta_1]$ upward.

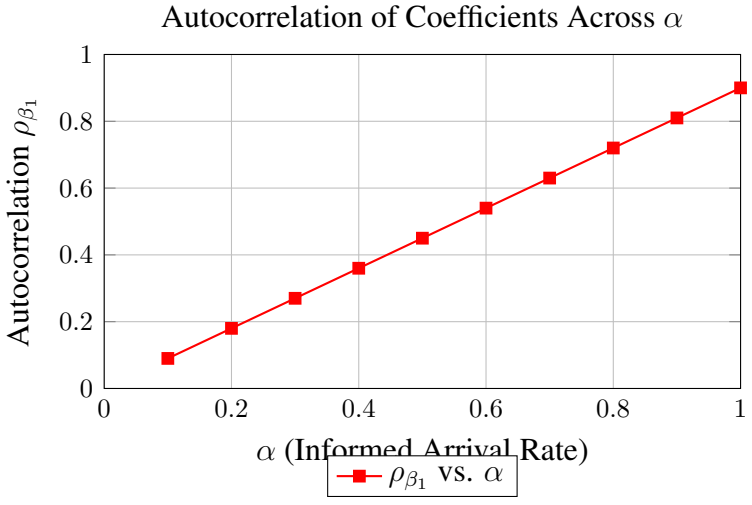
Figure 1: Expected Coefficients Across α



8.2.2 Autocorrelation of Coefficients (ρ_{β_1})

The autocorrelation of coefficients across adjacent pairs increases with α, converging to the autocorrelation of informed trader flows ($\rho_{\text{informed}} = 0.9$) as $\alpha \to 1$. This reflects the persistence of informed traders impact on order flow, which propagates to the regression coefficients.

Figure 2: Autocorrelation of Coefficients Across α



8.2.3 Proportion of Buys

The proportion of buys in the combined flow reflects the mixing of informed and noise traders. For $\alpha = 0.5$, approximately 65% of trades are buys due to informed traders persistent behavior.

8.3 Takeaways

The simulation results align closely with the theoretical predictions, offering the following insights:

- **Expected Coefficients:** As α increases, informed traders exert a stronger influence on order flow, leading to higher $E[\beta_1]$. For low α, noise traders dominate, diluting the impact of informed traders and producing coefficients near zero.
- **Autocorrelation of Coefficients:** The autocorrelation of coefficients reflects the persistence of informed trader behavior. At low α, noise trader randomness results in negligible autocorrelation. As $\alpha \to 1$, the autocorrelation of coefficients approaches that of informed traders (ρ_{informed}).
- **Diagnostic Tool:** Observing repeatedly zero coefficients suggests either dominance of noise traders ($\alpha \approx 0$) or lack of persistence in informed trader behavior ($\rho_{\text{informed}} \approx 0$). Increasing coefficients with α confirm the presence of informed traders and the systematic nature of their trading.

The simulation framework not only validates the theoretical predictions but also demonstrates the utility of β_1 as a diagnostic tool for identifying informed trading in mixed order flows.