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.

1 Introduction

Modern electronic financial markets centralize bilateral engagements between buyers and sellers by adopting a set of algorithmic rules that dictate how, when, and which orders are matched. These rules intend to guarantee efficiency and transparency in markets, so that prices accurately reflect the information available about underlying valuations of assets. With bids and offers being publicly displayed in a trading protocol known as limit order book, rumor-based attempts to manipulate prices have been replaced by automated strategies exploiting the infrastructure of algorithmic markets. These strategies intend to send demand signals perceived by other traders as legitimate — like a better price or larger quantity — and move prices away from the efficient value of the asset through the induced response of other traders.

Whether a manipulator can successfully distort prices, even if temporarily, is a crucial question for the functioning of financial markets. Establishing causality has remained a long-standing challenge in empirical work with high-frequency financial data — not only in manipulation studies, but in any context where researchers want to study the effect of a particular set of traders on prices. Greater time granularity helps separating 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. Confounding behavior can arise in equilibrium (Baruch and Glosten (2013), Hasbrouck (2018), Williams and Skrzypacz (2021)) and seemingly abnormal market activity may be due to information or liquidity needs. Ultimately, market conditions influence and are influenced by manipulation. To a large extent, we still lack a framework to address the fundamental identification problem of the empirical study of trading: "*what would have happened to prices without the trader's action?*"

I make progress on this front by introducing an empirical framework for causal identification in market microstructure with 3 new ingredients to study a form of manipulation that can be identified from the data. I study cases when a trader takes both sides of the same transaction but instead of letting orders cross — which could be considered illegal — uses a compliance tool to prevent legal exposure. This functionality is offered by every major exchange and in US futures markets, many of which are the largest and most traded assets in the world, 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.

Self-trading prevention (STP) functionalities are offered by exchanges because parallel management of multiple executing algorithms often puts a trader on both sides of the market, making unintended order crossing much more likely in modern financial markets than when humans traded at exchange floors. When passive and marketable orders from the same trader cross, the STP functionality must be instructed 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 removals at the CME that can only be triggered by the STP functionality, effectively revealing all self-trades, the lifetime of the passive orders canceled and of the marketable order that remains publicly displayed.

While self-trading may be accidental, there are material advantages to exploit the practice. Because the information necessary to traders to scan and learn about a self-trade is disseminated by the exchange in a message much heavier than regular order updates (e.g., a new quote, or a cancellation), it takes the average self-trader at least twice longer to be discovered than the average liquidity provider who cancels an order. The information is also not flagged or informed by the exchange — traders must be searching for self-trades and implement a sequence of steps to filter the data, effectively deploying the same strategy I use in the paper to identify when the STP tool is used. These informational advantages compound and may help the signals self-traders provide to the market appear more credible for a longer time.

I find that self-trading is pervasive. With a sample including some of the largest futures markets in the world — treasures, oil, gold, and E-mini S&P 500 — I show that at least 0.5% of posted liquidity self-trades without ever being filled by other traders' orders. This amounts to $50 billion/year of front-contract liquidity provision removed due to self-trading. As much as 4% 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. These findings are robust intraday and during long time series, over a variety of assets and underlying market regimes. These aggregate statistics match a 2013 remark by the Commodity Futures Trading Commission's (CFTC) 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]

Having showed that self-trading happens often, the second part of the paper introduces a new framework to estimate causal effects in high-frequency markets. But why should we expect self-trading to affect prices? Profitable instances of the manipulative practice involve two sequential types of market

[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). Because the report used confidential data with trader identifiers, rather than the market data approach I use, the congruence in both levels of self-trading reaffirms the validity of my approach.

effects. From the moment the self-trader's limit order enters the book until it is canceled by the STP functionality, it acts much like spoofing if the trader never intended to provide actual liquidity. Of course, "fake" liquidity can only exert the desired effect if it is perceived as a legitimate demand signal. The probability that this signal is internalized as credible by other traders increases in the price and size of the order. Better-priced quotes expose the trader to execution risk alone and larger orders augment execution risk at any price level. Thus, to entice other traders to provide liquidity in the a direction that is profitable to the self-trader, her orders must provide these signals.

Consistent with that, up to 40% of the limit orders that eventually self-trade improve quoted prices when entered in the book. At a baseline average of 10%, they are much more likely to improve prices than comparable regular orders from other traders. 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 and are consistent with credible demand signals to other traders. In line with the bite of this signal, following an order that eventually self-trades, more liquidity is added to that side and price of the order book than the average comparable order. Once more traders have created a stronger buy or sell position in the market, the self-trader triggers the STP tool and registers a new market price. Whether this causes other traders to follow suit and further move the price away is the type of question my framework can help answer.

My empirical strategy has three components. The first novel element provides a source of exogenous variation to market microstructure. I exploit a feature of electronic markets where traders can only learn about a new order with delay. When a trade (or any order) happens the exchange records the timestamp of the event in its matching engine, but then it needs to make that information public. To disseminate the information to market participants, the exchange faces an internal source of latency and records the timestamp when the first market participant could learn about the trade. Because this exchange latency is as-good-as random, I use it to assign an information treatment. The latency exposes trades occurring almost at the same time to the same information set, except the "news" about the reference trade.

The second ingredient in the empirical strategy involves a linear projection model where for an immediate price impact — that one for the first set of trades following a self-trade — the identifying assumptions are identical to a standard difference-in-differences: no anticipation and parallel trends. The specification also has as the causal estimand the average treatment effect on the treated. With market data containing the two timestamps and markets that are sufficiently liquid so that trades arrive during the exchange latency window, this approach can be applied in any high-frequency trading data, under this transparent and small set of assumptions.

It is important to stress the microstructure counterparties of these assumptions for empirical estimation. Idiosyncratic shocks (e.g., private information, liquidity needs), shocks correlated across market participants (e.g., cross-asset inventory management or arbitrage) are allowed to occur. Trades are also allowed to be placed strategically. The only implied assumption is that trades cannot be completely un-

correlated with the state of the order book. This orthogonality assumption (similar to a relevance-style test) can be tested, which together with other two tests — for violations of the Stable Unit Treatment Value Assumption (SUTVA) and potential selection on price trends by self-traders — closely connect the empirical strategy to widely used microstructure models of trading.

To illustrate the first test, in general we assume events happening after other events in the limit order book were aware of them and thus affected informationally. But a large class of microstructure models considers traders arriving randomly, or whose trade flow is uncorrelated over time. In my empirical strategy, the concern is that what I call impacted trades are orthogonal to the limit order book and thus to treatment. By writing out the empirical correspondent of a simple Glosten and Milgrom (1985) sequential trading model, I derive the conditions under which estimated price impacts are consistent with orthogonal trades — both in terms of expected coefficient and autocorrelation in the estimated dynamic effects. The other tests are developed in a similar spirit.

The third ingredient involves taking the immediate price impact estimator to dynamic estimates. This is challenging because it involves defining how to "pass on" the causality chain still relying on quasi-randomness introduced by exchange latency. This is not straightforward because unless one is willing to assume that information already incorporated into the limit order book from the first self-trade no longer impacts later trades used as controls, this dynamic approach could face the pitfalls in the recent difference-in-differences literature. Thus, dynamic treatments will not recover average treatment effects on the treated.

I build on the generalized linear projections setting in Dube et al. (2023) and impose an assumption in the context of high-frequency data where the "direct" effects from the reference trade to later trades outside the main causality chain are homogeneous for the same delay. That is, for treated and control trades at time $t+2$, whatever the effect of the original self-trade in t (other than the chain effect through the treatment in $t+1$), it must be the same on average for both groups of trades. If that is the case, this effect gets absorbed by time fixed effects and will not bias dynamic treatment effect estimates. Note that this does not impose homogeneous direct effects over time: trades used as treated and control with different delays $t+h$ can have different responses to the original shock, like a decaying informational effect over time.

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. These results are approximately symmetric for positive and negative price movements, and because of how often self-trading occurs — about 4% of trades in US treasuries futures for example — collectively amount to hundreds of millions of dollars in predictable short-term returns.

The final part of the paper plugs these causal price impacts, which are estimated over short lags, to market-wide phenomena. While the estimates of price impact are robust and non-negligible, at such high frequency one may wonder whether these even add up to meaningful aggregate effects. 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, this may not be the case if self-trading activity tends to cluster or intensify during periods with large market movements. In these situations, their contribution to price trends may add up and exacerbate fundamentals-driven responses.

I study this type of setting by focusing on flash events: brief moments of substantial price increases or declines. 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. This type of causal question can be tackled with a framework known as causal attribution. Rather than focusing on the consequences of an event — like the causal price effect estimation — it focuses on its causes.

I implement the model in Ganong and Noel (2022) for mortgage default and consider two potential causes for a binary event — the occurrence of a flash event. These are information shocks and self-trading. Because informed trading can only be approximated with noisy proxies, like the presence of large orders, I follow their strategy to use reverse regression, where the proxy (large trades) becomes the outcome. By comparing the change in large trades in flash events with self-trading to two benchmarks — change in large trades in flash events without self-trading and change in large trades around self-trading — I can estimate the contribution of self-trading without information episodes in driving flash events. 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.

The main contribution of this paper is to provide a new framework for causal identification with high-frequency trading data. Except for specific natural experiments at a more aggregate level (e.g., Bolandnazar et al. (2020), Kirilenko et al. (2017), Shkilko and Sokolov (2020), Eaton et al. (2022)), empirical work in microstructure builds on the vector autoregressive (VAR) tradition by Hasbrouck (1991) — see Brogaard et al. (2019) and Hirschey (2021) for recent applications. Subject to the usual caveats (Stock and Watson (2001)), VAR analysis has been attractive in market microstructure because greater data granularity typical in these data "orders" observations in time in a more precise way, a feature that macroeconomic data sampled at lower frequency lacks.

However, with trading speeds measured in millionths of a second, even tiny delays in reaction from traders can lead to violations in the underlying assumptions needed to identify responses in vector autoregressions. By simply bringing additional data — the timestamp of when traders learn about events in the exchange matching engine — we can see how temporal ordering alone is insufficient to build on for a causal framework. Even for a view of causality as simply predictive ability, an omitted delay (the timestamp used in this paper's data) may, and often does, put trades as having happened "after" other trades when in reality traders could not have yet learned about the "previous" orders.

My framework leverages variation exactly from such exceedingly high speeds in trading that generate arbitrary choke points in the exchange infrastructure and random latencies. By pinning down traders exposed by chance to information about trades, for every single order in the data, I use a linear projections approach with minimal identification assumptions and a precisely characterized causal estimand. The model is consistent with standard theoretical models of trading and therefore robust to the types

of shocks affecting markets — private information, strategic order placement, trend-chasing, as well as heterogeneous and dynamic responses.

I also 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 empirical literature on the effects of market manipulation 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. However, analyses based on investigations or prosecuted cases is not without its drawbacks: they require a careful consideration of selection (as in the paper by Kacperczyk and Pagnotta (2024)) — traders who were more easily detected may not represent the modal manipulator, just like regulators do not decide to pursue potential cases with same propensity.

Studying self-trading in electronic markets has two advantages over analyzing other types of market misconduct. First, I observe all instances when orders from the same trader cross and the prevention tool is used, sidestepping sampling bias and selection concerns. This also avoids relying 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 distinctive pattern of behavior that can be linked to its profitable use. My results show that the average self-trade generates price movements consistent with a profitable strategy, contrary to other attempts to "probe" the market which may result in small losses in exchange for increased information about latent demand (Clark-Joseph (2013)).

Turning to policy implications, while most forms of price manipulation are banned and offer legal exposure or at minimum hefty fines by regulatory agencies, successful enforcement depends on three sequential steps: (*i*) detect defined manipulative behavior, (*ii*) quantify its impact (harm), and (*iii*) prove intent (motive). Starting with establishing intent, meeting the legal burden of motive based on data and without "smoking gun" accessory information (i.e., a whistle-blower's cooperation) has been extremely challenging.[2] Without intent, regulators can at least hope to quantify adverse effects to other market participants from conduct that is consistent with manipulative behavior. The framework in my paper provides the necessary tools to estimate whether a group of trades with identifiable conduct — regardless of intent — have an effect on market prices.

More broadly, my analysis adds to the literature concerned with the rules organizing how traders interact in financial markets. Just like other markets, considerations about how to optimally design the "plumbing" in trading venues — rules that ensure efficiency through timely price discovery and a neutral

[2]Despite an increase in trading activity of several orders of magnitude in the last decades, 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 in 2010.

system of incentives — has been a central object of study in market microstructure. From how to conduct trading (Budish et al. (2015), Baldauf and Mollner (2020)), to limit pockets of instability (Chen et al. (2024)), to how should quoted prices and quantities be discretized (Li and Ye (2023)) and customer orders managed by intermediaries (Ernst and Spatt (2022)), research has investigated a host of issues in market design — from obscure rules to significantly altering how trading takes place (Christie and Schultz (1994)).

This paper focuses on how traders can exploit strategically even relatively peripheral functionalities in electronic markets. Self-trading prevention tools are offered by exchanges as a seemingly innocuous compliance resource, but as my results show, with non-negligible effects on price instability. While I do not argue that there exists an easy "fix" for the probability of orders crossing increasing in the intensity of liquidity provision, changes to the current design of the functionalities intended to protect traders who inadvertently self-trade seem warranted by my results. An option would be for exchanges to block orders from the same trader from crossing by setting the match allocation priority of limit orders to zero when a self-match happens, immediately restoring it after the trade. Instead of removing the limit order from the book, the exchange simply matches the trade to other limit orders, without any change in the status to the self-trader's orders. If anything, the results in this paper underscore the importance of taking into account potential loopholes and unfair advantages created by algorithms that organize trading.

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 (for immediate execution) on one side of the limit order book matches with one or more resting orders (i.e., liquidity providers) on the other side. Under normal circumstances, a match results in a trade, where the exchange sends out a 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 functionality exploited by traders in this paper prevents a match to turn into a trade. The tool automatically deletes one side of the matched orders when they come from the same trader in order to avoid a trade which could be illegal. Because exchanges and regulators refer to the practice that triggers the STP functionality as self-trading, even though the activity actually involves self-matching, I use the two terms interchangeably throughout the paper, only making a distinction when relevant. Moreover, even though the use of the tool is to prevent a self-trade, because of the default option leaving the marketable order alive in the book and deletes the trader's limit orders in a way difficult for the other market participants to learn, effectively the trader can derive all benefits from actually self-trading. For that reason, I refer to the manipulative behavior of the tool as simply self-trading.

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.*

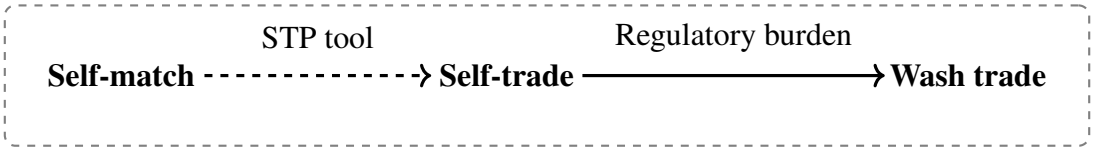
clearly states that both individuals engaging in self-trading and brokerage firms enabling the conduct are subject to regulatory consequences.

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 legal scrutiny in case their orders crossed. 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. For example, if a trader has a sell order at the best ask and submits a buy order for immediate execution, the tool automatically cancels the sell quote. The functionality allows users to change the default option, canceling the incoming (newest) order instead of the resting order (oldest). The tool 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 when incoming orders will still be queued up for execution.

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 interpretation in the industry is to consider orders submitted for the same account or different desks at the same proprietary high-frequency shop operating in a common market as the "same trader". 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. 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 form of market misconduct documented for centuries, successful prosecution of wash trading is relatively rare.[3] 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 STP tools can be exploited strategically — and their inadequate use offers negligible punishment risk to traders. Abuse of trading platform systems and functionalities is only considered a "violation" by exchanges. 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*

[3] During the US railroad expansion in the early 1900s, several prominent cases of wash trades 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)).

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*".[4]

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 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.

The discussion above underscores the cost-side incentives to take advantage of self-trading prevention tools. Triggering the STP tool imposes considerably less punishment risk to traders. I explore the potential economic benefits in the next section.

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.[5] This level of detail on market messages combined with the default option of CME's self-trading prevention

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

[5] 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.

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.[6]

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.

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.

[6]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.

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 that would only *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 baked in to the analysis.

3.2 General Patterns

I begin by establishing broad patterns of self-trading activity in futures markets.

Self-trading happens often. 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 the 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 the 6-month 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 slightly 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.

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.

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.

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 canceled 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. This is

a crucial demand signal in self-trading. Posted liquidity is partly cheap-talk, but self-trading imposes a direct cost to the trader. Albeit small, unless the price moves in her desired direction, self-trading may result in net losses, while posting liquidity faces only execution risk. 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.[7]

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 the flow of a trader with since intentions.

Figure 3 illustrates an example of this dynamics when the self-trader enters a passive buy order by setting a new price. Following the order 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 opposed 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

[7]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)). Because of that and other features of trading protocols, price and size are arguably the strongest demand signals observable from market data.

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 influence the price upward.

3.3.1 Empirical considerations

The dynamics I depict above is a general characterization of market manipulation in line with 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 prices without the self-trader's action? Before setting up a framework to answer this, I discuss some additional considerations regarding the conditions in which self-trading may be more os less likely to happen.

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.

From an identification perspective, the presence of a price trend acts as an omitted variable. Traders are likely to self-trade when they believe the potential price impact of their action is above a certain threshold. Because self-trading involves some inventory cost (the match against other traders), we should expect some sorting based on either observed price trends or unobserved market conditions that would amplify the demand signals from self-trading. This does not mean that self-trading cannot trigger price responses, even when no pre-existing price trend is ongoing. But a credible empirical strategy needs to take into account the fact that selection-on-gains will likely affect treatment effect magnitudes, and if not a threat to identification, at least make these effects more likely to be valid for a subset of treated units.

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.

3.3.2 Order flow buildp and market signals

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. This is an important result: on average 10% of orders improve bid or offer quotes (equivalent to 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, with an average of 623 contracts added up to one minute. This amounts to hundreds of millions of dollars in liquidity being supplied to the same price level of the self-trader's limit order in just one minute. The average self-trading strategy execution time across all futures markets is just over 1 minute.

These results speak to the ability of the spoofing component of a self-trading strategy in attracting order flow consistent with the direction of the profitable use of the strategy. Self-traders send credible demand signals — their limit orders are larger and better-priced than the average order.

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, we still lack a framework to address the fundamental identification problem of the empirical study of trading: "*what would have happened to prices without the trader's action?*"

This section makes progress on this front by introducing an empirical framework for causal identification in market microstructure with 3 new ingredients to study a form of manipulation that can be identified from the data. The approach can be applied to any "reference event" of choice where a group of market participants or order types meant to represent such group play the role of self-trading in my context. The only requirement for implementation is that the research has access to intraday data containing the timestamp of the transaction and another to when the information about that transaction becomes public.

The three components of the empirical framework are as follows. First, I introduce a novel source of variation to information exposure in high-frequency trading, where trades occurring almost at the same time have the same information set, except for the "news" about a reference trade which have arrived to some traders, but not yet the others. This delay in how the information about a trade becomes public — which I call exchange latency — is quasi-random and identifiable in the data.

The second ingredient involves a linear projection model where for an immediate price impact that one for the first set of trades following a self-trade, the identifying assumptions are identical to a standard difference-in-differences. With market data containing the two timestamps and markets that are sufficiently liquid so that trades arrive during this delay window, this approach can be applied in any high-frequency trading data, under this transparent and small set of assumptions.

The third ingredient involves taking the immediate price impact estimator to dynamic estimates. This is challenging because it involves defining how to "pass on" the causality chain still relying on quasi-randomness introduced by exchange latency. I show that with an additional assumption of homogeneity in responses from trades randomly picked by the delay window, dynamic treatment effects can be estimated causally.

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

In Figure 1, I illustrated the stylized routing system of CME's trading infrastructure. My data clocks two timestamps — a transaction and a sending time. 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. This exchange latency varies substantially — from 30 to 60,000 microseconds. More importantly, it is as-good-as random.

To probe whether exchange latency is as-good-as random, I run the following regression for every single trading message for all days and markets in my sample:

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

The parameters δ_k measure the importance of the aggregate quantity of inbound messages of type k (trade, new, update, cancel) in that minute. I also allow for potential persistence in latency, captured through ω_τ. The 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 dummies μ_q in the above specification net out those eventual effects.

Table 4 shows the results. Standard errors are clustered at market-and-minute. As expected given the enormous number of observations (over one billion), most coefficients are statistically significant. More importantly, the joint explanatory power is very low: in the specification with latency persistence, message traffic, as well as fixed effects, less than 10% of variation in exchange latency can be explained. Results in the Appendix replicate this specification by aggregating variables to the minute, separating latency in trade messages from limit order updates, replicates the same analysis by market-by-market, and considers even longer lags in persistence, as well as lags in the traffic variables. Across all specifications, the takeaway is the same: exchange latency is overwhelmingly "unpredictable" and therefore can be used as a source of exogenous variation in high-frequency data.

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 4), it can be used to assign treatment $z_{n,t}$ to trade n in moment t.

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.[8]

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.

[8] 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.

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 $z_{n,t}$:

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

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 $z_{n,t-1} = z_{n,t} = 0$.

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 (1) 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 (1) 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 (1), 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 $z_{n,t}$:

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

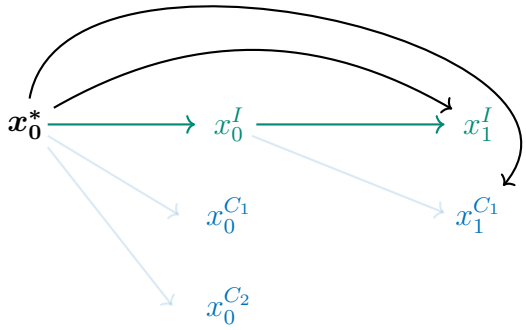
using as control units trades that have not experienced a change in treatment status at $t + h$.

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 (2). 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 (1) and (2)). 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 the Appendix, 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 (2) will absorb them. This is stated more formally by the assumption below.

ASSUMPTION 1: $\mathbb{E}\left[x_{t+h}^I | x_0^* = 1\right] - \mathbb{E}\left[x_{t+h}^I | x_0^* = 0\right] = \mathbb{E}\left[x_{t+h}^C | x_0^* = 1\right] - \mathbb{E}\left[x_{t+h}^C | x_0^* = 0\right]$, for all $h > 0$. Direct effects with same delay are homogeneous.

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 the direct effects outside the main causality chain are constant over time. Rather, it only imposes that for the same delay h, trades that are randomly picked as impacted or contemporaneous experience the same

effect from the self-trade.

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 displaying how many eligible contemporaneous and impacted trades we have at various h.

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 regression 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:

$$\begin{aligned} \beta_h = \quad & \alpha P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{informed}}) + (1-\alpha) P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 1, x_t^{\text{noise}}) \\ - \quad & \alpha P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{informed}}) - (1-\alpha) P(\mathbf{1}\{x_t > 0\} = 1 \mid z_{t+h} = 0, x_t^{\text{noise}}) \\ = \quad & \alpha \rho\left(x_t^{\text{informed}}\right) \end{aligned} \tag{3}$$

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 (3) 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 (3) 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 $\widehat{\beta}_{t+h}$.[9]

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 my 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

[9] 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.

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.

In this section, 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. Manipulative-like behavior is unlikely to be random — self-trades are more likely to happen when self-traders think the potential benefits of what could be perceived as market misconduct are larger. Besides this selection mechanism, there is a problem of attributing magnitudes: 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. This was first proposed by Pearl (1999) and Rosenbaum (2001) and focus on estimating the causes of an event, rather than its consequences (which is the usual causal estimand in economics).

For this analysis, I draw closely from the potential outcomes framework in Ganong and Noel (2022) which includes 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 "explanation" for large price movements in financial markets is the role of informed trading, or news being incorporated into prices more broadly. The empirical challenge with this motive is that informed trading 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 noisy 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.

5.1 Causal Attribution Framework

5.1.1 Setup

Let T^* be a dummy representing a private-information trade, S a dummy representing a self-trade, and the potential outcome Y is the occurrence of a flash event. T is the noisy proxy for informed trading, which I consider it to be the change in the average trade size. Larger order are more likely to be informed, with probability less than one. The potential outcome function $Y(T^*, S)$ maps to four potential outcomes, with $Y(1,1)$ representing a flash event with informed trading and self-trading and $Y(0,0)$ when a flash even does not happen (for example, a sustained price rally collapses after 5 minutes).

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

The assumption effectively rules out purely noise-driven sustained price movements. Self-trading boils down to attempting to appear as informed trading — when successfully disguised as a demand signal, it induces others to trade in the direction desired by the self-trader and become potentially profitable. Whether substantial price movements were of the "good" kind — price quickly incorporating shocks to asset value that happened to be large — may matter from a normative sense, from an econometric perspective one only needs to consider that truly informed trading or trading that wants to appear to be informed can have similar price effects.

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

Because successful self-trading appears to be informed, when there is actual informed trading the presence of self-traders determines potential outcomes as if there is more information-driven activity. This set of basic assumptions imply that flash events can be of one out of three types: purely information driven, $Y(1,0)$, purely manipulation driven, $Y(0,1)$, or caused by both informed trading and self-trading $Y(1,1)$. What fraction of flash events corresponds to each type is the question a causal attribution framework attempts to answer.

To do so, we need to construct counterfactuals that remove each of the potential causes of flash events. For example, the share of flash events that would be eliminated without informed trading measures its role: $\alpha_{\text{informed}} = \dfrac{\mathbb{E}\left[Y(T^*,1) \mid S=1\right] - \mathbb{E}\left[Y(0,1) \mid S=1\right]}{\mathbb{E}\left[Y(T^*,1) \mid S=1\right]}$ which is equivalent to the complement of the share of flash events caused *solely* by self-trading. That is because α_{informed} captures both $Y(1,0)$ and $Y(1,1)$ combined, thus $1 - \alpha_{\text{informed}}$ represents $Y(0,1)$.

5.1.2 The role of informed trading

Identifying assumptions. To estimate the role of informed trading on flash events using a proxy like trade size, we need two additional assumptions to employ reverse regression. Here I also follow Ganong and Noel (2022) and discuss how their identifying assumptions can be interpreted in the context of high-frequency trading data.

ASSUMPTION 3: Informed trading T^* is orthogonal to the potential outcome $Y(T^*, G)$ conditional on self-trading G.

Importantly, this conditional exogeneity assumption relates to the true information shock T^*. In the type of world where informed trading is assumed to arrive randomly — which includes the stylized trading model in Section 4 — innovations drive informed trading entirely and the assumption holds

unconditional on self-trading. But even when information is "assigned" to multiple traders, for example because several insiders learned about the same news, as long as the news does not include the occurrence of a flash event, the assumption also holds. In our setting, Assumption 3 then imposes that true information shocks are not based on public information already displayed in market data — it does not assume that informed traders do not place their orders unconditional on price paths. It also allows for self-trading being more likely to occur when informed trading is happening, $P(T^* = 1|S = 1) > P(T^* = 1|S = 0)$, which is consistent with the possibility of selection-on-gains by the self-trader. Assumption 3 also allows for heterogeneous effects $\mathbb{E}\left[Y(1,1) - Y(0,1)\right] > \mathbb{E}\left[Y(1,0) - Y(0,0)\right]$.

ASSUMPTION 4: $T\left(T^*, S, Y\right) = T(T^*)$ and $\{T(0), T(1)\}$ is orthogonal to (T^*, Y, S). The proxy for informed trading T increases on average when there is informed trading $\mathbb{E}\left[T(1)\right] \neq \mathbb{E}\left[T(0)\right]$.

This assumption imposes restrictions on the noisy proxy for informed trading. First, its potential-outcome function is independent on whether a flash event, self-trading, or an information shock happen. Suppose a trader receives a private information shock that is short-lived. This is a reasonable view of an informational advantage as an expiring option. If a flash event is more likely to happen (i.e., there is already a strong price trend in the market), and the price trend is in the direction of the private information — price is trending up and the trader received good news — she may decrease the size of her trades and increase execution speed. Conversely, if the trend is going against her news, the informed trader can execute larger quantities because they are easier to disguise, and therefore also trade faster. This implies that informed traders will on average change the average order size in the market the same way when flash events have self-trading or not.

Ganong and Noel (2022) show that with Assumptions 1, 2, 3, and 4 the causal estimand of interest α_{informed}, which measures the fraction of flash events that would be eliminated without informed trading, can be computed as:

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

where the second equality expresses the counterfactual in terms of the noisy proxy for informed trading T and can in turn be estimated from the data in the following way:

$$\alpha_{\text{informed}} = \frac{\overbrace{\mathbb{E}\left[T|Y = 1, S = 1\right]}^{\Delta_{\text{large trades in flash events with self-trading}}} - \mathbb{E}\left[T|S = 1\right]}{\underbrace{\mathbb{E}\left[T|Y = 1, S = 0\right]}_{\Delta_{\text{large trades in flash events without self-trading}}} - \underbrace{\mathbb{E}\left[T|S = 1\right]}_{\Delta_{\text{large trades around self-trading}}}} \tag{4}$$

Estimation. The correspondences above come from two different regressions. The first proxy I use for

informed trading is the average size of trades in a given minute. Conditioning T on $Y = 1$ implies estimating effects during the period leading up to a price movement being a flash event. I follow Ganong and Noel (2022) and consider that this is captured by comparing the last three minutes $\{-2, -1, 0\}$ of the 10-minute flash event to the beginning of the event, $\{-9, -8, ..., -3\}$. In my setting this means that the fact that the trend in price continued to rise over time should have been sustained by informed trading, proxied by the trade size outcome. This comparison can be made simply with a dummy $\mathbf{1}\{m \in \tau^{peak}\}$ where τ^{peak} is equal to one during the three final minutes leading up to the sustained 10-minute price trend. I consider alternative definitions for this temporal cutoff, as well as implement other robustness checks, later in this section.

The other comparison can be made with the use of another dummy, $\mathbf{1}\{STP\}$, which tracks whether a flash event had the presence of self-trading. To continue with the binary setting laid out in the causal attribution framework, I convert the (multi-valued) presence of self-trading in a minute to a dummy equal to one if they account for at least 5% of trades. I vary this cutoff to assess the robustness of estimated effects later. Estimating both time and self-trading indicators in the same regression gives the moment conditions in expression (4):

$$\frac{TradeSize_m}{TradeSize_{pre}} = a + \kappa \mathbf{1}\{STP\}_m + \gamma \mathbf{1}\{m \in \tau^{peak}\} + \beta \mathbf{1}\{m \in \tau^{peak}\} \times \mathbf{1}\{STP\}_m + \varepsilon_m \qquad (5)$$

where γ targets $\mathbb{E}[T \mid Y = 1, S = 0]$, and $\beta + \gamma$ gives the differential effect when self-trading is present $\mathbb{E}[T \mid Y = 1, S = 1]$. To translate the average trade size into relative changes over the duration of the flash event, I normalize the outcome by a pre-flash event period, $TradeSize_{pre}$, which is the average trade size in the three minutes prior to the beginning of the event. Note that even if the price was trending before the 10-minute flash event, specification (5) still captures changes in the proxy for informed trading during the period of strongest momentum period within a possibly longer directional trend.

The last ingredient needed to estimate α_{informed} is the change in the proxy value around self-trading, $\mathbb{E}[T \mid S = 1]$. Since this moment is obtained unconditional on flash events, we instead need only to focus on all qualified cases of self-trades and replicate the same analysis in regression (5) where $\mathbf{1}\{m \in \tau^{peak}\}$ tracks large orders during the 8th, 9th, and 10th minute following the self-trade and the baseline period $TradeSize_{pre}$ is defined like before, but relative to the transaction time of the self-trade: $\frac{TradeSize_m}{TradeSize_{pre}} = a + \omega \mathbf{1}\{m \in \tau^{peak}\} + \varepsilon_m$. Finally, the quantity $1 - \frac{\widehat{\gamma} + \widehat{\beta} - \widehat{\omega}}{\widehat{\gamma} - \widehat{\omega}}$ give the sole contribution of self-trading in flash events in US futures markets.

Results. The estimated values for the parameters are $\widehat{\gamma} = 0.0353$, $\widehat{\beta} = -0.0031$, and $\widehat{\omega} = -0.0011$. This implies that the average trade sizes increase by 3.3% in flash events with self-trades and 3.6% in flash events without the presence of self-trading. For a change in trade size following self-trades of just

0.1%, the estimated fraction of flash events caused by self-trading is

$$1 - \frac{0.0353 - 0.0031 - (-0.0011)}{0.0353 - (-0.0011)} = 8.5\%$$

which means that 91.5% of flash events in US futures markets are caused by informed trading alone or in combination with self-trading.

6 Conclusion

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.

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. These results are approximately symmetric for positive and negative price movements, and because of how often self-trading occurs — about 4% of trades in US treasuries futures for example — collectively amount to hundreds of millions of dollars in predictable short-term returns. These price impacts have aggregate consequences: almost 10% of flash events — brief moments of substantial price increases or declines — in futures markets can be causally attributed to self-trading alone.

References

Aggarwal, RajeshK. 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.

Bolandnazar, Mohammadreza, Robert J. Jackson Jr., Wei Jian, and Joshua Mitts (2020) "Trading Against the Random Expiration of Private Information: A Natural Experiment," *The Journal of Finance*, Vol. 75, No. 1, pp. 5–44.

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.

Chen, Hui, Anton Petukhov, Jian Wang, and Hao Xing (2024) "The Dark Side of Circuit Breakers," *The Journal of Finance*, Vol. 79, No. 2, pp. 1405–1455.

Christie, Willian G. and Paul H. Schultz (1994) "Why do NASDAQ Market Makers Avoid Odd-Eighth Quotes?" *The Journal of Finance*, Vol. 49, No. 5, pp. 1813–1840.

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.

Eaton, Gregory W., T. Clifton Green, Brian S. Roseman, and Yanbin Wu (2022) "Retail trader sophistication and stock market quality: Evidence from brokerage outages," *Journal of Financial Economics*, Vol. 146, No. 2, pp. 502–528.

Ernst, Thomas and Chester S Spatt (2022) "Payment for Order Flow And Asset Choice," Working Paper 29883, National Bureau of Economic Research.

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.

Hirschey, Nicholas (2021) "Do High-Frequency Traders Anticipate Buying and Selling Pressure?" *Management Science*, Vol. 67, No. 6, pp. 3321–3345.

Kacperczyk, Marcin and Emiliano S. Pagnotta (2024) "Legal Risk and Insider Trading," *The Journal of Finance*, Vol. 79, No. 1, pp. 305–355.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun (2017) "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance*, Vol. 72, No. 3, pp. 967–998.

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.

Li, Sida and Mao Ye (2023) "Discrete price, discrete quantity, and the optimal nominal price of a stock,"Technical report, Working paper.

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.

Shkilko, Andriy and Konstantin Sokolov (2020) "Every Cloud Has a Silver Lining: Fast Trading, Microwave Connectivity, and Trading Costs," *The Journal of Finance*, Vol. 75, No. 6, pp. 2899–2927.

Stock, James H. and Mark W. Watson (2001) "Vector Autoregressions," *Journal of Economic Perspectives*, Vol. 15, No. 4, p. 101115.

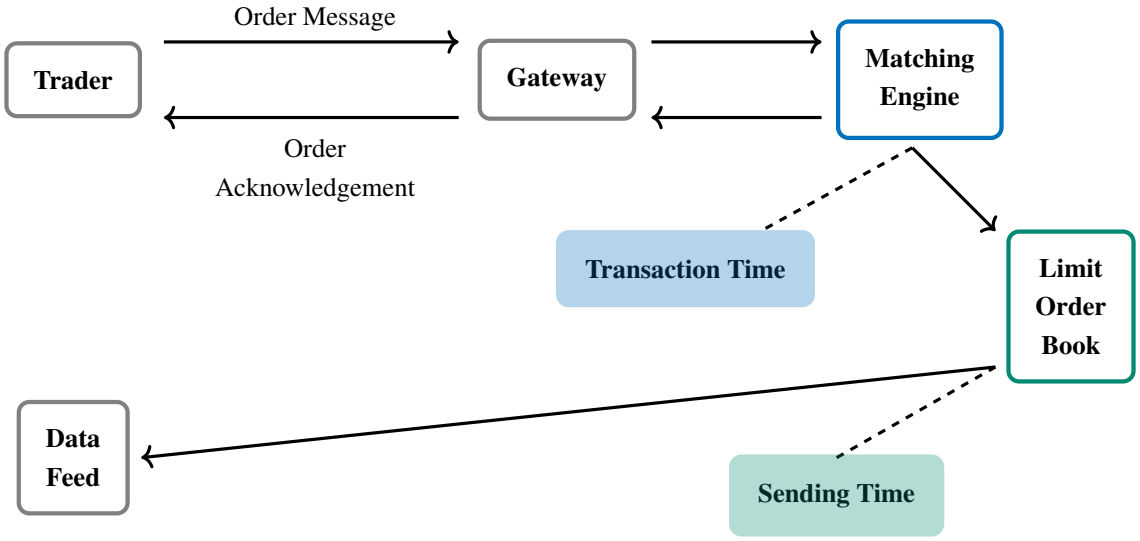
Tóth, Bence, Imon Palit, Fabrizio Lillo, and J. Doyne Farmer (2015) "Why is equity order flow so persistent?" *Journal of Economic Dynamics and Control*, Vol. 51, pp. 218–239.

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

BBO 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 Marketable Order

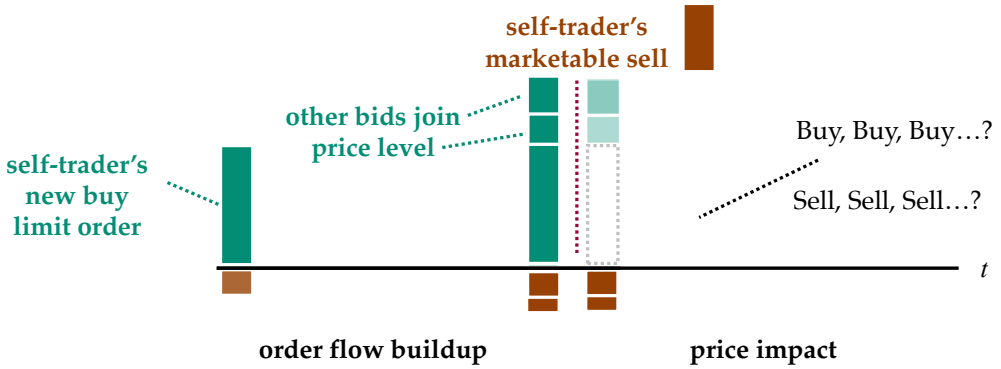
Activity	Quantity	Orders Matched Against	Price	Filled Quantity	Order ID	Timestamp (microsecond)	
Trade	**700**	**3**	**$100**	**160**	**13**	**8:45:23.587663**	***Crossing Trade***
Delete	0	1	$100	0	1	8:45:23.587663	*Self-Trade*
Delete	0	1	$100	0	2	8:45:23.587663	*Self-Trade*
Delete	0	1	$100	55	3	8:45:23.587663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

BBO 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

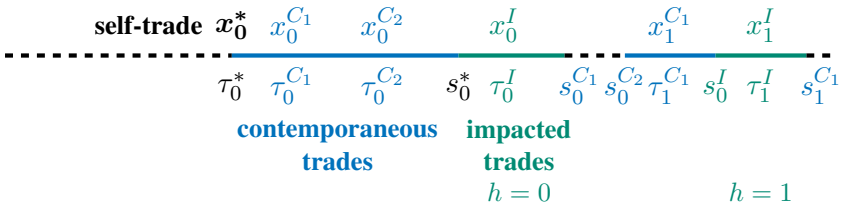
NOTES: This figure illustrates the algorithm that detects self-trading in futures data. **Panel A** shows the best bids and offers (BBO) outstanding. The best ask is $100 and the best bid $98. In the data, orders have an ID which is unique but does not identify market participants. A buy order for immediate execution at $100 arrives at 8:45:23.587663. The order is to sell 700 contracts. The exchange matches the order against liquidity provided at the best bid. **Panel B** shows the trade summary that describes the matches involved in this trade. 3 orders have a quantity filled of 0 and yet have been deleted by the exchange at the same exact instant as the trade. This is the data signature that the self-trading prevention (STP) tool was used: the functionality flagged that orders with ID 1, 2, and 6 all belonged to the same trader behind the buy order and deletes the three limit orders. The marketable buy is not modified and goes on to match against orders from other traders. **Panel C** shows the updated limit order book after the trade. Even though the trade only resulted in 160 contracts being filled, additional 475 contracts in liquidity have been removed by the STP tool. Because the marketable order had a pre-specified price to execute, it does not walk to book, i.e., trade against higher-priced sellers, and ends up establishing the new best bid. Because there are no longer offers at $100, the best ask also increases.

FIGURE 2: IDENTIFYING SELF-TRADES IN MESSAGE DATA



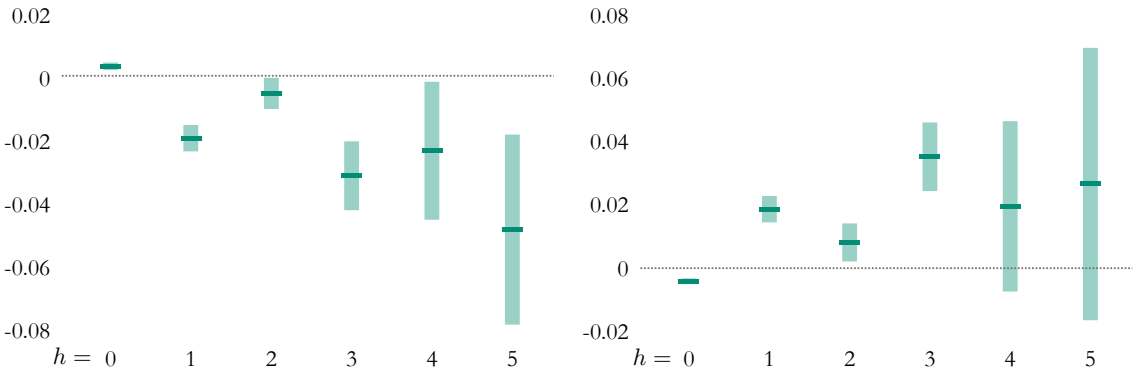
NOTES: This figure illustrates the two sequential types of demand signals associated with self-trading activity. In the example, the trader wants to drive up prices. From the moment the trader enters a limit order (a buy order in the case above) until forces a self-match with a marketable order (sell), the trader intends to entice other traders to also post buy orders at the same price level as her. If enough imbalance between buys and sells builds up, she may self-trade, removing more liquidity from the best bid level than the size of the trade. If this is followed by more buy orders than sales, the price will rise and the strategy could have been profitable.

FIGURE 3: STRATEGIC SELF-TRADING MECHANICS: TWO STEPS



NOTES: This figure illustrates the identification strategy in the paper for the first dynamic effect $h = 1$ following the immediate price impact ($h = 0$) of a self-trade. Transaction timestamps — when trades actually happen — are denoted by τ and sending timestamps — when the information about the trade becomes public — is denoted by s. The interval $s_0^* - \tau_0^*$ measures the exchange latency for the self-trade and randomly exposes a set of trades (x_0^I) to the information about x_0^*, but not others ($x_0^{C_1}, x_0^{C_2}$). The set of impacted trades comprises trades happening after the information on the self-trade becomes public, but before information about any other order becomes public. After information about the impacted trade becomes public and after the information on contemporaneous orders is also published, another set of trades becomes those impacted x_1^I. Contemporaneous orders in $h = 0$ and $h = 1$ are those happening during the exchange latency windows with published information after their respective windows. The immediate causal price impact is estimated by comparing changes in prices for x_0^I relative to $x_0^{C_1}, x_0^{C_2}$. The first dynamic price impact is estimated off of a comparison between x_1^I and $x_1^{C_1}$. The regression specification used to provide these treatment effects, as well as identifying assumptions and additonal details are described in the paper.

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, oild, gold, and E-mini S&P 500 from October 2019 to March 2020.

FIGURE 5: PRICE IMPACT ESTIMATES FOR SELF-TRADING

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

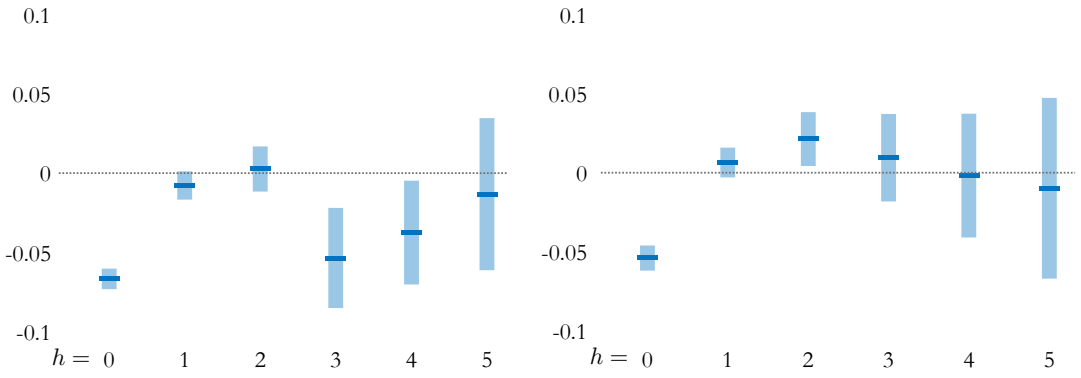
(b) Price change for $Sell_{\text{self-trade}}$

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

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

NOTES: This figure reports dynamic estimates $\widehat{\beta}_h$ of the 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 exchange latency $z_{n,t}$, where $\Delta z_{n,t+h} = 1$ for impacted trades and $z_{n,t+h} = 0$ for contemporaneous trades. Panels **(a)** and **(b)** 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. 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. 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.

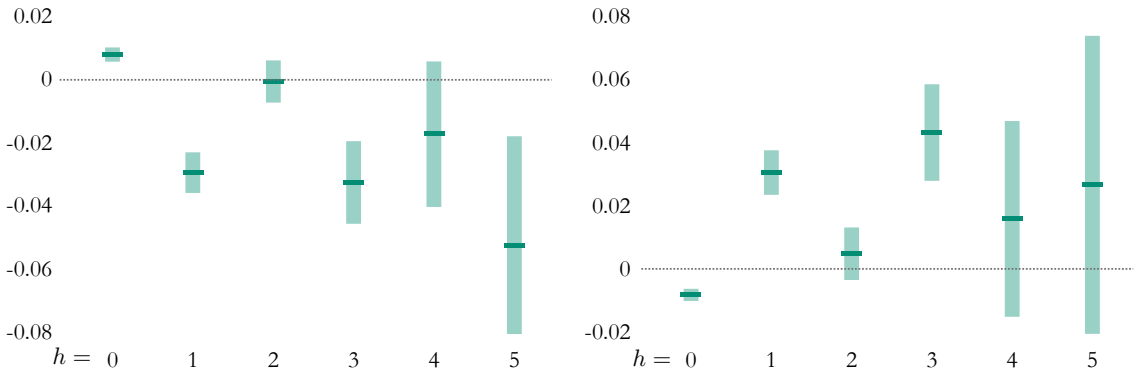
FIGURE 6: DETERMINANTS OF PRICE IMPACT ESTIMATES FOR SELF-TRADING



(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 $\textbf{outcome}_{n,t+h} = \delta_t + \beta_h \Delta z_{n,t+h} + \varepsilon_{n,t}$ 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: DIRECTIONAL TRADE IMPACT ESTIMATES FOR SELF-TRADING



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

NOTES: This figure reports dynamic estimates $\widehat{\beta}_h$ of the 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 exchange latency $z_{n,t}$, where $\Delta z_{n,t+h} = 1$ for impacted trades and $z_{n,t+h} = 0$ for contemporaneous trades. This is the same specification as the baseline model, but using a subset of contemporaneous trades (the control group) that includes all observations that could not have learned about the information shock. 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: REMOVING PRIVATE INFORMATION

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, and E-mini S&P 500, during the period 10/2019 to 03/2020. 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, and E-mini S&P 500, during the period 10/2019 to 03/2020. 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 canceled 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 mid-price 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.

TABLE 3: ORDER FLOW BUILDUP AROUND SELF-TRADING

	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: This table reports summary statistics and changes in the limit order book following the entry of a limit order that eventually self-trades. Times are regularized for easy of comparison. The column **Execution Frequency** gives the cumulative distribution for each time window; the other columns report averages within each time interval. Futures contracts are: 2-year and 10-year t-notes, gold, oil, and E-mini S&P 500, during the period 10/2019 to 03/2020. I use first-nearby contracts (nearest expiration) only for computations.

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.