Phantom Depth:
Theory and Evidence on Manipulable Market Signals*

Alex Boulatov
International College of Economics and Finance
Higher School of Economics, Moscow

Terrence Hendershott
Haas School of Business
University of California, Berkeley

Dmitry Livdan
Haas School of Business
University of California, Berkeley

Norman Schürhoff
Faculty of Business and Economics and Swiss Finance Institute
University of Lausanne

August 2, 2026

*Livdan and Schürhoff are Research Fellows of the CEPR. Schürhoff gratefully acknowledges support from the Swiss Finance Institute. Hendershott gratefully acknowledges support from the Norwegian Finance Initiative and provides expert witness services to a variety of clients. We thank seminar participants for helpful comments. All errors are our own.

Phantom Depth: Theory and Evidence on Manipulable Market Signals

Abstract

Spoofing manipulates markets by creating visible but withdrawable liquidity. A trader displays large orders that others treat as informative, inducing cross-side trading before the order's intent is verified, and then cancels after the display has affected behavior. We study this credibility problem in a Stackelberg model of manipulable displayed depth and test its sign restrictions using 90.6 million collapsed-depth spikes in S&P 500 securities from 2017 to 2023. Price reversals, often used to validate suspicious episodes, are economically zero unconditionally and become large only after conditioning on the realized post-event path. We instead identify the strategic state from within-episode behavior. Displayed depth raises cross-side trading by 77.4%, and cancellations respond to cross-side flow 4.5 times more strongly during the spike than before it. A cross-fitted trigger score predicts excluded outcomes with model-implied signs, and phantom depth is less prevalent when hidden-order activity is high.

JEL Classification: G14, G18, D82, C58, D47

Keywords: Market manipulation, phantom depth, displayed liquidity, spoofing, price impact, market microstructure, asymmetric information

1 Introduction

A limit order book is constructed from visible promises to trade. Some of these promises are costly to make and therefore credible. Others are informative but easy to withdraw. The displayed depth sits at this boundary. A large order at the best bid or offer appears to reveal trading interest, inventory pressure, or private information. Other traders see it and may trade before the order is executed. However, the trader who posts the order can often cancel before execution. The signal is useful because traders respond to it. It is not fully credible because the sender can withdraw it.

This paper studies how traders can manipulate markets by exploiting visible signals that are informative precisely because they are expected to be costly to fake, yet can be withdrawn before they are verified. The problem is a market-microstructure version of the credibility problem in Bénabou and Laroque (1992). In their setting, insiders and gurus influence beliefs through communication whose credibility is imperfect. Here, the communication is not a statement but an order-book action. A trader can create visible depth, induce others to react to that depth, and then remove the order before the market learns whether it reflected genuine willingness to trade.

The empirical challenge is that manipulation and legitimate liquidity provision can look almost identical at high frequency. A market maker may cancel a stale quote after observing adverse order flow. Rapid placement and cancellation of displayed liquidity is also a documented equilibrium feature of modern markets and need not reflect manipulation (Hasbrouck and Saar, 2009; Baruch and Glosten, 2013). A manipulator may follow a similar surface pattern, cancelling after cross-side traders respond to displayed depth. In both cases, trading activity arrives, and the displayed order disappears. Identification therefore turns on the source of the cancellation. The key distinction is whether the order is withdrawn to protect exposed liquidity, or whether it is withdrawn after a visible signal has induced the response it was designed to elicit.

A common way to make this distinction is to look for price reversals after suspicious orders disappear. The intuition is natural. If a trader displays artificial buy pressure, the price should rise while the display is live and fall after the display is removed. We show that this intuition is not a valid identification strategy when reversal enters the event definition. A mechanical detector that uses only displayed-depth features, such as size, duration, and collapse, produces a near-zero unconditional reversal of -0.13 basis points across 90.6 million collapsed-depth events in S&P 500 securities. Adding a directional-reversal condition produces the familiar large number, -11.66 basis points. The difference is mechanical. If the unconditional mean of R_i is close to zero, conditioning on $R_i < 0$ creates a large negative conditional mean even under a null with no manipulation. The observed reversal is therefore consistent with selection on the realized post-event price path rather than independent evidence of a strategic state.

This selection problem changes the object of identification. A reversal-based filter can find episodes that reverse, but once the post-event price path enters the sample definition, that same price path cannot serve as independent evidence of the strategic state. Manipulation must instead be identified from behavior observed before the outcome is realized. We therefore use timing inside the episode. Spoofing works as a display-and-trigger strategy in which a trader displays depth on

one side of the book, other traders respond while the display is live, and the trader cancels once the targeted response has arrived. The relevant object is the cancellation response to elicited cross-side flow rather than the post-cancellation reversal. A legitimate liquidity provider may also cancel after information arrives, but the displayed order itself should not make cross-side flow a much stronger cancellation trigger. That amplification is the distinctive prediction of the manipulation hypothesis.

We formalize this mechanism in a Stackelberg model of manipulable displayed depth. The manipulator posts a visible order before market makers and other traders respond. The order changes the perceived depth of the market and the informational content of order flow. The manipulator then keeps the option to cancel and trade after the market has reacted to the display. In equilibrium, the market can appear deeper than it truly is because part of displayed liquidity is supported by an order intended to be withdrawn. We call this state phantom depth. The model does not say that every cancelled depth spike is manipulation. It gives testable restrictions that separate display-and-trigger behavior from ordinary liquidity provision. We then endogenize the cancellation decision. At an interim stage, the manipulator observes noisy signals of the response elicited by the display and of the fill pressure that threatens execution. Cancellation becomes an optimal stopping decision rather than an exogenous event. The extended model derives the behavioral restrictions used in the empirical design, including the cross-side trigger, its dominance over fill avoidance, the logistic cancellation hazard, the lead-lag asymmetry with a signed and heterogeneous reverse relation, and the visibility comparative statics.

The restrictions follow the mechanism's sequence. Displayed depth should elicit cross-side trading. Cancellation should respond to that cross-side activity while the display is live. This response should be stronger than the response to fill-direction activity because fill-direction trades threaten execution but do not indicate that the targeted cross-side response has arrived. The cross-to-fill asymmetry should also be much larger during the depth spike than in the matched pre-spike window. The same behavioral state should predict excluded price, depth, variance, spread, timing, and price-impact outcomes in the directions implied by the model. Finally, phantom depth should be less common when displayed liquidity is less informative as a public signal. When hidden and reserve liquidity make the visible book a noisier measure of true trading interest, the display-and-trigger strategy should become less useful.

We test these predictions using TAQ data for S&P 500 securities from 2017 to 2023. The measurement begins with a deliberately mechanical detector that identifies anomalous displayed-depth spikes that collapse within two minutes, without imposing any reversal condition, trade-direction condition, or intent assumption. This separation is central to the design because detection only defines the candidate population, while identification comes from behavior inside the episode and from outcomes excluded from the construction of the behavioral measure.

The first empirical result is that reversal does not identify manipulation. In the full population of mechanically detected collapsed-depth events, the unconditional signed reversal is economically zero. The large reversal obtained from directional-reversal filters is explained by selection on the outcome variable. Moreover, the events most consistent with the display-and-trigger mechanism

are not the events that exhibit the canonical reversal. Episodes with cross-side activity during the displayed-depth spell generate a dislocation that persists, whereas episodes without cross-side activity largely revert. The reversal emphasized in prior work is therefore concentrated in the wrong part of the event population.

The second result is that the displayed depth changes the behavior of other traders before its credibility is resolved. In matched episodes, cross-side trade intensity rises sharply while the depth spike is live relative to the immediately preceding window. This is the first link in the mechanism: the displayed order elicits a response.

The third result is that cancellation is timed to that response. We estimate the Cross-Side Trigger Elasticity, the cancellation response to recent cross-side flow, controlling for fill-direction activity, spreads, volatility, lagged returns, and local trading conditions. The cross-side elasticity is positive and much larger than the fill-direction placebo. This comparison is central. A trader who only wants to avoid execution should respond most strongly to trades that threaten to fill the displayed order. A display-and-trigger strategy predicts the opposite. Cancellation should respond most strongly when the targeted cross-side flow arrives.

The fourth result addresses stale-quote sniping. Cross-side flow can be informative, and a legitimate liquidity provider may cancel after observing it even without any strategic display. We therefore ask whether the cross-to-fill asymmetry changes when the displayed-depth spike is present. During the spike, the asymmetry is about 4.5 times larger than in a matched pre-spike window for the same security, local time, and nearby market conditions. This amplification is the discriminating fact. The cancellation rule changes when displayed depth is live, consistent with a display-and-trigger mechanism rather than a generic response to informative order flow.

The fifth result reinforces the same interpretation. If cancellation were mainly a response to information shocks, the cross-side trigger should be strongest in information-driven episodes. Instead, it is stronger in calm episodes. That pattern is natural if the relevant object is not fundamental news, but the success of the displayed signal in inducing cross-side response.

We then ask whether the same within-episode behavior predicts outcomes that were not used to identify it. We construct a cross-fitted trigger score from cancellation timing variables alone. The score is estimated out of sample using episode-level sample splitting. It excludes post-event prices, post-event depth, reversals, dislocations, variance ratios, spreads, and price-impact estimates by construction. The score is therefore a behavioral summary of display-and-trigger timing, not a classifier trained to reproduce price outcomes.

The trigger score orders excluded outcomes in the directions predicted by the model. Higher-score episodes have larger signed dislocations, sharper depth withdrawal at collapse, lower post-collapse variance ratios, and wider subsequent effective spreads. In the high-score subset, the IV estimate of Kyle lambda during the live phantom-depth window is substantially below its pre-spike benchmark. We interpret this result as corroborating evidence for the depth-inflation channel, since the instrument relies on an exclusion restriction discussed in the appendix. The economic pattern is nevertheless consistent with the model. While phantom depth is live, signed order flow has lower

price impact, as if the market were deeper than it truly is.

We summarize the restrictions with a joint directional concordance test. The main object is the sign pattern, not the magnitude of an aggregate statistic. The six discriminating outcome families excluded from the trigger score all have the predicted sign. These are signed dislocation, variance ratios, effective spreads, depth withdrawal and recovery, and live-window price impact. Under a sign-only null in which each family is equally likely to be positive or negative, the probability of six correct signs is $2^{-6} \approx 0.016$. Because these outcomes are all ordered by the same cross-fitted trigger score, they are not independent. We therefore also report an episode-level bootstrap concordance measure that preserves their dependence. For transparency, we report the full set of seventeen oriented moments, including the first-stage cancellation moments and the lead-lag timing horizons. All seventeen have the model-implied sign. The economic content is that a single behaviorally identified state orders excluded outcomes in the pattern predicted by the model.

The final result studies where the strategy is most prevalent. If phantom depth works by manipulating a visible public signal, its prevalence should depend on the informativeness of the displayed liquidity. We test this prediction using MIDAS measures of hidden-order activity and cancellation intensity. Securities with more hidden-order activity have fewer phantom-depth episodes, both across securities and within securities over time. The cancel-to-trade ratio, by contrast, has little stable explanatory power once visibility is controlled for. This distinction is central to the interpretation. A pure cancellation-technology channel predicts that prevalence should load on cancellation intensity. The data instead point to visibility. Phantom depth is rarer when displayed liquidity is a weaker public signal.

The mid-2020 to early-2021 decline in phantom-depth frequency provides corroborating market-level evidence for the same channel. Per-security break dates bunch around changes in reporting and surveillance infrastructure, including the Consolidated Audit Trail (CAT) implementation phases, and around the JPMorgan spoofing settlement. The downward breaks are concentrated in high-visibility names, whereas upward breaks do not cluster in the same way. Because these events are close in calendar time to COVID normalization, retail-flow changes, enforcement, and market-structure developments, we do not treat the regime change as a stand-alone source of identification. Its role is to show that the same visibility channel that explains cross-sectional and within-security prevalence also organizes the time-series decline.

The paper connects to three literatures. First, it relates to the credibility literature beginning with Bénabou and Laroque (1992). We study a setting in which the signal is an action rather than an announcement, and the action's credibility comes from the possibility of execution. The paper also builds on Kyle (1985) and Rochet and Vila (1994) by embedding manipulable displayed depth in a linear trading environment. It relates to Allen and Gale (1992) by studying manipulation without requiring the manipulator to improve price discovery. Closely related theoretical work by Cartea, Jaimungal, and Wang (2020) studies a spoofer who manipulates limit-order-book imbalance while liquidating a position, trading off the gains from spoofing against expected regulatory penalties. Williams and Skrzypacz (2024) study spoofing in a Bayesian trading environment with endogenous

cancellation. Our point of departure is to model the displayed depth as a partially credible public signal in a Stackelberg timing game and to map the equilibrium into within-episode empirical restrictions.

The paper also contributes to the empirical literature on market manipulation. For surveys and taxonomies of manipulation, see Putniņš (2020). Prior work studies manipulation using suspicious trading patterns, account-level order sequences, prosecuted cases, detection-controlled samples, or surveillance-style classifiers. Lee, Eom, and Park (2013) use account-level order data from Korea to document spoofing sequences involving large displayed orders, opposite-side trading, and cancellation. Comerton-Forde and Putniņš (2014) study closing-price manipulation using prosecuted cases and detection-controlled estimation. Do and Putniņš (2023) hand-collect a global sample of prosecuted layering and spoofing cases and show that combinations of order-book imbalance, high quoting activity, abnormal cancellations, opposite-side trades, and cyclical depth and cancellation patterns detect spoofing with high out-of-sample accuracy. Chen, Hsieh, and Yang (2026) provide recent complementary evidence from Taiwan derivatives markets, showing that aggressive revised limit orders associated with potential spoofing move transaction prices, are followed by short-run reversals and order-imbalance decay, worsen market quality, and respond to a trading rule that restricts aggressive order placement.

Our contribution is to separate detection from identification. We show that price-based signatures can arise mechanically when post-event outcomes are used to define the event sample. We therefore construct the candidate set using only mechanical depth spikes, without using reversal, trade direction, account identity, or intent. The identification then comes from within-episode timing. Reversals, dislocations, depth dynamics, variance ratios, spreads, and price-impact estimates are excluded from the construction of the trigger score and are used only to test the model restrictions. In this sense, the paper complements the existing empirical literature by asking not only whether suspicious orders move prices, but also how to identify the strategic state before the post-event outcome is realized.

The paper also makes a methodological contribution by using cross-fitting and sign restrictions to test a latent strategic state. The econometric machinery is deliberately secondary to economics. Cross-fitting prevents the trigger score from being trained on the outcomes it is later used to predict. The joint sign test asks whether one behaviorally identified state orders several excluded outcomes in the directions implied by the model. The identifying power comes from the joint structure of the mechanism: the displayed depth elicits response, cancellation follows that response, excluded outcomes move in the predicted directions, and prevalence declines when the displayed liquidity becomes a weaker public signal.

The plan of the paper is as follows. Section 2 presents the model. Section 3 states the structural restrictions. Section 4 describes the data and the mechanical detection of phantom-depth episodes. Section 5 shows why reversal-based identification fails and develops the behavioral identification of the trigger score. Section 6 reports the overidentifying tests. Section 7 studies external validity and the visibility channel. Section 8 concludes.

2 A Model of Manipulable Market Signals

This section develops a model in which a strategic trader inflates displayed depth to manipulate the pricing rule. The model generates the behavioral predictions that the empirical tests evaluate, including the cross-side trigger, timing asymmetry, depth inflation, and the prediction that phantom depth is less prevalent when displayed liquidity is a weaker public signal. The derivations and proofs are in Appendix A.

2.1 Environment

We consider a one-period economy with a single risky asset whose fundamental value v is normally distributed,

$$v \sim \mathcal{N}(0, \sigma_v^2).$$

The intertemporal discount rate is set to zero. Four types of agents populate the market, each chosen to represent a distinct economic role observed in modern electronic markets.

Competitive market makers (MM) are risk-neutral and post a pricing rule that is informationally efficient with respect to total order flow. They infer the asset value from observed aggregate order flow and compete to earn zero expected profits. The reduced-form representation captures algorithmic liquidity provision that updates quotes in real time in response to the realized order flow.

Liquidity traders are risk-neutral, uninformed traders who submit aggregate market orders $u_t \sim \mathcal{N}(0, \sigma_u^2)$ at $t \in \{0, 1\}$, uncorrelated across time. These traders represent investors with exogenous trading motives, portfolio rebalancing, hedging, or fund flows, whose order flow is uninformative about fundamentals but is the source of noise that allows strategic traders to hide.

The perfectly informed strategic nonmanipulator (N) is a risk-neutral trader who observes the realization of v exactly and submits a demand schedule x_N. This trader represents participants with a strong informational advantage, proprietary trading desks, or sophisticated hedge funds, who trade on private information while taking into account the price impact of their trades and contribute to price discovery. Perfect information is an abstraction that places the nonmanipulator at the informational ceiling of the model. She is a useful benchmark because her optimal trading intensity isolates the residual information rents in the presence of the manipulator.

The strategic manipulator (M) is a privately informed trader who observes a noisy signal

$$\xi = v + \sigma_\xi \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, 1), \quad \varepsilon \perp v, \quad \varepsilon \perp u_t, \tag{1}$$

and, following Rochet and Vila (1994), the contemporaneous aggregate liquidity demand u_t. The manipulator maximizes the expected mean-variance utility of terminal wealth w,

$$\overline{U}(w) = \mathbb{E}[w \mid \Omega] - \tfrac{\alpha}{2}\mathrm{Var}[w \mid \Omega], \tag{2}$$

with risk aversion $\alpha > 0$ and information set Ω specified below. The manipulator's distinguishing

feature is the ability to submit, and possibly withdraw, a visible order at the top of the order book. We refer to her displayed order as the public signal. The signal is manipulative when it is submitted in part for its effect on the beliefs and actions of other agents rather than purely for execution.

The four agents capture the economic logic of the order book. Market makers price, the nonmanipulator informs, the liquidity traders provide noise, and the manipulator chooses how transparently to act on partial information. The distinction between a nonmanipulator and a manipulator is the central conceptual move. The nonmanipulator trades on information that improves prices, while the manipulator trades partly on information and partly on the impact of her displayed order on the pricing rule and on the responses of others. The former is welfare-improving; the latter is the object of regulatory concern.

The nonmanipulator is essential because phantom depth is not simply posting and canceling an order. It is the use of a visible order to change the strategic environment before the order is verified. Without another informed trader, the manipulator could still trade on private information and could still benefit from the option to cancel, but displayed depth would no longer operate as a manipulable public signal. The role of the nonmanipulator is to make this channel explicit. Her demand responds to the pricing rule and to the informational content of aggregate order flow. When the manipulator inflates displayed depth, the market appears deeper, the price-impact coefficient changes, and the nonmanipulator's trading intensity is distorted. The displayed order therefore affects prices not only through the manipulator's own demand but also through its effect on another strategic trader's response and on the amount of information incorporated into prices. In this sense, the nonmanipulator is not an auxiliary agent. She is the channel through which visible liquidity becomes payoff relevant and through which phantom depth crowds out price discovery. Without this channel, the model would collapse into informed trading with a cancellation option, rather than manipulation through strategic liquidity display.

2.2 Timing and Stackelberg structure

Time is discrete with two trading rounds at $t = 1$ and $t = 2$ and five event subperiods $t = 0^-, 0, 1^-, 1, 2$. Figure 1 summarizes the timing.

At time $t = 0$, the manipulator privately observes her signal ξ and the realized liquidity demand u_0 at $t = 0^-$. At $t = 0$, the liquidity traders submit u_0, the nonmanipulator observes v and submits x_N, and the manipulator submits her demand schedule $x_{M,0}$. MMs observe the total order y_0 and post the demand schedule. No trade occurs at $t = 0$.

In subsequent rounds, at $t = 1^-$, the manipulator cancels her $t = 0$ order with exogenous probability q. If cancellation fails (probability $1 - q$), the original order $x_{M,0}$ executes at $t = 1$, and the manipulator does not trade at $t = 2$. If cancellation is successful (probability q), the manipulator observes the second-round liquidity demand u_1 and submits a revised order $x_{M,1}$ to be executed at $t = 2$. The parameter q summarizes the effective manipulation technology, with a higher q corresponding to a better cancellation infrastructure, lower latency, and faster information processing.

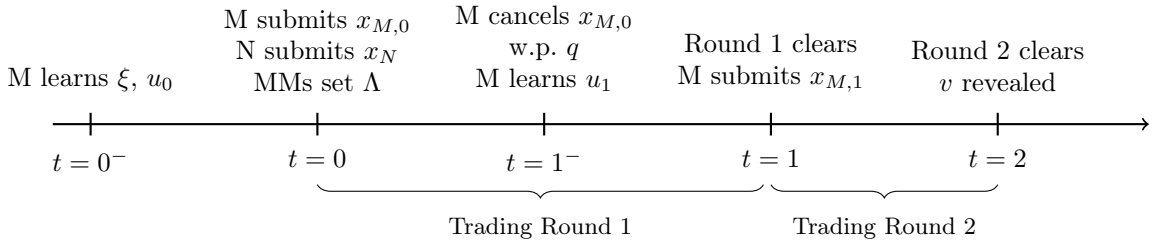


Figure 1: Timeline of the manipulation game.

This figure illustrates the five-event timeline of the manipulation game with one nonmanipulator and one manipulator across two trading rounds. At $t = 0^-$ the manipulator (M) observes her private signal ξ and the contemporaneous liquidity order u_0. At $t = 0$ M submits a demand schedule $x_{M,0}$, the nonmanipulator (N) submits x_N, and market makers (MM) set the inverse-depth coefficient Λ. At $t = 1^-$ the manipulator cancels her displayed order with exogenous probability q. At $t = 1$ the first round clears. If cancellation succeeds, M submits a revised order $x_{M,1}$, otherwise, the original order executes. At $t = 2$ the second round clears, and the fundamental value v is publicly revealed.

We assume that market makers cannot revise the pricing rule Λ in response to the manipulator's revision at $t = 1$. In the absence of the manipulator, holding Λ fixed across both rounds is consistent with zero expected profits because market makers face only liquidity traders at $t = 2$. The pricing rule is set once at $t = 0$ and applied to whichever realization of the manipulator's order ultimately clears.

The crucial timing assumption is that the manipulator is the Stackelberg leader. She commits to a demand schedule $x_{M,0}$ that the market makers and the nonmanipulator observe before they respond, and the leader internalizes the impact of her schedule on their reactions. The equilibrium concept is subgame-perfect Stackelberg rather than Bayesian Nash, matched to the institutional reality of the limit order book, where a visible order sits while other participants observe and react before the manipulator decides whether to withdraw.

The model is dynamic within an episode but static across episodes. This is deliberate. The empirical object is a short displayed-depth spell in which a visible order appears, other traders respond, and the displayed component is withdrawn. The data identify this local sequence at subsecond horizons. We therefore model the episode as the relevant strategic unit rather than embedding it in a repeated dynamic program across many possible spoofing attempts. Cross-episode forces surely matter, including surveillance, learning, and changes in enforcement risk, but we treat them as market-level state variables in Section 7 rather than as part of the within-episode game.

This modeling choice also clarifies the relation to Williams and Skrzypacz (2024). They study spoofing in a dynamic trading equilibrium with sequential order submission, cancellation, and later trading. Their model shows that spoofing can arise in equilibrium, slows price discovery, raises spreads and volatility, and has prevalence that varies with market conditions. Our model is dynamic in a different, more local sense. We focus on the timing inside a single displayed-depth episode. A trader displays depth, other traders respond while the signal is live, and the trader retains the option to withdraw before the signal is verified. We do not model a spoofer solving a repeated

dynamic program across episodes. Instead, the cross-episode environment enters through reduced-form market-level objects such as cancellation feasibility, signal visibility, surveillance technology, and enforcement risk. This keeps the model aligned with the empirical design, which identifies the strategic state from within-episode timing and then studies market-level variation separately.

2.3 Linear strategies and the pricing rule

Following Kyle (1985) and Rochet and Vila (1994), we restrict our attention to linear strategies. The manipulator's demand schedules in the two rounds are

$$x_{M,t} = \beta_{M,t}\,\xi + \gamma_{M,t}\,u_t, \qquad t \in \{0,1\}, \tag{3}$$

where $\beta_{M,t} > 0$ is the trading intensity on her private signal and $\gamma_{M,t}$ is the response to the contemporaneous liquidity order. The nonmanipulator submits

$$x_N = \beta_N\,v. \tag{4}$$

Market makers observe the total order flow $y_0 = x_N + x_{M,0} + u_0$ and post the linear price schedule[1]

$$P(y) = \mathbb{E}[v \mid y] = \Lambda y, \tag{5}$$

where $\Lambda > 0$ is the inverse market depth, analogous to Kyle (1985). All agents condition on this functional form at the outset.

Under the linear-quadratic structure of the model, optimization over the functional form reduces to optimization over the slope and intercept coefficients. The formal justification is that averaging the manipulator's expected utility over the realizations of (v, u_t) and optimizing the resulting coefficient-level objective recovers the correct first-order conditions, including the induced response of Λ, which is the "double expectation" argument of Rochet and Vila (1994), the details of which are deferred to Appendix A.

It is analytically convenient to work in dimensionless form. With $\beta_K \equiv \sigma_u/\sigma_v$ the Kyle benchmark intensity, define

$$\widehat{\beta}_{i,t} \equiv \beta_{i,t}\sigma_v/\sigma_u, \quad \widehat{\Lambda} \equiv \Lambda\sigma_u/\sigma_v, \quad a \equiv \alpha\sigma_u\sigma_v, \quad \tau \equiv \sigma_\xi^2/\sigma_v^2.$$

The product $a\tau$ scales the manipulator's inventory risk costs due to the imprecise signal, and the equilibrium depends on the primitives only through $(a\tau, q)$ and the auxiliary τ. The parameter τ is the noise-to-signal variance ratio of the manipulator's information,[2] so that $\tau = 0$ corresponds to a perfect signal and $a\tau \to 0$ to a manipulator who either bears no risk or faces none.

[1] The market makers also take into account that the initial manipulator's order may be canceled with a known probability $q \in [0;1]$.

[2] With $\xi = v + \sigma_\xi\varepsilon$ as in (1), $\mathrm{Var}(\widehat{\xi}) = 1+\tau$, $\mathbb{E}[\widehat{v} \mid \widehat{\xi}] = \widehat{\xi}/(1+\tau)$, and $\mathrm{Var}[\widehat{v} \mid \widehat{\xi}] = \tau/(1+\tau)$.

Market makers observe y_0 but do not know whether the manipulator's order will be canceled. Under Stackelberg they also observe the strategy parameters $(\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0})$ and set the informationally efficient slope $\widehat{\Lambda} = \mathrm{Cov}[\widehat{v}, \widehat{y}]/\mathrm{Var}[\widehat{y}]$.[3] The nonmanipulator's first-order condition yields the best response

$$\widehat{\beta}_N = \frac{1 - (1-q)\widehat{\Lambda}\widehat{\beta}_{M,0}}{2\widehat{\Lambda}}, \tag{6}$$

which collapses to the Kyle benchmark $1/(2\widehat{\Lambda})$ when $q = 1$ or $\widehat{\beta}_{M,0} = 0$, and is otherwise depressed by the manipulator's expected executed component. Substituting into the market makers' reaction function yields the inverse market depth as a function of the manipulator's strategy alone,

$$\widehat{\Lambda} = \frac{1}{2\sqrt{\varkappa\,\widehat{\beta}_{M,0}^2 + \overline{(1+\widehat{\gamma}_M)^2}}}, \qquad \varkappa \equiv (1-q)\big(1 + \tau - \tfrac{3}{4}(1-q)\big), \tag{7}$$

with $\overline{(1+\widehat{\gamma}_M)^2} \equiv (1-q)(1+\widehat{\gamma}_{M,0})^2 + q$. The coefficient $\varkappa$ encodes the joint effect of cancellation feasibility and signal imprecision, decreasing in q and increasing in τ. The manipulator can drive $\widehat{\Lambda}$ down by inflating $\widehat{\beta}_{M,0}^2$ in the denominator of (7). This is the phantom-depth mechanism in its simplest form. Appendix A contains the full derivation, including the two-scenario representation of $\widehat{y}$ and the explicit form of the reaction function.

2.4 Equilibrium and phantom depth

The manipulator's two-period problem is solved by backward induction. In the second round, the manipulator takes $\widehat{\Lambda}$ as given (the market makers cannot revise the pricing rule) and chooses $x_{M,1}$ to maximize her mean-variance utility. The first-order condition yields

$$\widehat{\beta}_{M,1}^* = \frac{1}{(1+\tau)(2\widehat{\Lambda} + a\tau)}, \qquad \widehat{\gamma}_{M,1}^* = -\frac{\widehat{\Lambda}}{2\widehat{\Lambda} + a\tau}. \tag{8}$$

The revised trading intensity $\widehat{\beta}_{M,1}^*$ is scaled down by the price-impact term $2\widehat{\Lambda}$ and the risk cost $a\tau$. The intercept $\widehat{\gamma}_{M,1}^*$ is negative, capturing the standard Rochet and Vila (1994) effect that an informed trader who observes the contemporaneous liquidity demand optimally trades against it. Substituting (8) into the manipulator's expected utility yields the indirect utility in the second round.

$$\overline{U}_2^* = \frac{\left(\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}\widehat{u}_1\right)^2}{4\left(\widehat{\Lambda} + \frac{a}{2}\tau\right)},$$

[3] The pricing rule is the best linear projection of value on order flow. The Internet Appendix compares this rule with the exact Bayesian posterior mean under the latent cancellation mixture and shows that the nonlinear residual is small in the moderate parameter region used for the numerical calculations; it also discusses the behavioral reading of the linear restriction. The within-episode trigger tests require only that the pricing rule be locally sticky inside the episode (Assumption 5), not exact linearity.

which is the continuation value that the manipulator carries into the first-round problem. The first-round problem is to choose $(\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0})$ to maximize the probability-weighted sum $(1-q)\overline{U}_1 + q\,\mathbb{E}[\overline{U}_2^*]$, where $\overline{U}_1$ is the mean-variance utility of the original order. The first-order conditions yield a system that has a closed-form expansion in the small-execution-risk limit $a\tau \ll 1$. The full derivation is in Appendix A. Here we state the asymptotic equilibrium.

Proposition 1 (Phantom depth) *In the small-execution-risk limit $a\tau \to 0$, on the bluffing branch — the equilibrium regime when the signal is precise enough for the displayed intensity to be worth inflating (Figure 2) — the equilibrium is characterized by the following four properties.*

(i) The manipulator's displayed trading intensity grows at rate

$$\widehat{\beta}_{M,0}^* = O\big((a\tau)^{-1}\big)\,.$$

The displayed order diverges as risk aversion or signal noise vanish.

(ii) The equilibrium inverse market depth satisfies

$$\widehat{\Lambda}^* = \frac{1}{2\sqrt{\varkappa}\,\widehat{\beta}_{M,0}^*}\,(1 + O(a\tau)) = O(a\tau), \tag{9}$$

so the market appears arbitrarily deep, $\widehat{\Lambda}^* \to 0$, as $a\tau \to 0$.

(iii) $\widehat{\Lambda}^*$ is increasing in a: greater risk aversion of the manipulator partially restores the price impact toward the no-withdrawal benchmark.

(iv) Phantom depth vanishes at both endpoints of the cancellation technology: at $q = 1$ market makers rationally ignore the never-executing displayed order and $\widehat{\Lambda}^* = 1/2$, the Kyle (1985) benchmark, exactly; at $q = 0$ the withdrawal option disappears, the bluffing component of the strategy shuts down, and $\widehat{\Lambda}^*$ is the $O(1)$ value of the two-informed-trader variant discussed in Section 2.5.

Two boundary points are useful for interpreting Proposition 1. First, item (iv)'s endpoints bracket the mechanism without implying a monotone comparative static in between: a higher cancellation probability strengthens the bluffing motive — the continuation weight is proportional to q, which raises $\widehat{\beta}_{M,0}^*$ and deepens the apparent market through $\varkappa\widehat{\beta}_{M,0}^2$ in (7) — but it also lowers the credibility weight $\varkappa$ itself, which works in the opposite direction. The net sign of $\partial\widehat{\Lambda}^*/\partial q$ changes over the interior, so Λ^* is not monotonic in q. The equilibrium is discontinuous at $q = 1$. When cancellation is certain, the displayed order never executes, so market makers rationally ignore it. The executed first-round order flow is then $\widehat{u}_0 + \widehat{\beta}_N\widehat{v}$, regardless of the manipulator's schedule, and $\widehat{\Lambda}^*$ returns to the Kyle-type benchmark $1/2$. For any $q < 1$, by contrast, the displayed order may execute with positive probability. That small execution risk disciplines beliefs and allows the bluffing component of the strategy to depress $\widehat{\Lambda}^*$. The discontinuity is economically meaningful.

Displayed depth affects the pricing rule only because it may have to stand behind trade. A purely cheap-talk display would be ignored.

Second, the small-$a\tau$ asymptotics in items (i) and (ii) are taken at fixed $q < 1$. They are not uniform as $q \to 1$. The two limits therefore do not commute. Taking $a\tau \to 0$ first at fixed $q < 1$ preserves a positive execution risk and generates phantom depth. Taking $q \to 1$ first eliminates execution risk, removes the credibility of the displayed order, and returns the market to the no-execution display benchmark. The empirical counterpart of q is also interior. Section 2.6 derives an equilibrium $q^\dagger \in (0, 1)$.

The following mechanism is behind the results of Proposition 1. The continuation value $\mathbb{E}[\overline{U}_2^*]$ is approximately $q/(4\widehat{\Lambda})$ when $a\tau$ is small, and $\widehat{\Lambda}$ is decreasing in $\widehat{\beta}_{M,0}$ through (7). The manipulator can therefore amplify her continuation value by submitting ever-larger displayed orders that drive $\widehat{\Lambda}$ toward zero. The only restraining force is the risk cost $\frac{a}{2}\tau\widehat{\beta}_{M,0}^2(1+\tau)$. The first-order condition balances these two forces, yielding $\widehat{\beta}_{M,0}^* \propto 1/(a\tau)$. The market appears arbitrarily deep as $a\tau \to 0$ because the manipulator inflates the visible component of order flow without bound, but the apparent depth is supported by an order she intends to cancel. This gives the model its name.

The closed-form expressions also have a simple visual interpretation. Figure 2 compares the asymptotic solution with the numerical solution for the manipulator's optimal first-round strategy. The approximation is nearly indistinguishable from the numerical solution over the plotted range. As the manipulator's signal becomes more precise, the optimal displayed intensity rises sharply, while the intercept component rises and then flattens. The figure therefore illustrates the two forces in the equilibrium: more precise information supports more aggressive displayed depth, and the option to withdraw sustains a separate bluffing component.

The figure makes clear that the model is not merely generating a technical existence result, but produces a concrete economic sequence. The manipulator displays more aggressively when private information is more precise, but the displayed order also contains a component whose value comes from being visible before it is verified. We now describe this sequence and the empirical restrictions it implies.

2.5 Economic mechanism and interpretation

The economic logic follows directly from the timing of the game. The manipulator first creates a visible signal by submitting the order $x_{M,0}$. Market makers observe the enlarged displayed book and set a lower inverse-depth coefficient $\widehat{\Lambda}$ through (7). The market therefore appears deeper than it truly is. Other traders respond to that apparent liquidity. The nonmanipulator reduces her own trading intensity in (6), because the displayed order changes the informational content of order flow, and aggressive traders on the opposite side are induced to trade. Once that response has arrived, the manipulator exercises the cancellation option. With probability q, she withdraws the displayed order and submits the revised order $x_{M,1}$, which trades at the pricing rule chosen while the phantom depth was still visible.

This sequence is the core of the mechanism. The manipulator does not need the displayed order

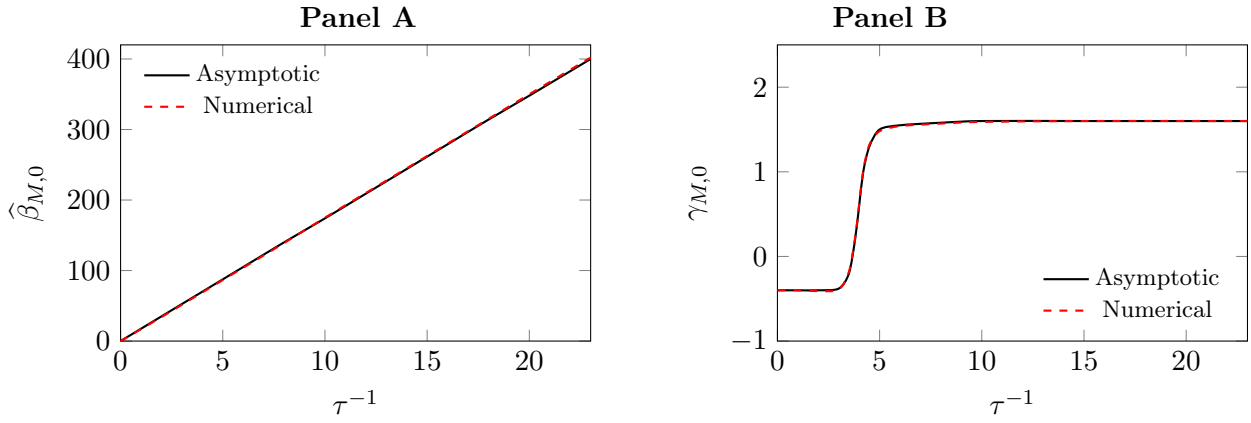


Figure 2: Manipulator's optimal first-round strategy

This figure plots the manipulator's optimal first-round displayed intensity $\widehat{\beta}^*_{M,0}$ and intercept component $\gamma^*_{M,0}$ as functions of signal precision τ^{-1}, the inverse of the normalized signal-noise ratio τ defined in Section 2.3, for the baseline parameterization used in Proposition 1. Solid lines show the closed-form asymptotic approximation, and dashed lines show the numerical solution. The two series are nearly indistinguishable over the plotted range. Panel A shows that the optimal displayed intensity rises approximately linearly with signal precision. Panel B shows that the intercept component rises sharply once the signal is sufficiently precise and then flattens. The figure illustrates the comparative statics behind the model that more precise private information supports more aggressive displayed depth, while the withdrawal option sustains a separate bluffing component of the strategy.

to execute. She needs it to be believed long enough to affect behavior. The value of the strategy comes from the gap between observation and verification. Other traders observe the displayed depth and react before learning whether the order represents genuine trading interest. The manipulator then withdraws the signal after it has done its work.

This also clarifies what the empirical signature of manipulation should be. Successful manipulation need not generate a large unconditional reversal after the collapse of displayed depth. The manipulator's payoff is earned while the signal is live, through the dislocation created by the display and the response it elicits. Any later reversal is observed after the manipulator has exited. It reflects the unwinding of the displayed component, the revised informed trade, ordinary liquidity provision, and subsequent information arrival. The more direct behavioral signature is therefore not reversal. It is the withdrawal of the visible order after the targeted cross-side response has arrived.

The mechanism also explains why phantom depth is socially costly. As the manipulator increases $\widehat{\beta}_{M,0}$, the nonmanipulator's optimal trading intensity $\widehat{\beta}_N$ falls,

$$\frac{\partial \widehat{\beta}_N}{\partial \widehat{\beta}_{M,0}} < 0 \quad \text{for any } q < 1.$$

Displayed liquidity rises, but informed trading is crowded out. The book looks deeper, yet prices become less informative. This is the central welfare distortion. The market appears more liquid precisely because part of the displayed liquidity is not meant to stand behind trade.[4]

[4]The crowding-out statement is relative to the apparent depth of the market. From (6), $\widehat{\beta}_N = \{1 - (1 -$

Signal precision matters for this welfare calculation. A manipulator with a weak signal can be especially distortionary, because she adds visible depth without adding much information to prices. A better-informed trader is still strategic, but her order flow is more closely connected to fundamentals. For this reason, observed prevalence alone is not a sufficient statistic for welfare loss, as shown in Figure 2. The social cost depends not only on how often phantom depth appears, but also on how much information is displaced when it appears.

Liquidity traders bear these costs in two ways. If the manipulator's first order executes, which happens with probability $1 - q$, liquidity traders face the usual adverse-selection cost of trading against informed demand. If the order is cancelled, they face a different cost. They trade under a pricing rule that was set for a market in which the displayed order was expected to execute, but the manipulator instead withdraws the signal and trades against the second-round liquidity shock. The loss is tied to the option to withdraw the public signal. In the model, aggregate loss is increasing and convex in q, which gives the efficiency rationale for anti-spoofing regulation.

The most delicate modeling assumption is that market makers do not fully revise Λ after the cancellation at $t = 1^-$. This assumption should be read as a reduced-form description of sluggish adjustment, not as a literal claim that quotes never update. The relevant horizon is very short. SIP consolidation latency, venue fragmentation, and message-processing delays mean that the effective pricing rule faced by traders may not adjust instantaneously to a cancellation inside the same sub-second event window. Institutionally, the displayed order sits in the book while other participants observe and react. The pricing rule governing that reaction is formed against the visible depth, not against the counterfactual book that would exist after the order is withdrawn.

The empirical strategy does not require the pricing rule to be literally fixed. What matters for identification is the sequence of display, response, and withdrawal. The tests in Section 5 use within-episode timing patterns, including the cross-side trigger, the cross-to-fill comparison, and the lead-lag asymmetry. These patterns distinguish display-and-trigger behavior from ordinary liquidity provision, whether Λ is fixed or merely adjusts with delay. The model gives the economic structure that organizes the tests; the identification does not require the model to be a literal description of every quote update.

The substantive assumption is the Stackelberg timing. A visible order is posted before other participants respond. That timing captures the institutional fact that displayed depth is observable before it is executed. It differs from the Bayesian Nash timing in Williams and Skrzypacz (2024), which studies a related spoofing environment with overlapping but distinct predictions. The remaining choices are mainly for tractability. Linear strategies and the normalized parameter $a\tau$ deliver a

$q)\widehat{\Lambda}\widehat{\beta}_{M,0}\}/(2\widehat{\Lambda})$. Holding the pricing rule fixed, a larger displayed loading $\widehat{\beta}_{M,0}$ lowers $\widehat{\beta}_N$. In the small-execution-risk limit, however, the level of $\widehat{\beta}_N$ diverges because $\widehat{\Lambda}^* \to 0$ while $\widehat{\Lambda}^*\widehat{\beta}^*_{M,0} \to 1/(2\sqrt{\varkappa})$. Hence $\widehat{\beta}_N \approx \{1 - (1-q)/(2\sqrt{\varkappa})\}/(2\widehat{\Lambda}^*) \to \infty$. The bracket is positive because $2\sqrt{\varkappa} > 1 - q$, which is equivalent to $\tau + q > 0$. This does not contradict crowding out. A risk-neutral nonmanipulator is constrained only by price impact, so her absolute trade size grows as the market appears deeper. The relevant benchmark is the Kyle-type intensity $1/(2\widehat{\Lambda})$ that she would choose as the sole informed trader at the same apparent depth. Relative to that benchmark, $\widehat{\beta}_N/\{1/(2\widehat{\Lambda})\} = 1 - (1-q)\widehat{\Lambda}\widehat{\beta}_{M,0} < 1$, and this ratio falls as the bluffing component grows. Crowding out therefore means less informed trading per unit of apparent depth, not necessarily a smaller absolute informed order.

closed-form characterization in the small-execution-risk limit and are internally consistent with the Stackelberg first-order conditions by the double-expectation argument in Appendix A.

Several extensions would change magnitudes without changing the main economic logic. Multiple manipulators would reduce the displayed intensity per manipulator. Dynamic learning by market makers would erode the bluffing channel over time. Heterogeneity in cancellation speed would change the effective value of the option to withdraw. In empirical tests, we therefore treat q as a stock-level reduced-form technology rather than as a literal primitive. The limiting case is useful. When $q = 0$, the model collapses to a variant of Kyle (1985) with two heterogeneously informed strategic traders. The entire manipulation channel comes from the option to withdraw the public signal.

The model therefore delivers a sequence of empirical implications. Phantom depth should elicit a cross-side response. Cancellation should be timed to that response. The response should be stronger than ordinary fill avoidance. High-trigger episodes should generate live dislocations and sharper depth withdrawal, but need not generate large post-collapse reversals. Finally, the strategy should be less useful when displayed liquidity is a weaker public signal. Section 3 turns these implications into structural restrictions that can be taken directly to the data.

The model is also intentionally narrow about the form of manipulation. It is a model of displayed-depth spoofing, not a model of every strategy that can distort prices. Other forms of manipulation may use layered orders across several price levels, cross-venue order placement, hidden or iceberg liquidity, coordinated trading across related assets, quote stuffing, momentum ignition, or slower sequences of trades and cancellations. Those strategies may have different timing signatures and different welfare implications. Our object is the class of episodes in which a large visible order changes behavior before execution and is then withdrawn. The empirical tests should be read with the same scope. They identify a display-and-trigger state inside mechanically detected collapsed-depth episodes, rather than a complete taxonomy of market manipulation.

2.6 Endogenous cancellation timing

The baseline model treats the cancellation probability q as an exogenous technology. This assumption makes the Stackelberg equilibrium tractable, but it leaves open the timing of cancellation, which is central to the empirical design. The cross-side trigger, the cross-to-fill comparison, and the lead-lag asymmetry are all statements about when the manipulator withdraws the displayed order. An exogenous cancellation probability cannot speak to those timing restrictions. This subsection endogenizes the cancellation decision and shows that the behavioral restrictions in Section 3 follow from the extended model. The display-response-withdrawal mechanism is therefore an equilibrium implication once the manipulator chooses when to cancel. Appendix B contains the derivations, proofs, and formal assumptions. Here we summarize the structure of the extension and its main results.

The timeline acquires an interim subperiod, $t = 0^+$, between order display and the cancellation decision. While the displayed order of size $\widehat{D} > 0$ is live, cross-side market-order flow contains

a reactive component $\nu\theta\widehat{D}$, where $\nu \in [0,1]$ is the visibility of displayed liquidity for the security and $\theta \sim \mathcal{N}(\bar{\theta}, \sigma_\theta^2)$, $\bar{\theta} > 0$, is the realized responsiveness of depth-conditioning traders, unknown to all players at $t = 0$; the reactive component stops accruing once the display is removed, and a fraction $\psi \in (0,1]$ of it persists into the next round, where the manipulator can trade against it. At $t = 1^-$ the manipulator observes noisy interim signals of the elicited response and of the fill pressure threatening execution,

$$\phi_c = \nu\theta\widehat{D} + e_c, \qquad \phi_f = \mu\,\zeta + e_f, \tag{10}$$

with ζ a latent execution-risk shock, and decides whether to cancel — withdrawing the entire first-round schedule, as in the baseline timeline — paying a cost $c+\varepsilon_L$, where $c \geq 0$ bundles queue-priority, message, and enforcement costs, and ε_L is an idiosyncratic latency shock with a logistic distribution of scale s. The configuration of interest is the canonical bluff where the manipulator's value estimate m is below the prior while the display advertises buying interest, so that elicited cross-side buying is precisely what she profits from selling into. Appendix B.6 shows this configuration arises optimally when visibility is high enough, rather than by assumption. Cancellation is then an optimal stopping decision given the interim posteriors, and its structure delivers the following results.

Proposition 2 (Logistic trigger hazard) *Under Assumptions 1–7, the probability of cancellation at $t = 1^-$ given the interim signals is*

$$\Pr\left(C = 1 \mid \phi_c, \phi_f\right) = \Lambda_C\left(\frac{\Delta\left(\phi_c, \phi_f\right)}{s}\right), \tag{11}$$

where Λ_C is the logistic link and Δ is the cancellation surplus derived in Appendix B.4, which is increasing in the posterior elicited response and in fill pressure. In the display-and-trigger configuration,

$$\text{CSTE} \equiv \frac{\partial \Pr\left(C = 1 \mid \cdot\right)}{\partial \phi_c} > 0, \qquad \text{FDE} \equiv \frac{\partial \Pr\left(C = 1 \mid \cdot\right)}{\partial \phi_f} > 0,$$

and CSTE $>$ FDE *if and only if the trigger condition (C2) of Appendix B.4 holds: the elicited response is sufficiently visible (ν), persistent (ψ), and informative, relative to the picking-off risk parameter that governs ordinary fill avoidance. Condition (C2) is exactly the boundary between the display-and-trigger strategy and ordinary liquidity provision.*

Proposition 2 derives the cross-side trigger in Restriction 1 from the manipulator's cancellation problem. It also gives the empirical hazard a direct structural interpretation. The logistic link is not an ad hoc specification. It is induced by the distribution of the latency shock. The episode fixed effect α_i absorbs the episode-level primitives that are fixed within an episode but vary across episodes, including the signal realization m_i, the displayed size $\widehat{D}_i$, the executed non-displayed component of the schedule, and the cancellation friction c.

The proposition also clarifies the role of fill-direction flow. The model does not predict FDE $= 0$. Even a manipulator using a display-and-trigger strategy still faces picking-off risk and therefore manages fill-direction pressure like any liquidity provider. The distinctive prediction is instead the

ordering CSTE $>$ FDE $>$ 0. Cancellation responds to both kinds of flow, but it responds more strongly to the cross-side activity that indicates the displayed signal has achieved its purpose.

Corollary 1 (Spike-conditional amplification) *The trigger component of the cross-side cancellation sensitivity exists only while a display is live: absent a displayed component ($\widehat{D} = 0$), cross-side flow is uninformative about the elicited response and the model-implied cross-to-fill gap is zero, so any pre-spike gap measures generic information-based repricing, and the within-episode gap must exceed it.*

Corollary 1 provides the model counterpart to the comparison between the in-spike and pre-spike cross-to-fill asymmetries. The amplification comes from the informational content of cross-side flow while the display is live. Cross-side trades are especially informative during the spike because they reveal that the displayed order has elicited the targeted response. Before the spike, the same flow may still matter for cancellation, but it does not carry the same information about the success of the displayed signal. The comparison therefore isolates the display-and-trigger channel rather than the mere presence of cancellable depth.

Proposition 3 (Timing asymmetry and the reverse lead) *Under the same assumptions: (i) cross-side flow observed at t raises the cancellation hazard at subsequent decision instants while the display is live; and (ii) cancellation does not raise future cross-side flow — conditional on the interim information, the difference in expected forward cross-side flow between cancellation and non-cancellation instants equals*

$$-\nu\widehat{D}\,[\kappa(t+h)-\kappa(t)]\,\widehat{\theta}_t,$$

where κ is the within-spell accrual profile of the reactive flow and $\widehat{\theta}_t$ is the interim posterior mean of the responsiveness; the difference is weakly negative on the event $\{\widehat{\theta}_t \geq 0\}$, which includes the mean posterior and all display-and-trigger states. The reverse lead is therefore weakly negative on average — unconditionally so under a mild lower bound on the latency dispersion given in Appendix B.5 — and more negative where the display-and-trigger state is stronger*, because the posterior $\widehat{\theta}_t$ is largest exactly in the episodes the trigger score identifies.*

Proposition 3 gives the model foundation for the lead-lag restriction. The first part delivers the forward timing prediction. Cross-side flow observed while the display is live raises the subsequent cancellation hazard because it signals that the displayed order has elicited the targeted response. The second part describes the reverse timing relation. Once cancellation occurs, the displayed order is no longer present to elicit additional cross-side flow. Conditional on the information available at the cancellation instant, expected future cross-side flow is therefore weakly lower after cancellation than after non-cancellation.

The signed expression in the proposition makes this intuition precise. The reverse effect is proportional to $-\nu\widehat{D}[\kappa(t+h)-\kappa(t)]\widehat{\theta}_t$. It is weakly negative because the remaining accrual of reactive flow, $\kappa(t+h)-\kappa(t)$, is nonnegative, displayed depth $\widehat{D}$ is positive, and the posterior

responsiveness $\widehat{\theta}_t$ is positive in display-and-trigger episodes. The effect is also heterogeneous. It is more negative when the display is more informative, when displayed depth is larger, and when the interim posterior assigns greater responsiveness to other traders. The model therefore predicts not only that cross-side flow leads cancellation, but also that cancellation should not symmetrically lead more cross-side flow. If anything, cancellation should reduce the future cross-side response, especially in episodes where the display-and-trigger state is strongest.

This is a useful sharpening of Restriction 2. The baseline restriction required the reverse relation to be approximately zero after conditioning on local information. The endogenous-cancellation model gives a stronger prediction. The reverse relation should be weakly negative on average, and the negative relation should be concentrated in high-trigger episodes. The asymmetry follows directly from the economics of the display. Cross-side flow can trigger withdrawal because it arrives while the signal is live. Withdrawal cannot create the same response afterward because it removes the signal that generated the response in the first place.

Proposition 4 (Equilibrium cancellation and the two market-level margins) *The aggregate cancellation probability is an equilibrium object: there exists $q^\dagger \in (0,1)$ such that market makers' pricing rule evaluated at $q^\dagger$ and the manipulator's optimal stopping rule are mutually consistent, and all baseline results hold verbatim with $q = q^\dagger$. At a stable interior equilibrium: (i) $q^\dagger$ is decreasing in the cancellation friction c and, provided the latency dispersion s is not too small (an explicit moment bound in Appendix B.6), increasing in visibility ν; (ii) the marginal value of posting a bluff display is increasing in ν and decreasing in c, and, when the posting cost is sufficiently convex, the optimal display is strictly positive if and only if ν exceeds a threshold $\nu^*(m,c)$ that falls with the strength of the trader's directional motive and rises with the friction. Integrating over the distribution of episode-level primitives, the* prevalence *of display-and-trigger episodes is a smooth, increasing function of visibility and a decreasing function of enforcement intensity.*

Proposition 4 turns the reduced-form visibility argument into equilibrium comparative statics. The aggregate cancellation probability is not a primitive. It is the fixed point between the pricing rule chosen by market makers and the manipulator's optimal stopping rule. At an interior fixed point, a more visible display makes cancellation more valuable because cross-side flow becomes more informative about whether the display has elicited the targeted response. A higher cancellation friction works in the opposite direction. The same forces determine entry into the display-and-trigger strategy. A trader posts a bluff display only when visibility is high enough relative to the friction and to the strength of her directional motive. After integrating over episode-level heterogeneity, prevalence is therefore increasing in visibility and decreasing in enforcement or other forces that raise the cost of cancellation.

The proposition also clarifies the scope of the empirical predictions. The entry threshold $\nu^*(m,c)$ varies across episodes because the signal realization m, the friction c, and the episode's price-level primitives vary across episodes. The market-level entry margin is therefore smooth. The model does not imply a sharp kink, nor does it require a hump-shaped response of prevalence across the visibility

distribution. It predicts a sign. More visible displayed liquidity should make the display-and-trigger strategy more attractive, while higher cancellation frictions should make it less attractive.

A second scope condition concerns magnitudes. Level predictions involving the product of visibility and displayed size depend on how the noise in interim order flow scales with the display. If the noise is stationary, larger and more visible displays make cross-side flow more informative in a strong quantitative sense. If instead larger displays sit in proportionally noisier flow, the informativeness of cross-side flow can saturate. In that case, the level predictions flatten, but the sign predictions remain. For this reason, the model is used to discipline signs rather than to impose tight magnitude restrictions. Appendix B.8 gives the formal version of this robustness argument.

A final measurement point links the model's latent signals to the empirical counts. In the model, ϕ_c is a latent summary of reactive flow. In the data, the corresponding variables are nonnegative and right-skewed trade counts. Under a standard log-normal arrival specification, the measured count is an exponential transformation of the latent signal. The cancellation-hazard index is then concave and saturating in measured counts even when it is linear or convex in the latent signal. This gives an economic reason to work with transformations such as $\log(1+x)$ and to keep the first stage flexible rather than imposing a linear index in raw counts. Appendix B.7 gives the measurement argument.

3 Structural Restrictions

The model in Section 2 turns the idea of phantom depth into restrictions on observable behavior. These restrictions do not ask whether a large order is canceled, since many legitimate orders are canceled. Instead, they ask whether cancellation is timed to the response that the displayed order was designed to elicit, whether this timing is stronger than ordinary fill avoidance, whether the same behavioral state predicts excluded price and depth outcomes, and whether the strategy is more common when displayed liquidity is more informative.

The sequence of events is as follows. A trader displays depth, and other traders observe the display and respond before the order is executed. The trader then withdraws the displayed order after the targeted response has arrived. The empirical content of the model is therefore not a post-collapse price reversal but rather a behavioral sequence of display, response, and withdrawal, together with the price and depth consequences of that sequence.

We organize the restrictions around this sequence. Restrictions 1 and 2 identify the behavioral mechanism inside the episode. Restrictions 3 and 4 assess the effects on price and liquidity using outcomes excluded from the trigger score's construction. Restriction 5 tests the cross-sectional implication that phantom depth depends on signal visibility rather than cancellation technology.

We start by introducing notations used in the empirical sections. Let i index a phantom-depth episode and let t index NBBO-update time within the spell window $[s_i, \tau_i]$ defined in Section 4. Let $T_i \in \{0, 1\}$ denote the latent display-and-trigger state. The empirical object used in later sections is the cross-fitted score $\widehat{T}_i^{(-k)}$, estimated from within-episode cancellation timing and excluding

post-event outcomes.

The cancellation hazard at the update time t is

$$\lambda_{it}^{C} = \Pr(C_{i,t+\delta} = 1 \mid \mathcal{F}_{it}),$$

where $C_{i,t+\delta}$ equals one if the displayed depth on the spike side is reduced by cancellation over the next δ milliseconds and $\mathcal{F}_{it}$ is the information available at time t. The variables ϕ_{it}^{cross} and ϕ_{it}^{fill} measure aggressive trade arrivals in the interval $[t-\Delta, t)$. Cross-side flow is aggressive trading on the side that benefits from the displayed signal. Fill-direction flow is aggressive trading that threatens execution against the displayed order. The signed dislocation D_i is the price movement from the pre-spike reference to the collapse instant. The signed reversal R_i is the price movement from collapse to five minutes after collapse, signed so that larger values correspond to less reversal against the phantom-depth side. Controls X_{it} include time-of-day, lagged returns, lagged total trade intensity, spread, and local volatility.

3.1 Behavioral restrictions inside the episode

The first two restrictions identify the manipulation-relevant state before any post-event price outcome is observed. This is a central departure from the reversal-based classification. The model does not ask whether the prices will eventually revert. It asks whether the trader withdraws the displayed signal after the signal has elicited the response it was meant to elicit.

Cross-side trigger. The key distinction is between cross-side and fill-direction flows. The cross-side flow informs the manipulator that the visible order has affected other traders' behavior. The fill-direction flow has a different interpretation. It threatens execution against the displayed order, but does not show that the targeted response has arrived.

This distinction separates manipulation from ordinary order fill avoidance. A trader who simply wants to avoid execution should react most strongly to fill-direction activity. A trader using a display-and-trigger strategy should react most strongly to cross-side activity. The displayed order is useful only if it causes other traders to move. Once that movement occurs, the manipulator wants to withdraw.

Restriction 1 (Cross-side trigger)

$$\frac{\partial \lambda_{it}^{C}}{\partial \phi_{it}^{\text{cross}}} > 0, \qquad \frac{\partial \lambda_{it}^{C}}{\partial \phi_{it}^{\text{cross}}} > \frac{\partial \lambda_{it}^{C}}{\partial \phi_{it}^{\text{fill}}}. \tag{12}$$

The first derivative is the Cross-Side Trigger Elasticity, or CSTE. It measures how the cancellation hazard responds when the targeted cross-side flow arrives while the displayed order is still live. The second derivative is the Fill-Direction Elasticity, or FDE. It measures the corresponding response to trades that threaten execution against the displayed order.

The comparison between the two elasticities is the key restriction. A positive CSTE alone is not sufficient evidence of manipulation. Cross-side trades may contain information, and an ordinary liquidity provider may cancel after observing informative order flow. The display-and-trigger mechanism makes a sharper prediction. Cancellation should respond more strongly to the flow that reveals success of the display than to the flow that merely raises execution risk. The inequality CSTE > FDE captures this distinction.

Proposition 2 derives both inequalities from the manipulator's cancellation problem. The first inequality, CSTE > 0, follows because cross-side flow is the response the manipulator wants to induce and then trade against. The second inequality follows under the trigger condition (C2), which is the model's formal boundary between display-and-trigger behavior and ordinary liquidity provision. The proposition also implies that the fill-direction coefficient should not be zero. Even a manipulator who uses displayed depth strategically still manages picking-off risk. The empirical prediction is therefore CSTE > FDE > 0, not CSTE > 0 and FDE $= 0$.

Stackelberg timing. The model also imposes an order of events. The displayed order appears first, other traders respond while it is visible, and the cancellation follows. This sequence implies an asymmetric lead-lag pattern. Cross-side activity should predict cancellation over short horizons. Once local market conditions are controlled for, cancellation should not symmetrically predict future cross-side activity.

This restriction is useful because many alternative mechanisms can generate contemporaneous correlation between trades and cancellations. A news shock, a volatility burst, or a liquidity shock can make both objects move at nearly the same time. Those common shocks do not, by themselves, imply manipulation. The Stackelberg mechanism is more specific since it predicts that cancellation is a downstream choice made after the cross-side response has arrived.

Restriction 2 (Stackelberg timing) *For $h > 0$,*

$$\frac{\partial \Pr(C_{i,t+h} = 1 \mid \mathcal{F}_{it})}{\partial \phi_{it}^{\text{cross}}} > 0, \qquad \frac{\partial \Pr(\phi_{i,t+h}^{\text{cross}} > 0 \mid \mathcal{F}_{it})}{\partial C_{it}} \approx 0. \tag{13}$$

Restriction 2 is a timing restriction, not an exclusion restriction. It does not require cross-side flow to be the only reason for cancellation. Orders are cancelled for many legitimate reasons, including adverse selection, inventory control, latency, and ordinary quote management. The restriction instead concerns the direction of the incremental relation. In the display-and-trigger mechanism, cross-side flow arrives while the displayed order is live and signals that the display has induced the intended response. Cancellation should therefore follow cross-side flow.

The reverse relation should not be symmetric. Once the order is cancelled, the displayed signal is removed. The cancellation itself should not generate additional future cross-side flow after conditioning on the information already available at time t. Proposition 3 sharpens this logic. Because cancellation removes the object that elicits cross-side flow, the reverse relation should be weakly negative rather than merely zero. The negative reverse effect should also be stronger in episodes

where the display-and-trigger state is stronger, since those are the episodes in which the displayed order was most successful in eliciting cross-side response.

Thus the prediction is not simply that trades and cancellations are correlated at high frequency. Many common shocks can generate such correlation. The model predicts an ordered sequence. Cross-side flow should predict later cancellation while the display is live, while cancellation should not predict more future cross-side flow once the display has been withdrawn.

3.2 Outcome restrictions excluded from the trigger score

The behavioral restrictions identify the state. The next restrictions ask whether that state has the consequences predicted by the model. Because these consequences are not used to construct the cross-fitted trigger score $\widehat{T}_i^{(-k)}$, they serve as overidentifying tests rather than as part of the detection rule.

Dislocation rather than reversal. In equilibrium, the manipulator earns the payoff while the displayed order is live. The displayed order changes how the market interprets order flow and creates a signed dislocation in the direction of the displayed signal. The model does not require a large unconditional reversal after collapse. By the time the post-collapse return is observed, the manipulator has already withdrawn the displayed order and traded. Once the displayed order is withdrawn, the subsequent price path reflects several forces simultaneously, including the manipulator's revised informed trade, ordinary liquidity provision, and later information arrival.

The model therefore makes a strong prediction for live dislocation and only a weak prediction for reversal.

Restriction 3 (Dislocation and reversal)

$$\mathbb{E}[D_i \mid T_i = 1] > \mathbb{E}[D_i \mid T_i = 0], \qquad \mathbb{E}[R_i \mid T_i = 1] \geq \mathbb{E}[R_i \mid T_i = 0]. \tag{14}$$

The first inequality is the main price implication of the model. When the display-and-trigger state is stronger, the displayed order should move prices farther in the direction induced by the signal while the signal is still live. This is the live dislocation D_i. It is the price effect created before the displayed order is withdrawn, and it is the object most directly tied to the manipulator's payoff.

The second inequality separates the depth-inflation mechanism from a simple pump-and-cancel story. In a pump-and-cancel story, the displayed order is interpreted as informed demand, prices overshoot, and removing the display should generate a reversal after collapse. In that case, high-trigger episodes would be followed by more negative post-collapse returns. The model here has a different logic. The displayed order makes the market appear deeper and lowers the sensitivity of prices to order flow while the signal is live. The manipulator's revised trade is therefore only partly incorporated before the display is withdrawn, and some of the adjustment can occur later as information is revealed. The prediction is therefore not an excess reversal. It is no stronger negative reversal in high-trigger episodes and possibly continuation.

The weak inequality for R_i captures this distinction. Since R_i is signed so that higher values mean fewer reversals on the phantom-depth side, the restriction $\mathbb{E}[R_i \mid T_i = 1] \geq \mathbb{E}[R_i \mid T_i = 0]$ says that the strategic state should not be identified by a larger post-collapse reversal. The mechanism is the live dislocation created by the displayed depth and the response it elicits, not a mechanically observed reversal after the display disappears.

The same trigger state should also leave traces in depth, volatility, and spreads. Let Depth_i^- denote displayed depth on the spike side immediately before the collapse update, and let Depth_i^+ denote displayed depth on the same side immediately after collapse. Define

$$\Delta\text{Depth}_i = \text{Depth}_i^+ - \text{Depth}_i^-.$$

If the trigger state corresponds to phantom depth, collapse should remove more displayed liquidity. Hence,

$$\mathbb{E}[\Delta\text{Depth}_i \mid T_i = 1] < \mathbb{E}[\Delta\text{Depth}_i \mid T_i = 0].$$

Next, let $RV_i(h)$ denote realized midquote variance accumulated over the first h seconds after collapse, and define the post-collapse variance ratio

$$VR_i(h) = \frac{RV_i(h)}{RV_i(5\text{s})}, \qquad h > 5\text{s}.$$

A lower value of $VR_i(h)$ means that price variation is more concentrated near collapse than continuing to accumulate over longer horizons. The model predicts

$$\mathbb{E}[VR_i(h) \mid T_i = 1] < \mathbb{E}[VR_i(h) \mid T_i = 0], \qquad h \in \{10, 30, 60, 300\}\text{s}.$$

The logic is that phantom depth distorts the signal content of order flow while it is live, so the price adjustment should be concentrated around withdrawal rather than followed by persistent post-collapse volatility.

Finally, let ES_i^- and ES_i^+ denote effective spreads before and after collapse, measured over the same event windows used in the empirical tests, and define

$$\Delta\text{ES}_i = \text{ES}_i^+ - \text{ES}_i^-.$$

If cancellation reveals that part of the displayed liquidity was not genuine, market makers should price greater adverse selection risk after the collapse. Thus,

$$\mathbb{E}[\Delta\text{ES}_i \mid T_i = 1] > \mathbb{E}[\Delta\text{ES}_i \mid T_i = 0].$$

All these predictions are different consequences of the same state. If the trigger score captures the display-and-trigger mechanism, it should order all of these excluded outcomes in the predicted directions.

Inverse market depth. The model's most direct depth-inflation prediction concerns the inverse market depth. In Proposition 1, the manipulator's displayed order increases the apparent variance of total order flow without proportionally increasing the covariance between order flow and fundamentals. The market, therefore, looks deeper while the phantom depth is live. The empirical implication of relation (9) for $\widehat{\Lambda}^*$ from Proposition 1 is that the price impact of the order flow should be lower during the live phantom-depth window than in nearby windows where the displayed signal is not present.

Restriction 4 (Inverse market depth)

$$\widehat{\Lambda}_i^{\text{live}} < \widehat{\Lambda}_i^{\text{pre}}, \qquad \widehat{\Lambda}_i^{\text{live}} < \widehat{\Lambda}_i^{\text{post}}. \tag{15}$$

Restriction 4 is not a restatement of the mechanical depth collapse used to define the event. It concerns price impact. If phantom depth makes the market appear deeper than it really is, then signed order flow should have a lower price impact while the phantom depth is live than before or after the episode.

3.3 Cross-sectional restriction on visibility

The final restriction moves from event time to the security-month level. The model contains a cancellation technology parameter q, but in modern electronic markets, the ability to cancel quickly is widespread. The more important cross-sectional question is whether displayed liquidity is informative as a public signal.

Write effective manipulation technology for stock j as

$$Q_j = q_j \times \nu_j,$$

where q_j captures cancellation feasibility and ν_j captures signal visibility. The hidden-order rate is an inverse proxy for ν_j. When hidden and reserve liquidity are more important, displayed depth is a noisier measure of true trading interest. A visible order should then have less influence on other traders' behavior, making the display-and-trigger strategy less useful.

Restriction 5 (Visibility) *The probability of the display-and-trigger state is increasing in signal visibility ν_j and decreasing in the hidden-order rate. Conditional on visibility, the cancel-to-trade ratio should not be the main predictor of phantom-depth prevalence.*

This restriction separates two margins that are closely related in practice, but conceptually distinct. One margin is cancellation feasibility. If phantom depth were mainly limited by the ability to cancel quickly, then securities with more cancellation activity should have more phantom-depth episodes. The other margin is signal visibility. If phantom depth works by manipulating displayed liquidity as a public signal, then the strategy should be more valuable when visible depth

is informative and less valuable when hidden and reserve liquidity make visible depth a noisy measure of true trading interest.

Proposition 4 derives this distinction. In the extended model, both the equilibrium cancellation probability and the value of posting a display are increasing in visibility and decreasing in cancellation frictions. A more visible display makes cross-side flow more informative about whether the order has elicited the targeted response. A higher friction reduces the value of withdrawing after that response arrives. Entry into the display-and-trigger strategy then depends on whether visibility is high enough relative to the friction and the trader's directional motive.

The restriction is therefore a monotone visibility prediction rather than a sharp threshold prediction at the market level. The entry threshold varies across episodes with the signal realization and other local primitives, so aggregation smooths the margin. Prevalence should decline when hidden-order activity is high because the displayed liquidity is then a weaker public signal. Conditional on this visibility channel, the cancel-to-trade ratio should not be the main predictor of phantom-depth prevalence. The issue is not simply that traders can cancel. It is that other traders condition on displayed liquidity before they know whether it is genuine.

3.4 Joint overidentification

Each restriction can be challenged in isolation. Cross-side flow may contain information. Dislocations may arise from ordinary liquidity shocks. Spreads may widen after cancellations for reasons unrelated to manipulation. Identifying content comes from the joint structure. A single state, estimated only from within-episode timing, must predict several outcomes that are excluded from its construction.

The joint moment vector is

$$g_i(\theta) = \begin{pmatrix} C_{it} - \Lambda_C\big(\alpha_i + f_c(\phi_{it}^{\text{cross}}) + f_f(\phi_{it}^{\text{fill}}) + X_{it}'\beta\big) \\ D_i - \alpha_s - \delta_D\, \widehat{T}_i^{(-k)} - X_i'\beta_D \\ R_i - \alpha_s - \delta_R\, \widehat{T}_i^{(-k)} - \rho D_i - X_i'\beta_R \\ \Delta\text{Depth}_{i,k} - \alpha_i - \sum_m \pi_m\, \mathbf{1}\{k=m\} - \sum_m \pi_m^*\mathbf{1}\{k=m\}\widehat{T}_i^{(-k)} - \varepsilon_{i,k} \\ VR_i(\ell) - \alpha_s - \delta_{VR}(\ell)\, \widehat{T}_i^{(-k)} - X_i'\beta_{VR} \\ \text{ES}_{i,t} - \alpha_s - \delta_{\text{ES}}\, \widehat{T}_i^{(-k)} - \eta\, \text{Post}_{i,t} - \xi\, \widehat{T}_i^{(-k)} \times \text{Post}_{i,t} - X_i'\beta_{\text{ES}} \end{pmatrix}. \tag{16}$$

Here $\widehat{T}_i^{(-k)}$ is the out-of-fold trigger score. The superscript indexes the fold used in cross-fitting. The score is estimated from within-episode timing variables and excludes the outcomes that appear in the later components of the moment vector.

The model predicts

$$f_c'(\cdot) > 0,\ f_c'(\cdot) > f_f'(\cdot),\ \delta_D > 0,\ \delta_R \geq 0,\ \delta_{VR}(\cdot) < 0,\ \delta_{\text{ES}} > 0,\ \pi_{-1}^* - \pi_0^* > 0,\ \pi_{+3}^* > 0. \tag{17}$$

The first two signs are the cross-side trigger restrictions. The next two are the price-outcome restric-

tions. The variance-ratio and spread signs capture the post-collapse market-quality implications. The depth restrictions say that high-trigger episodes should have sharper withdrawal at collapse and faster recovery afterward. Restriction 4 is tested separately in Section 6.3, and Restriction 5 is tested in Section 7. The inference framework follows Hansen (1982), Andrews and Soares (2010), Chernozhukov, Hong, and Tamer (2007), and Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018).

The purpose of the joint test is not to prove that every detected depth spike is a manipulation. It is to ask whether a single behaviorally identified state orders a set of excluded outcomes in the pattern implied by the Stackelberg equilibrium. The restrictions are credible as a group because they connect the whole economic sequence. The display elicits a cross-side response. Cancellation follows that response. The episode produces a live dislocation and sharper withdrawal of the displayed depth. The market appears deeper while the signal is live. The strategy is less common when displayed liquidity is a weaker public signal.

4 Data and Measurement

The empirical analysis uses three data sources, namely the NYSE Daily Trade and Quote (TAQ) database for quotes and trades, the SEC Market Information Data Analytics System (MIDAS) for daily security-level order-flow metrics, and a hand-coded panel of SEC and CFTC enforcement actions targeting spoofing-related conduct. The sample covers the 494 S&P 500 securities with continuous coverage between January 2017 and June 2023, a span of 1,632 trading days. The 494 securities comprise 465 common stocks and 29 exchange-traded products (ETPs) such as SPY, QQQ, and the sector ETFs. The Section 7.1 cross-sectional regressions report a robustness specification that excludes the 29 ETPs. The central design principle of the measurement strategy is that the mechanical detection of phantom-depth episodes uses only the statistical signature of an anomalous depth spike that subsequently collapses within two minutes. No directional filter, no intent assumption, and no condition on post-event price behavior is imposed at the detection stage. The behavioral content of the model is put to work later, in the within-episode tests of Section 5, where it serves to identify which of the mechanically detected episodes satisfy the model's behavioral restrictions. Separating detection from identification is essential because it allows subsequent reversal patterns, dislocation patterns, and cross-side timing patterns to be true findings rather than artifacts of selection.

4.1 Data and sample

The primary data are the Daily TAQ database (NYSE) covering January 2017 through June 2023. The NBBO file records the consolidated best bid and offer at nanosecond resolution under CAT reporting standards. The Trade file records every transaction with price, volume, exchange, correction indicator, and sale-condition codes. Nanosecond precision is essential because many phantom-depth episodes last only tens of milliseconds. Filters restrict to regular trading hours 9:35–16:00 ET (ex-

cluding the opening-auction window to avoid SIP flicker), drop quotes with non-positive bid or ask or negative spread or an invalid NBBO indicator, and drop trades with correction or special sale-condition codes. Appendix D documents the full filter list. Trade direction follows Lee and Ready (1991) with a 0-second lag, reflecting the SIP clock synchronization under CAT.[5] The midpoint is $\text{mid}_t = (\text{BestBid}_t + \text{BestOffer}_t)/2$, and quoted depth is NBBO size times the round-lot size of 100 shares.

The cross-sectional and time-series determinants come from the SEC's Market Information Data Analytics System (MIDAS), which reports daily order-flow metrics at the security-exchange level. Three MIDAS variables enter the analysis. The cancel-to-trade ratio is the number of order cancellations divided by the number of lit trades, the empirical proxy for the cancellation-feasibility parameter q_j in the model. The hidden-order rate is the fraction of executed volume filled against hidden or reserve orders, the inverse proxy for the visibility parameter ν_j. The trade-to-order volume ratio is the share of submitted order volume that executes. MIDAS variables are volume-weighted across exchanges to security-month frequency and lagged one month in all regressions.

A hand-coded panel of SEC and CFTC enforcement actions targeting spoofing and layering between 2010 and 2023 supplies the event-study shocks of Section 7. Four cases fall within the equity-market sample. SEC v. Citigroup (January 19, 2017) sanctioned an institutional desk for Treasury-spoofing conduct. SEC v. Taglich Brothers (October 17, 2019) charged an equity brokerage with layering in NASDAQ-listed securities. DOJ/CFTC v. JPMorgan Chase (September 29, 2020) settled for $920 million across precious-metals and Treasury futures, the largest spoofing penalty at the time. SEC v. Singh (September 19, 2022) charged a trader with layering in mid-cap U.S. equities. These dates serve as discrete shocks to the perceived cost of spoofing-related conduct in equities.

The analysis sample is the 494 S&P 500 securities with continuous TAQ coverage over the 1,632-trading-day window. The window lies entirely in the CAT-era reporting regime, spans three of the four equity-relevant enforcement actions, and brackets both the COVID-19 volatility shock and the SEC's December 2022 equity market structure proposals.

4.2 Mechanical detection of phantom-depth episodes

We begin with a deliberately mechanical definition of the event population. The goal at this stage is not to decide whether an episode is manipulative. It is only to identify large, short-lived disruptions to displayed depth that are candidates for the mechanism in the model. The behavioral content enters later, in Section 5, where we ask whether cancellation is timed to the response elicited by the display.

A depth spike on side s at time t is an NBBO update satisfying

$$z_{s,t} = \frac{D_{s,t} - \overline{D}_{s,t}}{\sigma_{D,s,t}} \geq K, \qquad \Delta D_{s,t} = D_{s,t} - \overline{D}_{s,t} \geq 1{,}000 \text{ shares}, \qquad D_{s,t}/\overline{D}_{s,t} \geq 2, \tag{18}$$

[5] Robustness with a 100-millisecond lag changes the trade classification on less than one percent of trades.

where $\overline{D}_{s,t}$ and $\sigma_{D,s,t}$ are the trailing 300-second mean and standard deviation of best-quote depth. The baseline threshold is $K = 3$. An episode begins at the spike and ends at the first subsequent update at which depth returns to within one rolling standard deviation of its pre-spike mean. The search window is capped at 120 seconds. Spikes that do not collapse within this window are treated as persistent depth placements rather than phantom-depth candidates. We also remove episodes shorter than 10 milliseconds to avoid SIP-consolidation flicker, and we keep only the first spike within a 120-second window on each side-symbol-day. Appendix D gives the step-by-step algorithm, the minimum-update requirement for the rolling window, the heavy-tail false-positive bound, and the threshold-sweep robustness.

The most important feature of the detector is what it does not use. It does not require a price reversal. It does not condition on post-collapse trade direction. It does not require prices to move in any particular direction. It does not use account identity, intent, or subsequent cross-side trading. It simply flags anomalous depth spikes that appear and then disappear. This separation is essential for identification. If the detector used the same outcomes that later tests are meant to explain, the empirical exercise would be circular.

The detector, therefore, defines a specific and limited class of events. It captures single-side displayed-depth spikes that exceed the $z = 3$, 1,000-share, and 10-millisecond thresholds and collapse within 120 seconds. It does not capture every possible form of manipulation. Layering strategies that spread orders across several price levels, hidden or iceberg strategies with only a small displayed tip, cross-venue strategies that split displayed size across exchanges, sub-10-millisecond message-timing strategies, and slow strategies that keep displayed depth in the book for more than two minutes are outside the detectable class. The paper's claims should be read accordingly. They apply to the class of displayed phantom-depth episodes that our detector is designed to measure.

A simple check suggests that this detectable class covers most cancel-heavy security-days in the S&P 500. In the MIDAS daily panel, restricted to S&P 500 securities, there are 41,813 security-days in the top five percent of the cancel-to-trade distribution. Of these, 92.9% contain at least one detected phantom-depth episode. The remaining 7.1% are cancel-heavy security-days with no detected episode. Some of them may contain manipulation outside our detectable class, while others likely reflect ordinary cancellation activity that never creates a large displayed-depth spike. This number is best interpreted as a floor on missed activity, not as a point estimate of hidden manipulation.

4.3 Threshold choices and sensitivity

The baseline thresholds are not chosen to generate a particular price pattern. We verify this by sweeping over the main detection parameters. The grid varies the depth-spike threshold,

$$K \in \{2.0, 2.5, 3.0, 3.5, 4.0, 5.0\},$$

Table 1: **Candidate phantom-depth events: full-sample descriptives.** This table reports descriptive statistics for the candidate phantom-depth event population. Stage-1 mechanical detection, $K = 3$, depth increase $\geq$ 1,000 shares, minimum duration 10 ms. Sample: 494 S&P 500 securities, January 2017–June 2023.

	Mean	Median	Std. dev.	P5	P95
Panel A. Event frequency and duration					
Events per security-day	14.2				
Duration (seconds)	11.19	1.80	21.46	0.02	62.43
Panel B. Depth spike characteristics					
Depth at spike (shares)	14,542	4,500	38,596	1,500	58,000
Depth increase (shares)	10,308	3,358	27,434	1,117	41,072
Depth ratio (spike / pre)	6.40	4.34	8.16	1.84	16.49
z-score at spike	5.58	4.17	4.11	3.07	12.86
Cross-side delay (s)	4.15	0.32	8.67	0.00	22.46
Panel C. Price dynamics					
Signed dislocation (bps)	3.05				
|Dislocation| (bps)	4.20	1.10			
Signed reversal (bps)	−1.29				
|Reversal| (bps)	23.27	10.33			
N events			10,643,414		
N security-days			750,821		
N securities			494		
Trading days			1,632		

the minimum absolute depth increase,

$$\{0, 500, 1{,}000, 2{,}000, 5{,}000\} \text{ shares},$$

and the minimum duration,

$$\{0, 0.01, 0.1, 1.0\} \text{ seconds}.$$

Across thresholds with $K \geq 2.5$ and a minimum increase of at least 500 shares, the central qualitative fact is unchanged. The unconditional signed reversal of mechanically detected events is economically negligible. A regression-discontinuity check at the baseline cutoff $K = 3$, reported in Appendix Table G.10, further shows that the results are not driven by a knife-edge threshold choice. The full grid is provided in the replication archive.

4.4 Descriptive statistics

Table 1 describes the event population produced by the baseline detector. The sample covers 494 S&P 500 securities from January 2017 through June 2023. The baseline parameters are $K = 3$, a minimum depth increase of 1,000 shares, and a minimum duration of 10 milliseconds. Appendix Table D.1 reports the corresponding annual counts, event rates, cross-side delays, and cancellation rates.

The detector identifies 10,643,414 eligible phantom-depth events across 494 securities and 1,632

trading days. The average security has 14.2 events per day, and 93.1% of security-days have at least one event. These episodes are therefore not confined to a small set of securities or unusual trading days. They are a regular feature of displayed depth in S&P 500 securities. Because the detector excludes smaller orders, hidden liquidity, cross-venue splitting, and slower display-and-cancel cycles, the count should be interpreted as a lower bound on the broader set of strategies that could manipulate displayed liquidity.

The episodes are short, and the response arrives quickly. The median episode lasts 1.80 seconds, with a mean of 11.19 seconds and a ninety-fifth percentile of 62.43 seconds. The median cross-side delay, measured from the spike to the first trade on the opposite side, is 0.32 seconds. Of all eligible episodes, 41.1% see a cross-side trade within 50 milliseconds, 52.1% within 500 milliseconds, and 58.0% within one second; the full distribution is reported in Appendix Table D.2. These numbers motivate the sub-second horizons used in the within-episode tests. The annual series in Appendix Table D.1 peaks in 2020, during the COVID volatility shock, and then declines to 11.1–11.7 events per security-day in 2021–2023.

The composition of the event population, summarized in Appendix Table D.2, is informative for the design of the later tests. The event population is symmetric across sides, with 50.1% of spikes on the bid and 49.9% on the ask. Most episodes end in cancellation rather than execution: 79.4% of the displayed depth is withdrawn before it can be hit, and only 20.6% of episodes end with execution against the displayed quote. About 52.8% of episodes are followed by a directional five-minute reversal against the spike side, and 47.2% are not. This near-even split is the basis for the unconditional and reversal-conditioned comparisons in Section 5.1. The execution margin is itself informative. The median executed share of displayed depth on the spike side is only 3.5%, the mean is 42.9% (heavily right-skewed), and 37.4% of episodes execute zero displayed volume. The displayed order behaves as a signal rather than as resting liquidity that intends to trade.

The depth spikes are economically large. The median depth at the spike is 4,500 shares, and the median increase relative to the pre-spike rolling mean is 3,358 shares. The median spike is more than four times the pre-spike depth, and the median z-score is 4.17, comfortably above the baseline threshold. At the moment of the spike, the median event replaces 77% of the displayed depth on the affected side with transient liquidity. Aggregated over the trading day, detected phantom depth is live for 159 seconds per security-day on average, or 0.68% of the regular session.

Figure 3 shows the same point in the event time. Quoted depth on the spike side rises sharply before the collapse, peaks just before the collapse update, and then falls immediately afterward. The figure is descriptive rather than identifying: the rise and collapse are features of the mechanical event definition. Its role is to show what the candidate episodes look like at NBBO-update resolution before we turn to price dynamics and behavioral identification.

The price moments are intentionally reported without conditioning on reversal. They are flat. The mean signed dislocation is 3.05 basis points, the mean signed reversal is -1.29 basis points, and the mean net price impact is 0.25 basis points with a median of zero. At the same time, the absolute reversal is much larger, with a mean of 23.27 basis points and a median of 10.33 basis

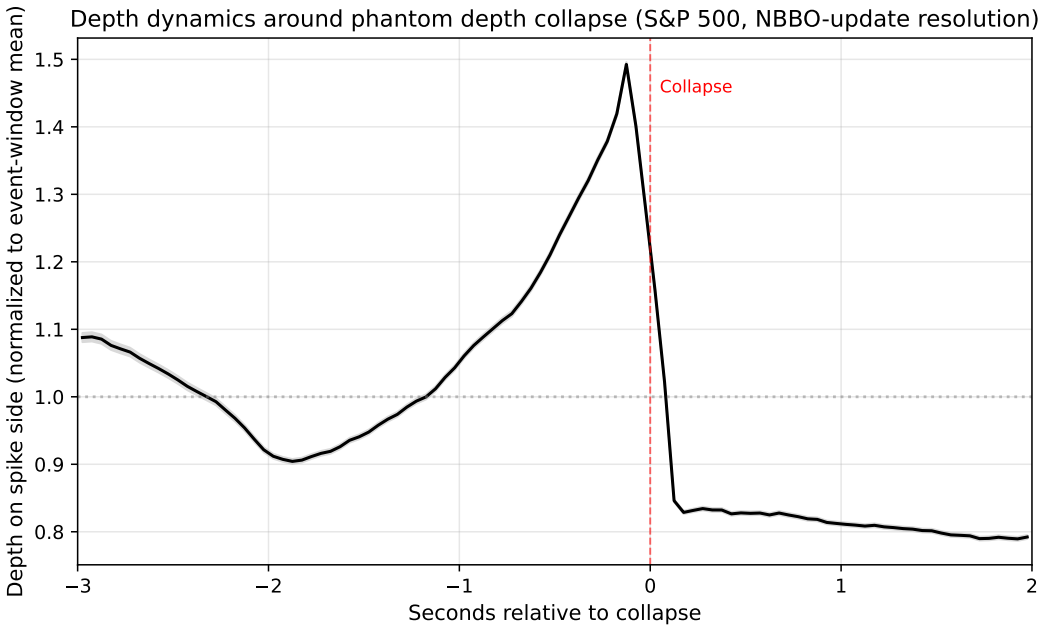


Figure 3: Depth dynamics around collapsed-depth episodes.
This figure plots quoted depth on the spike side, normalized within the event by its event-window mean, in 50 ms bins from $T-3$ s to $T+2$ s relative to collapse. Depth rises from about 0.91 times the event-window mean at $T-2$ s to about 1.50 times in the 100 ms before collapse, then falls to about 0.83 times immediately after collapse and continues drifting downward. The figure describes the mechanical anatomy of the candidate episodes. It does not use post-event prices, trigger scores, or reversal conditioning. Sample: 527,019 episodes and 36.2 million NBBO updates.

points. The large absolute movements show that these are not trivial events. The small signed mean shows that the unconditional event population does not contain the headline reversal often associated with manipulation.

A back-of-the-envelope calculation suggests that the dollar scale is material. Multiplying cross-side volume in cancelled-and-reversed episodes by the absolute reversal gives an annual aggregate transfer of approximately $887 million across the S&P 500. This number should not be interpreted as welfare loss or trading profit. It is a conservative scale calculation based on a narrow event class, and the average signed reversal in the full event population is close to zero. The more informative magnitude appears after the behavioral state is identified. Table 6 shows that the model-implied transfer per episode rises monotonically across quartiles of the cross-fitted trigger score. The dollar footprint is therefore tied to the display-and-trigger state rather than to the raw count of collapsed-depth events.

These descriptive facts motivate the identification strategy that follows. The event population is large, fast, and economically meaningful, but its unconditional price dynamics do not identify manipulation. The average signed reversal is close to zero, so identification cannot rest on the mean post-collapse price path. Section 5 shows that this small unconditional mean masks two distinct subpopulations and uses within-episode behavior to separate the display-and-trigger state from the rest of the mechanically detected events.

5 Behavioral Identification Beyond Reversal

Post-event reversal cannot serve as the main identifying evidence of phantom-depth manipulation. The reason is simple. Once the episode is selected using the post-collapse price path, the reversal is no longer independent evidence of the strategic state. A reversal-based filter is designed to find episodes that reverse, so it cannot then use the reversal as evidence that manipulation occurred. The empirical task is instead to identify the manipulation state before the post-event outcome is realized.

This section makes this point in two steps. First, it shows that large reversals arise mainly from conditioning on the realized post-collapse price path. Second, it develops an alternative identification strategy that uses within-episode cancellation timing to measure the display-and-trigger state predicted by the model.

5.1 Reversal selection

We begin with the unconditional population of mechanically detected collapsed-depth spikes. These events are selected only because the displayed depth rises sharply and then collapses within the detection window. The detector does not use reversal, trade direction, account identity, intent, or any subsequent price path.

Table 2 reports the signed five-minute post-collapse reversal for three nested samples drawn from 90.6 million collapsed-depth spikes in the S&P 500 panel. The broadest sample requires only an anomalous increase in displayed depth that collapses within 120 seconds. Its average reversal is -0.13 basis points. With more than 90 million observations, even this tiny number is statistically different from zero, with $t = -13.8$. Economically, however, it is close to zero relative to the cross-event standard deviation of roughly 92 basis points.

The baseline thresholds used in the main analysis, $z \geq 3$ and $\Delta D \geq 1{,}000$ shares, reduce the sample to 10.64 million eligible events. The mean reversal becomes -1.29 basis points, with $t = -19.2$. The sign is negative, but the magnitude remains small. Moving from all collapsed-depth spikes to the eligible-event spikes does not by itself produce the large reversals often associated with manipulation.

The large magnitudes appear only when the sample is conditioned on directional reversal. That restriction keeps 5.62 million events, about 53% of the eligible sample, and mechanically selects episodes whose post-collapse price path moves against the spike side. The mean reversal then falls to -11.66 basis points, with $t = -147$. This estimate is almost nine times larger than the mean of the eligible sample and roughly ninety times larger than the broad collapsed-depth mean. The comparison shows that the economically large reversal is produced mainly by conditioning on the realized post-event price path, not by the depth collapse alone.

The difference between the first and third rows is not, by itself, evidence of manipulation. It is the mechanical effect of conditioning on the sign of the outcome. If $\mathbb{E}[R_i]$ is close to zero, then conditioning on $R_i < 0$ necessarily produces a negative conditional mean. The magnitude of that

Table 2: **Selection by reversal conditioning.** This table reports the signed five-minute post-collapse reversal across three nested samples. Signed reversal is the midprice change from collapse to five minutes post-collapse, signed so that negative values are reversals against the phantom-depth side. The reversal-conditional mean is two orders of magnitude larger than the unconditional mean.

Sample	N	Mean reversal (bps)	t
All collapsed depth spikes	90,632,171	−0.13	−13.8
Eligible by (18) phantom-depth events	10,643,414	−1.29	−19.2
Directional-reversal subsample	5,620,597	−11.66	−147

conditional mean is governed by the dispersion and shape of the unconditional reversal distribution, not by a structural manipulation parameter. Appendix C formalizes this point and shows that the observed value of −11.66 basis points lies within the range implied by selection alone under standard null distributions.

Figure 4 illustrates this argument in the event time. It separates depth spike episodes by whether cross-side trading arrives while the displayed depth spike is live. Because the detector anchors the event at a selected point in the signed midprice path, both triggered and non-triggered episodes share a pre-event decline into the onset. That common pre-event slope is mechanical. It is not the diagnostic part of the figure.

The informative part of the figure is the post-onset divergence. Triggered episodes, the episodes in which displayed depth is followed by cross-side trading during the spell, settle at 1.96 basis points and hold the dislocation. Non-triggered episodes revert almost completely, ending at 0.18 basis points. The bottom panel removes the shared pre-event path by plotting the triggered-minus-non-triggered difference. That difference is essentially flat before the onset and then rises to about 2 basis points afterward.

This decomposition changes the interpretation of reversal. The episodes most consistent with the display-and-trigger mechanism are the triggered episodes, in which displayed depth is followed by the cross-side response emphasized by the model. Yet, these are not the episodes that generate the canonical reversal. They hold the dislocation. The reversal is concentrated in the non-triggered population, where displayed depth appears and disappears without the same cross-side response. A reversal filter therefore selects heavily from the wrong part of the event population.

The relevant price object in the model is the dislocation created while the displayed order is live. The manipulator earns the payoff before the post-collapse return is realized. A five-minute reversal may occur, but it is not the primitive source of the manipulator's gain and is not the right object for identification. The strategic state must be identified from behavior within the episode.

5.2 Within-episode cancellation timing

We now move from post-event prices to behavior inside the episode. The model implies a simple sequence. A trader displays depth to elicit a response. If the response arrives, the trader cancels. Strategic cancellation should therefore respond to the cross-side activity that the displayed order was meant to induce. The empirical task is to measure this response without using the post-collapse

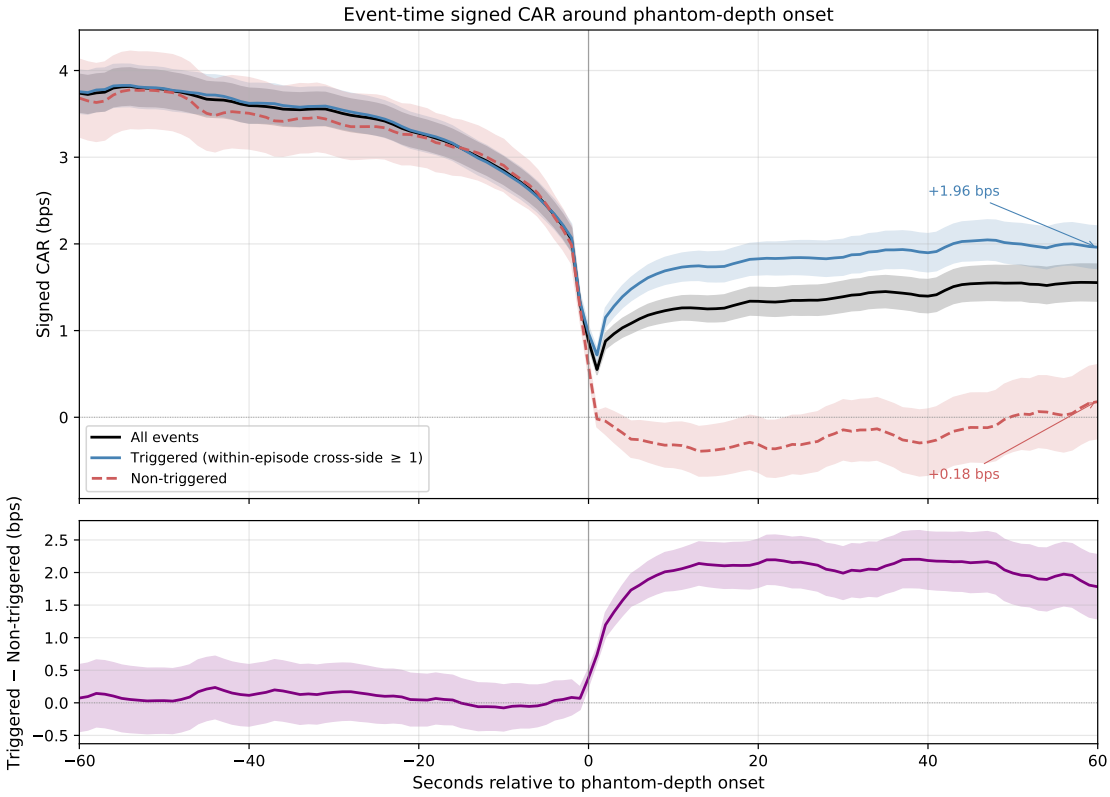


Figure 4: Event-time signed CAR around phantom-depth onset, by trigger status.
This figure plots the signed cumulative abnormal return around phantom-depth onset, separately by trigger status. Signed CAR is in basis points from $T - 60$ s to $T + 60$ s relative to phantom-depth onset. Top panel: all episodes, triggered episodes, and non-triggered episodes. Event time zero is the depth-spike onset. The common pre-event decline is a mechanical consequence of anchoring at the selected instant. The diagnostic content is the post-event divergence. Triggered episodes settle at 1.96 basis points and hold the dislocation. Non-triggered episodes revert to 0.18 basis points. Bottom panel: triggered minus non-triggered difference, which nets out the shared pre-event selection path.

price path.

Let episode i begin at spike onset s_i, end at collapse time τ_i, and occur on side $d_i \in \{\text{bid}, \text{ask}\}$. Let $\mathcal{T}_i$ denote the set of NBBO update times inside the spell $[s_i, \tau_i]$. At each update time $t \in \mathcal{T}_i$, we observe recent cross-side trade activity ϕ_{it}^{cross}, fill-direction activity ϕ_{it}^{fill}, displayed depth on the spike side, the spread, local volatility, adverse-selection return proxies, and total trade intensity. These variables form the information set $\mathcal{F}_{it}$.

Let $C_{i,t+\delta}$ equal one if displayed depth on the spike side falls by more than 100 shares for non-trade reasons over $(t, t+\delta]$, with baseline horizon $\delta = 100$ milliseconds. The first-stage object is the conditional cancellation probability

$$\Pr(C_{i,t+\delta} = 1 \mid \mathcal{F}_{it}).$$

The latent object is the trigger state T_i, which captures whether the episode follows the display-and-trigger behavior predicted by the model. We do not observe T_i directly. We infer it from

cancellation timing inside the spell.

The post-event outcomes are kept separate. Signed dislocation D_i, signed reversal R_i, depth dynamics, post-collapse variance ratios, effective spreads, and price-impact estimates do not enter the first-stage hazard. They are reserved for the overidentifying tests in Section 6. This separation is what prevents the score from being trained on the outcomes it is later used to explain.

One might instead try to condition the cancellation hazard on a decline in price sensitivity, since lower price impact is the manipulator's ultimate object in the depth-inflation mechanism. We do not do this. In the model, the pricing rule is set when the display appears and is sticky within the episode, as in Assumption 5. The interim information available to the manipulator is therefore not a newly observed change in price sensitivity. It is the realized response to the display. Cross-side flow measures precisely that response, and it is the flow the manipulator can trade against after withdrawing the displayed order. Conditioning the first stage on price-based objects would also weaken the exclusion discipline of the design. The price-sensitivity prediction is instead treated as an excluded implication of the model. Restriction 4 predicts lower price impact while the display is live, and this prediction is tested separately using the trigger score.

A related implication is also constrained by the data. In the bluff configuration of Section 2.6, the cancelled display should be followed by a genuine trade by the same trader, with the sign implied by the revised position. TAQ does not contain account identities that link a cancelled order to the owner's later trades. The feasible analogue is therefore aggregate post-collapse order flow in the direction implied by the model, rather than a literal owner-level test.

The relevant variation is high frequency. The median episode lasts 1.80 seconds, and many episodes last less than 100 milliseconds. We therefore work at the NBBO-update level. For each episode, we extract all NBBO updates and classified trades in the window

$$[s_i - 2\text{s},\ \tau_i + 2\text{s}],$$

and construct a spell panel that follows depth, trade flow, and local market conditions throughout the episode. For computational tractability, this panel is built on a 5% stratified random sample of the 10.6 million eligible episodes, stratified by year and event-day. The resulting sample contains 532,170 episodes. After merging with the NBBO timeline and removing incomplete records, the usable panel contains 527,019 episodes and 48,437,488 update-level observations. Distributional checks show that the sample closely matches the full event population. The median duration, spike size, pre-spike spread, signed dislocation, and signed reversal are all within 0.6% of their full-sample values, and the cross-side trade share and trigger frequency are within one percentage point.

The baseline cancellation hazard is

$$\Pr(C_{i,t+\delta} = 1 \mid \mathcal{F}_{it}) = \Lambda_C\Big(\alpha_i + f_c(\phi_{it}^{\text{cross}}) + f_f(\phi_{it}^{\text{fill}}) + X_{it}'\beta\Big), \tag{19}$$

where $\Lambda_C(\cdot)$ is the logistic link, α_i is an episode fixed effect, and X_{it} includes the spread, local volatility, adverse-selection return proxies, and contemporaneous total trade intensity. The baseline

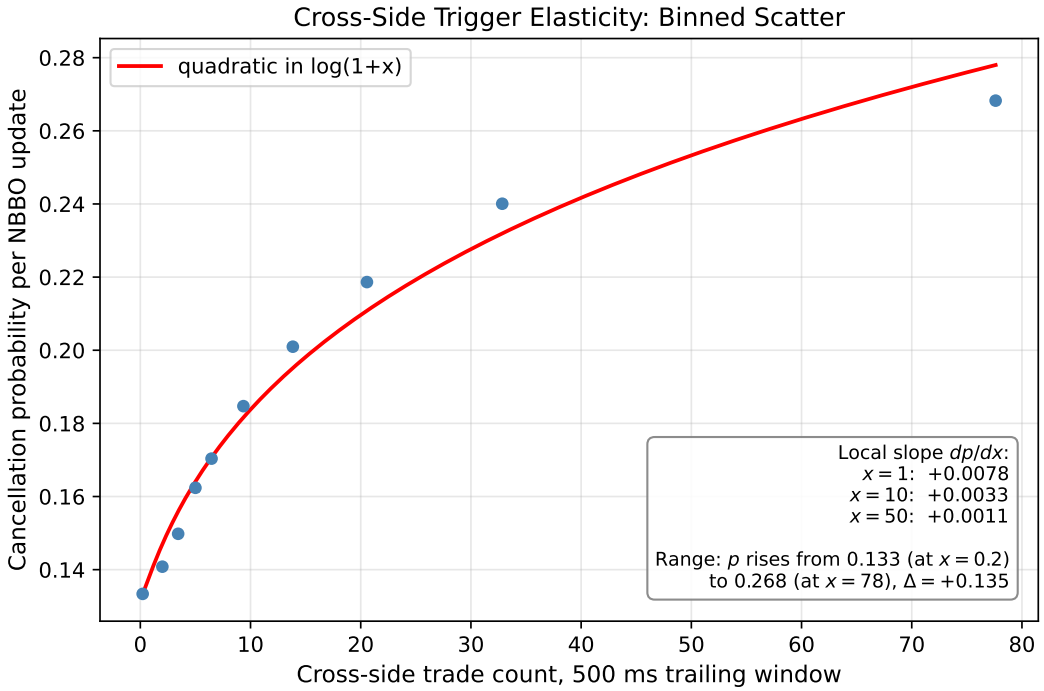


Figure 5: Cross-Side Trigger Elasticity. Binned scatter

This figure plots a binned scatter of the per-update cancellation indicator against the trailing 500 ms cross-side trade count, using in-episode NBBO updates from the S&P 500 spell panels. Equal-frequency quantile bins collapse to ten distinct points because cross-side counts are integer-valued and concentrated at zero. The solid line is a quadratic fit in $\log(1+x)$ weighted by bin size. The cancellation probability rises sharply when cross-side activity first appears and then flattens, consistent with a concave trigger response.

also reports the linear-probability version. In the baseline specification, f_c and f_f are linear in standardized log activity rates. In robustness checks, they are replaced by spline bases,

$$f_c(x) = \sum_{\ell=1}^{L} \theta_\ell^c B_\ell(\log(1+x)), \qquad f_f(x) = \sum_{\ell=1}^{L} \theta_\ell^f B_\ell(\log(1+x)),$$

with $L = 5$ knots placed at empirical quantiles.

The specification in (19) also has a direct interpretation in the endogenous cancellation extension. The logistic link is induced by the latency shock in Proposition 2. The episode fixed effect α_i absorbs primitives that are fixed within a spell but vary across spells, including the trader's value estimate, the display size, the executed non-displayed position, and the withdrawal friction. Omitting the fixed effect would risk attributing this episode-level heterogeneity to the flow coefficients. The $\log(1+x)$ basis is consistent with the measurement mapping in Appendix B.7. The model's interim signal is a latent summary of reactive flow, while the data record skewed trade counts. Under the resulting count mapping, the hazard index is concave and saturating in measured counts, which is the shape seen in the binned estimates. We do not estimate the model structurally. The first stage remains a reduced form, with the model supplying the link, the fixed effect, and the predicted signs.

The Cross-Side Trigger Elasticity is the average derivative of the cancellation hazard with respect

to cross-side activity,

$$\text{CSTE} = \mathbb{E}\left[\frac{\partial}{\partial \phi_{it}^{\text{cross}}} \Lambda_C\left(\alpha_i + f_c(\phi_{it}^{\text{cross}}) + f_f(\phi_{it}^{\text{fill}}) + X_{it}'\beta\right)\right]. \tag{20}$$

The Fill-Direction Elasticity, or FDE, is defined analogously. In the linear-probability specification, the two elasticities are the slope coefficients on cross-side and fill-direction activity.

Restriction 1 predicts two inequalities. First, CSTE > 0, because cancellation should become more likely when the targeted response arrives. Second, CSTE $>$ FDE, because a display-and-trigger strategy should respond more to the cross-side activity that reveals the success of the display than to trades that merely threaten execution of the displayed order.

Figure 5 plots the cancellation indicator against the trailing 500 millisecond cross-side trade count. The relationship is concave in $\log(1 + x)$. The marginal effect is largest when cross-side activity first appears and declines as the count grows. Moving from no cross-side activity to the highest activity bins raises the cancellation probability by 13.5 percentage points. Among 146,578 transitions from zero to positive cross-side activity, the cancellation rate rises from 3.9% to 20.4%, a 16.6 percentage point increase. The increase is already 15.1 percentage points at the 50 millisecond horizon, where most observations are. The pattern is visible before imposing the full hazard specification and before using any post-event price outcome.

5.3 Cross-side elasticity and the fill-direction comparison

Table 3 reports the baseline estimates. Column (1) estimates the reduced-form response of cancellation to cross-side activity. Column (2) adds fill-direction activity in the same specification and includes the full set of controls. The reduced-form CSTE is 0.123, with a t-statistic of 7.21. The estimate is almost identical to the 10-stock pilot estimate of 0.124, and it is stable under alternative clustering. The t-statistic is 6.93 with standard errors clustered by symbol and 7.22 with two-way clustering by episode and date.

The horse race in column (2) is the main test. The cross-side coefficient is 0.121, with a t-statistic of 7.13, while the fill-direction coefficient is 0.018, with a t-statistic of 2.57. Cross-side activity is therefore about seven times as important as fill-direction activity in the linear-probability metric. This ordering is difficult to reconcile with a pure fill-avoidance story. A trader whose only objective is to avoid execution should be most sensitive to trades that threaten to hit the displayed order. Instead, cancellation responds much more strongly to the arrival of the cross-side flow that the display was designed to elicit.

The positive fill-direction coefficient is also informative. The model does not predict that fill-direction flow should be irrelevant. A trader using a display-and-trigger strategy still faces picking-off risk and therefore still manages exposure to trades that threaten execution. Proposition 2 predicts the ordered pair CSTE $>$ FDE > 0. The estimates in column (2), 0.121 and 0.018, match this joint prediction. The fill-direction coefficient should therefore be read as part of the mechanism rather than as a failed placebo.

Table 3: **Cross-Side Trigger Elasticity and channel decomposition.** This table reports the Cross-Side Trigger Elasticity (CSTE) and the horse-race decomposition against the fill-direction channel. The dependent variable is an indicator for a non-trade depth reduction greater than 100 shares within 100 ms. Trade counts are winsorized at the 99th percentile and standardized. Controls include spread, local volatility, and four adverse-selection return horizons. Standard errors are clustered at the episode level.

	(1) Reduced form	(2) Horse race
$\hat{\gamma}^{\text{cross}}$ (CSTE)	0.123^{***}	0.121^{***}
	(0.017)	(0.017)
$\hat{\gamma}^{\text{fill}}$ (FDE)		0.018^{**}
		(0.007)
Spread, σ, AS returns	Yes	Yes
R^2	0.063	0.063
N observations	16,455,380	16,455,380
N episodes	518,911	518,911
Cancel rate	0.492	0.492
t statistic, cross	7.21	7.13
t statistic, fill		2.57

The same inequality appears in the logistic specification used to construct the trigger score. The logit coefficient is 1.048 on standardized cross-side activity and 0.118 on standardized fill-direction activity, yielding a cross-to-fill ratio of 8.9. Thus CSTE > FDE holds in both linear and nonlinear specifications.

The result is also broad across securities. Estimating the hazard separately for each of the 494 securities, the CSTE estimate is positive for 448 securities, or 90.7% of the sample. It is statistically positive at the 5% level for 381 securities, or 77.1%. Only 19 securities have a significantly negative coefficient. Figure 6 shows the full cross-sectional distribution. The figure makes clear that the pooled estimate is not driven by a small number of very active securities. The sign is positive across almost the entire panel.

The pooled and per-stock estimates answer different questions. The pooled estimate is event-weighted and is therefore influenced by liquid securities with many episodes. The per-stock distribution asks whether the sign is broadly present across securities. The important fact is that the sign is positive in almost all securities, rather than that every security has the same elasticity.

A remaining concern is that active markets generate trades and cancellations at the same time. If a common shock raises both, a positive cross-side coefficient could appear without strategic monitoring. The horse race already addresses the symmetric version of this concern by comparing cross-side activity with fill-direction activity and total intensity. We also use a within-episode permutation test. For each episode, we randomly swap the cross-side and fill-direction labels with probability one-half, preserving the episode, total trade count, duration, spread path, volatility, and adverse-selection returns. We repeat the procedure 200 times within each of five sample years. Every year, the true CSTE lies 12 to 34 permutation standard deviations above the permuted distribution, and no permutation draw reaches the true value.

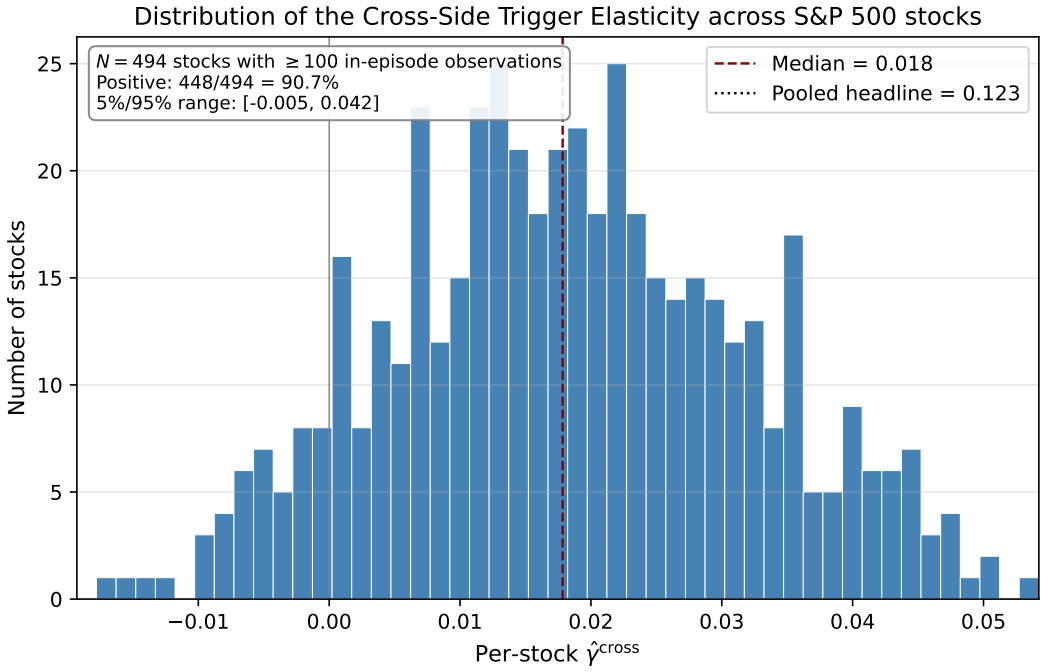


Figure 6: Distribution of the Cross-Side Trigger Elasticity across S&P 500 securities

This figure plots the cross-sectional distribution of per-stock Cross-Side Trigger Elasticity (CSTE) estimates. The reduced-form regression is run separately on each security's spell panel. 448 of 494 securities have a positive point estimate, and 381 are statistically positive. The key feature is the near universality of the sign.

5.4 Timing and alternative explanations

The CSTE establishes that cross-side activity and cancellation are strongly related. The model requires a stronger timing pattern. The response to the displayed order should arrive first, and cancellation should follow. We therefore ask whether cross-side activity predicts later cancellation, rather than merely coinciding with it.

We estimate this directionality at horizons

$$h \in \{50, 100, 200, 500, 1{,}000\} \text{ ms}.$$

Let $\widetilde{x}_{it}^{\text{cross}}$ denote the standardized cross-side trade count in the backward 500 millisecond window ending at t. The forward equation is

$$\Pr(C_{i,t+h} = 1 \mid \mathcal{F}_{it}) = \alpha + \gamma^{\rightarrow}(h)\widetilde{x}_{it}^{\text{cross}} + X_{it}'\beta + \varepsilon_{it}. \tag{21}$$

The reverse equation is

$$\Pr(\widetilde{x}_{i,t+h}^{\text{cross}} > 0 \mid \mathcal{F}_{it}) = \alpha + \gamma^{\leftarrow}(h)C_{it} + X_{it}'\beta + \varepsilon_{it}. \tag{22}$$

Both equations use the same controls. Restriction 2 predicts $\gamma^{\rightarrow}(h) > 0$. The reverse coefficient

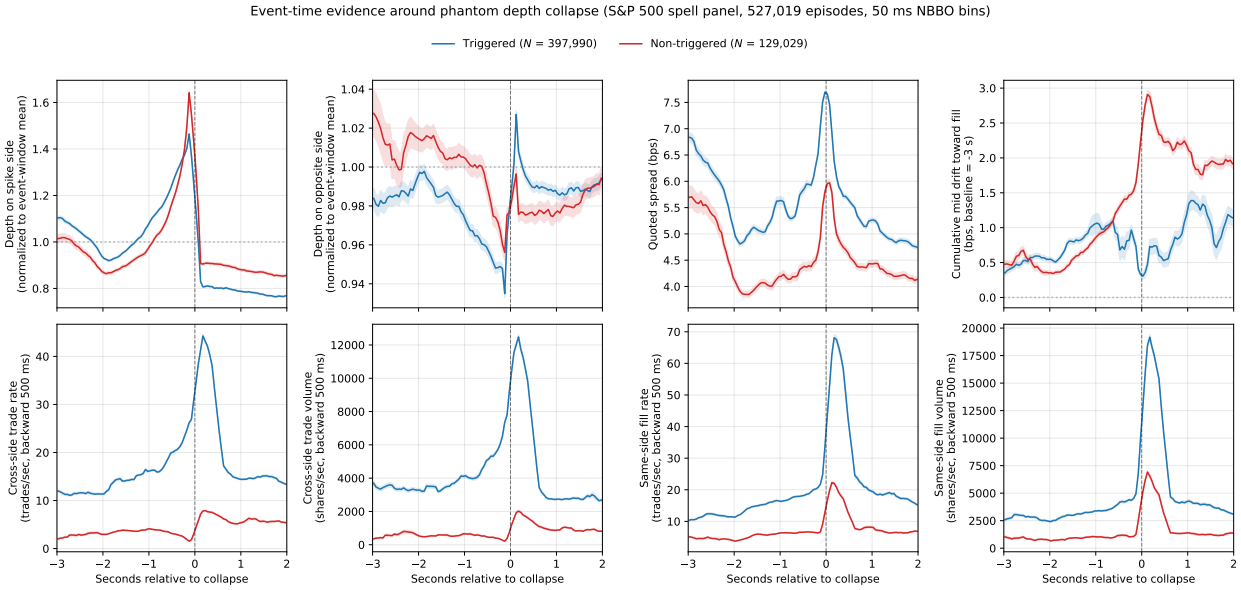


Figure 7: Event-time evidence around phantom-depth collapse by trigger status

This figure plots cross-event means of cross-side trade rate and cancellation hazard at 50 ms NBBO-update resolution around phantom-depth collapse, separately by trigger status. The window is $[-3, 2]$ seconds relative to collapse, computed from the S&P 500 spell panel. Triggered episodes have at least one cross-side trade during the episode. Non-triggered episodes have none. The cross-side rate rises before the cancellation cluster, with a lead of about 200 ms.

should be close to zero, or weakly negative, once cancellation removes the displayed order that elicited the cross-side response.

Figure 7 gives the event-time anatomy of this sequence. Around the collapse of displayed depth, the cross-side trade rate rises from roughly 12 trades per second to roughly 44 trades per second. The increase begins about 200 milliseconds before the collapse. Cancellation rises after the cross-side spike has already begun. Cross-side trading comes first, and cancellation follows.

Timing evidence is necessary, but is not sufficient by itself. A benign stale-quote-sniping story can also generate a positive relation between cross-side flow and later cancellation. A liquidity provider may see cross-side flow as informative, infer that the quote has become exposed, and cancel quickly. This behavior is rational and legitimate. It can produce both a positive CSTE and a lead from cross-side activity to cancellation.

A related benign explanation is fleeting-order quoting. Liquidity providers may post and cancel displayed liquidity rapidly as part of normal competition for the spread (Hasbrouck and Saar, 2009; Baruch and Glosten, 2013). Such behavior can generate fast cancellations and short-lived displayed depth without manipulation. The identifying issue is whether the displayed-depth spike changes the cancellation rule. Under ordinary fleeting-order quoting, cancellation responds to exposure, information, and queue-management incentives that would matter even without the spike. Under display-and-trigger manipulation, cancellation becomes especially sensitive to cross-side flow while the displayed order is live, because that flow reveals that the display has induced the intended response.

Table 4 summarizes four comparisons that separate display-and-trigger behavior from ordinary sniping and fleeting-order quoting. The first comparison asks whether displayed depth elicits cross-side trading. Relative to a matched pre-spike window in the same stock, day, and local minute, the cross-side trade rate rises from 8.91 to 15.81 trades per second during the episode. The increase is 6.90 trades per second, with $t = 63$, across 264,896 matched episodes. Fill-direction activity moves in the opposite direction, falling by 3.46 trades per second. Cross-side activity is therefore not merely an outside signal to which a market maker reacts. It rises when the displayed-depth spike is present.

The second comparison asks whether the cross-to-fill asymmetry is amplified by the display. During the spike, the estimated gap between the cross-side and fill-direction cancellation responses is

$$0.144 - 0.025 = 0.119.$$

In the matched pre-spike window, the same gap is only 0.026. The asymmetry is therefore about 4.5 times larger when the displayed-depth spike is live. This amplification is not a mechanical consequence of larger displayed depth during the spike. Within the live window, the cross-to-fill cancellation gap does not increase with available depth. Sorting episodes into terciles of spike-side depth gives gaps of 0.203, 0.125, and 0.075. A fixed cancellation threshold that is easier to trigger when displayed depth is large therefore cannot account for the asymmetry.

The third comparison uses information regimes. If the pattern were mainly stale-quote sniping, the cross-side trigger should be strongest when information shocks are most important. The data show the opposite. The CSTE is 0.158 in calm episodes and 0.117 in high-information episodes. The trigger is stronger when the information-shock interpretation is weaker.

The fourth comparison uses market-level variation in visibility. If the binding margin were cancellation speed, phantom-depth frequency should load on cancellation intensity. It does not. The MIDAS cancel-to-trade ratio has little stable explanatory power. Hidden-order activity, by contrast, strongly predicts lower phantom-depth prevalence, both across securities and within securities over time. This is the cleanest market-level distinction. Sniping operates through speed and information. Display-and-trigger manipulation operates through the informativeness of visible liquidity.

The table separates necessary timing evidence from discriminating evidence. The Stackelberg lead is consistent with the model, but it does not rule out stale-quote sniping because informative flow can also precede legitimate cancellation. The other rows are more restrictive. Displayed depth is followed by more cross-side trading. The cancellation response to cross-side flow becomes much stronger while the depth spike is live. The trigger is stronger when information shocks are less prominent. Phantom-depth prevalence falls when displayed liquidity is a weaker public signal.

A second benign alternative is cancel-and-reprice, or quote chasing. A liquidity provider with a resting order on the displayed side may cancel and repost at a better price once cross-side trading moves the market. This behavior can generate a positive cross-side trigger, a lead from cross-side flow to cancellation, and a live dislocation without manipulation. We test this alternative using the spike-side best-price path after collapse, re-extracted from the NBBO for all 1,632 trading days.

Table 4: **Sniping discrimination diagnostics.** This table reports diagnostics that separate strategic display-and-trigger behavior from ordinary stale-quote sniping. The first three rows report the sniping-discrimination tests described in this section. The visibility row reports the within-stock panel coefficient on MIDAS hidden-order share in the log-events specification with stock and month fixed effects. The Stackelberg-lead row reports the $h = 200$ ms coefficient from Section 6.4. It is consistent with both manipulation and sniping and is included as a necessary timing condition.

Diagnostic	Estimate	Model	Sniping	Reading
Elicitation. In-episode vs. pre-spike cross-side rate	6.90 trades/s ($t = 63$)	> 0	≈ 0	Manipulation
Cross-to-fill asymmetry. Spike vs. pre-spike ratio	4.5×	> 1	≈ 1	Manipulation
CSTE in calm (c) vs. information (i) regime	0.158 vs. 0.117	$c > i$	$c < i$	Manipulation
Visibility loading. Hidden share on log event count	-2.77 ($t = -8.77$)	< 0	≈ 0	Manipulation
Stackelberg lead at 200 ms	0.191 ($t = 556$)	> 0	> 0	Necessary

Each episode is classified by whether the spike-side best price improves relative to its pre-spike level within a short post-collapse window. Improvement is the signature of repricing. Failure to improve is the signature of withdrawal.

The test asks whether the trigger score predicts the live-window dislocation differently across the two classes. Cancel-and-reprice predicts that the trigger-dislocation relation should be concentrated in the reprice class. The evidence goes in the opposite direction. The trigger-dislocation slope is stronger in the withdrawal class, with an interaction t-statistic between -4.7 and -5.4 across one-, two-, and five-second windows. This cuts against the strong form of the repricing alternative, although the evidence should be read as suggestive rather than decisive. The trigger predicts dislocation in both classes and is only about one-fifth weaker under repricing, so the distinction is a gradient rather than a clean separation. A classification based on whether displayed depth reappears at the touch does not separate the two classes. Across terciles of the reprice signal, the slope is not strictly monotone. A sharper test would require order-level data that identify the reposting order directly. Appendix Table G.9 reports the interactions across windows and both classification axes.

The joint pattern is the source of identification. Each fact on its own has a possible non-manipulation explanation. Together, they point to a narrower mechanism. Displayed depth elicits cross-side trading. Cancellation responds much more strongly to cross-side activity than to fill-direction activity. The asymmetry is much larger when the displayed order is present. The trigger is stronger when the information-shock explanation is weaker. The strategy is less common when displayed liquidity is a weaker public signal.

5.5 Trigger score

The hazard model produces an episode-level trigger score, which is a summary of within-episode behavior, rather than a regulatory label, an intent measure, or a classifier trained on price outcomes. Episodes receive high scores when their cancellation timing resembles the display-and-trigger pattern.

We construct the score with five-fold cross-fitting at the episode level. All NBBO-update rows for a given episode are assigned to the same fold. For each fold k, we estimate the hazard model on the other four folds and predict cancellation probabilities for episodes in the held-out fold. The baseline score is the average predicted cancellation probability over the episode

$$\widehat{T}_i^{(-k)} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \widehat{\Pr}^{(-k)}(C_{i,t+\delta} = 1 \mid \mathcal{F}_{it}). \tag{23}$$

As a robustness check, we also use the maximum predicted probability within the episode,

$$\widehat{T}_i^{(-k,\max)} = \max_{t \in \mathcal{T}_i} \widehat{\Pr}^{(-k)}(C_{i,t+\delta} = 1 \mid \mathcal{F}_{it}). \tag{24}$$

The exclusion property is the central feature of the score. The first stage uses only within-episode variables, including cross-side and fill-direction trade counts, spreads, local volatility, adverse-selection returns, total trade intensity, and displayed depth. It does not use signed dislocation, signed reversal, post-collapse depth dynamics, post-collapse variance ratios, post-collapse effective spreads, or price-impact estimates. Those variables enter only as outcomes in Section 6.

The first stage has meaningful but imperfect predictive power. The out-of-fold AUC is 0.732, the Brier score is 0.218, and calibration is close. In each fold, the mean predicted cancellation probability is within 0.005 of the realized cancellation rate. The distribution of scores is right-skewed. The median is 0.42, the 90th percentile is 0.63, the 95th percentile is 0.74, and the 99th percentile is 0.93. About 6.4% of episodes have a score above 0.7. In the main second-stage tests, we use the continuous score. Binary cutoffs are used only for description.

Because the score is noisy, the second-stage tests are conservative in the usual measurement-error sense. If $\widehat{T}_i^{(-k)}$ is an imperfect proxy for the latent trigger state, the slopes on excluded outcomes are attenuated toward zero. Finding the predicted signs despite this noise is therefore more informative than finding them with a score trained directly on the outcomes.

The evidence in this section identifies a behavioral state from within-episode timing. Section 6 now asks whether that state has the consequences predicted by the model. It takes the cross-fitted trigger score and tests whether it orders excluded price, depth, variance, spread, timing, and price-impact outcomes in the model-implied directions.

6 Overidentifying Tests

The trigger score in Section 5 is constructed only from within-episode cancellation timing. It does not use post-event prices, post-collapse depth, variance ratios, effective spreads, or price-impact estimates. This section asks whether that behaviorally identified state orders these excluded outcomes in the directions implied by the model.

The logic is deliberately overidentified since, in isolation, any outcome could have another explanation. A dislocation may reflect ordinary information arrival. A variance ratio may move because

Table 5: **Dislocation and reversal as functions of the cross-fitted trigger score.** This table reports second-stage regressions of signed dislocation and signed reversal on the cross-fitted trigger score. The score $\widehat{T}_i^{(-k)}$ comes from equation (23), standardised to unit variance. Dependent variables in basis points. Controls include pre-event spread, pre-event local volatility, log pre-event volume, and $\log \text{Duration}_i$. Standard errors clustered by symbol in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$ in a two-sided test. Restriction 3 predicts $\delta_D > 0$ and $\delta_R \geq 0$.

	D_i (Dislocation)		R_i (Reversal)	
	(1)	(2)	(3)	(4)
$\widehat{T}_i^{(-k)}$	2.12***	1.85***	0.084	0.165**
	(0.131)	(0.113)	(0.072)	(0.083)
D_i				0.0387*
				(0.021)
Symbol FE	No	Yes	No	Yes
Controls	Yes	Yes	Yes	Yes
N	340,884	340,884	340,884	340,884
t-statistic on $\widehat{T}_i^{(-k)}$	16.2	16.4	1.2	2.0

of transient liquidity shocks. Spreads may widen for reasons unrelated to phantom depth. The test is whether the same trigger score, built without these outcomes, predicts all of them in the pattern implied by the display-and-trigger mechanism.

6.1 Price outcomes

We begin with the model's central price implication. The manipulator profits through the price dislocation created while the displayed order is live. The model does not predict that high-trigger episodes should be followed by especially negative post-collapse price reversals. Once the cross-side response has arrived and the order has been withdrawn, the later price path reflects ordinary information and liquidity provision, as well as any unwinding of the displayed component.

We estimate

$$D_i = \alpha_s + \delta_D \widehat{T}_i^{(-k)} + X_i'\beta_D + \varepsilon_i^D, \tag{25}$$

$$R_i = \alpha_s + \delta_R \widehat{T}_i^{(-k)} + \rho D_i + X_i'\beta_R + \varepsilon_i^R, \tag{26}$$

where D_i is the signed dislocation and R_i is the signed post-collapse reversal, both measured in basis points. The fixed effect α_s is at the symbol level. The vector X_i includes the pre-event spread, pre-event local volatility, log pre-event volume, and log duration.

Restriction 3 predicts $\delta_D > 0$ and $\delta_R \geq 0$. The first inequality is the main price-outcome restriction. Episodes with a higher trigger score should exhibit larger live dislocations in the direction induced by the displayed order. The second inequality is just as important because it separates our test from a reversal-based interpretation. A higher trigger score should not identify episodes by making their post-collapse reversals more negative. The mechanism is about the price dislocation created while the displayed signal is live, not about selecting episodes that happen to reverse afterward.

Table 6: **Outcomes by trigger-score quartile.** This table reports mean dislocation and reversal across within-symbol-year quartiles of the cross-fitted trigger score. Quartiles are formed within symbol-year to absorb cross-stock heterogeneity in the level of $\widehat{T}_i^{(-k)}$. The ratio of Q_4 to Q_1 dislocation is 5.6. Reversal becomes less negative monotonically with the trigger score, consistent with $\delta_R \geq 0$. The descriptive transfer equals the absolute live-window dislocation in dollars per share multiplied by cross-side volume during the spell. It sets the manipulator's counterparty share to one and should be read as an upper-bound scale calculation, not as profit or welfare loss.

Quartile of $\widehat{T}_i^{(-k)}$	Q_1	Q_2	Q_3	Q_4
Mean dislocation D_i (bps)	1.12	1.53	3.16	6.30
Mean reversal R_i (bps)	−1.95	−1.28	−1.07	−0.92
Cross-side fill probability	0.246	0.437	0.467	0.581
Mean depth spike (shares)	9,010	9,253	9,976	13,068
Mean cross-side volume (shares)	15,685	18,603	21,862	29,520
Descriptive transfer ($ per episode)	79	114	157	272

The non-negative reversal prediction should therefore be read as a selection restriction rather than as a causal claim about post-collapse prices. High-score episodes are those in which cancellation occurs after the displayed order has elicited the predicted response and before the price path has necessarily retraced. Conditional on the live dislocation D_i, this selection should not produce a more negative reversal. If anything, the high-score subsample should be less negatively skewed in the post-collapse window than a sample defined directly by reversal.

Table 5 reports the estimates. The trigger score strongly predicts dislocation. Without symbol fixed effects, the slope is 2.12 basis points per standard deviation of the score, with $t = 16.2$. With symbol fixed effects, the slope is 1.85 basis points, with $t = 16.4$. The within-symbol estimate is slightly smaller, as expected if securities differ in their baseline dislocation levels, but the sign and precision are essentially unchanged.

The reversal estimates look very different. The slope is small and weakly positive. Without conditioning on dislocation, the estimate is 0.084 basis points with $t = 1.2$. With symbol fixed effects and D_i included, the estimate is 0.165 basis points with $t = 2.0$. Thus, high-trigger episodes do not produce the large negative reversals emphasized in reversal-based tests. They produce larger live dislocations and, if anything, less negative post-collapse reversals.

The quartile evidence makes the mechanism visible in levels. Table 6 sorts episodes into trigger-score quartiles within symbol-year. Mean dislocation rises monotonically from 1.12 basis points in the bottom quartile to 6.30 basis points in the top quartile. The top-to-bottom ratio is 5.6. Cross-side fill probability also rises sharply, from 0.246 to 0.581, and the mean depth spike increases from 9,010 to 13,068 shares. Reversal moves in the opposite direction from a reversal-only interpretation. It becomes less negative as the trigger score rises, from −1.95 basis points in the bottom quartile to −0.92 basis points in the top quartile.

The first four rows describe a coherent display-and-trigger state. High-score episodes display more depth, attract more cross-side fills, and generate larger live dislocations. They are not followed by stronger reversals. This is the price mechanism in the model. The relevant price movement is the displacement created while the displayed signal is live, not a post-collapse reversal selected after the fact.

The last two rows give a descriptive measure of economic scale. For each episode, we multiply the absolute live-window dislocation in dollars per share by the cross-side volume that trades during the spell. This is the dollar amount transacted at the dislocated price if the trader posting the display is the counterparty to all induced cross-side flow. That counterparty assumption is deliberately generous. The number is therefore an upper-bound scale calculation, in the same spirit as the aggregate ceiling in Section 4.4, not a profit or welfare estimate. The transfer rises monotonically with the trigger score, from about 79 dollars per episode in the bottom quartile to about 272 dollars in the top quartile, a factor of 3.5. The behavioral state identified from cancellation timing is therefore associated with a larger dollar footprint as well as a larger price dislocation.

Two qualifications are important. First, the displayed order's execution probability does not fall with the trigger score, so high-trigger episodes are not simply episodes in which the trader more reliably escapes execution. Second, when the displayed order does execute, its mark-to-collapse return is positive rather than negative, averaging about 7 basis points. The data therefore do not show a live-window execution loss on the displayed order. A positive mark on executed displayed liquidity is consistent with the manipulation mechanism, but it is also consistent with benign informed liquidity provision. This ambiguity is exactly why the identification relies on the joint within-episode timing restrictions rather than on realized execution outcomes alone.

The directional structure is also visible in aggregate order flow, even without account identities. In the post-collapse window, net signed trade volume runs opposite to the displayed side. Episodes with a displayed bid are followed by net selling of about 1,500 shares, while episodes with a displayed ask are followed by net buying of about 1,100 shares. This is the aggregate signature of displaying liquidity on one side while trading pressure moves on the other. Because TAQ does not link cancelled orders to the owner's later executions, this is not a direct account-level measure of the manipulator's position. It is a descriptive pattern consistent with the display-and-trigger mechanism.

Table 7 extends the dislocation regression to longer horizons. The slope on the trigger score remains stable from the instant of collapse through the end of the day, ranging from 1.84 to 1.99 basis points per standard deviation. The t-statistics decline at longer horizons because return variance increases with horizon, but the point estimates do not disappear.

These longer-horizon estimates should not be read as evidence that the displayed order permanently moves fundamentals. The event population is selected on cancellation and, therefore, on states in which the price has moved in the spike direction and has not immediately reverted. This is the symmetric counterpart of the reversal-selection problem in Section 5.1. The persistence is therefore descriptive of the selected population. Identification comes from within-episode behavior and the joint sign pattern, not from persistence alone.

6.2 Depth, variance, and spread outcomes

We next use the trigger score to test the model's predictions of sharper depth withdrawal at collapse, lower post-collapse variance ratios, and wider subsequent spreads. These outcomes are not used to build the trigger score.

Table 7: **Multi-horizon dislocation as a function of the cross-fitted trigger score.** This table reports trigger-score slope estimates on signed dislocation across five horizons. Each column reports the slope $\hat{\delta}_D^{(h)}$ on the standardized cross-fitted trigger score $\widehat{T}_i^{(-k)}$ in a regression of signed price displacement from the pre-spike midprice to the midprice at horizon h on the score, symbol fixed effects, and the controls of Table 5. The instant column reproduces Table 5, column (2); the five-minute column uses the sum of signed dislocation and signed reversal. The thirty-minute, sixty-minute, and end-of-day columns are estimated on a same-day NBBO extraction covering the full January 2017 through June 2023 sample (527,067 merged episodes, all 1,632 trading days). End-of-day is the displacement from the pre-spike midprice to the last consolidated NBBO at or before 16:00 ET on the event day. Standard errors clustered by symbol.

Horizon h	Instant	5 min	30 min	60 min	EOD
$\hat{\delta}_D^{(h)}$ (bps)	1.84	2.12	1.91	1.92	1.99
t-statistic	16.3	13.6	9.4	7.1	5.5
N (episodes)	340,884	340,884	340,906	340,906	340,906

We first estimate the event-time depth panel

$$\text{Depth}_{i,k} = \alpha_i + \sum_{m \neq -1} \beta_m \mathbf{1}\{k = m\} + \sum_{m \neq -1} \pi_m^* \mathbf{1}\{k = m\} \widehat{T}_i^{(-k)} + \varepsilon_{i,k}, \tag{27}$$

where k indexes event-time bins around collapse and α_i is an episode fixed effect. The sample contains 5,215,847 depth-panel observations from 474,169 scored episodes.

The main effects trace the average shape of a depth spike. Normalized depth rises modestly before collapse, jumps to 1.76 in the pre-collapse bin, falls to 1.03 at the collapse bin, and then settles slightly below its pre-event level. Trigger-score interactions show how this profile changes with the behavioral state. The peak depth is almost unchanged across trigger levels. The interaction at the pre-collapse bin is small, $\pi_{-1}^* = -0.003$ with $t = -2.0$. The collapse-bin interaction is sharply negative, $\pi_0^* = -0.019$ with $t = -22.4$. Thus, high-trigger episodes do not simply have a higher normalized peak depth. They have a sharper withdrawal at collapse. The implied withdrawal interaction $\pi_{-1}^* - \pi_0^*$ is positive and strongly significant, with the joint-test estimate reported in Table 10.

The post-collapse depth interactions are positive. They rise from $\pi_2^* = 0.004$, with $t = 5.1$, to $\pi_4^* = 0.009$, with $t = 12.5$, and then stabilize. High-trigger episodes therefore rebuild displayed depth more quickly after the cancellation cluster. This is the depth pattern implied by the model. The transient component is withdrawn sharply, while genuine resting liquidity returns after the signal has disappeared.

We next estimate variance-ratio regressions,

$$VR_i^{\text{rel}}(\ell/5\text{s}) = \alpha_s + \delta_{VR}(\ell) \widehat{T}_i^{(-k)} + X_i' \beta_{VR} + \varepsilon_i^{VR}, \tag{28}$$

for horizons $\ell \in \{10, 30, 60, 300\}$ seconds, where

$$VR_i^{\text{rel}}(\ell/5\text{s}) = \frac{\text{Var}(r_{i,t}^{(\ell)})}{(\ell/5)\text{Var}(r_{i,t}^{(5\text{s})})}.$$

Table 8: **Variance ratios as functions of the cross-fitted trigger score.** This table reports trigger-score slope estimates on relative variance ratios across four horizons. Relative variance ratios are anchored at five seconds and are defined as $VR_{\text{rel}}(\ell/5\text{s}) = \text{Var}(r_t^{(\ell)})/\{(\ell/5)\text{Var}(r_t^{(5\text{s})})\}$. Under a random walk, the ratio equals one. The regression includes symbol fixed effects and the standard controls of Section 6.1. Standard errors are clustered by symbol. Restriction 3 predicts $\delta_{VR}(\ell) < 0$ at all horizons.

Horizon ℓ (s)	10	30	60	300
Population mean $VR_{\text{rel}}(\ell/5\text{s})$	1.144	0.659	0.375	0.083
Symbol FE specification				
$\widehat{\delta}_{VR}(\ell)$	−0.089	−0.069	−0.043	−0.0092
t-statistic	−7.73	−6.52	−6.16	−5.64
Quartile means				
Q_1			0.410	0.092
Q_2			0.374	0.082
Q_3			0.355	0.077
Q_4			0.314	0.069

Under a random walk, the relative variance ratio equals one. Restriction 3 predicts $\delta_{VR}(\ell) < 0$, because high-trigger episodes should concentrate more of their price movement near the collapse rather than generating persistent price variation over longer horizons.

Table 8 shows that the variance-ratio slopes are negative at every horizon. The estimates are −0.089 at ten seconds, −0.069 at thirty seconds, −0.043 at sixty seconds, and −0.0092 at five minutes. All four estimates are strongly significant. The quartile means show the same monotonic pattern. At the 60-second horizon, the variance ratio falls from 0.410 in the bottom quartile to 0.314 in the top quartile. At the five-minute horizon, it falls from 0.092 to 0.069.

There is no contradiction between lower variance ratios and the weak-reversal result in Table 6. The variance ratio is a second-moment object. A lower value means that price variation is concentrated near the collapse rather than continuing to accumulate over longer horizons. Reversal is a first-moment object. It asks whether the post-collapse drift runs against the live-window dislocation. A high-trigger episode in which the price adjusts quickly around collapse and then remains near the new level has a low variance ratio and no reversal. That is the pattern in the data. A temporary-impact story would instead pair the low variance ratio with negative signed drift, which is exactly what the reversal restriction tests and does not find.

A Hasbrouck-style permanent-transitory decomposition gives the same qualitative result. The within-symbol slope for the information-share component is −0.00854, with $t = -6.20$, so high-trigger episodes have proportionally more transitory return variance. This evidence is measured in the cancellation-selected sample and is consistent with the model's prediction that phantom depth creates a temporary distortion in price discovery.

The spread evidence points in the same direction. Effective spreads widen by 0.291 basis points from before to after the event, with $t = 14.84$. The trigger score also predicts larger spread increases. The slope is 0.164 in the pooled specification, with $t = 8.2$, and 0.080 with symbol fixed effects, with $t = 4.8$.

The within-symbol estimate is the more useful benchmark because ΔES_i has heavy tails. We

therefore winsorize the spread change at the 0.5/99.5 percentiles, matching the treatment of dislocation and reversal in Section 6.1. Without this adjustment, a small number of episodes with extreme post-event midpoint behavior dominate the per-symbol estimates. With the same treatment applied across outcomes, the spread moment reinforces the depth-withdrawal and variance-ratio evidence.

6.3 Corroborating evidence from IV price impact

The model also predicts that the market appears deeper while phantom depth is live. In a Kyle-style pricing rule, inverse market depth is the price impact per unit of signed order flow. If displayed depth inflates apparent liquidity, price impact should be lower during the live window than before the spike.

We find evidence consistent with this depth-inflation channel. We sort episodes by the cross-fitted trigger score and estimate price impact at the 250 millisecond frequency, instrumenting stock-level signed flow with contemporaneous SPY midquote returns. The live-window price impact falls below its pre-spike benchmark in the quartiles where the behavioral evidence is strongest. In the top trigger-score quartile, the ratio of live to pre-spike price impact is 0.53, with $t = -4.23$ against a null of one. In the third quartile, the ratio is 0.58, with $t = -3.21$. The drop is absent in the bottom two quartiles.

This evidence should be read as corroborating rather than central to identification. The IV estimate requires the assumption that index-level returns affect the individual-stock midpoint only through that stock's signed flow. That exclusion restriction may be imperfect if market makers reprice individual securities directly from index futures or ETF quotes. For this reason, the price-impact test is not used to define the trigger state. It is an excluded outcome that moves in the direction predicted by the model. Appendix G.1 reports the full specification, the quartile estimates, and the associated caveats.

6.4 Stackelberg timing

Restriction 2 is a timing restriction. The model does not merely predict that cross-side activity and cancellation are correlated. It predicts an order of events. The displayed order is posted first. Cross-side traders respond while the display is live. The manipulator then observes that response and cancels. Thus, cross-side activity should arrive before cancellation. Cancellation should not, in a symmetric way, predict future cross-side activity.

This distinction matters because contemporaneous correlation is easy to generate. A news shock, volatility burst, or liquidity shock can produce both aggressive trades and cancellations in the same short interval. Such common shocks do not by themselves imply manipulation. The model requires a directional pattern: the relevant cross-side flow arrives first, and the cancellation follows.

We test this directionality using the lead-lag specifications introduced in Section 5.4. The lead specification asks whether cross-side activity at time t predicts cancellation at $t + h$:

$$\Pr(C_{i,t+h} = 1 \mid \mathcal{F}_{it}) = \alpha + \gamma^{\rightarrow}(h)\widetilde{x}^{\text{cross}}_{it} + X'_{it}\beta + \varepsilon_{it}.$$

The reverse specification asks whether cancellation at time t predicts cross-side activity at $t + h$:

$$\Pr(\widetilde{x}^{\text{cross}}_{i,t+h} > 0 \mid \mathcal{F}_{it}) = \alpha + \gamma^{\leftarrow}(h)C_{it} + X'_{it}\beta + \varepsilon_{it}.$$

The controls are the same in both equations. Under the Stackelberg timing of the model, $\gamma^{\rightarrow}(h)$ should be positive at the relevant sub-second horizons, while the reverse relation should be weak. The identifying content is not simply that trades and cancellations occur near each other. It is that cross-side activity leads cancellation rather than the other way around.

Table 9 reports the lead coefficients for

$$h \in \{50, 100, 200, 500, 1000\} \text{ milliseconds.}$$

The forward relation is positive at every horizon. The coefficient is 0.178 at 50 milliseconds, 0.184 at 100 milliseconds, 0.191 at 200 milliseconds, 0.187 at 500 milliseconds, and 0.175 at 1,000 milliseconds. The corresponding t-statistics range from 498 to 556 on the cancel-event panel. These estimates formalize the event-time pattern in Figure 7. Cross-side activity rises before the cancellation cluster, and cancellation follows shortly afterward.

The reverse relation goes in the opposite direction. We compare the future cross-side arrival rate after cancellation instants with the corresponding rate after non-cancellation instants in the same episode, controlling for elapsed time, remaining time in the episode, and the local market state. The reverse coefficient $\gamma^{\leftarrow}(h)$ is negative at every horizon. It equals -0.032 at 100 milliseconds, -0.031 at 500 milliseconds, and -0.029 at one second, with t-statistics around -10. Cancellation is followed by less cross-side activity, not more. The forward lead is therefore much larger than the reverse relation and has the opposite sign. The asymmetry reflects a directional sequence rather than common high-frequency comovement.

The negative reverse relation is concentrated in the episodes identified by the trigger score. Splitting the reverse coefficient by trigger-score quartile, the future cross-side response after cancellation is close to zero in the lowest-trigger quartile and becomes progressively more negative in higher quartiles. It reaches about -0.09 at 100 milliseconds in the top quartiles. Interacting cancellation with the out-of-fold trigger score gives coefficients of -0.278 at 100 milliseconds, -0.256 at 500 milliseconds, and -0.233 at one second, each with a t-statistic near -20. This pattern is exactly what the display-removal interpretation predicts. Cancellation removes the object that was eliciting cross-side flow, and the removal matters most in episodes where the display was doing the most work.

The heterogeneity is not mechanical mean reversion in cross-side flow. Adding an interaction between cancellation and the contemporaneous cross-side level leaves the trigger-score interaction intact and slightly larger, at -0.343 at 100 milliseconds. The contemporaneous-flow interaction enters with the opposite sign. The concentration of the reverse effect therefore tracks the behaviorally identified state rather than the level of flow itself.

Table 9: **Lead-lag asymmetry: cross-side activity precedes cancellation.** This table reports lead-lag regression coefficients of cancellation on cross-side activity across five horizons. The lead specification (21) regresses cancellation at $t+h$ on cross-side activity at t. Estimated on the full cancel-event panel of $N = 2{,}678{,}888$ rows. Pre and post rates are the cross-side trade-arrival indicators in the symmetric windows $[t-h, t)$ and $(t, t+h]$, respectively. The arrival flag is the per-tick increment in the 100 ms rolling cross-trade count, so pre and post are matched-width forward and backward windows on the same flag. Standard errors are clustered at the episode level. Restriction 2 predicts $\gamma^{\rightarrow}(h) > 0$ at the relevant Stackelberg horizon.

Horizon h (ms)	50	100	200	500	1000
Pre rate (per second)	0.470	0.499	0.533	0.581	0.628
Post rate (per second)	0.292	0.316	0.342	0.394	0.452
Lead $\gamma^{\rightarrow}(h)$	0.178	0.184	0.191	0.187	0.175
t-statistic	547	549	556	535	498

This is the form derived in Proposition 3. The conditional reverse lead equals

$$-\nu\widehat{D}\,[\kappa(t+h) - \kappa(t)]\,\widehat{\theta}_t.$$

It is proportional to the interim posterior on the elicited response, which is largest in the episodes selected by the trigger score. It vanishes when there is no response for the removal of the display to switch off. The near-zero reverse lead in the lowest-score quartile and the strongly negative reverse lead in the top quartiles are therefore the two empirical ends of the same model expression.

The coefficient $\gamma^{\rightarrow}(h)$ is remarkably stable across horizons, ranging from $\gamma^{\rightarrow}(1000) = 0.175$ at one second to $\gamma^{\rightarrow}(200) = 0.191$ at 200 milliseconds, with similarly large t-statistics throughout. The timing pattern is therefore not an artifact of a particular window length. Cross-side activity consistently predicts subsequent cancellation, and the relation is already visible at the finest horizon the consolidated feed can resolve. This is the same ordering shown in Figure 7, where the cross-side trade rate rises before displayed depth collapses.

The reverse direction completes the asymmetry. As reported above, the reverse coefficient $\gamma^{\leftarrow}(h)$ is negative at every horizon — cancellation is followed by *less* cross-side activity, not more — and the negative relation is concentrated in high-trigger episodes, which is the display-removal pattern derived in Proposition 3. Cross-side activity is the leading variable, cancellation is the responding variable, and the removal of the display switches the elicited flow off.

The lead-lag evidence should be read together with the cross-to-fill comparison in Section 5. Timing alone would not rule out every rational liquidity-provision story. A fast market maker may also cancel after informative flow. What the timing result adds is directionality. Combined with the fact that the response is much stronger for cross-side than fill-direction flow and is amplified during the displayed-depth spell, the lead-lag pattern supports the display-and-trigger interpretation rather than a generic contemporaneous trading shock.

6.5 Joint test and robustness

The preceding subsections examine the restrictions one at a time. That is useful for interpretation, but it is not the main test of the model. The model does not predict only that high-trigger episodes

have larger dislocations, or only that variance ratios fall, or only that cross-side activity leads cancellation. It predicts that all of these patterns should be organized by the same state. A single trigger score, estimated from within-episode cancellation timing alone, should order the excluded outcomes in the directions implied by the equilibrium.

This is the role of the joint test. It asks whether the sign pattern generated by the model holds as a system. We orient each moment so that the model predicts a positive standardized value and test whether the moments are jointly far from the null, in which case they carry no coordinated information on the model's signs.

The directional concordance statistic is

$$S_n = \sum_{j=1}^{J} z_j, \qquad z_j = \text{sign}_j \, \widehat{m}_j(\theta)/\widehat{\omega}_j(\theta),$$

where $\widehat{m}_j$ is the sample moment in component j of (16), $\widehat{\omega}_j$ is its standard error, and $\text{sign}_j = +1$ if the model predicts that moment to be positive and $\text{sign}_j = -1$ if the model predicts it to be negative. The null is $\mathbb{E}[z_j] = 0$ for all j, that is, the sample moments carry no coordinated information on the model's sign pattern. The alternative is $\mathbb{E}[z_j] > 0$ jointly. To address the concern that the test could be dominated by a single high-precision moment, we also report the weakest-moment statistic $\min_j z_j$, which is positive whenever every restriction is at least weakly in the predicted direction.

The sampling uncertainty of the supplementary concordance statistic S_n is estimated with an episode-level block bootstrap. We resample the 527,019 episodes with replacement 999 times, refit each second-stage regression on the resampled panel, and recompute the oriented statistics z_j in each draw. This preserves the cross-moment dependence induced by the shared cross-fitted trigger score. The first-stage hazard moments and the lead-lag cancellation moments are estimated on very large update-level panels and are computationally too costly to refit in each bootstrap replication. For the bootstrap calculation of S_n only, these moments enter through Gaussian draws using the standard errors from the original regressions.

Our main inference does not rely on the magnitude of S_n. The first-stage and lead-lag moments are estimated with very high precision, so any aggregate sum of oriented z-statistics is mechanically dominated by those components. We therefore treat S_n as a supplementary concordance measure. The headline evidence comes from a sign-only test on the discriminating outcomes that are excluded from the trigger score. We orient each moment so that the model predicts a positive sign and compute the benchmark probability 2^{-6} that the six discriminating outcome families would all be correctly signed under independent fair-coin signs. As a transparency check on the full moment vector, we also report a sign-randomization test that fixes the magnitudes $|z_j|$ and draws 9,999 random sign vectors over all seventeen oriented moments, following the moment-inequality logic of Andrews and Soares (2010).

Table 10 reports the component moments grouped by their role in the empirical design. The headline test uses the six discriminating outcome families that are excluded from the trigger score. All six have the predicted sign, and each is individually significant at the one-sided five percent

Table 10: **Joint directional concordance test of model-implied sign restrictions.** The table reports the model-implied sign restrictions grouped by their role in the test. The statistic z_j is the sign-oriented standardized moment, $z_j = \text{sign}_j \widehat{m}_j / \widehat{\omega}_j$, so the model predicts $z_j > 0$. The headline sign test is computed on the six discriminating outcome families excluded from the trigger score. These are signed dislocation, the variance-ratio family, effective spreads, depth withdrawal, depth recovery, and live-window price impact. The four variance-ratio horizons are treated as one family. Rows marked "defn." are first-stage hazard coefficients. Rows marked "nec." are necessary timing moments that are also consistent with stale-quote sniping. Rows marked "diag." report the weak reversal inequality. Rows marked "ref." are alternative specifications shown for transparency. The supplementary statistic S_n sums the oriented z_j's over all seventeen non-reference moments and is reported only for completeness.

Moment	Sign restriction	Estimate	SE	t-stat	z_j	Role
First-stage and definitional moments						
Hazard cross-side coefficient	> 0	1.048	0.003	328.9	328.9	defn.
Hazard cross-side minus fill-direction	> 0	0.930	0.003	282.2	282.2	defn.
Necessary timing condition						
$\gamma^{\rightarrow}$(50ms)	> 0	0.178	0.0003	547	547	nec.
$\gamma^{\rightarrow}$(100ms)	> 0	0.184	0.0003	549	549	nec.
$\gamma^{\rightarrow}$(200ms)	> 0	0.191	0.0003	556	556	nec.
$\gamma^{\rightarrow}$(500ms)	> 0	0.187	0.0004	535	535	nec.
$\gamma^{\rightarrow}$(1000ms)	> 0	0.175	0.0004	498	498	nec.
Discriminating overidentifying outcomes excluded from the trigger score						
δ_D pooled	> 0	2.12	0.131	16.2	—	ref.
δ_D symbol FE	> 0	1.85	0.113	16.4	16.4	Yes
δ_{VR}(10s/5s)	< 0	-0.089	0.011	-7.7	7.7	Yes
δ_{VR}(30s/5s)	< 0	-0.069	0.011	-6.5	6.5	fam.
δ_{VR}(60s/5s)	< 0	-0.043	0.007	-6.2	6.2	fam.
δ_{VR}(300s/5s)	< 0	-0.0092	0.002	-5.6	5.6	fam.
$\delta_{\text{ES},\Delta}$ pooled	> 0	0.164	0.020	8.2	8.2	Yes
$\pi^*_{-1} - \pi^*_0$ withdrawal	> 0	0.0146	0.0016	9.1	9.1	Yes
π^*_{+3} recovery	> 0	0.0068	0.0007	9.8	9.8	Yes
$\widehat{\Lambda}^{\text{pre}} - \widehat{\Lambda}^{\text{live}}$ in Q_4	> 0	2.56	1.08	2.4	2.4	Yes
Diagnostic weak inequality						
δ_R pooled, no D_i	≥ 0	0.084	0.072	1.2	—	diag.
δ_R symbol FE, with D_i	≥ 0	0.165	0.083	2.0	2.0	diag.
Headline sign test on the six discriminating outcome families						
Discriminating families correctly signed			6 of 6			
Sign-only probability, 2^{-6}			0.016			
Weakest discriminating moment, $\min_j z_j$			2.4			
Full exhibit and supplementary measures						
All 17 oriented moments correctly signed			Yes			
Sign-randomization p-value on the 17, $B = 9{,}999$			$< 1/9{,}999$			
S_n, sum of oriented z_j			3,370.1			
$S_n / \widehat{\text{SE}}(S_n)$, episode bootstrap			733.8			

level. Under a sign-only null in which each family is equally likely to be positive or negative, the probability of six correct signs is $2^{-6} \approx 0.016$. This fair-coin calculation is a benchmark rather than an exact joint p-value because the outcomes are predicted by the same cross-fitted trigger score and are not fully independent. The point is nevertheless transparent. The result does not rest on one high-precision moment. The weakest discriminating moment is the live-window price-impact estimate, with $z_j = 2.4$.

We separate these six families from the rest of the moment vector. The first two rows are first-stage hazard coefficients and are close to definitional because the trigger score is the fitted cancellation hazard. The five lead-lag horizons are a necessary timing condition, but they are not by themselves discriminating because stale-quote sniping can also produce flow before cancellation. The reversal slope δ_R is a diagnostic for the weak inequality $\delta_R \geq 0$. A non-negative estimate is useful because it rules out the reversal-only interpretation, but it is not treated as a risky confirming sign. For completeness, the full exhibit shows that all seventeen oriented moments have the model-implied sign. A sign-randomization test that fixes the magnitudes and randomly assigns signs places the observed concordance beyond all 9,999 random draws.

The supplementary sum of oriented statistics is very large, with $S_n = 3{,}370.1$ and a bootstrap standard error of 4.59. The resulting concordance statistic is 733.8. We report this number for completeness, but it should not be the main object of interpretation. Its magnitude reflects the very large samples used to estimate the first-stage hazard and lead-lag moments. The economically meaningful fact is simpler. The same behaviorally identified state orders excluded outcomes in the directions implied by the model.

The interpretation is straightforward. The trigger score is not merely picking up large or unusual events. It is picking up a particular behavioral state. High-score episodes have larger signed dislocations, do not have more negative post-collapse reversals, have lower variance ratios, exhibit sharper depth withdrawal and faster depth recovery, and show lower live-window price impact in the high-score subset. They also display the predicted Stackelberg timing, with cross-side activity arriving before cancellation. None of these outcomes is used to construct the score. Their joint ordering is therefore evidence that the within-episode timing state has the structure predicted by the model.

We close this section with three robustness checks. First, the cross-fitting design is not driving the results. Changing the number of folds from the baseline $K = 5$ to $K = 3$ or $K = 10$ leaves the main estimates essentially unchanged. The cross-to-fill semi-elasticity ratio remains close to 8.9, second-stage slopes move by less than five percent, and the joint sign pattern is unchanged.

Second, the results do not depend on averaging predicted cancellation probabilities over the episode. Replacing the average score in (23) with the maximum predicted cancellation probability in (24) gives second-stage estimates within one standard error of the baseline. The two scores are highly correlated, with $\rho \approx 0.9$. The maximum-score version is useful because it emphasizes the peak moment of strategic cancellation rather than the average hazard throughout the spell. The fact that both measures produce the same sign pattern suggests that the results are not an artifact of the aggregation rule.

Third, the trigger score is not just a proxy for spike size. We residualize $\widehat{T}_i^{(-k)}$ with respect to spike magnitude, duration, and pre-event volatility. These variables explain only 3.97% of the score variation. The residualized score, $\widehat{T}_i^{(-k),\perp}$, reproduces the full sign pattern in the second stage. By contrast, a placebo based on raw spike size reproduces the dislocation result but fails to reproduce the reversal direction and variance-ratio predictions. Large depth spikes can mechanically

be associated with larger price movements. They do not by themselves generate the full pattern of excluded outcomes. The overidentifying evidence comes from behavioral timing rather than from displayed-depth size.

Two-way clustering by symbol and date also leaves the inference essentially unchanged, with t-statistics moving by less than five percent. The conclusion is robust to alternative cross-fitting folds, alternative score aggregation, residualization against observable spike characteristics, placebo spike-size tests, and alternative clustering.

The evidence in this section is internal to the detected episodes. It shows that a state identified from within-episode timing predicts excluded price, depth, variance, spread, timing, and price-impact moments in the directions implied by the model. The next question is whether the same mechanism appears in market-level variation. If phantom depth works by manipulating the informativeness of displayed liquidity, then its prevalence should vary with the informational content of visible depth.

7 External Validity and the Visibility Channel

The within-event evidence shows that a trigger state identified from cancellation timing predicts the price, depth, variance, spread, and price-impact patterns implied by the model. This section turns to the market-level prediction. The model separates two margins. The first is cancellation feasibility, denoted by q_j. The second is signal visibility, denoted by ν_j. Fast cancellation is a necessary technology for phantom-depth manipulation, but it is widely available in modern electronic markets. The more discriminating margin is whether displayed liquidity remains informative enough to move other traders.

The evidence supports this visibility channel. Section 7.1 shows that phantom-depth frequency loads strongly on hidden-order activity, both across securities and within securities over time, while the cancel-to-trade ratio has little stable explanatory power. Section 7.2 then studies the mid-2020 to early-2021 decline in phantom-depth frequency. The break evidence is not a stand-alone identification strategy since CAT implementation, enforcement, COVID normalization, and changes in market activity are close in calendar time. Instead, we use it as corroborating evidence. The decline is concentrated among high-visibility names and around changes in surveillance and reporting infrastructure, consistent with the same channel that appears in the cross-sectional and within-security evidence.

7.1 Visibility is the binding margin

The model writes the effective manipulation technology for security j as

$$Q_j = q_j \times \nu_j,$$

Table 11: **Cross-sectional determinants: cancellation feasibility vs. visibility.** This table reports cross-sectional regressions of average phantom-depth frequency on the cancel-to-trade ratio and the hidden-order rate. The dependent variable is average phantom-depth events per security-day. The hidden-order rate is an inverse proxy for signal visibility. The cancel-to-trade ratio is statistically insignificant and changes sign with controls. The hidden-order rate is strongly negative in all specifications. Standard errors are HC1 robust.

	(1) Baseline	(2) + ETP control	(3) Stocks only
log(CancelRatio)	298	−79	−112
	(437)	(426)	(493)
Hidden rate (ν^{-1})	−34,258***	−46,002***	−45,527***
	(8,148)	(8,605)	(9,204)
log Volume	2,459***	3,038***	3,283***
	(441)	(471)	(511)
Mcap rank	−569***	−851***	−873***
	(138)	(150)	(164)
ETP indicator		−5,047***	
		(1,217)	
N	494	494	465

where q_j captures cancellation feasibility and ν_j captures signal visibility. Hidden and reserve liquidity provide a natural inverse proxy for ν_j. When more trading interest is non-displayed, the visible book becomes a noisier measure of true supply, demand, and private information. A large displayed order should then have less influence on other traders, making phantom-depth manipulation less useful. The empirical prediction is that phantom-depth frequency should be lower when hidden-order activity is high, and should not be explained mainly by cancellation intensity once visibility is controlled for.

We begin with the cross-section by estimating the following panel specification

$$\text{Freq}_j = \alpha + \beta_q \log(\text{CancelRatio}_j) + \beta_\nu \text{HiddenRate}_j + X_j'\Gamma + \varepsilon_j. \tag{29}$$

The dependent variable is the average number of phantom-depth events per security-day across the 494 securities in the sample. The controls are log daily share volume, market-capitalization rank, and an exchange-traded-product (ETP) indicator. The ETP control is included because exchange-traded products can have elevated cancellation rates for reasons related to authorized-participant arbitrage rather than displayed-depth manipulation.

Table 11 shows the contrast. The hidden-order rate has a strongly negative coefficient in every specification, with t-statistics between -4.20 and -5.35. Securities in which more trading interest is non-displayed have fewer phantom-depth episodes. The cancel-to-trade ratio is statistically indistinguishable from zero in every specification, changes sign when the ETP control is added, and remains null when ETPs are excluded. Even though the cancel-to-trade ratio ranges from 9.2 to 609.9 across the cross-section, there is no evidence that more cancellation-intensive securities have more phantom-depth episodes.

The evidence points to signal visibility as the binding economic margin. Interventions that change how much weight traders place on displayed liquidity therefore bear more directly on the

mechanism than interventions that treat spoofing only as a problem of cancellation speed.

The cross-section is informative but does not remove permanent differences across high-hidden and low-hidden securities. We therefore test the same prediction within securities over time. The security-month panel is

$$\text{Freq}_{s,m} = \alpha_s + \beta Z_{s,m-1} + X'_{s,m}\Gamma + \varepsilon_{s,m}, \tag{30}$$

where s indexes securities and m indexes calendar months. The dependent variable is phantom-depth frequency for security s in month m, and $Z_{s,m-1}$ is either the one-month lagged log cancel-to-trade ratio or the lagged hidden-order rate. The lag prevents contemporaneous phantom-depth episodes from mechanically entering the right-hand-side market-quality measure. Controls include monthly volume, realized volatility, and quoted spreads. Security fixed effects absorb permanent cross-security differences. Month fixed effects absorb aggregate market conditions, enforcement climate, technology trends, and the common component of the regime change studied in Section 7.2. Appendix Table G.5 reports the estimates.

The two right-hand-side variables behave very differently, and the distinction between specifications is important. Without the visibility control, the cancel-to-trade ratio is unstable. It is positive in pooled specifications, but it can become strongly negative once fixed effects are included and the variable is used on its own (Appendix Table G.5). Once hidden-order activity is included, however, the cancel-to-trade ratio has no positive explanatory power (Appendix Table G.7). The hidden-order rate, by contrast, is negative and significant throughout. Its coefficient is -7.84 in the pooled specification, with $t = -13.18$; -10.85 with security fixed effects; and -2.77 with both security and month fixed effects, with $t = -8.77$. When a given security becomes more hidden over time, its phantom-depth frequency falls. The within-security evidence therefore reinforces the cross-section. The evidence does not support a pure cancellation-feasibility channel; phantom depth is more prevalent when displayed liquidity is a stronger public signal.

A natural concern is that raw event frequency counts every mechanically detected episode, even though not every detected episode is manipulation. If the visibility loading reflected generic displayed-depth dynamics rather than the strategic state, it should weaken once the dependent variable focuses on episodes most consistent with display-and-trigger behavior. We therefore re-estimate both the cross-section and the within-stock panel using two trigger-weighted outcomes. The first is the count of episodes whose cross-fitted trigger score falls in the top quartile within symbol-year. The second is the sum of trigger scores.

The hidden-order rate remains strongly negative in all specifications. In the cross-section, the hidden-order coefficient is significant with t-statistics between -4.1 and -5.9 across the baseline, ETP-dummy, and ETP-excluded specifications. The cancel-to-trade ratio remains insignificant, as in the raw-count regression. In the within-stock panel with security and month fixed effects, the hidden-order coefficient is -1.28 for the top-quartile-trigger count, with $t = -6.4$, and -1.96 for the sum of trigger scores, with $t = -7.9$. It is also negative and significant in the pooled and security-fixed-effects specifications.

The visibility result is therefore not an artifact of counting all mechanically detected events. It

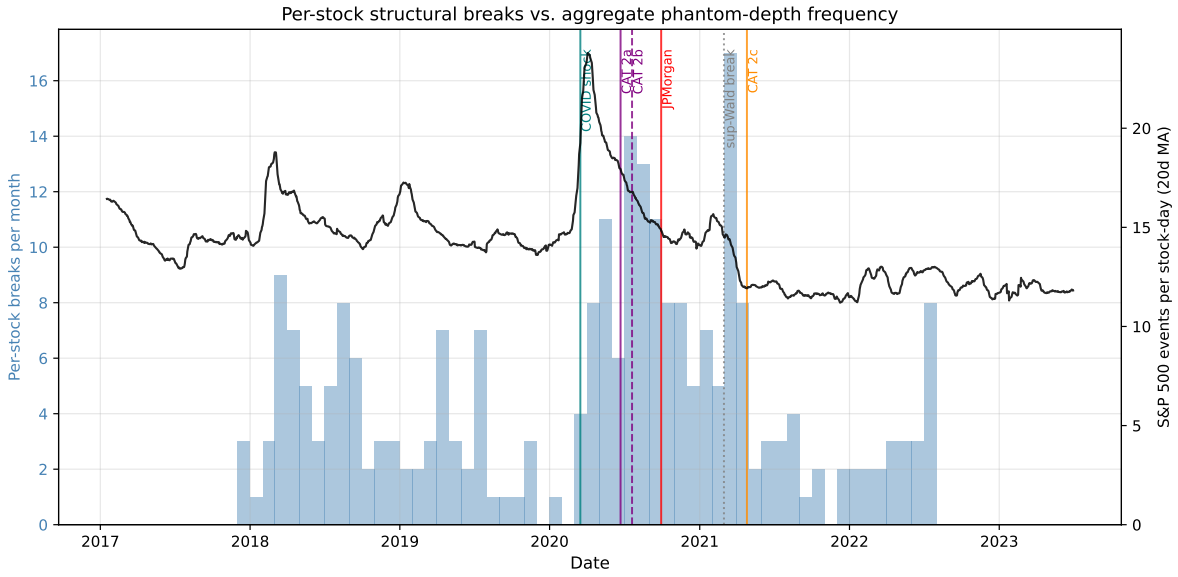


Figure 8: Phantom-depth event frequency and per-stock structural break dates

The figure plots the 20-day moving average of S&P 500 events per security-day (black line, right axis) and the monthly histogram of per-stock Quandt-Andrews sup-Wald break dates for securities with statistically significant downward breaks at $p < 0.05$ (blue bars, left axis). Vertical lines mark the COVID-19 volatility peak on March 16, 2020, CAT Phase 2a on June 22, 2020, CAT Phase 2b on July 20, 2020, the JPMorgan settlement on September 29, 2020, CAT Phase 2c on April 26, 2021, and the aggregate sup-Wald break date in March 2021. The histogram shows that the aggregate decline is the pooled effect of staggered security-level transitions concentrated in mid-2020 through spring 2021.

survives when prevalence is measured using the episodes that the trigger score identifies as most consistent with the strategic state. By contrast, the cancel-to-trade ratio never takes the positive sign predicted by a pure cancellation-feasibility channel. It is insignificant in the cross-section under every dependent variable. In the panel, it is null on the full detected set and weakly negative on the scored sample. This change tracks the move to the smaller sample on which the trigger score is defined rather than the trigger weighting itself. Appendix Tables G.6 and G.7 report the estimates. Section 7.2 applies the same reweighting to the time-series break test and reaches the same conclusion.

7.2 Time-series evidence from the 2020 regime change

Aggregate phantom-depth frequency declines sharply between mid-2020 and early 2021. The 20-day moving average in Figure 8 falls from roughly sixteen events per security-day before the decline to roughly twelve afterward. A Quandt-Andrews sup-Wald test for an unknown aggregate break locates the break in March 2021, with statistic 43.6 against an Andrews five-percent critical value of about 8.9, $p \approx 2 \times 10^{-9}$, and a level shift of -4.1 events per security-day. The result is robust to dropping calendar year 2020, which gives a sup-Wald statistic of 35.4 and a level shift of -2.6. Thus, the break is not simply the COVID-volatility spike of March and April 2020.

The aggregate break date masks substantial heterogeneity. We estimate a Quandt-Andrews

Table 12: **Direction-differential bunching of per-stock break dates at CAT and JPMorgan.** The table reports the share of per-stock structural-break dates within a ±60-day window of each CAT and enforcement anchor, separately for securities with downward and upward breaks. Uniform-random-anchor null shares come from 9,999 Monte Carlo draws over the eligible window. The down-minus-up differential is significant at $p < 10^{-4}$ for every anchor. Because several anchor windows overlap, the rows should be read as descriptive bunching diagnostics rather than independent tests.

Anchor	Down share	Up share	Null share	Down p	Up p	Differential p down − up
CAT Phase 2a (22 Jun 2020)	0.172	0.047	0.072	$< 10^{-4}$	0.97	$< 10^{-4}$
CAT Phase 2b (20 Jul 2020)	0.172	0.030	0.072	$< 10^{-4}$	1.00	$< 10^{-4}$
JPMorgan (29 Sep 2020)	0.156	0.021	0.072	$< 10^{-4}$	1.00	$< 10^{-4}$
CAT Phase 2c (26 Apr 2021)	0.117	0.021	0.072	0.004	1.00	$< 10^{-4}$

sup-Wald break on each security's monthly phantom-depth-frequency series, using 15% trimming and a linear-trend baseline. Of the 494 securities, 256 have a statistically significant downward break at $p < 0.05$, 236 have a statistically significant upward break, and only 2 have no significant break. Because the procedure finds breaks of both signs in almost the entire panel, the raw count of downward breaks is not by itself the key evidence. The key evidence is directional. Downward break dates bunch near the CAT reporting phases and the JPMorgan settlement, whereas break dates do not.

The downward cohort's break-month distribution peaks in calendar 2020, with 89 breaks and a median month of July 2020. This timing is close to CAT Phase 2a, which began large-broker-dealer equity reporting on June 22, 2020, and CAT Phase 2b, which began large-broker-dealer options reporting on July 20, 2020. A second cluster of 54 breaks occurs in spring 2021, close to CAT Phase 2c, which added rejected-route reporting on April 26, 2021. Figure 8 makes this staggered structure visible.

Table 12 reports bunching shares within ±60-day windows around each anchor. The benchmark is a uniform-random-anchor null computed from 9,999 Monte Carlo permutations over the eligible window. The down-cohort share is 17.2% around Phase 2a, 17.2% around Phase 2b, 15.6% around JPMorgan, and 11.7% around Phase 2c, compared with a uniform null of 7.2%. The up-cohort shares are much smaller, between 2.1% and 4.7%. Because the Phase 2a, Phase 2b, and JPMorgan windows overlap, the anchor-specific shares should be read as descriptive bunching diagnostics rather than independent tests.

Two within-data comparisons help interpret the bunching. The first uses the upward-break cohort. The same calendar shocks, including COVID volatility, retail-flow growth, and market-structure changes, affect both cohorts. If the downward clustering near CAT and JPMorgan dates were only a generic calendar artifact, upward breaks should cluster at the same dates. They do not. At the ±60-day window, the upward-break share near each anchor is below the uniform null, while the down-minus-up differential is between 0.096 and 0.142, with permutation p-values below 10^{-4}.

The second comparison uses independent market-quality measures. Re-estimating the per-stock sup-Wald test on monthly MIDAS cancel-to-trade ratios gives 205 securities with significant downward breaks, but those breaks peak in calendar 2019, with median month July 2019, and show no

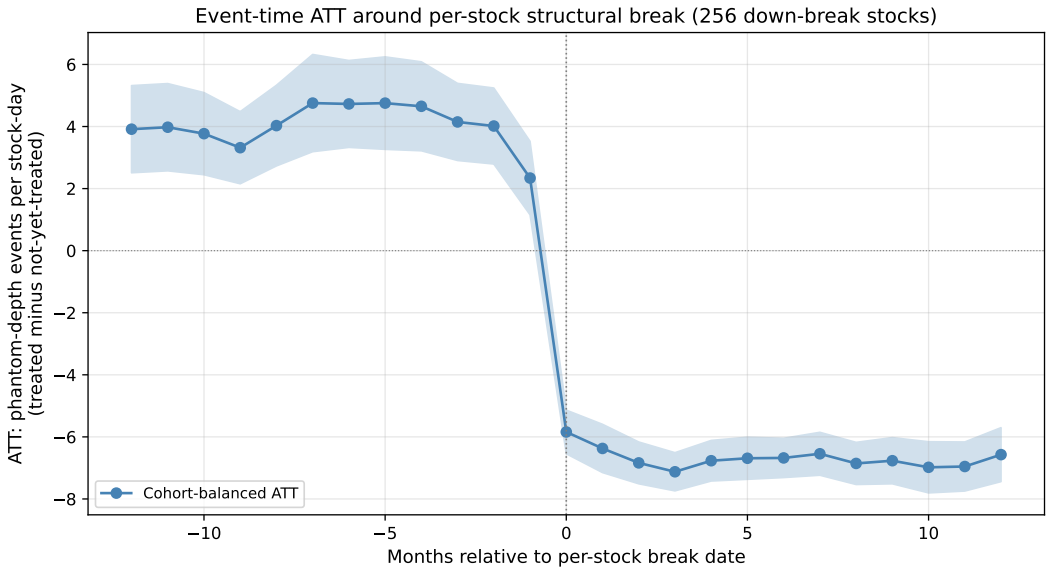


Figure 9: Cohort-balanced event-time ATT around per-stock structural breaks

The figure plots the cohort-balanced average treatment effect on the treated in event time around each security's own structural break. The treated population is the 256 securities with a significant downward structural break at $p < 0.05$. The ATT is the difference between the mean events per security-day for treated securities at event time τ and the calendar-matched mean for not-yet-treated securities, following the Callaway-Sant'Anna cohort-balanced estimator. The shaded band is a 95% confidence interval from a stratified-by-security bootstrap with $B = 300$. Treated securities have higher pre-break levels, so the design should be read as an event-time comparison of changes rather than equality of levels. The pre-period ATT slope is 0.07 per month, with $F = 2.55$ and $p = 0.11$.

clustering at any of the four anchors. The hidden-order-rate downward breaks peak in 2018 and cluster only around Phase 2a. Thus, the 2020 phantom-depth cluster is measure-specific. It does not simply reproduce the break timing of cancellation intensity or hidden-order activity.

A final check weights the break series toward the episodes most consistent with the strategic state. The baseline break series counts every detected episode, even though many detected episodes are ordinary displayed-depth events rather than manipulation. If the clustering reflected generic depth dynamics, it should weaken when the series is restricted to high-trigger episodes. We therefore re-estimate the per-stock sup-Wald test using the daily count of episodes whose cross-fitted trigger score falls in the top quartile within symbol-year, leaving the rest of the procedure unchanged.

The directional clustering survives. The downward-break cohort bunches at 0.156 within the ±60-day window around each of the three 2020 anchors and at 0.110 around Phase 2c, compared with the uniform-null benchmark of 0.072. The down-minus-up differential remains significant at every anchor, with permutation p-values of 0.002 or below. The 2020 to 2021 decline is therefore concentrated among the episodes the trigger score identifies as most consistent with display-and-trigger behavior, rather than among mechanically detected depth spikes in general. Appendix Table G.8 reports the trigger-weighted shares.

Figure 9 plots the cohort-balanced average treatment effect on the treated (ATT) around each security's own break. The ATT at $\tau = 0$ is -5.84 events per security-day, and the post-break ATT averages -6.7 over $\tau \in [+1, +12]$, with t-statistics below -15 throughout. Treated securities have

higher pre-break event frequency than not-yet-treated securities, so the identifying comparison is not equality of pre-period levels. It is the absence of a differential pre-trend before the break and the discrete drop afterward. The pre-period ATT slope is 0.07 per month, with $F = 2.55$ and $p = 0.11$. A placebo that shifts each security's break date back by 24 months gives an essentially null post-event ATT of +1.79, compared with an actual ATT difference of −10.95. The 2020 subcohort alone has an upward pre-period slope, consistent with the selection of high-frequency names into the break cohort.

The composition of the cohort that transitions closest to the aggregate break provides the link to the visibility channel. Among downward-break securities with complete pre-2021 covariates and break dates in the binning window, we classify 253 into three calendar bins: BEFORE, for breaks in 2018–2020, with $n = 176$; AROUND, for breaks in January–June 2021, with $n = 42$; and AFTER, for breaks in 2021H2–2022, with $n = 35$. The AROUND bin straddles the aggregate sup-Wald date and CAT Phase 2c. A multinomial logit of cohort membership on ex-ante characteristics, sector indicators, and an ETP indicator gives a joint Wald statistic of 29.34 for the AROUND-versus-BEFORE block, with $p = 0.006$. Three predictors are individually significant: lower hidden-order rate, with $z = -2.97$ and $p = 0.003$; financial-sector membership, with $z = 2.61$ and $p = 0.009$; and ETP status, with $z = 2.60$ and $p = 0.009$. The pairwise comparison gives the same result. The AROUND cohort has an average hidden-order rate of 0.108, compared with 0.128 in the BEFORE cohort, with $t = -3.86$ and $p = 0.0002$.

The economic interpretation is that the securities transitioning near the aggregate break and Phase 2c are a high-visibility slice of the sample. They include financial stocks with deep institutional trading, exchange-traded products linked to cross-market arbitrage, and large-cap securities with active displayed order books. These are precisely the settings in which displayed liquidity is likely to be a salient public signal. Phase 2c's rejected-route reporting plausibly raised the cost of displayed-depth tactics in such names, but the evidence should be read as corroborating rather than conclusive. The timing is consistent with the visibility channel, and the composition of the affected cohort matches the cross-sectional visibility result.

Enforcement actions are part of the same regime change, but they are not the binding source of identification. The sample contains four relevant actions: SEC v. Citigroup in January 2017, SEC v. Taglich Brothers in October 2019, DOJ/CFTC v. JPMorgan Chase in September 2020, and SEC v. Singh in September 2022. We estimate security-day panel coefficients on pooled $\text{Post}_{e,t}$ indicators around each event:

$$\text{Freq}_{s,t} = \alpha_s + \beta_e \text{Post}_{e,t} + \delta \text{Trend}_t + X'_{s,t}\Gamma + \varepsilon_{s,t}, \tag{31}$$

using a balanced security-day panel with security fixed effects, a linear trend, and controls for VIX, own realized volatility, and log volume. Standard errors are two-way clustered by symbol and date (Cameron, Gelbach, and Miller, 2011). Appendix Table G.4 reports the estimates. JPMorgan and Singh have deterrence-signed coefficients of −1.94 events per security-day, with $t = -7.26$, and −0.77, with $t = -3.46$. Citigroup is essentially null. Taglich is positive, consistent with heightened

detection or other contemporaneous changes rather than immediate deterrence. The JPMorgan effect is the largest and most precisely estimated and coincides with the September 2020 portion of the bunching evidence in Table 12. These enforcement coefficients align with the regime-change interpretation, but they cannot separately identify enforcement from CAT implementation at the resolution of the monthly break-date data.

Taken together, the cross-sectional, within-security, and break-date evidence point to the same channel. Phantom-depth prevalence is lower when displayed liquidity is a weaker public signal. The mid-2020 to early-2021 decline is concentrated among securities whose break dates bunch around changes in surveillance and reporting infrastructure, and among high-visibility names identified by the static evidence. We therefore interpret the regime change as corroborating evidence for the visibility channel rather than as a stand-alone source of identification.

8 Conclusion

This paper reframes spoofing as a problem of endogenous credibility. Displayed depth matters because traders treat visible orders as evidence of trading interest, inventory pressure, or private information. That is what makes the limit order book useful. It is also what makes it vulnerable. A visible order can affect beliefs before it is executed, and it can often be cancelled before its credibility is tested. Spoofing exploits this interval between observation and verification. The manipulator does not need irrational counterparties. She needs a market in which rational traders cannot ignore displayed liquidity without also discarding useful information.

We formalize this logic in a Stackelberg model of manipulable displayed depth. A trader posts visible depth, other agents respond to the display, and the trader retains the option to cancel. In equilibrium, the displayed order changes how the market interprets order flow while the order is live. The market can therefore appear deeper than it really is because part of the displayed liquidity is supported by an order that may not stand behind the trade. Phantom depth is not simply a large order that disappears. It is a temporary distortion in a public signal on which other traders condition.

The empirical analysis separates detection from identification. Across 90.6 million mechanically detected collapsed-depth events, the unconditional signed reversal is economically close to zero; the large reversal often associated with manipulation appears only after conditioning on the realized post-event price path, which builds the outcome into the event definition and mechanically creates the pattern it is meant to validate. We therefore identify the strategic state from behavior inside the episode: displayed depth elicits cross-side trading, cancellation follows that elicited flow, and the cancellation response is far stronger during the depth spike than in the matched pre-spike window — the amplification that separates display-and-trigger behavior from ordinary stale-quote avoidance.

The overidentifying evidence supports the same mechanism. A trigger score constructed only from within-episode cancellation timing predicts outcomes excluded from its construction — larger signed dislocations, sharper depth withdrawal, lower post-collapse variance ratios, wider subsequent

effective spreads, and lower live-window price impact — different traces of one sequence: the display elicits a response, withdrawal follows the response, and the market appears deeper while the signal is live. The post-event price path is not the source of identification; it is one of the outcomes the behavior predicts.

The market-level evidence points to the same channel. Phantom depth is not primarily explained by the mechanical ability to cancel quickly; what matters is whether displayed liquidity remains informative enough to move other traders. Prevalence is lower when hidden-order activity is higher, both across securities and within securities over time, while cancellation intensity has little stable explanatory power once visibility is controlled for. The mid-2020 to early-2021 decline in frequency is concentrated among securities whose break dates bunch around changes in reporting and surveillance infrastructure and among high-visibility names, consistent with the visibility channel rather than a stand-alone source of identification.

This distinction matters for market design. Policies that focus only on cancellation costs treat spoofing as a problem of withdrawal technology. The evidence points to a deeper issue, which is the informational weight that market participants place on visible liquidity. When displayed depth is a strong public signal, manipulating it is valuable. When hidden and reserve liquidity make displayed depth less informative, the same strategy becomes less useful. The relevant policy margin is therefore not only how quickly traders can cancel, but also how much the market learns from what is displayed before cancellation occurs.

The same logic extends beyond financial markets. Many economic and social environments rely on public signals that are useful before they are fully verified. A seller lists inventory that may not be available. A platform displays popularity metrics that can be revised. A firm announces capacity or demand before commitment is complete. An online actor creates visible engagement before the underlying interest is known to be genuine. In each case, the signal has value because others respond to it. Manipulation becomes possible when the sender can benefit from that response and then withdraw, revise, or abandon the signal before credibility is tested. Phantom depth is therefore one instance of a broader problem. Public signals are most vulnerable when they are both informative and reversible.

The broader lesson is that the value of displayed liquidity depends on credibility, and credibility is determined in equilibrium. Visible orders matter because traders respond to them, but the same response that makes display useful also makes it vulnerable to manipulation. Phantom depth is the order-book version of a more general mechanism in which a public signal affects behavior before it is verified. When the sender can benefit from that response and withdraw the signal before credibility is tested, manipulation can arise even in markets populated by rational traders.

References

Aggarwal, R. and G. Wu (2006): "Stock Market Manipulations," Journal of Business, 79(4), 1915–1953.

Allen, F. and D. Gale (1992): "Stock-Price Manipulation," Review of Financial Studies, 5(3), 503–529.

Andrews, D. W. K. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," Econometrica, 78(1), 119–157.

Baruch, S. and L. R. Glosten (2013): "Fleeting Orders," Working paper, Columbia Business School.

Bertrand, M., E. Duflo, and S. Mullainathan (2004): "How Much Should We Trust Differences-in-Differences Estimates?" Quarterly Journal of Economics, 119(1), 249–275.

Bénabou, R. and G. Laroque (1992): "Using Privileged Information to Manipulate Markets: Insiders, Gurus, and Credibility," Quarterly Journal of Economics, 107(3), 921–958.

Biais, B., T. Foucault, and S. Moinas (2015): "Equilibrium Fast Trading," Journal of Financial Economics, 116(2), 292–313.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference with Multiway Clustering," Journal of Business & Economic Statistics, 29(2), 238–249.

Cartea, Á., S. Jaimungal, and J. Penalva (2015): Algorithmic and High-Frequency Trading. Cambridge University Press.

Cartea, Á., S. Jaimungal, and Y. Wang (2020): "Spoofing and Price Manipulation in Order-Driven Markets," Applied Mathematical Finance, 27(1–2), 67–98.

Chen, J., P.-F. Hsieh, and J. J. Yang (2026):"Order Spoofing, Price Impact, and Market Quality," Pacific-Basin Finance Journal, 97, 103077.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," The Econometrics Journal, 21(1), C1–C68.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," Econometrica, 75(5), 1243–1284.

Comerton-Forde, C. and T. J. Putniņš (2014): "Stock Price Manipulation: Prevalence and Determinants," Review of Finance, 18(1), 23–66.

Do, B. L. and T. J. Putniņš (2023): "Detecting Layering and Spoofing in Markets," SSRN working paper, doi:10.2139/ssrn.4525036.

Fabre, T. and D. Challet (2025): "Learning the Spoofability of Limit Order Books With Interpretable Probabilistic Neural Networks," arXiv preprint arXiv:2504.15908.

Goldstein, I. and A. Guembel (2008): "Manipulation and the Allocational Role of Prices," Review of Economic Studies, 75(1), 133–164.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," Econometrica, 50(4), 1029–1054.

Hasbrouck, J. (1993): "Assessing the Quality of a Security Market: A New Approach to Transaction-Cost Measurement," Review of Financial Studies, 6(1), 191–212.

Hasbrouck, J. and G. Saar (2009): "Technology and Liquidity Provision: The Blurring of Traditional Definitions," Journal of Financial Markets, 12(2), 143–172.

Jiang, G., P. G. Mahoney, and J. Mei (2005): "Market Manipulation: A Comprehensive Study of Stock Pools," Journal of Financial Economics, 77(1), 147–170.

Kumar, P. and D. J. Seppi (1992): "Futures Manipulation with 'Cash Settlement,'" Journal of Finance, 47(4), 1485–1502.

Kyle, A. S. (1985): "Continuous Auctions and Insider Trading," Econometrica, 53(6), 1315–1335.

Lee, E. J., K. S. Eom, and K. S. Park (2013): "Microstructure-based Manipulation: Strategic Behavior and Performance of Spoofing Traders," Journal of Financial Markets, 16(2), 227–252.

Lee, C. M. C. and M. J. Ready (1991): "Inferring Trade Direction from Intraday Data," Journal of Finance, 46(2), 733–746.

Menkveld, A. J. and B. Z. Yueshen (2019): "The Flash Crash: A Cautionary Tale About Highly Fragmented Markets," Management Science, 65(10), 4470–4488.

Putniņš, T. J. (2020): "An Overview of Market Manipulation," in C. Alexander and D. Cumming, eds., *Corruption and Fraud in Financial Markets: Malpractice, Misconduct and Manipulation*, Wiley, 13–44.

Rochet, J.-C. and J.-L. Vila (1994): "Insider Trading without Normality," Review of Economic Studies, 61(1), 131–152.

Tao, X., A. Day, L. Ling, and S. Drapeau (2022): "On Detecting Spoofing Strategies in High-Frequency Trading," Quantitative Finance, 22(8), 1405–1425.

van Kervel, V. and A. J. Menkveld (2019): "High-Frequency Trading around Large Institutional Orders," Journal of Finance, 74(3), 1091–1137.

Williams, B. and A. Skrzypacz (2024): "Spoofing in Equilibrium," Journal of Finance, forthcoming.

Appendix

Appendix A Equilibrium Derivations

This appendix derives the equilibrium of Section 2. The full setup, normalization, market makers' problem, nonmanipulator's best response, and the manipulator's backward-induction solution are reproduced here.

A.1 Strategies and prices

At $t = 0$, liquidity traders submit the liquidity market order u_0, a nonmanipulator privately observes the realization of the fundamental asset value v and submits the informed order $x_N = \beta_N\, v$, and the manipulator submits her demand schedule, $x_{M,0} = \beta_{M,0}\xi + \gamma_{M,0}u_0$. MMs observe the total order flow, y_0, equal to

$$y_0 = \beta_N v + \beta_{M,0}\xi + (1 + \gamma_{M,0})\, u_0,$$

and set the informationally efficient price schedule (5) used in both trading rounds.

At $t = 1^-$, the manipulator cancels her initial order $x_{M,0}$ to be executed at $t = 1$ with an exogenous probability q. If the order cancellation fails with probability $1 - q$, then $x_{M,0}$ is executed at $t = 1$, and the manipulator does not trade at $t = 2$. Therefore, the order executed at $t = 1$ is

$$y_{-1} = \begin{cases} u_0 + \beta_N v, & \text{prob} = q \\ y_0, & \text{prob} = 1 - q \end{cases}, \tag{A1}$$

and the price obtained by the nonmanipulator and liquidity traders is

$$P_1 \equiv P(y_{-1}) = \begin{cases} \Lambda\left[u_0 + \beta_N v\right], & \text{prob} = q \\ \Lambda\, y_0, & \text{prob} = 1 - q \end{cases}. \tag{A2}$$

After the market clears at $t = 1$, the manipulator who cancelled at $t = -1$ learns the liquidity demand u_1 and submits a revised order $x_{M,1} = \beta_{M,1}\xi + \gamma_{M,1}u_1$ to be executed in the second trading round at $t = 2$. Therefore, the total order flow at $t = 2$ takes the form

$$y_2 = \begin{cases} \beta_{M,1}\xi + (1 + \gamma_{M,1})\, u_1, & \text{prob. } q \\ u_1, & \text{prob. } 1 - q \end{cases},$$

and the execution price is

$$P_2 \equiv P(y_2) = \begin{cases} \Lambda\left[\beta_{M,1}\xi + (1 + \gamma_{M,1})\, u_1\right], & \text{prob} = q \\ \Lambda u_1, & \text{prob} = 1 - q \end{cases}. \tag{A3}$$

We can represent the manipulator's strategies as a combination of market and limit orders as follows

$$x_{M,0} = \begin{cases} 0, & \text{prob} = q \\ \kappa_{M,0} + \frac{1}{\lambda_{M,0}} P_1, & \text{prob} = 1 - q \end{cases}, \tag{A4}$$

with the intercept $\kappa_{M,0}$ and slope $\frac{1}{\lambda_{M,0}}$ defined as

$$\begin{aligned} \kappa_{M,0} &\equiv \frac{1}{1+\gamma_{M,0}}\left(\beta_{M,0}\xi - \gamma_{M,0}\beta_N v\right), \\ \frac{1}{\lambda_{M,0}} &\equiv \frac{\gamma_{M,0}}{1+\gamma_{M,0}}\frac{1}{\Lambda}, \end{aligned} \tag{A5}$$

with $\beta_N = \frac{1-2\Lambda\beta_{M,0}}{2\Lambda}$. Importantly, since the optimal inverse market depth Λ depends on $\beta_{M,0}$ and $\gamma_{M,0}$, (A5) defines a unique relation between the "Rochet-Vila coefficients" $\{\beta_{M,0}, \gamma_{M,0}\}$ and the parameters $\left\{\kappa_{M,0}, \frac{1}{\lambda_{M,0}}\right\}$ characterizing the price contingent ("limit") order of the manipulator. Note that the manipulator's order (A4) can still be treated as a "price contingent" market order when we construct the equilibrium.

A.2 Rescaling

We introduce the rescaled parameters

$$\widehat{\beta}_{i,t} \equiv \beta_{i,t}\frac{\sigma_v}{\sigma_u}, \qquad \widehat{\Lambda} \equiv \Lambda\frac{\sigma_u}{\sigma_v}, \qquad a \equiv \alpha\sigma_u\sigma_v, \qquad \tau \equiv \frac{\sigma_\xi^2}{\sigma_v^2}, \tag{A6}$$

with the rescaled signals and wealth

$$\widehat{\xi} \equiv \xi/\sigma_v = \widehat{v} + sqrt\tau\,\varepsilon, \qquad \widehat{v} = v/\sigma_v, \qquad \widehat{w}_t = w_t/(\sigma_u\sigma_v). \tag{A7}$$

Let $\widehat{u}_t = \frac{\widetilde{u}_t}{\sigma_u}$, $t \in \{0,1\}$, be the rescaled aggregate uninformed liquidity order. Under parameterization (A6) the rescaled strategy of manipulator $\widehat{x}_t \equiv \frac{x_t}{\sigma_u}$ can be written as

$$\widehat{x}_{M,0} = \widehat{\beta}_{M,t}\widehat{\xi} + \gamma_{M,t}\widehat{u}_0 = \widehat{\kappa}_{M,0} + \frac{1}{\widehat{\lambda}_{M,0}}\widehat{P}_1. \tag{A8}$$

and rescaled prices $\widehat{P}_t \equiv \frac{P_t}{\sigma_v}$, $t \in \{1,2\}$, are equal to

$$\widehat{P}_1 = \begin{cases} \widehat{\Lambda}\left[\widehat{u}_0 + \widehat{\beta}_N\widehat{v}\right], & \text{prob} = q \\ \widehat{\Lambda}\left[\widehat{\beta}_N\widehat{v} + \widehat{\beta}_{M,0}\widehat{\xi} + (1+\widehat{\gamma}_{M,0})\,\widehat{u}_0\right], & \text{prob} = 1-q \end{cases} \tag{A9}$$

$$\widehat{P}_2 = \begin{cases} \widehat{\Lambda}\left[\widehat{\beta}_{M,1}\widehat{\xi} + (1+\gamma_{M,1})\,\widehat{u}_1\right], & \text{prob} = q \\ \widehat{\Lambda}\widehat{u}_1, & \text{prob} = 1-q \end{cases}, \tag{A10}$$

where

$$\begin{aligned} \widehat{\kappa}_{M,0} &= \frac{1}{1+\widehat{\gamma}_{M,0}}\left(\widehat{\beta}_{M,0}\widehat{\xi} - \widehat{\gamma}_{M,0}\widehat{\beta}_N\widehat{v}\right), \\ \frac{1}{\widehat{\lambda}_{M,0}} &= \frac{\widehat{\gamma}_{M,0}}{1+\widehat{\gamma}_{M,0}}\frac{1}{\widehat{\Lambda}}. \end{aligned} \tag{A11}$$

A.3 Market makers' problem

In the first trading round, market makers and the informed nonmanipulator play the role of followers to the informed manipulator, who is the Stackelberg leader. By construction of the Stackelberg

equilibrium, market makers *directly observe* the strategy of the informed manipulator before constructing their reaction function. Effectively, market makers presume that with probability $1-q$ they are playing against two informed traders, the nonmanipulator and a manipulator who failed to revise her time-zero order, and with probability q the manipulator successfully revises, and market makers trade against the nonmanipulator only. The total order flow at $t=0$, therefore, takes the two-scenario form

$$\widehat{y} = \begin{cases} \widehat{y}_0 = \widehat{\beta}_N \widehat{v} + \widehat{\beta}_{M,0}\widehat{\xi} + (1+\widehat{\gamma}_{M,0})\widehat{u}_0, & \text{prob } 1-q, \\ \widehat{y}_1 = \widehat{u}_0 + \widehat{\beta}_N \widehat{v}, & \text{prob } q. \end{cases} \tag{A12}$$

Expression (A12) makes clear what market makers observe and what they do not observe. They observe the manipulator's displayed strategy, summarized by $(\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0})$, and they know the cancellation probability q. They do not observe the realized cancellation state when forming the pricing rule.

Because the total order flow is a mixture over the cancellation state, the exact Bayesian price,

$$\mathbb{E}[\widehat{v} \mid \widehat{y} = y],$$

is generally nonlinear in y. Conditional on either cancellation state, the inference problem is linear Gaussian. Unconditionally, the posterior weights on the two states depend on the realized order flow. The market maker's exact conditional mean is therefore a weighted average of two linear prices, with weights that vary with y.

In the model, we work within the linear-pricing class in the spirit of Kyle (1985) and use the best linear projection of $\widehat{v}$ on $\widehat{y}$,

$$\widehat{\Lambda} = \frac{\mathrm{Cov}[\widehat{v}, \widehat{y}]}{\mathrm{Var}[\widehat{y}]}, \tag{A13}$$

which is the slope coefficient in the linear regression of $\widehat{v}$ on $\widehat{y}$. This is the standard linear minimum mean-square-error (LMMSE) approximation used when the joint distribution is non-Gaussian but the first and second moments are well defined. Internet Appendix G.1 shows how this linear rule relates to the exact Gaussian-mixture posterior mean and measures the incremental mean-square error from imposing linearity:

$$\eta_0^2 = \mathbb{E}\left[\left\{\mathbb{E}[\widehat{v} \mid \widehat{y}] - \widehat{\Lambda}\widehat{y}\right\}^2\right].$$

In the moderate parameter region used for the numerical equilibrium calculations, this residual is very small, about 0.0035% of the variance of the exact Bayesian price or less. In more aggressive spoofing regions, the residual can become economically meaningful, so we report it as a diagnostic for the quality of the linear approximation.

The inverse market depth parameter $\widehat{\Lambda}$ can be calculated from (A13) as

$$\widehat{\Lambda} = \frac{\mathrm{Cov}[\widehat{v}, \widehat{y}]}{\mathrm{Var}[\widehat{y}]} = \frac{(1-q)\widehat{\beta}_{M,0} + \widehat{\beta}_N}{(1-q)(1+\tau)\widehat{\beta}_{M,0}^2 + 2(1-q)\widehat{\beta}_{M,0}\widehat{\beta}_N + \widehat{\beta}_N^2 + \overline{(1+\widehat{\gamma}_M)^2}}, \tag{A14}$$

with the simplifying notation

$$\overline{(1+\widehat{\gamma}_M)^2} = (1-q)(1+\widehat{\gamma}_{M,0})^2 + q.$$

To derive (A14), compute the covariance using the two-scenario representation. Since $\mathbb{E}[\widehat{v}\widehat{\xi}] =$

$\mathbb{E}[\widehat{v}(\widehat{v}+\sqrt{\tau}\,\varepsilon)]=1$ and $\mathbb{E}[\widehat{v}\widehat{u}_0]=0$, the numerator is

$$\mathrm{Cov}[\widehat{v},\widehat{y}]=(1-q)\big(\widehat{\beta}_N+\widehat{\beta}_{M,0}\big)+q\,\widehat{\beta}_N=\widehat{\beta}_N+(1-q)\widehat{\beta}_{M,0}.$$

For the denominator, use $\mathbb{E}[\widehat{\xi}^2]=1+\tau$, $\mathbb{E}[\widehat{u}_0^2]=1$, $\mathbb{E}[\widehat{v}^2]=1$, and $\mathbb{E}[\widehat{\xi}\widehat{u}_0]=0$ to compute the variance under each scenario and take the probability-weighted sum:

$$\begin{aligned}\mathrm{Var}[\widehat{y}]&=(1-q)\big[\widehat{\beta}_N^2+(1+\tau)\widehat{\beta}_{M,0}^2+2\widehat{\beta}_{M,0}\widehat{\beta}_N+(1+\widehat{\gamma}_{M,0})^2\big]+q\big[\widehat{\beta}_N^2+1\big]\\&=(1-q)(1+\tau)\widehat{\beta}_{M,0}^2+2(1-q)\widehat{\beta}_{M,0}\widehat{\beta}_N+\widehat{\beta}_N^2+\overline{(1+\widehat{\gamma}_M)^2}.\end{aligned}$$

Equation (A14) follows.

The economics of (A14) are best understood through the lens of signal extraction. The numerator measures the informativeness of the order flow, the covariance between $\widehat{v}$ and $\widehat{y}$, while the denominator measures the total variability of $\widehat{y}$. The manipulator affects both: her informed component $\widehat{\beta}_{M,0}\widehat{\xi}$ increases the informativeness (raising the numerator), while her order as a whole increases total variability (raising the denominator). Whether $\widehat{\Lambda}$ rises or falls depends on which effect dominates. When the manipulator's bluffing component $\gamma_{M,0}\widehat{u}_0$ is large, the denominator effect dominates, and $\widehat{\Lambda}$ declines—this is the phantom-depth channel. The expected order flow $\overline{(1+\widehat{\gamma}_M)^2}$ in the denominator is probability-weighted across the two scenarios (manipulation succeeds or fails), reflecting the market makers' rational expectation over the manipulator's eventual action.

A.4 Nonmanipulator's best response

Using the linear strategy $\widehat{x}_N=\widehat{\beta}_N\widehat{v}$ and the two-scenario pricing rule, the expected terminal wealth of the nonmanipulator in the first trading round is

$$\begin{aligned}\widehat{w}_N&=\mathbb{E}\big[\widehat{\beta}_N\widehat{v}\big(\widehat{v}-\widehat{P}_1\big)\big]\\&=q\,\mathbb{E}\big[\widehat{\beta}_N\widehat{v}\big(\widehat{v}-\widehat{\Lambda}(\widehat{u}_0+\widehat{\beta}_N\widehat{v})\big)\big]\\&\quad+(1-q)\,\mathbb{E}\big[\widehat{\beta}_N\widehat{v}\big(\widehat{v}-\widehat{\Lambda}(\widehat{\beta}_N\widehat{v}+\widehat{\beta}_{M,0}\widehat{\xi}+(1+\widehat{\gamma}_{M,0})\widehat{u}_0)\big)\big]\\&=q\,\widehat{\beta}_N(1-\widehat{\Lambda}\widehat{\beta}_N)+(1-q)\,\widehat{\beta}_N\big(1-\widehat{\Lambda}(\widehat{\beta}_N+\widehat{\beta}_{M,0})\big),\end{aligned}\tag{A15}$$

where the last line uses $\mathbb{E}[\widehat{v}^2]=1$, $\mathbb{E}[\widehat{v}\widehat{\xi}]=1$, and $\mathbb{E}[\widehat{v}\widehat{u}_0]=0$. Simplifying,

$$\widehat{w}_N=\widehat{\beta}_N\big(1-\widehat{\Lambda}\widehat{\beta}_N\big)-(1-q)\widehat{\beta}_N\widehat{\Lambda}\widehat{\beta}_{M,0}.\tag{A16}$$

Optimizing $\widehat{w}_N$ with respect to $\widehat{\beta}_N$ for given $\widehat{\Lambda}$ and $\widehat{\beta}_{M,0}$, the first-order condition yields the best-response function

$$\widehat{\beta}_N=\frac{1-(1-q)\widehat{\Lambda}\widehat{\beta}_{M,0}}{2\widehat{\Lambda}},$$

which is equation (6) of the main text.

The structure of this best response reveals the competitive dynamics. The term $1/(2\widehat{\Lambda})$ is the familiar Kyle (1985) intensity—what the nonmanipulator would choose as the sole informed trader. The correction $-(1-q)\widehat{\beta}_{M,0}/2$ captures the competitive pressure from the manipulator. When the manipulator's order is more likely to be executed (low q) or when it is larger (high $\widehat{\beta}_{M,0}$), the nonmanipulator pulls back. Note that the correction depends on $(1-q)\widehat{\beta}_{M,0}$: it is the *expected* manipulator order in the first trading round that matters for the nonmanipulator's strategy, since the nonmanipulator conditions on the probability-weighted exposure.

Substituting the nonmanipulator's best response into (A14) and simplifying yields the inverse market depth as a function of the manipulator's strategy alone,

$$\widehat{\Lambda} = \frac{1}{2}\,\frac{1}{\sqrt{\varkappa\widehat{\beta}_{M,0}^2 + \overline{(1+\widehat{\gamma}_M)^2}}}, \qquad \varkappa = (1-q)\big(1+\tau-\tfrac{3}{4}(1-q)\big)\,. \tag{A17}$$

Equation (A17) encapsulates the central tension in the model. The inverse market depth is inversely related to the total variability of order flow: more variable flow dilutes the signal per unit, so market makers reduce the price sensitivity. The manipulator can exploit this by increasing $\widehat{\beta}_{M,0}$, which inflates the denominator and pushes $\widehat{\Lambda}$ down. However, the coefficient $\varkappa$ ensures that the manipulator's informed component partially offsets the dilution. The parameter $\varkappa$ is decreasing in q: when the manipulator is more likely to cancel, her order contributes less to the expected information content, making it easier to inflate the denominator without proportionally increasing the numerator.

A.5 Manipulator's problem

The manipulator's problem is solved by backward induction. We first characterise the revised strategy in the second trading round, taking the pricing rule as given, and then solve for the initial strategy at $t = 0$.

Final period $t = 1$

The manipulator's revision optimization problem takes the form

$$\widehat{x}_{M,1}^*(\cdot) = \underset{\widehat{x}_{M,1}(\cdot)}{\arg\max}\, \mathbb{E}_{\widehat{v}|\widehat{\xi},\widehat{u}_1}\big[U(\widehat{w}_2)\big]\,, \tag{A18}$$

where the terminal wealth is $\widehat{w}_2 = \widehat{x}_{M,1}(\widehat{v} - \widehat{P}_2)$. Making use of the linear strategy and the pricing rule for the successful-revision scenario, $\widehat{P}_2 = \widehat{\Lambda}[\widehat{x}_{M,1} + \widehat{u}_1]$, the manipulator's expected utility takes the form

$$\begin{aligned}\overline{U}_2 &= \mathbb{E}_{\widehat{v}|\widehat{\xi},\widehat{u}_1}[\widehat{w}_2] - \tfrac{a}{2}\operatorname{Var}_{\widehat{v}|\widehat{\xi},\widehat{u}_1}[\widehat{w}_2] \\ &= \widehat{x}_{M,1}\left(\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}[\widehat{x}_{M,1} + \widehat{u}_1]\right) - \frac{a}{2}\tau\,\widehat{x}_{M,1}^2,\end{aligned} \tag{A19}$$

where we use the conditional moments $\mathbb{E}[\widehat{v}|\widehat{\xi}] = \widehat{\xi}/(1+\tau)$ and $\operatorname{Var}[\widehat{v}|\widehat{\xi}] = \tau/(1+\tau)$, with the rescaling absorbing the factor of $(1+\tau)$ in the variance term. When optimizing (A19) with respect to $\widehat{x}_{M,1}$, the manipulator keeps in mind that the market makers do not adjust the market depth to the revised order flow, so $\widehat{\Lambda}$ does not depend on $\widehat{x}_{M,1}$ in the second round.

The first-order condition yields

$$\widehat{x}_{M,1} = \frac{\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}\widehat{u}_1}{2\widehat{\Lambda} + a\tau}. \tag{A20}$$

Recalling the linear parametrisation $\widehat{x}_{M,1} = \widehat{\beta}_{M,1}\widehat{\xi} + \widehat{\gamma}_{M,1}\widehat{u}_1$, we obtain the optimal revised-round parameters

$$\widehat{\beta}_{M,1}^* = \frac{1}{1+\tau}\,\frac{1}{2\widehat{\Lambda} + a\tau}, \qquad \widehat{\gamma}_{M,1}^* = -\frac{\widehat{\Lambda}}{2\widehat{\Lambda} + a\tau}. \tag{A21}$$

The second-period strategy (A21) has a transparent economic interpretation. The trading intensity $\widehat{\beta}^*_{M,1}$ is the manipulator's optimal response to her private signal, scaled down by both the market's price impact ($2\widehat{\Lambda}$) and the risk cost of imprecise information ($a\tau$). When $a\tau \to 0$, the manipulator trades aggressively on her signal: $\widehat{\beta}^*_{M,1} \to 1/(2\widehat{\Lambda})$, matching the intensity of a risk-neutral perfectly informed trader. The intercept $\widehat{\gamma}^*_{M,1}$ is always negative, capturing the standard Rochet and Vila (1994) effect: by observing u_1, the manipulator trades against the liquidity order, extracting rents from liquidity traders. The magnitude $|\widehat{\gamma}^*_{M,1}|$ is bounded above by $1/2$, approached in the limit $a\tau \to 0$.

Crucially, the revised strategy (A21) takes $\widehat{\Lambda}$ as given—the manipulator cannot influence the pricing rule in the second round. This is what makes the first-period problem strategically rich: the manipulator's initial order shapes the pricing rule that she will later face as a price-taker.

Alternatively, one can verify (A21) using the double-expectation method. Construct the "double expected" utility $\overline{\overline{U}}_2 = \mathbb{E}_{\widehat{\xi},\widehat{u}_1}[\overline{U}_2]$, which does not depend on realizations of $\widehat{\xi}$ and $\widehat{u}_1$ and depends only on the parameters $\{\widehat{\beta}_{M,1}, \widehat{\gamma}_{M,1}\}$:

$$\begin{aligned}\overline{\overline{U}}_2 &= \mathbb{E}_{\widehat{\xi},\widehat{u}_1}\left[(\widehat{\beta}_{M,1}\widehat{\xi} + \widehat{\gamma}_{M,1}\widehat{u}_1)\left(\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}(\widehat{\beta}_{M,1}\widehat{\xi} + (1+\widehat{\gamma}_{M,1})\widehat{u}_1)\right)\right] \\ &\quad - \mathbb{E}_{\widehat{\xi},\widehat{u}_1}\left[\frac{a}{2}\tau(\widehat{\beta}_{M,1}\widehat{\xi} + \widehat{\gamma}_{M,1}\widehat{u}_1)^2\right] \\ &= \widehat{\beta}_{M,1}\big(1 - (1+\tau)\widehat{\Lambda}\widehat{\beta}_{M,1}\big) - \widehat{\Lambda}\widehat{\gamma}_{M,1}(1+\widehat{\gamma}_{M,1}) - \frac{a}{2}\tau\big((1+\tau)\widehat{\beta}^2_{M,1} + \widehat{\gamma}^2_{M,1}\big), \end{aligned} \tag{A22}$$

where we use $\mathbb{E}[\widehat{\xi}^2] = 1+\tau$, $\mathbb{E}[\widehat{u}_1^2] = 1$, and $\mathbb{E}[\widehat{\xi}\widehat{u}_1] = 0$. The first-order conditions of (A22) with respect to $\widehat{\beta}_{M,1}$ and $\widehat{\gamma}_{M,1}$ are

$$1 - 2(1+\tau)\widehat{\Lambda}\widehat{\beta}_{M,1} - a\tau(1+\tau)\widehat{\beta}_{M,1} = 0, \tag{A23}$$

$$-\widehat{\Lambda}(1 + 2\widehat{\gamma}_{M,1}) - a\tau\widehat{\gamma}_{M,1} = 0. \tag{A24}$$

Solving (A23)–(A24) recovers (A21).

Substituting (A21) into (A19) yields the manipulator's optimal indirect utility in the second round:

$$\overline{U}^*_2 = \frac{\left(\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}\widehat{u}_1\right)^2}{4\left(\widehat{\Lambda} + \frac{a}{2}\tau\right)}. \tag{A25}$$

Initial period $t = 0$

The manipulator's overall strategy across both rounds is

$$\widehat{x}_M = \begin{cases} \widehat{\beta}_{M,0}\widehat{\xi} + \widehat{\gamma}_{M,0}\widehat{u}_0, & \text{prob } 1-q, \\ \widehat{\beta}^*_{M,1}\widehat{\xi} + \widehat{\gamma}^*_{M,1}\widehat{u}_1, & \text{prob } q, \end{cases} \tag{A26}$$

with $\widehat{\beta}^*_{M,1}$ and $\widehat{\gamma}^*_{M,1}$ defined by (A21). The first-round objective of the manipulator is

$$\overline{U}_1 = \mathbb{E}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}\left[(1-q)\,U(\widehat{w}_1) + q\,\overline{U}^*_2\right], \tag{A27}$$

where $\overline{U}_2^*$ is the continuation value from (A25) and the first-round terminal wealth is

$$\begin{aligned}\widehat{w}_1 &= \widehat{x}_{M,0}\Big(\widehat{v} - \widehat{\Lambda}[\widehat{x}_{M,0} + \widehat{x}_N + \widehat{u}_0]\Big)\\ &= \widehat{x}_{M,0}\Big(\widehat{v} - \widehat{\Lambda}[\widehat{x}_{M,0} + \widehat{u}_0] - \tfrac{1}{2}\widehat{v}(1-(1-q)\widehat{\Lambda}\widehat{\beta}_{M,0})\Big)\\ &= \widehat{x}_{M,0}\Big(\tfrac{1}{2}\widehat{v}(1+(1-q)\widehat{\Lambda}\widehat{\beta}_{M,0}) - \widehat{\Lambda}[\widehat{x}_{M,0} + \widehat{u}_0]\Big)\,, \end{aligned} \tag{A28}$$

where we substitute the nonmanipulator's optimal $\widehat{x}_N = \widehat{\beta}_N\widehat{v}$ from (6). Since the manipulator does not observe $\widehat{u}_1$ at $t=0$, the continuation value must be "expected out." The expectation of the indirect utility (A25) over $\widehat{u}_1$ gives

$$\mathbb{E}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}\Big[\overline{U}_2^*\Big] = \frac{\frac{1}{(1+\tau)^2}\widehat{\xi}^2 + \widehat{\Lambda}^2}{4\big(\widehat{\Lambda} + \frac{a}{2}\tau\big)}. \tag{A29}$$

Performing the expectations in (A27) yields

$$\begin{aligned}\overline{U}_1 &= (1-q)\Big(\mathbb{E}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}[\widehat{w}_1] - \tfrac{a}{2}\operatorname{Var}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}[\widehat{w}_1]\Big) + q\,\mathbb{E}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}\Big[\overline{U}_2^*\Big]\\ &= (1-q)\left(\widehat{x}_{M,0}\left(\frac{1}{1+\tau}\widehat{\xi} - \widehat{\Lambda}[\widehat{x}_{M,0}+\widehat{u}_0]\right) - \frac{a}{2}\tau\,\widehat{x}_{M,0}^2\right) + q\,\mathbb{E}_{\widehat{v},\widehat{u}_1|\widehat{\xi},\widehat{u}_0}\Big[\overline{U}_2^*\Big]\,.\end{aligned}$$

Following the double-expectation approach, we construct $\overline{\overline{U}}_1 = \mathbb{E}_{\widehat{\xi},\widehat{u}_0}[\overline{U}_1]$, which is a function of the two parameters $\{\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0}\}$ that characterize the manipulator's strategy in the first trading round:

$$\begin{aligned}\overline{\overline{U}}_1 &= (1-q)\Big(\tfrac{1}{2}\widehat{\beta}_{M,0} - \widehat{\Lambda}\widehat{\beta}_{M,0}^2\Big(1+\tau-\tfrac{1-q}{2}\Big) - \widehat{\Lambda}\widehat{\gamma}_{M,0}(1+\widehat{\gamma}_{M,0})\Big)\\ &\quad - (1-q)\tfrac{a}{2}\tau\Big(\widehat{\beta}_{M,0}^2(1+\tau) + \widehat{\gamma}_{M,0}^2\Big) + q\,\frac{(1+\tau)^{-1} + \widehat{\Lambda}^2}{4(\widehat{\Lambda} + \frac{a}{2}\tau)},\end{aligned} \tag{A30}$$

with $\widehat{\Lambda}$ given by (A17). The derivation uses $\mathbb{E}[\widehat{\xi}^2] = 1+\tau$, $\mathbb{E}[\widehat{u}_0^2] = 1$, $\mathbb{E}[\widehat{\xi}\widehat{u}_0] = 0$, and the identity $\mathbb{E}[\widehat{\beta}_{M,0}\widehat{\xi}\cdot\widehat{v}] = \widehat{\beta}_{M,0}\,\mathbb{E}[\widehat{v}\widehat{\xi}]/(1+\tau) = \widehat{\beta}_{M,0}/(1+\tau)\cdot(1+\tau) = \widehat{\beta}_{M,0}$ after rescaling. Note that (A30) with (A17) gives the double expected utility as a function of the two parameters $\{\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0}\}$. The last term on the right-hand side of (A30), the continuation value of trading in the second round, is approximately $q/(4\widehat{\Lambda})$ when $a\tau$ is small. Since $\widehat{\Lambda}$ is decreasing in $\widehat{\beta}_{M,0}$ through (A17), the manipulator can increase her continuation value by submitting ever-larger initial orders that drive $\widehat{\Lambda}$ toward zero. The only check on this bluffing incentive is the risk cost $\frac{a}{2}\tau\widehat{\beta}_{M,0}^2(1+\tau)$, which penalizes large orders through the variance of first-period wealth. The first-order condition balances these two forces, yielding $\widehat{\beta}_{M,0}^* \propto 1/(a\tau)$.

In principle, the problem reduces to optimizing (A30) with respect to $\{\widehat{\beta}_{M,0}, \widehat{\gamma}_{M,0}\}$. In general, this must be done numerically. However, in the limit of small $a\tau$, we obtain the following closed-form asymptotic expansion.

Proposition 5 (Asymptotic equilibrium) *In the limit* $a\tau \ll 1$,

$$\widehat{\beta}^{*}_{M,0} = (1-q)\widehat{\beta}^{*(0)}_{M,0} + q\,\frac{\sqrt{\varkappa}}{1+\tau}\,\frac{1}{2a\tau(1+\tau)A^2(q,\tau)},$$
$$\gamma^{*}_{M,0} = -\frac{1-\frac{1-q}{\varkappa}(1+\tau-\frac{1-q}{2})-\frac{q}{1+\tau}\frac{1}{A^2(q,\tau)}}{2A(q,\tau)-\frac{1-q}{\varkappa}(1+\tau-\frac{1-q}{2})-\frac{q}{1+\tau}\frac{1}{A^2(q,\tau)}},$$

with $\widehat{\beta}^{*(0)}_{M,0} = \frac{1-\frac{1+\tau-\frac{1-q}{2}}{\sqrt{\varkappa}}}{2a\tau(1+\tau)}$ *and* $A(q,\tau) = 1 + a\tau\sqrt{\varkappa}\widehat{\beta}^{*(0)}_{M,0}$.

Proof. In the limit $a\tau \to 0$, the manipulator's trading intensity diverges, $\widehat{\beta}^{*}_{M,0} \to \infty$. This makes sense: in this limit the manipulator is not risk-averse, and she may take infinite positions because the continuation value in the last term of (A30) diverges as $\widehat{\Lambda} \to 0$. The manipulator can drive $\widehat{\Lambda}$ toward zero by submitting exceedingly large displayed orders, as follows from (A17). However, since the cancellation may not succeed and the manipulator may have to trade in the first round with probability $1-q$, she still submits finite orders for any finite risk aversion. In the limit of large $\widehat{\beta}_{M,0}$, it follows from (A17) that

$$\widehat{\Lambda} = \frac{1}{2\sqrt{\varkappa}\widehat{\beta}_{M,0}}\left(1-\frac{\alpha^2}{2\varkappa\widehat{\beta}^2_{M,0}}\right), \tag{A31}$$

and the utility takes the form

$$\begin{aligned}\overline{\overline{U}}_1 \quad &= (1-q)\left(\frac{1}{2}\widehat{\beta}_{M,0} - \left(1+\tau-\frac{1-q}{2}\right)\frac{\widehat{\beta}_{M,0}}{2\sqrt{\varkappa}}\left(1-\frac{\alpha^2}{2\varkappa\widehat{\beta}^2_{M,0}}\right) - \frac{\widehat{\gamma}_{M,0}\left(1+\widehat{\gamma}_{M,0}\right)}{2\sqrt{\varkappa}\widehat{\beta}_{M,0}}\right)\\
&- (1-q)\,\frac{a}{2}\tau\left(\widehat{\beta}^2_{M,0}\left(1+\tau\right)+\widehat{\gamma}^2_{M,0}\right) + \frac{q}{2}\frac{1}{(1+\tau)}\frac{\sqrt{\varkappa}\widehat{\beta}_{M,0}}{A\left(q,\tau\right)}\\
&+ \frac{q}{4}\frac{1}{(1+\tau)}\frac{\alpha^2}{\sqrt{\varkappa}\widehat{\beta}_{M,0}}\frac{1}{A^2\left(q,\tau\right)} + \frac{q}{8}\frac{1}{\sqrt{\varkappa}\widehat{\beta}_{M,0}}\frac{1}{A\left(q,\tau\right)}.\end{aligned} \tag{A32}$$

The FOC for the utility (A32) w.r.t. $\widehat{\beta}_{M,0}$ and $\widehat{\gamma}_{M,0}$ yields

$$0 = (1-q)\left(\frac{1}{2}-\frac{1}{2\sqrt{\varkappa}}\left(1+\tau-\frac{1-q}{2}\right)-a\tau\left(1+\tau\right)\widehat{\beta}_{M,0}\right)+\frac{q}{2}\frac{1}{(1+\tau)}\frac{\sqrt{\varkappa}}{A^2\left(q,\tau\right)}, \tag{A33}$$

and

$$\begin{aligned}0 \quad &= \quad (1-q)\left(\frac{(1-q)\left(1+\widehat{\gamma}_{M,0}\right)}{2\sqrt{\varkappa^3}\widehat{\beta}_{M,0}}\left(1+\tau-\frac{1-q}{2}\right)-\frac{\left(1+2\widehat{\gamma}_{M,0}\right)}{2\sqrt{\varkappa}\widehat{\beta}_{M,0}}-a\tau\widehat{\gamma}_{M,0}\right)\\
&\quad + \frac{q}{2}\frac{(1-q)\left(1+\widehat{\gamma}_{M,0}\right)}{\sqrt{\varkappa}\widehat{\beta}_{M,0}\left(1+\tau\right)}\frac{1}{A^2\left(q,\tau\right)},\end{aligned} \tag{A34}$$

respectively. Solving (A33) and (A34), we obtain

$$\widehat{\beta}^{*}_{M,0} = (1-q)\,\widehat{\beta}^{*(0)}_{M,0} + q\frac{\sqrt{\varkappa}}{1+\tau}\frac{1}{2a\tau\,(1+\tau)\,A^{2}\,(q,\tau)}, \tag{A35}$$

$$\gamma^{*}_{M,0} = -\frac{1-\frac{1-q}{\varkappa}\left(1+\tau-\frac{1-q}{2}\right)-\frac{q}{1+\tau}\frac{1}{A^{2}(q,\tau)}}{2A\,(q,\tau)-\frac{1-q}{\varkappa}\left(1+\tau-\frac{1-q}{2}\right)-\frac{q}{1+\tau}\frac{1}{A^{2}(q,\tau)}}, \tag{A36}$$

with $\widehat{\beta}^{*(0)}_{M,0} = \frac{1-\frac{1+\tau-\frac{1-q}{2}}{\sqrt{\varkappa}}}{2a\tau(1+\tau)}$. ■

The manipulator's first-period trading intensity $\widehat{\beta}^{*}_{M,0}$ has two components. The first, proportional to $(1-q)$, reflects genuine trading on the private signal: with probability $1-q$ the order will be executed, and the manipulator trades in the right direction at a rate governed by $\widehat{\beta}^{*(0)}_{M,0}$. The second, proportional to q, reflects the bluffing motive: with probability q, the order will be cancelled and the manipulator inflates her order to manipulate the pricing rule. The bluffing component is proportional to $1/(a\tau)$, confirming that it diverges as execution risk vanishes.

The intercept $\gamma^{*}_{M,0}$ is equally revealing. In the standard Rochet and Vila (1994) setting, $\gamma < 0$ reflects the informed trader's optimal front-running of the liquidity order. In our model, $\gamma^{*}_{M,0}$ is determined by a more complex interaction. When the manipulator submits a large bluffing order, she also needs to absorb the market maker's supply. The intercept adjusts so that the manipulator's total order—combining the signal component $\widehat{\beta}_{M,0}\widehat{\xi}$ and the liquidity component $\gamma_{M,0}\widehat{u}_0$—optimally consumes the available depth. In the high-bluffing regime (low $a\tau$), $|\gamma^{*}_{M,0}|$ becomes large, meaning the manipulator's demand schedule is steep and highly sensitive to the liquidity demand.

Proposition 6 (Inverse market depth) *In the limit $a\tau \ll 1$, on the bluffing branch,*

$$\widehat{\Lambda}^{*} = \frac{1}{2\sqrt{\varkappa}\,\widehat{\beta}^{*}_{M,0}}\,(1+O(a\tau)) = O(a\tau),$$

with $\widehat{\beta}^{*}_{M,0}$ given by Proposition 5, and (i) $\widehat{\Lambda}^{*}$ is increasing in a; (ii) at $q=1$, $\widehat{\Lambda}^{*} = \frac{1}{2}$ exactly, while at $q=0$ the bluffing component vanishes and $\widehat{\Lambda}^{*}$ is the $O(1)$ value of the no-withdrawal two-informed-trader benchmark.

Proof. Substituting the asymptotic expansion of $\widehat{\beta}^{*}_{M,0}$ from Proposition 5 into (A17), and making use of the large-$\widehat{\beta}_{M,0}$ approximation (A31), we obtain $\widehat{\Lambda}^{*} = \left(2\sqrt{\varkappa}\widehat{\beta}^{*}_{M,0}\right)^{-1}(1+O(a\tau))$; since $\widehat{\beta}^{*}_{M,0} = O((a\tau)^{-1})$, $\widehat{\Lambda}^{*} = O(a\tau)$. For part (i), $\widehat{\beta}^{*}_{M,0}$ is proportional to $1/(a\tau)$ with a coefficient free of a, so $\widehat{\Lambda}^{*}$ is proportional to $a\tau$ and increasing in a. For part (ii), at $q=1$ equation (A17) gives $\widehat{\Lambda} = \frac{1}{2}$ for *any* strategy, because $\varkappa = 0$ and $\overline{(1+\widehat{\gamma}_M)^2} = 1$ since the displayed order never executes and market makers price only $\widehat{\beta}_N\widehat{v} + \widehat{u}_0$. At $q=0$ the continuation term in (A30) vanishes, the bluffing component of $\widehat{\beta}^{*}_{M,0}$ in Proposition 5 is proportional to q and shuts down, and the model reduces to the no-withdrawal benchmark with two heterogeneously informed traders, whose $\widehat{\Lambda}^{*}$ is $O(1)$. ■

Remark 1 (No monotone comparative static in q) *$\partial\widehat{\Lambda}^{*}/\partial q$ changes sign in the interior. A higher q strengthens the bluffing motive — the continuation weight in (A30) is proportional to q — which raises $\widehat{\beta}^{*}_{M,0}$ and lowers $\widehat{\Lambda}$ through $\varkappa\widehat{\beta}^{2}_{M,0}$ in (A17); but it also lowers the credibility weight $\varkappa$ itself, which raises $\widehat{\Lambda}$ directly. Numerically, the net effect changes sign: $\widehat{\Lambda}^{*}$ falls and then rises in q*

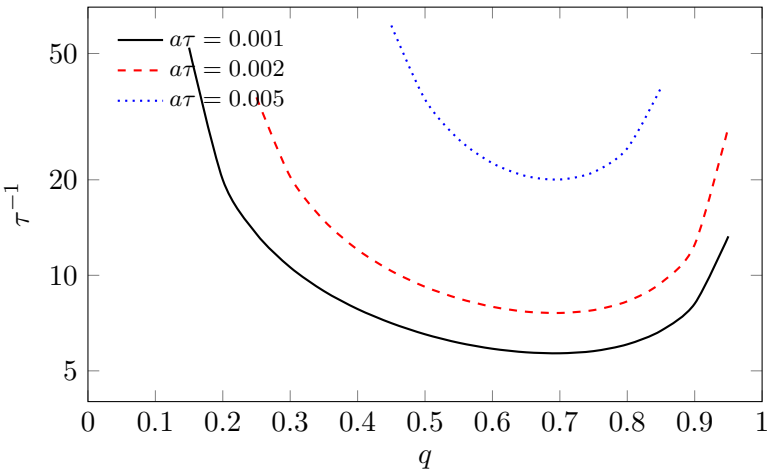


Figure A.1: The bluffing region

This figure maps the region of the parameter space in which the bluffing branch characterized by Proposition 5 is the global optimum of the manipulator's problem, computed by direct numerical maximization of (A30) under the reaction function (A17). Each curve plots, for a fixed value of the composite risk parameter $a\tau$, the critical signal precision $1/\tau^*(q, a\tau)$ above which inflating the displayed intensity dominates ordinary informed trading; the bluffing region lies *above* each curve. Below the curve the global optimum is the genuine-trading branch, with bounded $\widehat{\beta}^*_{M,0}$ and $\widehat{\Lambda}^* = O(1)$.

on the bluffing branch, returning to the $O(1)$ *benchmark values of part (ii) at the endpoints, where the bluffing branch ceases to be optimal.*

Figure A.1 makes the scope of Propositions 5–6 precise. Three features stand out. First, the bluffing region requires an *interior* cancellation probability: each boundary curve is U-shaped in q and closes toward both endpoints, the numerical counterpart of part (ii) of Proposition 6 — at $q = 0$ there is no withdrawal option to exploit, and near $q = 1$ the credibility weight $\varkappa$ collapses, so driving $\widehat{\Lambda}$ down requires unboundedly large displays. Second, the region expands as $a\tau$ falls: lower execution risk makes the display-inflation strategy viable at coarser signals. Third, the parameterization of Figure 2 (precisions τ^{-1} up to 23 at small $a\tau$) lies inside the region, so the asymptotic characterization applies throughout the range plotted there; outside the region the manipulator trades as an ordinary informed trader and $\widehat{\Lambda}^*$ remains between the two-informed-trader value and the ignore-the-display benchmark $\frac{1}{2}$. Replication code and the underlying grids are in `numerics_regime_071926.py` and the accompanying `csv` files.

A.6 Double-expectation approach

We start with a variant of Kyle'85 model where the manipulator plays the role of the leader. The optimization problem is still given by (A18), but the manipulator does not observe the liquidity order flow. The manipulator moves first; therefore, she has to take into account her impact on the pricing rule in her optimization problem. For simplicity, we will also assume that the informed trader directly observes the fundamental, or $\xi = v$. Since the point is general and is not limited to

manipulation strategies, we will drop the subscripts. The expected utility of the informed trader is

$$\begin{aligned}\overline{U} &= \mathbb{E}_{\widehat{v},\widehat{u}|\widehat{\xi}}[\widehat{w}_2] - \frac{a}{2}Var_{\widehat{v},\widehat{u}|\widehat{\xi}}[\widehat{w}_2] = \mathbb{E}_{\widehat{u}}\left[\widehat{x}\left(\widehat{v} - \widehat{P}\left(\widehat{x}+\widehat{u}\right)\right) - \frac{a}{2}\tau\widehat{x}^2\right] \\ &= \widehat{x}\left(\widehat{v} - \widehat{\overline{P}}\left(\widehat{x}\right)\right) - \frac{a}{2}\tau\widehat{x}^2, \end{aligned} \tag{A37}$$

where $\widehat{\overline{P}}\left(\widehat{x}\right) = \mathbb{E}_{\widehat{u}}\left[\widehat{P}\left(\widehat{x}+\widehat{u}\right)\right]$. Following the above logic, we introduce the informed trader's strategy as a function of the signal $\widehat{x} = \widehat{X}\left(\widehat{v}\right)$, and the optimization problem with "double expected" utility takes the form

$$\widehat{X}\left(\cdot\right) = \arg\max_{\widehat{X}(\cdot)} \overline{\overline{U}}, \tag{A38}$$

with

$$\overline{\overline{U}} = \mathbb{E}_{\widehat{v}}\left[\overline{U}\right] = \mathbb{E}_{\widehat{v}}\left[\widehat{X}\left(\widehat{v}\right)\left(\widehat{v}\widehat{\overline{P}}\left(\widehat{X}\left(\widehat{v}\right)\right)\right) - \frac{a}{2}\tau\widehat{X}\left(\widehat{v}\right)^2\right]. \tag{A39}$$

The FOC yields

$$\begin{aligned}\delta\overline{\overline{U}} &= \mathbb{E}_{\widehat{v}}\left[\delta\widehat{X}\left(\widehat{v}\right)\left(\widehat{v} - \widehat{\overline{P}}\left(\widehat{X}\left(\widehat{v}\right)\right) - \widehat{X}\left(\widehat{v}\right)\widehat{\overline{P}}'\left(\widehat{X}\left(\widehat{v}\right)\right) - a\tau\widehat{X}\left(\widehat{v}\right)\right)\right] \\ &\quad -\mathbb{E}_{\widehat{v}}\left[\widehat{X}\left(\widehat{v}\right)\delta\widehat{\overline{P}}\left(\widehat{X}\left(\widehat{v}\right)\right)\right], \end{aligned} \tag{A40}$$

where $\delta\widehat{X}\left(\widehat{v}\right)$ is an arbitrary variation of the informed trader's strategy $\widehat{X}\left(\widehat{v}\right)$. The last term in the r.h.s. of (A40) represents the "structural" variation of the double expected utility $\delta\overline{\overline{U}}$ arising due to the variation of the *functional form* $\delta\widehat{\overline{P}}\left(\cdot\right)$ of the pricing rule as a response to the variation of the *functional form* of the informed trader's strategy $\delta\widehat{X}\left(\cdot\right)$. Namely, we have

$$\begin{aligned}\delta\widehat{P}\left(y\right) &= \mathbb{E}_{\widehat{v}|y}\left[\left(\widehat{v} - \widehat{P}\left(y\right)\right)\left(y - \widehat{X}\left(\widehat{v}\right)\right)\delta\widehat{X}\left(\widehat{v}\right)\right] \\ &= \frac{\int_{-\infty}^{+\infty} e^{-\frac{\widehat{v}^2}{2}} e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}}\left(\widehat{v} - \widehat{P}\left(y\right)\right)\left(y - \widehat{X}\left(\widehat{v}\right)\right)\delta\widehat{X}\left(\widehat{v}\right)d\widehat{v}}{\int_{-\infty}^{+\infty} e^{-\frac{\widehat{v}^2}{2}} e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}} d\widehat{v}}, \end{aligned} \tag{A41}$$

and hence

$$\begin{aligned}\delta\widehat{\overline{P}}\left(\widehat{X}\left(\widehat{v}\right)\right) &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}} \delta\widehat{P}\left(y\right)dy \\ &= \mathbb{E}_u\mathbb{E}_{\widehat{v}'|y=\widehat{X}(\widehat{v})+u}\left[\left(\widehat{v}' - \widehat{P}\left(y\right)\right)\left(y - \widehat{X}\left(\widehat{v}'\right)\right)\delta\widehat{X}\left(\widehat{v}'\right)\right]. \end{aligned} \tag{A42}$$

Taking into account (A40), (A41) and (A42), we obtain the structural variation component $J\left(\widehat{X}(\cdot)\right) = \mathbb{E}_{\widehat{v}}\left[\widehat{X}(\widehat{v})\,\delta\widehat{\overline{P}}\left(\widehat{X}(\widehat{v})\right)\right]$ of the double expected utility[6] in the form

$$\begin{aligned} J &= \mathbb{E}_{\widehat{v},\widehat{u}}\left[\widehat{X}(\widehat{v})\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}}\delta\widehat{P}(y)\,dy\right] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}e^{-\frac{\widehat{v}^2}{2}}e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}}\widehat{X}(\widehat{v})\,\delta\widehat{P}(y)\,dyd\widehat{v} \\ &= \mathbb{E}_{\widehat{v},\widehat{u}}\left[\delta\widehat{X}(\widehat{v})\frac{\partial}{\partial y}\left(\left(\widehat{v}-\widehat{P}(y)\right)\widehat{Q}(y)\right)\right], \end{aligned} \tag{A43}$$

where

$$\widehat{Q}(y) \equiv \mathbb{E}_{\widehat{v}|y}\left[\widehat{X}(\widehat{v})\right] = \frac{\int_{-\infty}^{+\infty}e^{-\frac{\widehat{v}^2}{2}}e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}}\widehat{X}(\widehat{v})\,d\widehat{v}}{\int_{-\infty}^{+\infty}e^{-\frac{\widehat{v}^2}{2}}e^{-\frac{\left(y-\widehat{X}(\widehat{v})\right)^2}{2}}d\widehat{v}}. \tag{A44}$$

Combining (A40) and (A43) yields

$$\begin{aligned} \delta\widehat{\overline{U}} &= \mathbb{E}_{\widehat{v}}\left[\delta\widehat{X}(\widehat{v})\left(\widehat{v}-\widehat{\overline{P}}\left(\widehat{X}(\widehat{v})\right)-\widehat{X}(\widehat{v})\,\widehat{\overline{P}}'\left(\widehat{X}(\widehat{v})\right)-a\tau\widehat{X}(\widehat{v})\right)\right] \\ &\quad -\mathbb{E}_{\widehat{v},\widehat{u}}\left[\delta\widehat{X}(\widehat{v})\frac{\partial}{\partial y}\left(\left(\widehat{v}-\widehat{P}(y)\right)\widehat{Q}(y)\right)\right]. \end{aligned} \tag{A45}$$

Now, consider a linear equilibrium with $\widehat{X}(\widehat{v}) = \widehat{\beta}\widehat{v}$ and $\widehat{P}(y) = \mathbb{E}_{\widehat{v}|y}[\widehat{v}] = \widehat{\Lambda}y$. Then we have $\widehat{Q}(y) = \mathbb{E}_{\widehat{v}|y}\left[\widehat{\beta}\widehat{v}\right] = \widehat{\Lambda}\widehat{\beta}y$, and hence all terms in the r.h.s. of (A45) are linear in $\widehat{X}$ [7]. Therefore, the FOC can be satisfied by linear strategies. Since the SOC is also satisfied, we conclude that the linear equilibrium in the Stackelberg setting exists.

Now, since we know that the linear equilibrium with $\widehat{X}(\widehat{v}) = \widehat{\beta}\widehat{v}$ exists, it is sufficient to consider only *linear deviations* of the form $\delta\widehat{X}(\widehat{v}) = \delta\widehat{\beta}\widehat{v}$. Substituting this along with $\widehat{Q}(y) = \widehat{\Lambda}\widehat{\beta}y$ and $\widehat{P}(y) = \widehat{\Lambda}y$ into (A45), we obtain

$$\begin{aligned} \delta\widehat{\overline{U}} &= \mathbb{E}_{\widehat{v}}\left[\widehat{v}^2\delta\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}-a\tau\widehat{\beta}\right)-\widehat{v}^2\delta\widehat{\beta}\widehat{\Lambda}\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}\right)\right] \\ &= \delta\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}-a\tau\widehat{\beta}\right)-\delta\widehat{\beta}\widehat{\Lambda}\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}\right), \end{aligned} \tag{A46}$$

where the last term in the r.h.s. is coming from the structural variation of the utility.

On the other hand, the approach described above yields the double expected utility in the form

$$\widehat{\overline{U}}_2 = \widehat{\beta}\left(1-\widehat{\Lambda}\widehat{\beta}\right)-\frac{a}{2}\tau\widehat{\beta}^2,$$

where, in the Stackelberg setting, the informed traders take into account the impact of their own trading on the inverse market depth $\widehat{\Lambda}$ and hence treat it as a function of their trading intensity:

[6] Note that J is a *functional* of the trading strategy X depending only on the *functional form* of this strategy $X(\cdot)$.

[7] Note that the only questionable term is the last term in the r.h.s. of (A45) which only arises in the Stackelberg setting.

$\widehat{\Lambda}=\widehat{\Lambda}\left(\widehat{\beta}\right)=\frac{\widehat{\beta}}{1+\widehat{\beta}^2}$. Therefore, the FOC is

$$\delta\overline{\overline{U}}=\delta\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}-a\tau\widehat{\beta}\right)-\widehat{\beta}^2\delta\widehat{\Lambda}\left(\widehat{\beta}\right). \tag{A47}$$

Comparing (A46) and (A47), we observe that the last term in the r.h.s. of (A47) corresponds to the structural variation of the expected utility. Taking into account the dependence $\widehat{\Lambda}\left(\widehat{\beta}\right)$, we obtain $\widehat{\beta}^2\delta\widehat{\Lambda}\left(\widehat{\beta}\right)=\widehat{\beta}^2\frac{\partial\widehat{\Lambda}}{\partial\widehat{\beta}}\delta\widehat{\beta}=\delta\widehat{\beta}\widehat{\Lambda}\widehat{\beta}\left(1-2\widehat{\Lambda}\widehat{\beta}\right)$. Taking this into account, we observe that the structural variation is correctly represented by the last term in the r.h.s. of (A47). Since the other terms are also identical, we conclude that the optimization problem with "double expected" utility is equivalent to the initial optimization problem. As stated above, this requires that the linear equilibrium exists, but we already know that this is the case.

Now, we analyze the Stackelberg setting in the linear version of Rochet and Vila (1994) model. Similar to the above, we show that the linear equilibrium still exists in this case, and then show that it can be derived using a "double expected" utility approach. Note that this is even more technically difficult than the problem considered above, while the application of "double expected" utility leads to a tractable analytic problem, simplifying the calculations even more significantly.

Namely, assume that the informed trader observes both the realization of the fundamental $\widehat{v}$ and the liquidity demand $\widehat{u}$. The informed trader optimizes

$$U=\widehat{X}\left(\widehat{v},\widehat{u}\right)\left(\widehat{v}-\widehat{P}\left(y\right)\right)-\frac{a}{2}\tau\widehat{X}\left(\widehat{v}\right)^2, \tag{A48}$$

with the total order flow $y=\widehat{X}\left(\widehat{v},\widehat{u}\right)+\widehat{u}$. Note that since now the informed trader observes both $\widehat{v}$ and $\widehat{u}$, in her optimization problem enters the price functional $\widehat{P}\left(y\right)$ instead of the expected price functional $\widehat{\overline{P}}\left(x\right)=\mathbb{E}_{\widehat{u}}\left[\widehat{P}\left(x+\widehat{u}\right)\right]$ arising in the Kyle (1985) setting considered above.

To simplify the analysis, we will conjecture a linear equilibrium $\widehat{X}\left(\widehat{v},\widehat{u}\right)=\widehat{\beta}\widehat{v}+\widehat{\gamma}\widehat{u}$ with some coefficients $\left\{\widehat{\beta},\widehat{\gamma}\right\}$ that have to be determined in equilibrium, and the corresponding information efficient linear pricing rule $\widehat{P}\left(y\right)=\widehat{\Lambda}y$ with $y=\widehat{\beta}\widehat{v}+\left(1+\widehat{\gamma}\right)\widehat{u}$ and

$$\widehat{\Lambda}=\frac{Cov_{\widehat{v},\widehat{u}}\left[\widehat{v},y\right]}{Var_{\widehat{v},\widehat{u}}\left[y\right]}=\frac{\widehat{\beta}}{\widehat{\beta}^2+\left(1+\widehat{\gamma}\right)^2}. \tag{A49}$$

Since we need to prove that the conjectured linear equilibrium is a true equilibrium, we allow for arbitrary deviations $\delta\widehat{X}\left(\widehat{v},\widehat{u}\right)$ from the informed trader's linear strategy and show that for the equilibrium choice of the coefficients $\left\{\widehat{\beta},\widehat{\gamma}\right\}$ uniquely characterizing linear strategy, the resulting equilibrium is stable. In doing so, we adopt the concept of the "double expected" utility optimization, consistent with the above discussion.

The FOC for the "double expected" utility yields

$$\delta\overline{\overline{U}}=\mathbb{E}_{\widehat{v},\widehat{u}}\left[\delta\widehat{X}\left(\widehat{v},\widehat{u}\right)\left(\widehat{v}-\widehat{\Lambda}\left(\widehat{X}+\widehat{u}\right)-\widehat{\Lambda}\widehat{X}-a\tau\widehat{X}\right)-\widehat{X}\delta\widehat{P}\left(y\right)\right], \tag{A50}$$

where the last term in the r.h.s. comes from the structural variation of the pricing rule. Note that all the terms in the r.h.s. lead to the linear equilibrium. In particular, in the absence of the last term, we obtain a linear strategy leading to the linear Bayesian Nash equilibrium in our model. Therefore, we need to focus on the last term in the r.h.s. and show that it is also consistent with a linear equilibrium.

This "structural" term arises because the variation of the informed trader's strategy affects the market maker's beliefs on the fundamental conditional on the observation of the total order flow. In other words, variation of the informed trader's strategy affects the market maker's conditional distribution of the fundamental when they observe the total order flow. Clearly, this effect is more subtle than the standard "sliding along the residual supply curve" coming from the variation of the magnitude of the informed demand. In contrast to this, the structural variation depends on the functional form of the strategy variations, and affects the shape of the residual supply curve. In principle, this reaction even may be non-local, i.e. a variation at some point may affect the entire supply schedule (representing the limit order book in this model).

The structural variation is given by, analogous to (A41)

$$\begin{aligned}
\delta\widehat{P}\left(y\right) &= \mathbb{E}_{\widehat{v}|y}\left[\left(\widehat{v}-\widehat{\Lambda}y\right)\frac{\left(y-\widehat{\beta}\widehat{v}\right)}{\left(1+\widehat{\gamma}\right)^{2}}\delta\widehat{X}\left(\widehat{v},\frac{y-\widehat{\beta}\widehat{v}}{1+\widehat{\gamma}}\right)\right] \\
&= \frac{\int_{-\infty}^{+\infty}e^{-\frac{\widehat{v}^{2}}{2}}e^{-\frac{\left(y-\widehat{\beta}\widehat{v}\right)^{2}}{2\left(1+\widehat{\gamma}\right)^{2}}}\left(\widehat{v}-\widehat{\Lambda}y\right)\left(y-\widehat{\beta}\widehat{v}\right)\delta\widehat{X}\left(\widehat{v},\frac{y-\widehat{\beta}\widehat{v}}{1+\widehat{\gamma}}\right)d\widehat{v}}{\left(1+\widehat{\gamma}\right)^{2}\int_{-\infty}^{+\infty}e^{-\frac{\widehat{v}^{2}}{2}}e^{-\frac{\left(y-\widehat{\beta}\widehat{v}\right)^{2}}{2\left(1+\widehat{\gamma}\right)^{2}}}d\widehat{v}} \\
&= \frac{1}{\left(1+\widehat{\gamma}\right)^{2}}\int_{-\infty}^{+\infty}\frac{e^{-\frac{\left(\widehat{v}-\widehat{\Lambda}y\right)^{2}}{2\Sigma_{M}}}}{\sqrt{2\pi\Sigma_{M}}}\left(\widehat{v}-\widehat{\Lambda}y\right)\left(y-\widehat{\beta}\widehat{v}\right)\delta\widehat{X}\left(\widehat{v},\frac{y-\widehat{\beta}\widehat{v}}{1+\widehat{\gamma}}\right)d\widehat{v},
\end{aligned}$$

with the residual uncertainty in the market makers' information set

$$\Sigma_{M}=\mathbb{E}_{\widehat{v}|y}\left[\left(\widehat{v}-\widehat{\Lambda}y\right)^{2}\right]=\frac{\left(1+\widehat{\gamma}\right)^{2}}{\widehat{\beta}^{2}+\left(1+\widehat{\gamma}\right)^{2}}. \tag{A51}$$

Next we calculate

$$\begin{aligned}
\mathbb{E}_{\widehat{v},\widehat{u}}\left[\widehat{X}\delta\widehat{P}\left(y\right)\right] &= \mathbb{E}_{\widehat{v},\widehat{u}}\left[\left(\widehat{\beta}\widehat{v}+\widehat{\gamma}\widehat{u}\right)\delta\widehat{P}\left(y\right)\right] \\
&= \int_{-\infty}^{+\infty}\frac{d\widehat{v}}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{2\pi}}e^{-\frac{\widehat{v}^{2}}{2}}e^{-\frac{\left(y-\widehat{\beta}\widehat{v}\right)^{2}}{2\left(1+\widehat{\gamma}\right)^{2}}}\frac{\left(\widehat{\beta}\widehat{v}+\widehat{\gamma}\widehat{u}\right)}{\left(1+\widehat{\gamma}\right)^{3}}\times \\
&\quad\times\int_{-\infty}^{+\infty}\frac{d\widehat{v}'}{\sqrt{2\pi\Sigma_{M}}}e^{-\frac{\left(\widehat{v}'-\widehat{\Lambda}y\right)^{2}}{2\Sigma_{M}}}\left(\widehat{v}'-\widehat{\Lambda}y\right)\left(y-\widehat{\beta}\widehat{v}'\right)\delta\widehat{X}\left(\widehat{v}',\widehat{u}^{*}\right),
\end{aligned} \tag{A52}$$

with $\widehat{u}^{*}=\frac{y-\widehat{\beta}\widehat{v}'}{1+\widehat{\gamma}}$. Changing the order of integration and taking into account that

$$\frac{\widehat{v}^{2}}{2}+\frac{\left(y-\widehat{\beta}\widehat{v}\right)^{2}}{2\left(1+\widehat{\gamma}\right)^{2}}+\frac{\left(\widehat{v}'-\widehat{\Lambda}y\right)^{2}}{2\Sigma_{M}}=\frac{\widehat{v}'^{2}}{2}+\frac{\left(y-\widehat{\beta}\widehat{v}'\right)^{2}}{2\left(1+\widehat{\gamma}\right)^{2}}+\frac{\left(\widehat{v}-\widehat{\Lambda}y\right)^{2}}{2\Sigma_{M}},$$

and $X = \frac{\widehat{\beta}\widehat{v}+\widehat{\gamma}y}{1+\widehat{\gamma}}$, we obtain

$$\begin{aligned}\mathbb{E}_{\widehat{v},\widehat{u}}\left[\widehat{X}\delta\widehat{P}\left(y\right)\right] &= \int_{-\infty}^{+\infty}\frac{d\widehat{v}'}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{dy}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\frac{d\widehat{v}}{\sqrt{2\pi\Sigma_M}}e^{-\frac{\widehat{v}'2}{2}}e^{-\frac{\left(y-\widehat{\beta}\widehat{v}'\right)^2}{2\left(1+\widehat{\gamma}\right)^2}}e^{-\frac{\left(\widehat{v}-\widehat{\Lambda}y\right)^2}{2\Sigma_M}}\times \\ &\times\frac{\left(\widehat{\beta}\widehat{v}+\widehat{\gamma}y\right)}{\left(1+\widehat{\gamma}\right)^3}\left(\widehat{v}'-\widehat{\Lambda}y\right)\left(y-\widehat{\beta}\widehat{v}'\right)\delta\widehat{X}\left(\widehat{v}',\widehat{u}^*\right) \\ &= \frac{\left(\widehat{\Lambda}\widehat{\beta}+\widehat{\gamma}\right)}{\left(1+\widehat{\gamma}\right)^3}\mathbb{E}_{\widehat{v}',y}\left[y\left(\widehat{v}'-\widehat{\Lambda}y\right)\left(y-\widehat{\beta}\widehat{v}'\right)\delta\widehat{X}\left(\widehat{v}',\widehat{u}^*\right)\right]. \end{aligned} \tag{A53}$$

Clearly, the above variation vanishes if $\widehat{\Lambda}\widehat{\beta}+\widehat{\gamma}=0$ and $\widehat{\gamma}\neq -1$. This is the case in the standard Rochet and Vila (1994) model and its direct extension with the non-zero CARA risk aversion. Let's show how it works in this case.

The "double expected" utility takes the form

$$\begin{aligned}\overline{\overline{U}} &= \mathbb{E}_{\widehat{v},\widehat{u}}\left[\left(\widehat{\beta}\widehat{v}+\widehat{\gamma}\widehat{u}\right)\left(\widehat{v}-\widehat{\Lambda}\left(\widehat{\beta}\widehat{v}+\left(1+\widehat{\gamma}\right)\widehat{u}\right)\right)-\frac{a}{2}\tau\left(\widehat{\beta}\widehat{v}+\widehat{\gamma}\widehat{u}\right)^2\right] \\ &= \frac{\widehat{\beta}\left(1+\widehat{\gamma}\right)}{\widehat{\beta}^2+\left(1+\widehat{\gamma}\right)^2}-\frac{a}{2}\tau\left(\widehat{\beta}^2+\widehat{\gamma}^2\right), \end{aligned} \tag{A54}$$

The FOC with respect to $\widehat{\beta}$ and $\widehat{\gamma}$ yields

$$\left(1+\widehat{\gamma}\right)\frac{\left(1+\widehat{\gamma}\right)^2-\widehat{\beta}^2}{\left(\widehat{\beta}^2+\left(1+\widehat{\gamma}\right)^2\right)^2}=a\tau\widehat{\beta}, \tag{A55}$$

and

$$\widehat{\beta}\frac{\widehat{\beta}^2-\left(1+\widehat{\gamma}\right)^2}{\left(\widehat{\beta}^2+\left(1+\widehat{\gamma}\right)^2\right)^2}=a\tau\widehat{\gamma}, \tag{A56}$$

respectively. Also, the inverse market depth $\widehat{\Lambda}$ is given by (A49). Combining (A55) and (A56) yields

$$a\tau\left(\frac{\widehat{\beta}}{1+\widehat{\gamma}}+\frac{\widehat{\gamma}}{\widehat{\beta}}\right)=0, \tag{A57}$$

and since $a\tau\neq 0$, it means that $\frac{\widehat{\beta}^2+\widehat{\gamma}\left(1+\widehat{\gamma}\right)}{\widehat{\beta}\left(1+\widehat{\gamma}\right)}=0$, or equivalently

$$\widehat{\beta}^2+\widehat{\gamma}\left(1+\widehat{\gamma}\right)=0. \tag{A58}$$

On the other hand, we have, making use of (A49),

$$\widehat{\Lambda}\widehat{\beta}+\widehat{\gamma}=\left(1+\widehat{\gamma}\right)\frac{\widehat{\beta}^2+\widehat{\gamma}\left(1+\widehat{\gamma}\right)}{\widehat{\beta}^2+\left(1+\widehat{\gamma}\right)^2}=0, \tag{A59}$$

where the last equality follows from (A58) which directly follows from the first order conditions (A55) and (A56). Therefore, we confirm that the linear equilibrium exists.

Similar to the above, once we know that the linear equilibrium exists, we may consider only linear deviations $\left\{\delta\widehat{\beta}, \delta\widehat{\gamma}\right\}$ from the conjectured linear strategy, i.e., optimize only w.r.t. $\left\{\widehat{\beta}, \widehat{\gamma}\right\}$. Then, similar to the Kyle (1985) model setting considered above, we can show that $\delta\widehat{P}(y) = y\delta\widehat{\Lambda}$, which means that the structural variation reduces to changing the inverse market depth represented by the slope of the pricing rule. This justifies the short-hand approach to the derivation of the linear equilibria presented above. In short, the point is that since we know that the equilibrium in question is linear characterized by the slopes $\left\{\widehat{\beta}, \widehat{\gamma}\right\}$, these coefficients do not depend on realizations of $\{\widehat{v}, \widehat{u}\}$, and hence we can expect them out and construct a target functional which is simply a function of the parameters $\left\{\widehat{\beta}, \widehat{\gamma}\right\}$, and needs to be optimized w.r.t. them.

This procedure is unique since if the initial target functional has a linear optimum, the "double expected" will reveal it. Also, if the result of the "double expected" optimization is unique, it will mean that the linear optimum of the initial problem is also unique. The reason is that the "double expected" target is an average of the initial one. Therefore, if some solution delivers an optimum to the initial problem, it will also deliver an optimum to the "double expected" problem, and hence the linear solutions of the initial problem will be a subset of those generated by the "double expected" optimization procedure. In other words, we do not "lose" any linear solutions by applying the short-hand approach, but may get some "extra" solutions. However, if the resulting solution is unique, the initial problem should also have a unique linear solution once we know that it exists.

A.7 Existence and uniqueness within the linear class

Existence within the linear class is established by the double-expectation method of A.6: we verify that the FOC for the linear coefficients is consistent with a linear equilibrium, and that the second-order condition holds. Uniqueness within the linear class follows because the asymptotic expansion is invertible and the resulting fixed-point map is a contraction in the relevant parameter region.

Appendix B Endogenous Cancellation Timing

This appendix contains the assumptions, derivations, and proofs for the extension of Section 2.6. It takes the baseline equilibrium objects — the pricing rule $\widehat{\Lambda}$, the round-two continuation problem, and the normalization $(a\tau, q)$ of Appendix A — as inputs, grafts an interim observation-and-cancellation stage onto the timeline, and closes the model by replacing the exogenous q with the equilibrium cancellation probability. All variables are in the dimensionless units of A.2: $\widehat{v} = v/\sigma_v$, $\widehat{u}_t = u_t/\sigma_u$, $\widehat{\xi} = \xi/\sigma_v$, $\widehat{\Lambda} = \Lambda\sigma_u/\sigma_v$, and $m \equiv \mathbb{E}[\widehat{v} \mid \widehat{\xi}] = \widehat{\xi}/(1+\tau)$ is the manipulator's interim value estimate; $\widehat{X} \equiv \widehat{\beta}_{M,0}\widehat{\xi} + \widehat{\gamma}_{M,0}\widehat{u}_0$ denotes the non-displayed component of the manipulator's first-round schedule, known to her at $t = 0$. Throughout, "cross-side" flow means aggressive trading on the side the display encourages, "fill-direction" flow means aggressive trading that executes against the display, and we orient signs so the display is at the bid, $\widehat{D} > 0$, without loss of generality.

B.1 Timeline and assumptions

The timeline adds one event subperiod, $t = 0^+$, to the baseline five-subperiod structure of Section 2.2: at $t = 0$ the manipulator's schedule includes a displayed component of size $\widehat{D} \geq 0$; at $t = 0^+$ reactive cross-side flow and fill-direction flow arrive while the display is visible and the manipulator observes the noisy summaries (10); at $t = 1^-$ she chooses whether to cancel, paying $c + \varepsilon_L$ if she does; as in the baseline timeline, cancellation withdraws the *entire* first-round schedule

— the displayed component $\widehat{D}$ and the non-displayed component $\widehat{X}$ alike — and execution clears the schedule in full; the rest of the timeline is unchanged.

Assumption 1 (Reactive cross-side flow) *While the displayed order is live, cross-side market-order flow contains a component that responds to displayed depth. Cumulative reactive cross-side volume over $t = 0^+$ equals $\nu\theta\widehat{D}$, where $\nu \in [0, 1]$ is the visibility of displayed liquidity for the security and $\theta \sim \mathcal{N}(\bar{\theta}, \sigma_\theta^2)$, $\bar{\theta} > 0$, is the realized responsiveness of depth-conditioning traders, unknown to all players at $t = 0$. The reactive component is zero when no display is live and stops accruing once the display is removed.*

Discussion. This is the central new primitive. In the baseline model, displayed depth affects behavior only by lowering $\widehat{\Lambda}$; Assumption 1 adds agents who condition order submission on displayed imbalance, as execution algorithms and queue-reactive order-flow models do, and as Section 5.4 documents in reduced form (+6.90 cross-side trades per second when a spike is live relative to a matched pre-spike window). The randomness of θ is essential rather than decorative: with a deterministic response there would be nothing to learn from the tape, the cancellation decision would be made at $t = 0$, and there would be no within-episode timing to test. The stopping property — reactive flow ceases when the display is removed — is the natural consequence of the flow being a *response* to the display, and it is what generates the lead-lag asymmetry below.

Assumption 2 (Display-and-trigger configuration) *The displayed component may be posted on the side opposite to the manipulator's intended trade. The configuration of interest is $m < 0$ with a displayed bid $\widehat{D} > 0$; the baseline same-direction strategy is nested as $\widehat{D} = 0$.*

Discussion. Under the baseline loading $\beta_{M,0} > 0$, elicited cross-side buying would make the manipulator's subsequent buy more expensive and could never produce a positive cross-side trigger. The canonical spoof has the opposite geometry: display buy-side depth, induce buying, sell into the elevated price. Appendix B.6 shows that choosing $\widehat{D} > 0$ in this configuration is optimal when visibility is high enough, so the configuration arises endogenously. Conditioning on one orientation is without loss; the mirror case is symmetric, matching the empirical signing convention.

Assumption 3 (Interim information structure) *At $t = 1^-$ the manipulator observes $\phi_c = \nu\theta\widehat{D} + e_c$ and $\phi_f = \mu\zeta + e_f$, where $e_c \sim \mathcal{N}(0, \sigma_c^2)$ and $e_f \sim \mathcal{N}(0, \sigma_f^2)$ are noise, $\zeta \sim \mathcal{N}(0, \sigma_\zeta^2)$ is a latent execution-risk shock, $\mu > 0$, and $(\theta, \zeta, e_c, e_f)$ are mutually independent and independent of (v, ξ, u_0, u_1).*

Discussion. ϕ_c is the model counterpart of the trailing cross-side trade count and ϕ_f of the fill-direction count; Appendix B.7 maps the latent Gaussian signal into observed counts. Orthogonality of e_c to fundamentals is a deliberate simplification: aggressive buying is in reality good news about value, which creates an offsetting force *lowering* the cancellation response; Proposition 7 treats the contaminated case $\phi_c = \nu\theta\widehat{D} + \chi\widehat{v} + e_c$ in full and shows the trigger survives an explicit threshold on χ. The separation of ζ from θ is the formal statement of the paper's distinction between fill pressure (execution risk) and cross-side flow (evidence the signal worked).

Assumption 4 (Cancellation friction and latency shock) *Cancelling at $t = 1^-$ costs $c + \varepsilon_L$, where $c \geq 0$ is deterministic and ε_L, independent of everything else, has the logistic distribution $\Pr(\varepsilon_L \leq x) = \Lambda_C(x/s) = (1 + e^{-x/s})^{-1}$ with scale $s > 0$.*

Discussion. Without friction, the manipulator would cancel with probability one, and the hazard would be uninformative. The deterministic part bundles queue-priority loss, message costs, and expected enforcement costs; surveillance and enforcement changes are comparative statics in c. The logistic latency shock converts the threshold rule into exactly the link function Λ_C of the empirical hazard (19); any continuous strictly increasing shock distribution delivers the same qualitative results with a different link.

Assumption 5 (Sluggish pricing rule) *Market makers set $\widehat{\Lambda}$ once at $t = 0$, using the correct equilibrium cancellation probability $q^\dagger$, and do not revise it at $t = 1^-$; they do not observe (ϕ_c, ϕ_f) separately from total order flow.*

Assumption 6 (Persistent component of the elicited flow) *A fraction $\psi \in (0, 1]$ of the reactive response survives into round two: if the display was live at $t = 0^+$, round-two order flow contains the predictable component $\psi\nu\theta\widehat{D}$ in addition to $x_{M,1}$ and u_1. By $t = 2$ the manipulator has observed the realized round-one reactive flow and therefore knows θ.*

Discussion. This is the channel through which the elicited response is payoff-relevant: imbalance-conditioning algorithms split parent orders and momentum responses decay over seconds, so the manipulator who cancels and turns seller faces predictable uninformed buying pressure. Selling into predictable uninformed flow is profitable for exactly the reason in Rochet and Vila (1994): the baseline manipulator already trades against the observed liquidity order ($\gamma^*_{M,1} < 0$), and the persistent component is formally an observable addition to u_1, so no new profit mechanism is introduced. Setting $\psi = 0$ kills the trigger and collapses the extension to a fill-avoidance model with CSTE $= 0$; through the lens of the model, the empirical CSTE > 0 is direct evidence that $\psi\nu > 0$.

Assumption 7 (Modularity of the first-round strategy) *The manipulator's $t = 0$ schedule retains the baseline linear form in (ξ, u_0), augmented by the displayed component $\widehat{D}$ of Appendix B.6. The interim stage affects the $t = 0$ problem only through the equilibrium cancellation probability $q^\dagger$ replacing q and the additive option value of the display.*

Discussion. This keeps the extension tractable. The formal justification is an envelope argument: the interim signals are independent of (v, ξ, u_0, u_1) by Assumption 3, so learning at $t = 0^+$ does not update beliefs about fundamentals or liquidity, and the marginal conditions pinning down the baseline linear coefficients are unaffected up to $q \to q^\dagger$. The approximation is exact for the interim stage and first-order accurate for the $t = 0$ stage. The linear price-impact cross term between $\widehat{D}$ and the executed non-displayed component $\widehat{X}$ is carried *explicitly* in the branch values below (the $\widehat{\Lambda}\widehat{X}\widehat{D}$ term of (A69)); the remaining feedback from $\widehat{D}$ into the equilibrium coefficients is the depth-inflation channel already characterized by Proposition 1 in the small-$a\tau$ limit.

B.2 Interim beliefs

All interim inference is Gaussian projection. From Assumptions 1 and 3, (θ, ϕ_c) is bivariate normal with $\mathbb{E}[\phi_c] = \nu\widehat{D}\bar{\theta}$, $\mathrm{Cov}(\theta, \phi_c) = \nu\widehat{D}\sigma^2_\theta$, and $\mathrm{Var}(\phi_c) = \nu^2\widehat{D}^2\sigma^2_\theta + \sigma^2_c$, so

$$\widehat{\theta}(\phi_c) \equiv \mathbb{E}[\theta \mid \phi_c] = \bar{\theta} + \rho_c\left(\phi_c - \nu\widehat{D}\bar{\theta}\right), \qquad \rho_c = \frac{\nu\widehat{D}\,\sigma^2_\theta}{\nu^2\widehat{D}^2\sigma^2_\theta + \sigma^2_c}, \tag{A60}$$

$$\omega^2 \equiv \mathrm{Var}(\theta \mid \phi_c) = \sigma_\theta^2 - \frac{\left(\nu \widehat{D} \sigma_\theta^2\right)^2}{\nu^2 \widehat{D}^2 \sigma_\theta^2 + \sigma_c^2} = \frac{\sigma_\theta^2 \sigma_c^2}{\nu^2 \widehat{D}^2 \sigma_\theta^2 + \sigma_c^2}, \tag{A61}$$

and, by identical algebra with $(\bar{\theta}, \nu\widehat{D}, \sigma_\theta^2)$ replaced by $(0, \mu, \sigma_\zeta^2)$,

$$\widehat{\zeta}(\phi_f) \equiv \mathbb{E}[\zeta \mid \phi_f] = \rho_f \, \phi_f, \qquad \rho_f = \frac{\mu \, \sigma_\zeta^2}{\mu^2 \sigma_\zeta^2 + \sigma_f^2}. \tag{A62}$$

Substituting $\phi_c = \nu\theta\widehat{D} + e_c$ into (A60) gives the state-space representation in the primitives (θ, e_c), which do not depend on $(\nu, \widehat{D}, c)$,

$$\widehat{\theta} = \bar{\theta} + \rho_c \left(\nu \widehat{D} \left(\theta - \bar{\theta}\right) + e_c\right), \qquad \widehat{\theta} \sim \mathcal{N}\left(\bar{\theta}, \ \sigma_\theta^2 - \omega^2\right), \tag{A63}$$

the unconditional variance being $\rho_c^2 \mathrm{Var}(\phi_c) = \sigma_\theta^2 - \omega^2$ (the law of total variance). Two properties drive the results below and are recorded as a lemma.

Lemma 1 (Projection coefficients) *Write $g \equiv \nu\widehat{D} \geq 0$. (i) $\rho_c(g) = g\sigma_\theta^2/(g^2\sigma_\theta^2 + \sigma_c^2)$ satisfies $\rho_c(0) = 0$ and $\partial\rho_c/\partial g = \sigma_\theta^2(\sigma_c^2 - g^2\sigma_\theta^2)/(g^2\sigma_\theta^2 + \sigma_c^2)^2$, strictly positive iff $g < \sigma_c/\sigma_\theta$. (ii) The posterior signal-to-noise ratio $h(g) \equiv \rho_c \, g = g^2\sigma_\theta^2/(g^2\sigma_\theta^2 + \sigma_c^2) \in [0, 1)$ is strictly increasing in g on $(0, \infty)$, and $\omega^2 = (1 - h)\sigma_\theta^2$ is strictly decreasing. (iii) $\lambda \equiv \widehat{\Lambda}\psi\nu\widehat{D} = \widehat{\Lambda}\psi g$ is linear and strictly increasing in g.*

Proof. (i) is the quotient rule; (ii) $h'(g) = 2g\sigma_\theta^2\sigma_c^2/(g^2\sigma_\theta^2 + \sigma_c^2)^2 > 0$; (iii) by inspection. ■

Part (i) — *without a display there is nothing for cross-side flow to be informative about* — is the source of the spike-conditional amplification in Corollary 1. Part (ii) matters for Appendix B.8: the informativeness ρ_c per unit of flow is hump-shaped in g, but the total resolved information h is globally monotone under fixed noise, and saturates once the noise scales with the display.

B.3 Branch values at the cancellation stage

Fix the configuration of Assumption 2: $m < 0$, displayed bid $\widehat{D} > 0$. At $t = 1^-$ the manipulator compares the value of cancelling, V_C, with the value of letting the schedule stand, V_E. Cancellation withdraws the entire first-round schedule, so the execution branch clears both the displayed component $\widehat{D}$ and the non-displayed component $\widehat{X}$, while the cancellation branch clears neither and opens the round-two continuation. We evaluate the comparison risk-neutrally; Remark 2 restores the mean-variance penalty.

The cancellation branch. If she cancels, she enters round two facing the flow of Assumption 6. Knowing θ at $t = 2$ and observing $\widehat{u}_1$, she solves

$$\max_x \ x\left(m - \widehat{\Lambda}\left(x + \widehat{u}_1 + \psi\nu\theta\widehat{D}\right)\right), \tag{A64}$$

a strictly concave quadratic (second derivative $-2\widehat{\Lambda} < 0$), with first-order condition $m - 2\widehat{\Lambda}x - \widehat{\Lambda}\widehat{u}_1 - \lambda\theta = 0$, optimum and indirect profit

$$x^* = \frac{N}{2\widehat{\Lambda}}, \qquad \pi^* = x^*\left(N - \widehat{\Lambda}x^*\right) = \frac{N^2}{4\widehat{\Lambda}}, \qquad N \equiv m - \widehat{\Lambda}\widehat{u}_1 - \lambda\theta, \quad \lambda \equiv \widehat{\Lambda}\psi\nu\widehat{D} > 0. \tag{A65}$$

In the configuration $m < 0$ with $\theta > 0$, $x^* < 0$: she *sells* into the buying pressure her display elicited — the canonical spoof, produced by the model rather than asserted. Taking interim expectations with $\widehat{u}_1 \sim \mathcal{N}(0,1)$ and $\theta \mid \phi_c \sim \mathcal{N}(\widehat{\theta}, \omega^2)$ independent,

$$\begin{aligned}\mathbb{E}\left[N^2 \mid \phi_c\right] &= \mathbb{E}\left[(m-\lambda\theta)^2 \mid \phi_c\right] - 2\widehat{\Lambda}\,\mathbb{E}\left[m-\lambda\theta \mid \phi_c\right]\underbrace{\mathbb{E}[\widehat{u}_1]}_{=0} + \widehat{\Lambda}^2\underbrace{\mathbb{E}[\widehat{u}_1^2]}_{=1} \\ &= \left(m-\lambda\widehat{\theta}\right)^2 + \lambda^2\omega^2 + \widehat{\Lambda}^2,\end{aligned}$$

so that

$$V_C(\phi_c) = \frac{\left(m-\lambda\widehat{\theta}(\phi_c)\right)^2 + \widehat{\Lambda}^2 + \lambda^2\omega^2}{4\widehat{\Lambda}} \;-\; c \;-\; \varepsilon_L, \qquad \frac{\partial V_C}{\partial \phi_c} = \frac{\psi\nu\widehat{D}\,\rho_c\left(\lambda\widehat{\theta} - m\right)}{2} \;>\; 0 \tag{A66}$$

on $\widehat{\theta} > m/\lambda$, a strictly negative threshold (in particular whenever $\widehat{\theta} \geq 0$): a stronger observed response raises the posterior on θ, hence the predictable pressure she can sell against, hence the value of getting out of the displayed bid. V_C does not depend on ϕ_f: once the order is cancelled, fill pressure is irrelevant.

The execution branch. If she does not cancel, the schedule executes at the round-one clearing. The expected clearing price is pinned down by the baseline pricing rule: the clearing flow is $\widehat{y}_0 + \widehat{D}$, the manipulator knows $(\widehat{\xi}, \widehat{u}_0)$, the interim signals carry no information about $\widehat{v}$, and $\mathbb{E}[\widehat{v} \mid \mathcal{F}_{1^-}] = m$, so

$$\bar{p}(\widehat{D}) = \widehat{\Lambda}\left(\widehat{\beta}_N m + \widehat{\beta}_{M,0}\widehat{\xi} + (1+\widehat{\gamma}_{M,0})\widehat{u}_0\right) + \widehat{\Lambda}\widehat{D} \;\equiv\; \bar{p}_0 + \widehat{\Lambda}\widehat{D}, \tag{A67}$$

the second term being the display's own price impact. The interim value of the executed position — the display plus the non-displayed component — is

$$V_E(\phi_f) = \left(\widehat{D} + \widehat{X}\right)\left(m - \bar{p}(\widehat{D})\right) - k_\zeta\,\widehat{D}\,\widehat{\zeta}(\phi_f), \qquad \frac{\partial V_E}{\partial \phi_f} = -k_\zeta\widehat{D}\rho_f < 0, \qquad \frac{\partial V_E}{\partial \phi_c} = 0, \tag{A68}$$

where $k_\zeta > 0$ converts the execution-risk shock into a per-share adverse-selection loss on the displayed order (fills arrive disproportionately from informed sellers), and $\bar{p}_0 + \widehat{\Lambda}\widehat{X} > m$ is maintained so the bluff is costly in expectation. Since $\bar{p}(\widehat{D})$ and $\widehat{X}$ are $\mathcal{F}_{1^-}$-measurable, they affect no signal derivative: the $\widehat{X}$ terms shift the level of the cancellation surplus episode by episode — and are absorbed by the episode fixed effect in the empirical hazard — but contribute nothing to the flow sensitivities.

Remark 2 (Risk aversion) *With the baseline mean-variance penalty, the round-two problem* (A64) *gains* $-\frac{a}{2}\tau x^2$*; the first-order condition gives* $x^*_{\mathrm{MV}} = N/(2\widehat{\Lambda} + a\tau)$ *and*

$$\pi^*_{\mathrm{MV}} = x^*_{\mathrm{MV}}N - \left(\widehat{\Lambda} + \tfrac{a}{2}\tau\right)\left(x^*_{\mathrm{MV}}\right)^2 = \frac{N^2}{4\left(\widehat{\Lambda} + \frac{a}{2}\tau\right)},$$

so every formula below holds with $4\widehat{\Lambda}$ *replaced by* $4(\widehat{\Lambda} + \frac{a}{2}\tau)$ *in the denominators: magnitudes shrink, no sign changes. On the execution branch the penalty is quadratic in the fixed executed position* $\widehat{D}+\widehat{X}$ *and free of the signals, so* (A68) *is unchanged.*

B.4 The cancellation hazard: proof of Proposition 2

Define the cancellation surplus net of the latency shock,

$$\Delta(\phi_c,\phi_f) \equiv \frac{\left(m-\lambda\widehat{\theta}(\phi_c)\right)^2+\widehat{\Lambda}^2+\lambda^2\omega^2}{4\widehat{\Lambda}} - c - \left(\widehat{D}+\widehat{X}\right)(m-\bar{p}_0) + \widehat{\Lambda}\widehat{D}\left(\widehat{D}+\widehat{X}\right) + k_\zeta \widehat{D}\rho_f\phi_f, \tag{A69}$$

so that $V_C - V_E = \Delta - \varepsilon_L$ and the manipulator cancels iff $\varepsilon_L \le \Delta(\phi_c,\phi_f)$.

Proof of Proposition 2. By Assumption 4, $\Pr(C=1 \mid \phi_c,\phi_f) = \Pr(\varepsilon_L \le \Delta) = \Lambda_C(\Delta/s)$, which is (11). By the chain rule and (A66)–(A68),

$$\text{CSTE} = \Lambda_C'\left(\tfrac{\Delta}{s}\right)\frac{\psi\nu\widehat{D}\rho_c\left(\lambda\widehat{\theta}-m\right)}{2s} \;>\; 0, \tag{A70}$$

$$\text{FDE} = \Lambda_C'\left(\tfrac{\Delta}{s}\right)\frac{k_\zeta\widehat{D}\rho_f}{s} \;>\; 0, \tag{A71}$$

the first strictly positive on the branch $\widehat{\theta} > m/\lambda$, which has probability at least $\Phi\left(\bar{\theta}/\sqrt{\sigma_\theta^2-\omega^2}\right) > \frac{1}{2}$ by (A63) (the exact probability is $\Phi\left((\bar{\theta}-m/\lambda)/\sqrt{\sigma_\theta^2-\omega^2}\right)$, larger still since $m/\lambda < 0$) and contains all nonnegative posteriors. Dividing (A70) by (A71) and cancelling the common factor $\Lambda_C'(\Delta/s)\widehat{D}/s$, CSTE > FDE if and only if

$$\frac{\psi\nu\,\rho_c\left(\lambda\widehat{\theta}-m\right)}{2} \;>\; k_\zeta\rho_f, \tag{C2}$$

which is the trigger condition of the proposition: it holds when the display is visible (ν), the elicited response persistent (ψ) and informative (ρ_c), and the directional motive strong ($|m|$), and it fails for a trader whose displayed order is genuine exposed liquidity ($\psi\nu \approx 0$, k_ζ large). ■

Remark 3 (The backfire branch) *For $\widehat{\theta} < m/\lambda < 0$ — a posterior so negative that the display is estimated to have elicited* selling *— the cancellation value rises again, because the manipulator can then buy against elicited selling: V_C is U-shaped in $\widehat{\theta}$ with minimum at $\widehat{\theta} = m/\lambda$, and the hazard inherits that shape. The display-and-trigger prediction concerns the empirically relevant increasing branch; the complementary event is the model's counterpart of a display that simply fails.*

Remark 4 (A scale-free ratio) *At any interim state,* $\text{CSTE}/\text{FDE} = \psi\nu\rho_c(\lambda\widehat{\theta}-m)/(2k_\zeta\rho_f)$*, free of the latency scale s, of the link function (any continuous positive latency density gives the same ratio), and of additively separable controls. In principle this is an overidentifying restriction — the ratio should be stable across specifications and enforcement regimes even when the hazard level moves — but because the estimated fill coefficient is small, the empirical ratio is a low-powered statistic in this sample, and we do not lean on it.*

Proof of Corollary 1. At $\widehat{D} = 0$ both (A70) and (A71) carry an explicit factor $\widehat{D}$, and additionally $\rho_c(0) = 0$ and $\lambda = 0$ by Lemma 1(i), so the model-implied cross-to-fill gap is identically zero before the spike and strictly positive during it under (C2). Any cross-side cancellation sensitivity measured in a matched pre-spike window therefore reflects channels outside the display-and-trigger mechanism, and the within-episode gap must exceed the pre-spike gap. ■

Proposition 7 (Common-value contamination) *Suppose the cross-side signal is* $\phi_c = \nu\theta\widehat{D} + \chi\widehat{v} + e_c$ *with* $\chi \geq 0$. *Let* $\sigma_m^2 \equiv \mathrm{Var}[\widehat{v} \mid \widehat{\xi}] = \tau/(1+\tau)$, $\Sigma_\chi \equiv \nu^2\widehat{D}^2\sigma_\theta^2 + \chi^2\sigma_m^2 + \sigma_c^2$, *and* $\widetilde{\phi} \equiv \phi_c - \nu\widehat{D}\bar{\theta} - \chi m$. *Then the interim posteriors are* $\widehat{\theta} = \bar{\theta} + \rho_\theta\widetilde{\phi}$ *and* $\widetilde{m} \equiv \mathbb{E}[\widehat{v} \mid \widehat{\xi}, \phi_c] = m + \rho_v\widetilde{\phi}$ *with* $\rho_\theta = \nu\widehat{D}\sigma_\theta^2/\Sigma_\chi$ *and* $\rho_v = \chi\sigma_m^2/\Sigma_\chi$; *up to terms of order* χ^2 *and a signal-independent constant, the trigger derivative becomes*

$$\mathrm{CSTE}^\chi = \Lambda_C'\left(\tfrac{\Delta}{s}\right)\frac{1}{s}\left[\frac{\left(\lambda\widehat{\theta} - \widetilde{m}\right)(\lambda\rho_\theta - \rho_v)}{2\widehat{\Lambda}} - \left(\widehat{D} + \widehat{X}\right)\rho_v\left(1 - \widehat{\Lambda}\widehat{\beta}_N\right)\right],$$

FDE *is unchanged, and* $\mathrm{CSTE}^\chi > 0$ *at the mean interim state whenever*

$$\psi\,\nu^2\widehat{D}^2\sigma_\theta^2 \;>\; \frac{\chi\sigma_m^2}{\widehat{\Lambda}}\left(1 + \frac{2\widehat{\Lambda}\left(\widehat{D} + \widehat{X}\right)^+}{\lambda\bar{\theta} + |m|}\right), \tag{A72}$$

where $x^+ \equiv \max\{x, 0\}$.

Proof. Conditional on $\widehat{\xi}$, $(\theta, \widehat{v}, \phi_c)$ is jointly normal with $\theta \perp \widehat{v}$ and

$$\mathrm{Cov}(\theta, \phi_c) = \nu\widehat{D}\sigma_\theta^2, \qquad \mathrm{Cov}(\widehat{v}, \phi_c) = \chi\sigma_m^2, \qquad \mathrm{Var}(\phi_c) = \Sigma_\chi;$$

the projection theorem gives the posteriors, and $\mathrm{Cov}(\theta, \widehat{v} \mid \phi_c, \widehat{\xi}) = -\nu\widehat{D}\sigma_\theta^2\chi\sigma_m^2/\Sigma_\chi < 0$ (response and value compete to explain the same flow). At $t = 2$ the manipulator knows θ; the further update of $\widehat{v}$ from $\phi_c - \nu\theta\widehat{D} = \chi\widehat{v} + e_c$ has regression coefficient $O(\chi)$ and value $O(\chi^2)$, and is dropped. The round-two problem is (A64) with $m \to \widetilde{m}$; the interim expectation gains the deterministic covariance term $\lambda\nu\widehat{D}\sigma_\theta^2\chi\sigma_m^2/(2\widehat{\Lambda}\Sigma_\chi)$ (she sells when θ is high, and high θ means low $\widehat{v}$ given the flow), which shifts the level of Δ but no derivative. Hence $\partial V_C/\partial\phi_c = (\lambda\widehat{\theta} - \widetilde{m})(\lambda\rho_\theta - \rho_v)/(2\widehat{\Lambda})$. On the execution branch both the position value and the clearing price (A67) use $\widetilde{m}$, so $\partial V_E/\partial\phi_c = (\widehat{D} + \widehat{X})\rho_v(1 - \widehat{\Lambda}\widehat{\beta}_N)$: good news moves the worth of the standing executed position, net of the part the clearing price absorbs, in the direction of the net exposure $\widehat{D} + \widehat{X}$ (a net-sell exposure makes the term *help* the trigger). Subtracting and applying the chain rule gives CSTE^χ; $\zeta \perp v$ leaves FDE unchanged. At the mean state $(\widehat{\theta}, \widetilde{m}) = (\bar{\theta}, m)$, positivity is implied by $(\lambda\bar{\theta} + |m|)(\lambda\nu\widehat{D}\sigma_\theta^2 - \chi\sigma_m^2) > 2\widehat{\Lambda}(\widehat{D} + \widehat{X})^+\chi\sigma_m^2$ (the common factor Σ_χ^{-1} cancels and $0 < 1 - \widehat{\Lambda}\widehat{\beta}_N < 1$ by (6)); moving the $\chi\sigma_m^2$ term to the right-hand side and dividing by $\widehat{\Lambda}(\lambda\bar{\theta} + |m|)$, with $\lambda = \widehat{\Lambda}\psi\nu\widehat{D}$, delivers exactly (A72). ■

Proposition 7 is the formal version of the information-based alternative: the contamination is a race between squared display visibility and the value content of flow, which is why the amplification test and the visibility tests are the right places to look for the trigger even if short-horizon flow carries value information, and why the adverse-selection return proxies belong in the control set of (19) — they are the data analogue of partialling out $\chi\widehat{v}$.

B.5 Timing: proof of Proposition 3

To state the result precisely, make the within-spell accrual explicit. Let $\kappa : [0,1] \to [0,1]$ be strictly increasing with $\kappa(0) = 0$: if the display is live throughout $[0, t]$, cumulative reactive cross-side volume is $\nu\theta\widehat{D}\kappa(t)$. Fix a monitoring time t and horizon $h > 0$ inside the spell, write $\kappa_1 \equiv \kappa(t)$,

$\kappa_2 \equiv \kappa(t+h) - \kappa(t) > 0$, and decompose the signals into the level and the increment,

$$S_1 \equiv \phi_{c,t} = \nu\theta\widehat{D}\,\kappa_1 + \epsilon_1, \qquad S_2 \equiv \phi_{c,t+h} - \phi_{c,t} = \nu\theta\widehat{D}\,\kappa_2\,\mathbf{1}\{\text{display live on } (t,t+h]\} + \epsilon_2, \quad \text{(A73)}$$

with $\epsilon_i \sim \mathcal{N}(0,\sigma_i^2)$ and $(\theta,\epsilon_1,\epsilon_2,\varepsilon_L,\zeta,e_f)$ mutually independent. (We write S_i for the signal level and increment to avoid collision with the persistence parameter ψ.) The indicator is the stopping property of Assumption 1 in accrual form.

Lemma 2 (Two-signal projection) *If the display is live on $(t,t+h]$, then with $g_i \equiv \nu\widehat{D}\kappa_i$,*

$$\widehat{\theta}_{t+h} \equiv \mathbb{E}[\theta \mid S_1,S_2] = \bar{\theta} + w_1\left(S_1 - g_1\bar{\theta}\right) + w_2\left(S_2 - g_2\bar{\theta}\right), \qquad w_i = \frac{g_i/\sigma_i^2}{\sigma_\theta^{-2} + g_1^2/\sigma_1^2 + g_2^2/\sigma_2^2} > 0,$$

and $\mathbb{E}[\widehat{\theta}_{t+h} \mid \mathcal{F}_t] = \widehat{\theta}_t \equiv \mathbb{E}[\theta \mid S_1]$.

Proof. With $S = (S_1,S_2)'$, $g = (g_1,g_2)'$, $D = \mathrm{diag}(\sigma_1^2,\sigma_2^2)$: $\mathrm{Cov}(\theta,S) = \sigma_\theta^2 g'$, $\mathrm{Var}(S) = \sigma_\theta^2 gg' + D$, and by Sherman–Morrison the projection weights are $w' = \sigma_\theta^2 g' D^{-1}/(1+\sigma_\theta^2 g' D^{-1} g)$, the Bayesian precision-weighting formula; positivity is immediate and the tower property is the law of iterated projections. ■

Proof of Proposition 3. (i) Conditional on the display being live at $t+h$, the hazard there is $\Lambda_C(\Delta(\widehat{\theta}_{t+h},\phi_{f,t+h})/s)$, and by the chain rule and Lemma 2,

$$\frac{\partial}{\partial\phi_{c,t}}\Pr\left(C_{t+h}=1 \mid \mathcal{F}_{t+h}\right) = \Lambda_C'\left(\tfrac{\Delta}{s}\right)\frac{1}{s}\cdot\frac{\lambda\left(\lambda\widehat{\theta}_{t+h} - m\right)}{2\widehat{\Lambda}}\cdot w_1 \;>\; 0$$

on the branch $\widehat{\theta}_{t+h} > m/\lambda$; integrating over $(S_2,\phi_{f,t+h})$ gives the $\mathcal{F}_t$-conditional statement. (ii) Given $\mathcal{F}_t$ and a live display, $C_t = \mathbf{1}\{\varepsilon_L \le \Delta_t\}$ with Δ_t $\mathcal{F}_t$-measurable and $\varepsilon_L \perp (\theta,\epsilon_2)$, so C_t carries *no information* about S_2 beyond $\mathcal{F}_t$: the entire effect of conditioning runs through the causal switch in (A73). Computing both sides,

$$\mathbb{E}\left[S_2 \mid C_t=1,\mathcal{F}_t\right] - \mathbb{E}\left[S_2 \mid C_t=0,\mathcal{F}_t\right] = 0 - \nu\widehat{D}\kappa_2\,\widehat{\theta}_t = -\,\nu\widehat{D}\kappa_2\,\widehat{\theta}_t, \qquad \text{(A74)}$$

weakly negative on $\{\widehat{\theta}_t \ge 0\}$ and strictly so on $\{\widehat{\theta}_t > 0\}$, which is the display of the proposition. (Equation (A74) treats the display as live throughout $(t,t+h]$ on the no-cancellation branch; allowing further decision instants inside the window scales the second term by the survival probability, which attenuates the magnitude but cannot reverse the sign.) The heterogeneity statement is (A74) read across episodes: the reverse lead is proportional to the interim posterior on the elicited response, which is highest exactly in the episodes the trigger score identifies and near zero where there is no response for the removal to switch off. For the unconditional version, $\mathbb{E}[\widehat{\theta}_t \mid C_t=0] = \bar{\theta} - \mathrm{Cov}(\widehat{\theta}_t,\Lambda_C(\Delta_t/s))/(1-\bar{q}_t)$ with $\bar{q}_t \equiv \mathbb{E}[\Lambda_C(\Delta_t/s)]$; since $x \mapsto \Lambda_C(x/s)$ is $(4s)^{-1}$-Lipschitz, $|\mathrm{Cov}| \le \mathrm{sd}(\widehat{\theta}_t)\,\mathrm{sd}(\Delta_t)/(4s)$, so the unconditional reverse lead is strictly negative whenever $s > \mathrm{sd}(\widehat{\theta}_t)\,\mathrm{sd}(\Delta_t)/\big(4\bar{\theta}(1-\bar{q}_t)\big)$, a mild condition requiring only that the latency dispersion not be small relative to the dispersion of the surplus. ■

B.6 Equilibrium and the display decision: proof of Proposition 4

The fixed point. Write $L(q) \equiv \widehat{\Lambda}(q)$ for the pricing slope delivered by the baseline equilibrium, equation (7) evaluated at the equilibrium strategy coefficients, viewed as a function of the cancellation probability. We require only the following regularity: L is continuous on $[0,1)$ with

$0 < L_{\min} \leq L(q) \leq L_{\max} < \infty$ there. This holds at fixed primitives because the equilibrium coefficients are finite for any finite risk aversion and vary continuously with q; no monotonicity of L is used, and nothing is assumed at the boundary $q = 1$, where the equilibrium correspondence is discontinuous (Section 2.4). The extension supplies the reverse map: integrating the hazard over the interim signals and the manipulator's information,

$$q^*\left(\widehat{\Lambda};\nu,c\right) \equiv \mathbb{E}_{\widehat{\xi},\widehat{u}_0,\phi_c,\phi_f}\left[\Lambda_C\left(\frac{\Delta\left(\phi_c,\phi_f;\widehat{\Lambda},\nu,c\right)}{s}\right)\right] \in (0,1), \tag{A75}$$

strictly interior because $\Lambda_C \in (0,1)$ and Δ is finite a.s., and continuous in $\widehat{\Lambda}$ on the compact interval $[L_{\min}, L_{\max}] \subset (0,\infty)$ by dominated convergence (the integrands are polynomials in Gaussian variables over denominators bounded away from zero). Continuity on a compact interval gives $\bar{q} \equiv \max_{\widehat{\Lambda}\in[L_{\min},L_{\max}]} q^*(\widehat{\Lambda};\nu,c) < 1$.

Proof of Proposition 4, existence. $F(q) \equiv q^*(L(q);\nu,c)$ is continuous on $[0,\bar{q}] \subset [0,1)$ and maps it into $(0,\bar{q}]$: values of F never exceed $\bar{q}$ by construction. Hence $G(q) \equiv F(q) - q$ satisfies $G(0) = F(0) > 0$ and $G(\bar{q}) = F(\bar{q}) - \bar{q} \leq 0$, and the intermediate value theorem gives a fixed point $q^\dagger \in (0,\bar{q}] \subset (0,1)$ — the boundary $q = 1$ is never approached, so the discontinuity there is immaterial. Given $q^\dagger$, the baseline equilibrium objects evaluated at $q = q^\dagger$ and the cancellation rule (11) are mutually consistent by Assumption 7; if $|\partial_\Lambda q^* \cdot L'(q)| < 1$ on $[0,\bar{q}]$, F is a contraction there and the fixed point is unique. ■

Lemma 3 (Surplus moments) *Fix the episode-level primitives $(m, \widehat{X}, \bar{p}_0, \widehat{D}, \widehat{\Lambda})$ with $\widehat{D} > 0$. Then (i) $\mathbb{E}[(m - \lambda\widehat{\theta})^2 + \lambda^2\omega^2] = (m - \lambda\bar{\theta})^2 + \lambda^2\sigma_\theta^2$, so the ex-ante mean surplus*

$$\mathbb{E}[\Delta] = \frac{\left(m - \lambda\bar{\theta}\right)^2 + \lambda^2\sigma_\theta^2 + \widehat{\Lambda}^2}{4\widehat{\Lambda}} - c - \left(\widehat{D} + \widehat{X}\right)(m - \bar{p}_0) + \widehat{\Lambda}\widehat{D}\left(\widehat{D} + \widehat{X}\right) \tag{A76}$$

depends on the interim information structure only through λ: the resolved-variance gain in $\mathbb{E}[(m - \lambda\widehat{\theta})^2]$ is exactly offset by the flexibility term $\lambda^2\omega^2$, by the law of total variance. (ii) $\partial\mathbb{E}[\Delta]/\partial c = -1$. (iii) $\partial\mathbb{E}[\Delta]/\partial\nu = \frac{\psi\widehat{D}}{2}\left[\bar{\theta}\left(\lambda\bar{\theta} + |m|\right) + \lambda\sigma_\theta^2\right] > 0$.

Proof. (i) By (A63), $\mathrm{Var}(\widehat{\theta}) = \sigma_\theta^2 - \omega^2$, so $\mathbb{E}[(m - \lambda\widehat{\theta})^2] = (m - \lambda\bar{\theta})^2 + \lambda^2(\sigma_\theta^2 - \omega^2)$; adding $\lambda^2\omega^2$ cancels ω^2. Take expectations in (A69) with $\mathbb{E}[\phi_f] = 0$. (ii) is immediate; (iii) differentiates (A76) with $\partial\lambda/\partial\nu = \widehat{\Lambda}\psi\widehat{D}$ and $-m = |m|$. ■

Proof of Proposition 4, comparative statics (i). For the friction: $\partial\Delta/\partial c = -1$ pointwise, so differentiating (A75) under the integral, $\partial q^*/\partial c = -s^{-1}\mathbb{E}[\Lambda_C'(\Delta/s)] < 0$ exactly. For visibility: parameterize by the primitives $(\theta, e_c, \zeta, e_f)$ and let $G \equiv \partial\Delta/\partial\nu$ pointwise; G is a Gaussian polynomial with $\mathbb{E}[G] = \partial\mathbb{E}[\Delta]/\partial\nu > 0$ by Lemma 3(iii). G is not signed pointwise (for realizations revealing a weak response, more visibility lowers the surplus), so control the link curvature: since $\Lambda_C'(0) = \frac{1}{4}$, $\Lambda_C' \leq \frac{1}{4}$, and $\sup_x |\Lambda_C''(x)| = 1/(6\sqrt{3})$,

$$\frac{\partial q^*}{\partial\nu} = \frac{1}{s}\left\{\frac{1}{4}\mathbb{E}[G] - \mathbb{E}\left[\left(\tfrac{1}{4} - \Lambda_C'\left(\tfrac{\Delta}{s}\right)\right)G\right]\right\} \geq \frac{1}{s}\left\{\frac{1}{4}\mathbb{E}[G] - \frac{\mathbb{E}\left[|\Delta||G|\right]}{6\sqrt{3}\,s}\right\} > 0$$

whenever $s > s^* \equiv 2\mathbb{E}[|\Delta||G|]/(3\sqrt{3}\,\mathbb{E}[G])$, an explicit moment bound requiring the hazard not to be a deterministic threshold. In both cases the equilibrium effect follows from the implicit function

theorem applied to $q - q^*(L(q); \nu, c) = 0$ under the contraction condition, which guarantees a positive denominator. ■

The display decision. At $t = 0$, conditional on $m < 0$, the manipulator chooses $\widehat{D} \geq 0$ to maximize the net option value of the display,

$$W(\widehat{D}) = M(\widehat{D}) - M(0) - \kappa_m \widehat{D}^2, \qquad M(\widehat{D}) \equiv \mathbb{E}\left[\max\left\{V_C(\phi_c),\, V_E(\phi_f)\right\}\right], \tag{A77}$$

where $\kappa_m \widehat{D}^2$ is a convex posting cost (detection risk and other convex posting frictions; Assumption 7 — the *linear* price-impact externality of the display on the executed non-displayed component is not folded in here but carried explicitly by the $\widehat{\Lambda}\widehat{X}\widehat{D}$ term of (A69)). The logistic latency shock makes the option value available in closed form: using $\max\{a, b\} = b + (a - b)^+$, $V_C - V_E = \Delta - \varepsilon_L$, and $\mathbb{E}[(x - \varepsilon_L)^+] = \int_{-\infty}^{x} \Lambda_C(u/s)\, du = s \log(1 + e^{x/s})$,

$$W(\widehat{D}) = \widehat{D}\left(m - \bar{p}_0 - \widehat{\Lambda}\widehat{X}\right) - \widehat{\Lambda}\widehat{D}^2 + s\,\mathbb{E}\left[\mathrm{sp}\left(\frac{\Delta(\widehat{D})}{s}\right)\right] - s\,\mathrm{sp}\left(\frac{\Delta_0}{s}\right) - \kappa_m \widehat{D}^2, \qquad \mathrm{sp}(x) \equiv \log\left(1 + e^x\right), \tag{A78}$$

with $\mathrm{sp}' = \Lambda_C$. At $\widehat{D} = 0$ every signal-dependent term in (A69) carries a factor $\widehat{D}$ or $\rho_c(0) = 0$, so

$$\Delta_0 = \frac{m^2 + \widehat{\Lambda}^2}{4\widehat{\Lambda}} - c - \widehat{X}\,(m - \bar{p}_0) \quad \text{is signal-independent, and} \quad q_0 = \Lambda_C\left(\frac{\Delta_0}{s}\right) \tag{A79}$$

exactly; signal-independence of Δ_0 ($\mathcal{F}_0$-measurable, hence deterministic within the episode) is what makes the cancellation event independent of (ϕ_c, ϕ_f) at zero display, so cross terms such as $\mathbb{E}[\mathbf{1}\{\text{execute}\}\widehat{\zeta}]$ vanish. Expanding the surplus around $\widehat{D} = 0$ with $\ell \equiv \widehat{\Lambda}\psi\nu$, $a_1 \equiv \psi\nu\bar{\theta}|m|/2$, $b_1 \equiv |m|\psi\nu^2\sigma_\theta^2/(2\sigma_c^2)$, and $\bar{\kappa} \equiv \widehat{\Lambda}\psi^2\nu^2(\bar{\theta}^2 + \sigma_\theta^2)/4$,

$$\Delta(\widehat{D}) = \Delta_0 + \left(a_1 - \left(m - \bar{p}_0 - \widehat{\Lambda}\widehat{X}\right) + k_\zeta \rho_f \phi_f\right)\widehat{D} + \left(b_1 e_c + \bar{\kappa} + \widehat{\Lambda}\right)\widehat{D}^2 + O(\widehat{D}^3),$$

and differentiating (A78) under the integral sign gives the exact first two derivatives at zero,

$$W'(0) = q_0\, a_1 + (1 - q_0)\left(m - \bar{p}_0 - \widehat{\Lambda}\widehat{X}\right), \tag{A80}$$

$$W''(0) = -2\kappa_m - 2(1 - q_0)\widehat{\Lambda} + 2q_0\bar{\kappa} + \frac{\Lambda_C'(\Delta_0/s)}{s}\left[\left(a_1 + |m| + \bar{p}_0 + \widehat{\Lambda}\widehat{X}\right)^2 + k_\zeta^2 \rho_f \mu \sigma_\zeta^2\right]. \tag{A81}$$

The first displayed share is worth posting iff $W'(0) > 0$, i.e. iff

$$\nu \;>\; \nu^*(m, c) \equiv \frac{2\,(1 - q_0)\left(|m| + \bar{p}_0 + \widehat{\Lambda}\widehat{X}\right)}{q_0\, \psi\, \bar{\theta}\, |m|} \tag{A82}$$

(the dependence of ν^* on the episode's price-level primitives $(\bar{p}_0, \widehat{X})$ is suppressed in the notation; the numerator is positive under the maintained condition $\bar{p}_0 + \widehat{\Lambda}\widehat{X} > m$).

Proof of Proposition 4, display and prevalence (ii). *Existence of an optimal display:* sp is increasing and 1-Lipschitz, so $s\mathbb{E}[\mathrm{sp}(\Delta/s)] - s\,\mathrm{sp}(\Delta_0/s) \leq \mathbb{E}|\Delta(\widehat{D}) - \Delta_0| \leq c_1\widehat{D} + (\bar{\kappa} + \widehat{\Lambda})\widehat{D}^2$ with $c_1 \equiv \psi\nu|m|(\bar{\theta} + \sigma_\theta)/2 + |m| + \bar{p}_0 + \widehat{\Lambda}|\widehat{X}| + k_\zeta\rho_f\mathbb{E}|\phi_f|$, using $\mathbb{E}[\widehat{\theta}^2] + \omega^2 = \bar{\theta}^2 + \sigma_\theta^2$; substituting into (A78), the own-impact terms cancel and $W(\widehat{D}) \leq (c_1 + m - \bar{p}_0 - \widehat{\Lambda}\widehat{X})\widehat{D} - (\kappa_m - \bar{\kappa})\widehat{D}^2 \to -\infty$ whenever $\kappa_m > \bar{\kappa}$, so a maximizer exists on a compact set $[0, \bar{D}]$ and, if $W'(0) > 0$, every maximizer is strictly

positive. *Comparative statics:* by (A79), Δ_0 contains neither ν nor $\bar{\theta}$, so q_0 is free of ν and (A82) is a genuine single-crossing condition in ν; moreover $\partial W'(0)/\partial\nu = q_0\psi\bar{\theta}|m|/2 > 0$ and $\partial W'(0)/\partial c = (\partial q_0/\partial c)(a_1 + |m| + \bar{p}_0 + \widehat{\Lambda}\widehat{X}) < 0$ under the maintained condition $\bar{p}_0 + \widehat{\Lambda}\widehat{X} > m$, so for large κ_m (where $\widehat{D}^* = W'(0)/(2\kappa_m)(1 + O(\kappa_m^{-1}))$ by the maximizer bound and the first-order condition) the optimal display rises with visibility and falls with the friction. For κ_m large enough that $2\kappa_m$ exceeds the bounded curvature of the option-value term on $[0, \bar{D}]$, W is strictly concave there and the entry characterization is two-sided: $\widehat{D}^* > 0$ if and only if (A82) holds; for smaller κ_m, (A82) remains sufficient for entry. *The threshold:* $\partial\nu^*/\partial q_0 < 0$ and $\partial q_0/\partial c < 0$ give $\partial\nu^*/\partial c > 0$; writing $\nu^* = \big(2(1-q_0)/(q_0\psi\bar{\theta})\big)\big(1 + (\bar{p}_0 + \widehat{\Lambda}\widehat{X})/|m|\big)$ and holding the episode's price-level primitives $(\bar{p}_0, \widehat{X})$ fixed, both factors fall with $|m|$ (the second directly whenever $\bar{p}_0 + \widehat{\Lambda}\widehat{X} > 0$, the first because Δ_0 rises with $|m|$ provided $|m| > 2\widehat{\Lambda}\max\{-\widehat{X}, 0\}$), so stronger directional motives bluff at lower visibility. *Smoothness of the market-level margin:* the episode-level entry decision is a threshold in ν, but m (and, across securities and days, c, the price-level primitives, and the posting-cost environment) varies across episodes, so prevalence at the security level is $\Pr(\nu_j > \nu^*(m,c)) \times \mathbb{E}[\text{episodes} \mid \widehat{D}^* > 0]$, and since $\nu^*(m,c)$ has a continuous distribution induced by m and the episode primitives, prevalence is a smooth, strictly increasing function of ν_j and strictly decreasing in c — a monotone gradient, not a kink. The model therefore does *not* predict a sharp threshold or a hump-shaped response of the enforcement break across the visibility distribution, and no such shape is imposed or tested in Sections 7.1–7.2. ■

B.7 Measurement: from latent signals to counts

The model's interim signal ϕ_c is a latent Gaussian summary of reactive flow; the data record nonnegative, heavily right-skewed trade counts n. The standard bridge is the log-normal arrival specification: the measured count and the latent signal are linked by

$$\phi_c \;=\; \mu_0 + \sigma_0 \log(1+n), \qquad \sigma_0 > 0, \tag{A83}$$

so that $\log(1+n)$ is Gaussian, matching the skewness of observed counts. Both (μ_0, σ_0) are free parameters of the measurement map; we normalize the location to the reactive-flow mean, $\mu_0 = \nu\widehat{D}\bar{\theta}$, so that the coefficients below are exact (a general μ_0 shifts the linear coefficient without affecting its sign near the mean interim state). Substituting (A83) into the surplus (A69), the cancellation-hazard index as a function of the *measured* count is

$$\Delta(n) = \text{const} + \underbrace{\frac{\lambda\rho_c\left(\lambda\bar{\theta} - m\right)\sigma_0}{2\widehat{\Lambda}}}_{>0}\log(1+n) + \underbrace{\frac{\lambda^2\rho_c^2\sigma_0^2}{4\widehat{\Lambda}}}_{\geq 0}\left[\log(1+n)\right]^2 + k_\zeta\widehat{D}\rho_f\phi_f,$$

increasing in n throughout the empirically relevant range and *concave and saturating in the count itself* wherever the linear term dominates, since $\log(1+n)$ is concave and the quadratic coefficient is second order (it carries ρ_c^2). Three consequences. First, the concavity of the estimated hazard in trade counts — the binscatter of Section 5.2 — is the model's own prediction on the measured scale, with no appeal to link saturation. Second, the flexible spline basis in $\log(1+x)$ used in (19) is the correct reduced form: the model rationalizes the specification without replacing it, and no curvature or vertex-type restriction is testable on the count scale, because the location and scale (μ_0, σ_0) of the measurement map are free. Third, the sign predictions of Proposition 2 are unaffected, since (A83) is a monotone reparameterization.

B.8 Robust signs versus fragile levels

The extension's *sign* predictions — CSTE > 0, FDE > 0, CSTE $>$ FDE under (C2), the zero pre-spike gap of Corollary 1, the signed and heterogeneous reverse lead of Proposition 3, and the monotone prevalence gradient of Proposition 4 — require only that $\rho_c > 0$ when a display is live, $\rho_c = 0$ when it is not, and $k_\zeta \rho_f > 0$. Its *level* predictions in the visibility-size product $\nu\widehat{D}$ (how steeply the trigger strengthens with display size or visibility) additionally require two auxiliary stationarity assumptions, and both are strong.

First, fixed signal noise. If instead the interim noise scales with the display — larger displays sit in proportionally noisier flow, as when reactive and unrelated flow grow together with attention — then with $\sigma_c^2(\widehat{D}) = \sigma_{c0}^2 + \bar{\sigma}^2\nu^2\widehat{D}^2$ the signal-to-noise ratio of Lemma 1(ii) becomes

$$h = \frac{\nu^2\widehat{D}^2\sigma_\theta^2}{\nu^2\widehat{D}^2\left(\sigma_\theta^2 + \bar{\sigma}^2\right) + \sigma_{c0}^2} \nearrow \frac{\sigma_\theta^2}{\sigma_\theta^2 + \bar{\sigma}^2} < 1 :$$

h still vanishes at $\widehat{D} = 0$ (so Corollary 1 is untouched) but *saturates* rather than approaching one, flattening the informativeness channel at empirically relevant display sizes. Second, unbounded tradable persistence. The payoff channel $\lambda = \widehat{\Lambda}\psi\nu\widehat{D}$ grows linearly only if the manipulator can trade against the elicited flow without limit; with any execution-capacity or participation constraint, λ is increasing and concave in $\nu\widehat{D}$ and likewise saturates. Under either relaxation the cross-side sensitivity is increasing but flat in $\nu\widehat{D}$ over most of its range, while every sign prediction above is unchanged.

This is why the empirical strategy of Sections 5–7 takes signs, not magnitudes, to the data. The within-episode and cross-sectional tests are reduced forms whose signs the model supplies; monotonicity of trigger *strength* in display size or visibility is not among the maintained restrictions, and the model, under the realistic noise and capacity regimes above, does not demand it.

Appendix C Selection Theorem and Reversal-Based Tests

This appendix supplies the formal content underlying the unconditional-null diagnostic of Section 4. The selection theorem characterizes the algebraic relationship between unconditional and reversal-conditional means of a signed reversal variable. The wrong-population lemma shows that under the model's outcome restriction, a reversal-based filter selects against rather than for the manipulation population.

C.1 The selection identity

We formalize the mechanical relationship between the unconditional mean and the reversal-conditional mean. Let R_i denote the signed reversal of episode i, oriented so that the reversal emphasized in prior work corresponds to $R_i < 0$. Let

$$D_i = \mathbf{1}\{R_i < 0\}$$

denote the directional-reversal classifier. Write

$$\mu = \mathbb{E}[R_i], \qquad \sigma_R^2 = \mathrm{Var}(R_i), \qquad p = \Pr(D_i = 1),$$

and assume $p \in (0,1)$ and $\sigma_R^2 > 0$. Define

$$\mu_1 = \mathbb{E}[R_i \mid D_i = 1], \qquad \mu_0 = \mathbb{E}[R_i \mid D_i = 0].$$

Theorem 1 (Reversal selection identity) *Under the conditions stated above,*

$$\mu_1 - \mu = (1-p)(\mu_1 - \mu_0) = \frac{1-p}{p}(\mu - \mu_0). \tag{A84}$$

Moreover,

$$|\mu_1 - \mu| \le \sigma_R \sqrt{\frac{1-p}{p}}. \tag{A85}$$

Thus, conditioning on the realized sign of the post-collapse return mechanically moves the selected mean away from the unconditional mean. The magnitude of this movement is governed by the scale and shape of the unconditional distribution of R_i, not by a structural manipulation parameter.

Proof. The law of total expectation gives

$$\mu = p\mu_1 + (1-p)\mu_0.$$

Subtracting μ from μ_1 and rearranging yields

$$\mu_1 - \mu = (1-p)(\mu_1 - \mu_0) = \frac{1-p}{p}(\mu - \mu_0),$$

which proves (A84).

For the envelope, decompose the variance into within-cell and between-cell components:

$$\sigma_R^2 = p \operatorname{Var}(R_i \mid D_i = 1) + (1-p) \operatorname{Var}(R_i \mid D_i = 0) + p(1-p)(\mu_1 - \mu_0)^2.$$

Since the within-cell variances are nonnegative,

$$p(1-p)(\mu_1 - \mu_0)^2 \le \sigma_R^2.$$

Using $\mu_1 - \mu = (1-p)(\mu_1 - \mu_0)$ gives

$$|\mu_1 - \mu| \le \sigma_R \sqrt{\frac{1-p}{p}},$$

which proves (A85). ■

The theorem should be read as an identity and an envelope, not as a lower bound. It does not say that every reversal-conditioned sample must have a large negative mean. It says that once the sample is selected on $R_i < 0$, the selected mean is mechanically shifted below the unconditional mean, and the size of that shift is a property of the unconditional reversal distribution.

In the eligible sample of Table 2, the unconditional mean is

$$\mu = -1.29 \text{ bps},$$

the reversal-conditioned mean is

$$\mu_1 = -11.66 \text{ bps},$$

and

$$p = \frac{5{,}620{,}597}{10{,}643{,}414} = 0.528.$$

The identity then implies

$$\mu_0 = \frac{\mu - p\mu_1}{1-p} = 10.3 \text{ bps}.$$

Thus the reversal-conditioned mean is paired mechanically with a positive mean in the complementary sample. The selected negative mean is not independent evidence of a strategic state. It is the arithmetic consequence of sorting the same event population by the realized sign of the post-collapse return.

The envelope is also consistent with the observed magnitudes. Using $\sigma_R = 218$ bps gives

$$\sigma_R\sqrt{\frac{1-p}{p}} = 218\sqrt{\frac{0.472}{0.528}} \approx 206 \text{ bps}.$$

The observed shift,

$$\mu_1 - \mu = -10.37 \text{ bps},$$

lies well inside this envelope. The point is not that selection must generate a reversal of any particular size. The point is that a reversal obtained by conditioning on $R_i < 0$ is an outcome of the selection rule itself. It cannot then be used as independent validation of manipulation.

A threshold version of the same argument is immediate. Let

$$D_i^c = \mathbf{1}\{R_i < c\}, \qquad p_c = \Pr(R_i < c) \in (0,1),$$

and define

$$\mu_c^- = \mathbb{E}[R_i \mid R_i < c], \qquad \mu_c^+ = \mathbb{E}[R_i \mid R_i \geq c].$$

Then

$$\mu = p_c\mu_c^- + (1-p_c)\mu_c^+,$$

so

$$\mu_c^- - \mu = (1-p_c)(\mu_c^- - \mu_c^+). \tag{A86}$$

The corresponding envelope is

$$|\mu_c^- - \mu| \leq \sigma_R\sqrt{\frac{1-p_c}{p_c}}. \tag{A87}$$

If $c < \mu$, then the selected conditional mean is mechanically below the unconditional mean:

$$\mathbb{E}[R_i \mid R_i < c] \leq c < \mu.$$

As the threshold moves farther into the left tail, p_c decreases and the envelope in (A87) expands.

This is the selection problem in its most general form. Any rule that selects events using the realized post-collapse price path will mechanically alter the selected mean of that same price path. A useful diagnostic is the symmetric classifier $D_i' = \mathbf{1}\{R_i > 0\}$. Applied to the eligible sample, it produces a positive conditional mean of comparable magnitude and opposite sign. The symmetry is mechanical. It reflects the sorting rule, not a separate economic mechanism.

C.2 What the conditioning identifies under the alternative

The selection theorem describes the null. The conditioning is also uninformative under the alternative that some episodes are genuine manipulation, in that the object the conditioning identifies is not the object the theory predicts. Let $T_i \in \{0,1\}$ denote the trigger score, with $T_i = 1$ for episodes generated by the display-and-trigger strategy of Section 2. The model's outcome-linkage restriction implies two ordinal predictions,

$$\mathbb{E}[\text{Dislocation}_i \mid T_i = 1] > \mathbb{E}[\text{Dislocation}_i \mid T_i = 0], \tag{A88}$$

$$\mathbb{E}[R_i \mid T_i = 1] \geq \mathbb{E}[R_i \mid T_i = 0]. \tag{A89}$$

The first is sharp, in that the manipulator profits during the dislocation. The second is the empirically counterintuitive prediction that R_i is less negative under manipulation because the manipulator exists.

Lemma 4 (Tail draws the wrong population) *Let $T_i \in \{0,1\}$ with $\Pr(T_i = 1) \in (0,1)$, and suppose the conditional distributions $F(\cdot \mid T_i = 1)$ and $F(\cdot \mid T_i = 0)$ of R_i are ordered by first-order stochastic dominance, with $F(\cdot \mid T_i = 1)$ dominating. Then for every c in the interior of the support,*

$$\Pr(T_i = 1 \mid R_i < c) \;\leq\; \Pr(T_i = 1), \tag{A90}$$

with strict inequality whenever the dominance is strict at c. The lower tail of R_i is enriched with $T_i = 0$ episodes relative to the unconditional sample.

Proof. By Bayes' rule, $\Pr(T_i = 1 \mid R_i < c) = \Pr(T_i = 1)\, F(c \mid T_i = 1)/F(c)$, where $F(c) = \pi_T F(c \mid T_i = 1) + (1-\pi_T)F(c \mid T_i = 0)$. First-order stochastic dominance gives $F(c \mid T_i = 1) \leq F(c \mid T_i = 0)$, hence $F(c \mid T_i = 1) \leq F(c)$, hence $\Pr(T_i = 1 \mid R_i < c) \leq \Pr(T_i = 1)$. ■

A worked mixture example clarifies the mechanism. Consider a stylized DGP with a fraction π_T of manipulation events with $R_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and a fraction $1 - \pi_T$ of non-manipulation events with $R_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$. The model predicts $\mu_1 > \mu_0$, with $\mu_1 \approx 0$ and $\mu_0 < 0$. Setting $\pi_T = 0.76$, $\mu_1 = 0$, $\mu_0 = -4$, $\sigma_1 = \sigma_0 = 200$ bps reproduces an unconditional mean near -1 bps. A back-of-envelope calculation yields $\Pr(T_i = 1 \mid R_i < 0) \approx 0.74 < 0.76 = \Pr(T_i = 1)$, confirming Lemma 4. The conditional mean reversal in this DGP is approximately -160 bps, dominated by mechanical truncation and not by a shift in mixture composition.

Appendix D Detection Algorithm and Data Filters

D.1 Quote filters

Restrict to regular trading hours (9:35–16:00 ET) and exclude quotes with non-positive bid/ask, negative spread, or invalid NBBO indicator. Exclude trades with non-zero correction indicator or sale-condition codes for contingent/stopped/out-of-sequence transactions. Lee-Ready zero-lag classification.

D.2 Rolling depth statistics

For each side s, compute $\overline{D}_{s,t}$ and $\sigma_{D,s,t}$ over a trailing 300 s window with at least 30 updates. Spike: $z_{s,t} \geq K$, $D_{s,t} - \overline{D}_{s,t} \geq 1{,}000$ shares, $D_{s,t}/\overline{D}_{s,t} \geq 2$. Baseline $K = 3$.

D.3 Collapse detection

Forward search over 120 s for the first update where depth returns within one standard deviation of pre-spike mean. Events that do not collapse within the window are discarded.

D.4 Duration filter and de-duplication

Minimum duration 10 ms (SIP consolidation flicker). Minimum inter-event gap 120 s within each side-symbol-day.

D.5 Threshold sweep

The detection algorithm is robust to threshold choices. Sweeping $K \in \{2.0, 2.5, 3.0, 3.5, 4.0, 5.0\}$, minimum depth $\in \{0, 500, 1{,}000, 2{,}000, 5{,}000\}$ shares, and minimum duration $\in \{0, 0.01, 0.1, 1.0\}$ s, the mean signed reversal remains negligible and the CSTE results stable across specifications.

D.6 Annual event breakdown

Table D.1: **Phantom-depth events by year.** This table reports the annual breakdown of the eligible event population. The 2020 peak coincides with the COVID volatility shock operating through the z-score detector. The sustained decline to 11.1–11.7 events per security-day in 2021–2023 reflects the March 2021 regime change documented in Section 7.2.

Year	Days	Symbols	N events	Events/security-day	Med. CS delay (s)	% Cancel
2017	249	471	1,621,599	13.8	0.22	78.6
2018	251	490	1,803,121	14.7	0.57	81.6
2019	252	494	1,727,433	13.9	0.56	80.5
2020	253	494	2,013,377	16.1	0.28	77.3
2021	252	493	1,457,851	11.7	0.15	77.6
2022	251	481	1,385,971	11.5	0.28	80.7
2023	124	462	634,062	11.1	0.21	79.8
Total	1,632	494	10,643,414	14.2	0.32	79.4

D.7 Categorical composition

Table D.2 describes the eligible event population along five dimensions. Panel A separates episodes that end through cancellation from episodes that end through execution against the displayed quote. Panel B reports whether the spike occurs on the bid or ask side. Panel C separates episodes by whether the five minute signed reversal moves against the spike side. Panel D measures how quickly the market responds on the opposite side of the book, using the delay from spike onset to the first cross-side trade. Panel E reports how much of the displayed depth on the spike side is executed during the live window.

Table D.2 shows that the eligible event population is balanced across sides, dominated by cancellations but not mechanically defined by them, and accompanied by substantial market response while the displayed depth is live. Cancellations account for 79.4% of episodes, but 20.6% end through execution against the displayed quote, so the sample is not simply a collection of cancelled orders. The bid and ask sides are almost exactly balanced, which helps rule out a directional asymmetry in the detector itself.

The reversal split is also informative. Directional reversals occur in 52.8% of episodes, only modestly above one half. Thus, reversal is common, but it is not a defining feature of the eligible

Table D.2: **Composition of eligible phantom-depth events.** This table reports the categorical composition of the eligible phantom-depth event population across five dimensions. The sample consists of the 10,643,414 episodes satisfying the baseline detection thresholds: collapsed depth, $z \geq 3$, and $\Delta D \geq 1{,}000$ shares. The sample covers 494 S&P 500 securities and 1,632 trading days from January 2017 through June 2023. Panel A separates events that end in cancellation from events that end in execution against the displayed quote. Panel B reports the bid and ask balance. Panel C reports the share of episodes whose five minute post-collapse signed reversal R_i is negative, corresponding to a directional reversal against the spike side. This subsample appears as the third row of Table 2. Panel D reports the distribution of cross-side response delays, measured from spike onset to the first trade on the opposite side. The row "No cross-side trade in spell" covers episodes with no opposite-side trade during the live window. Panel E reports the executed share of displayed depth on the spike side over the live window.

	N		Share
Panel A. Collapse type			
Cancel	8,447,479		79.4%
Fill	2,195,935		20.6%
Panel B. Side of the spike			
Bid	5,337,058		50.1%
Ask	5,306,356		49.9%
Panel C. Five minute signed reversal			
$R_i < 0$ directional reversal	5,620,597		52.8%
$R_i \geq 0$ no reversal	5,022,817		47.2%
Panel D. Cross-side response delay			
No cross-side trade in spell	190,852		1.8%
$[0, 50]$ ms	4,374,198		41.1%
$(50, 100]$ ms	254,287		2.4%
$(100, 500]$ ms	912,187		8.6%
$(500, 1{,}000]$ ms	629,343		5.9%
$(1, 5]$ s	2,011,066		18.9%
> 5 s	2,271,481		21.3%
Panel E. Executed share of displayed depth			
Zero fill	3,980,637		37.4%
Partial fill, $\in (0, 1)$	6,662,777		62.6%
Mean executed share		42.9%	
Median executed share		3.5%	

population. This reinforces the point in Section 5.1: large reversals arise when the sample is conditioned on the post-event price path, not because collapsed depth mechanically identifies a manipulation state.

The cross-side response distribution shows that most episodes are accompanied by trading on the opposite side of the book while the display is live. Only 1.8% of episodes have no cross-side trade during the spell, and 41.1% have a cross-side trade within 50 milliseconds. At the same time, the long right tail in response delays shows that the sample contains slower episodes as well, which is useful because the timing tests can compare rapid display-and-trigger episodes to otherwise similar but slower episodes.

Panel D is a waiting-time distribution, so the bins should not be compared only by their raw shares. The first bin is narrow and nevertheless contains 41.1% of episodes, showing that cross-side activity often arrives almost immediately after the depth spike. The $(1, 5]$ second bin is also large in levels, but it spans four seconds, so its implied arrival rate is much lower. It captures the slower

tail of episodes in which cross-side trading eventually occurs, but not at the subsecond horizon emphasized by the timing tests.

Finally, the executed-share distribution shows that exposure to execution risk is highly skewed. A large fraction of episodes have no fill, but the mean executed share is 42.9% while the median is only 3.5%. This gap indicates that most episodes involve little execution of the displayed depth, while a smaller number experience substantial fills. That pattern is consistent with the idea that displayed depth is often withdrawn before much of it trades, but it also makes clear why the empirical design needs to distinguish strategic withdrawal from ordinary fill avoidance.

Appendix E Additional details of empirical tests

E.1 Construction of Spell Panels

For each detected episode, extract NBBO updates and trades in $[s_i - 2\text{s}, \tau_i + 2\text{s}]$. At each update, record displayed depth on the spike side together with same-side and opposite-side total depth, the bid-ask spread in basis points, mid-quote returns at horizons $\{100, 500, 1{,}000, 5{,}000\}$ ms, a local volatility estimator constructed as the rolling standard deviation of mid-quote returns, rolling counts of fill-direction, cross-side, and total trades over windows $\{100, 500, 1{,}000, 5{,}000\}$ ms, non-trade depth reductions exceeding 100 shares as cancel events, and forward cancel indicators at horizons $\{50, 100, 500, 1{,}000\}$ ms.

Trade classification uses Lee and Ready (1991). The 5% stratified sample retains 532,170 events, of which 527,019 are usable after merging panel data. Sample stratification is by year and event-day to ensure representativeness, with medians within 0.6% of the full population on key variables.

E.2 Cross-Fitting Procedure

The cross-fitted trigger score $\widehat{T}_i^{(-k)}$ is constructed as follows.

1. Partition 527,019 episodes into $K = 5$ disjoint folds at the episode level (not the row level), preserving within-episode dependence.
2. For each fold k, estimate the hazard (19) on the union of folds $\{1, \ldots, K\} \setminus \{k\}$, yielding $\widehat{\theta}^{(-k)}$.
3. For each episode i in fold k, compute either the average score $\widehat{T}_i^{(-k)}$ in (23) or the maximum in (24).
4. In second-stage regressions for the model-implied moments, $\widehat{T}_i^{(-k)}$ is an excluded regressor (not used in the hazard estimation that produced it and not constructed from any post-event outcome).

Cross-fitting controls own-observation bias, in that the score for episode i does not use i's own data. The choice of $K = 5$ trades off bias (smaller for larger K) against variance in the estimated $\widehat{\theta}^{(-k)}$. Robustness to $K \in \{3, 10\}$ has been confirmed in pilot exercises and produces second-stage coefficients within one standard error of the $K = 5$ baseline reported in the main text.

E.3 Permutation Inference

The within-episode permutation test randomly swaps cross-side and fill-direction labels within each episode with probability one-half, preserving total activity and within-episode time structure. The estimated CSTE under the permutation should equal the FDE in expectation, and the gap between true and permuted CSTE quantifies behavioral specificity.

For each of 200 draws per sample year (1,000 total), the CSTE was re-estimated. Across every year, the true CSTE exceeded the permuted distribution by 12 to 34 permutation standard deviations. The mean gap between true and permuted CSTE is 0.028.

Appendix F Additional Figures

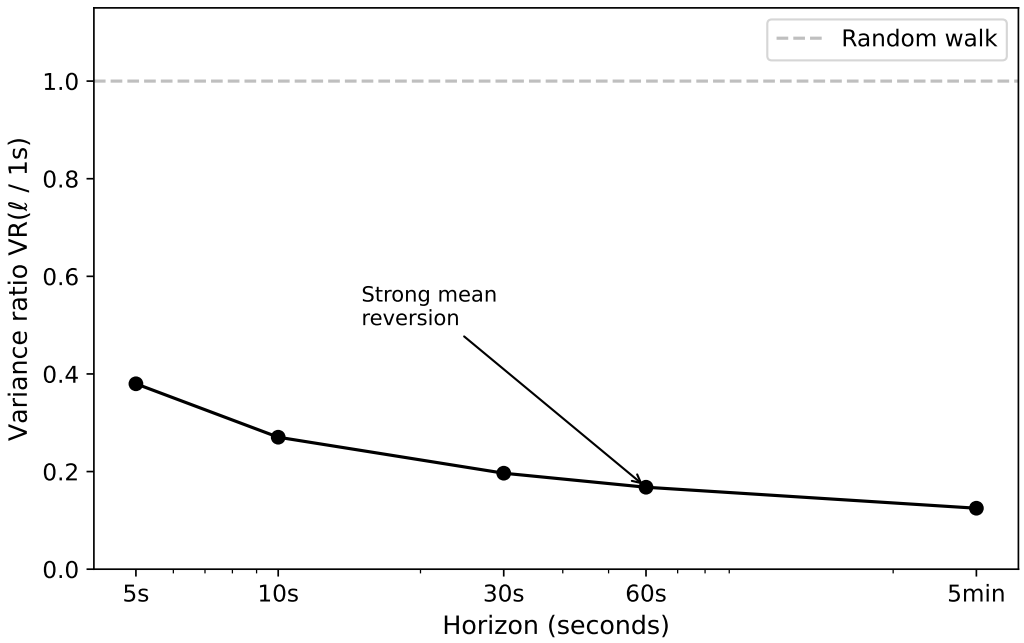


Figure F.2: Post-collapse variance ratios.
This figure plots post-collapse variance ratios $VR(\ell/1\text{s})$ at horizons $\ell \in \{5, 10, 30, 60, 300\}$ s. Under a random walk, $VR = 1$. All horizons show strong mean reversion ($VR \ll 1$), declining from 0.38 at 5 s to 0.13 at 300 s. $N = 532{,}170$ events.

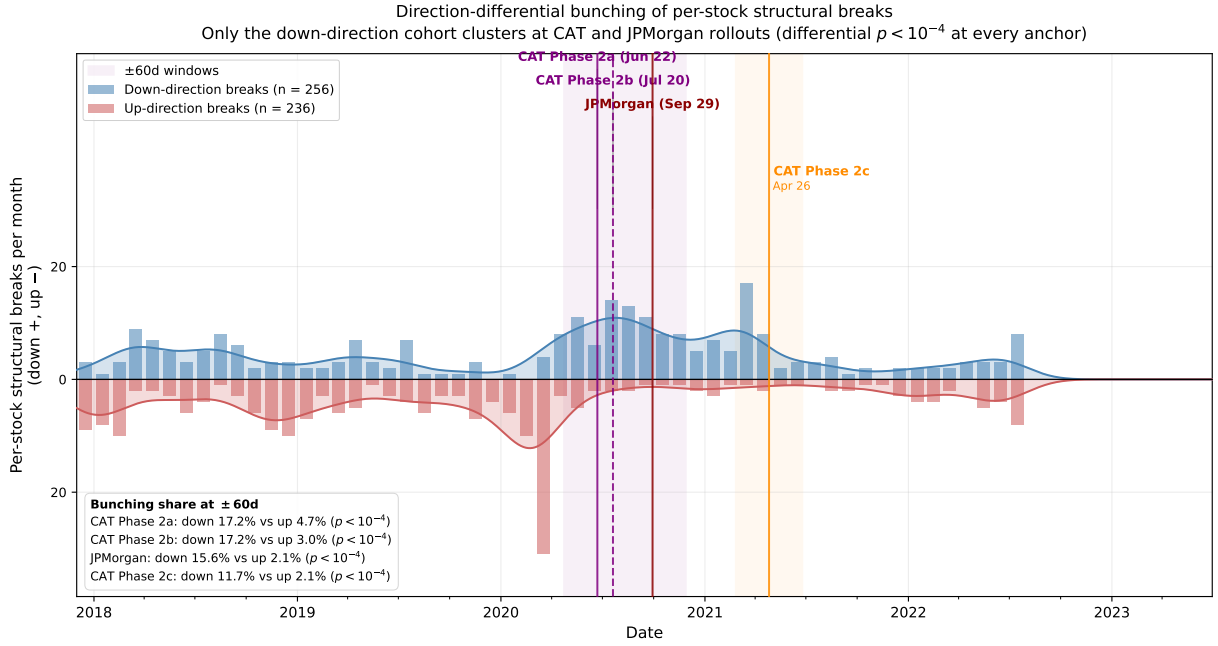


Figure F.3: Direction-differential bunching of per-stock structural breaks

This figure plots a mirror histogram of monthly per-stock break dates separately for the down- and up-direction cohorts. Down-direction breaks (256 securities, blue) are plotted upward; up-direction breaks (236 securities, red) are plotted mirrored downward. Smooth curves are Gaussian-kernel density estimates scaled to expected counts per month. Vertical lines mark CAT Phase 2a (22 June 2020), Phase 2b (20 July 2020), the JPMorgan settlement (29 September 2020), and CAT Phase 2c (26 April 2021); ±60-day windows are shaded. The down-direction cohort visibly bunches at the 2020 events and around Phase 2c; the up-direction cohort is anti-clustered around the 2020 anchors (with no countervailing peak elsewhere in 2020). Bunching shares and differential permutation p-values are in the lower-left annotation and in Table 12.

Table G.3: **IV-identified $\widehat{\Lambda}$ in the live phantom-depth window.** This table reports IV-identified price-impact estimates across regimes (pre, live, post) and trigger-score quartiles. Estimates are 2SLS or IVGMM versions of equation (A91) with SPY mid-quote return as the instrument for stock-level signed order flow at 250-millisecond bin frequency. Standard errors in parentheses, clustered at the episode level. *Panel A.* By trigger-score quartile $\widehat{T}_i^{(-k)}$ from Section 5.5, formed at the bin level. The t-statistic tests $\widehat{\Lambda}^{\text{live}}/\widehat{\Lambda}^{\text{pre}} = 1$ against the model's prediction of a ratio below one, with standard error computed by the delta method under approximate independence of the two regression estimates. The depth-inflation channel is identified in Q_3 ($t = -3.21$) and Q_4 ($t = -4.23$), where the model identifies manipulation, and is absent in Q_1 and Q_2. *Panel B.* Pooled across quartiles with one and two lags of the SPY return as additional instruments. The Hansen J overidentification test fails to reject in both overidentified specifications. First-stage F-statistics on the excluded instruments range from 3.3×10^3 to 4.3×10^8 across the three endogenous regressors. Sample is the 235-day intersection of the spell panel with the SPY-NBBO extraction window, $n = 300{,}291$ bins after winsorization in Panel B.

Panel A. By trigger-score quartile

Quartile of $\widehat{T}_i^{(-k)}$	Q_1 (low)	Q_2	Q_3	Q_4 (high)
$\widehat{\Lambda}^{\text{pre}}$ (bps/SD)	39.98	11.11	10.94	5.39
	(27.00)	(1.71)	(2.03)	(1.05)
$\widehat{\Lambda}^{\text{live}}$ (bps/SD)	39.50	16.71	6.29	2.83
	(11.04)	(2.82)	(0.86)	(0.25)
$\widehat{\Lambda}^{\text{live}}/\widehat{\Lambda}^{\text{pre}}$	0.99	1.50	**0.58**	**0.53**
t-statistic (ratio vs. 1)	−0.02	1.47	**−3.21**	**−4.23**
N pre bins	12,793	10,750	10,965	11,798
N live bins	35,640	37,684	37,462	36,634

Panel B. Pooled with Hansen overidentification test

Specification	Instr.	$\widehat{\Lambda}^{\text{pre}}$	$\widehat{\Lambda}^{\text{live}}$	$\widehat{\Lambda}^{\text{post}}$	J (df)	p
Just-identified	3	15.12 (1.94)	7.63 (0.74)	5.14 (0.41)	—	—
Overid., one lag	6	16.00 (2.02)	7.22 (0.58)	5.00 (0.35)	6.43 (3)	0.09
Overid., two lags	9	16.19 (2.02)	6.86 (0.51)	4.94 (0.33)	8.93 (6)	0.18

Appendix G Additional Robustness Tables

G.1 IV-identified price impact in the live window

This appendix reports the instrumental-variables price-impact estimates summarized in Section 6.3. The model predicts that the market appears deeper while phantom depth is live, so inverse market depth, the price impact per unit of signed flow, should be lower during the live window than before the spike. A simple OLS regression of price changes on signed flow is not the right object because the live window contains the manipulator's own taker trade, which is aligned with the price movement the manipulator anticipates and confounds price impact with endogenous trading. We therefore instrument stock-level signed flow with contemporaneous SPY midquote returns at the 250 millisecond frequency and estimate

$$\Delta P_{i,t}^{w} = \alpha^{w} + \widehat{\Lambda}^{w}\,\text{SignedFlow}_{i,t}^{w} + \varepsilon_{i,t}^{w}, \quad w \in \{\text{pre}, \text{live}, \text{post}\}. \tag{A91}$$

The pre-window is $[s_i - 2\text{s}, s_i - 100\text{ms})$, the live window is the spell itself, and the post-window is $(\tau_i, \tau_i + 60\text{s}]$. Signed flow is instrumented with the SPY midquote return in the same 250 millisecond bin.

The strongest test sorts episodes by the cross-fitted trigger score and estimates (A91) within trigger-score quartiles. The model predicts the live-window drop in $\widehat{\Lambda}$ where the behavioral evidence identifies phantom depth. That is what Panel A of Table G.3 shows. In the top quartile, the ratio

Table G.4: **Enforcement shocks and phantom-depth frequency.** This table reports security-day panel coefficients on $\text{Post}_{e,t}$ indicators around the four enforcement announcements. The estimating equation is (31), with security fixed effects, a linear trend, and pooled $\text{Post}_{e,t}$ indicators for the four events. Standard errors two-way clustered by symbol and date (Cameron, Gelbach, and Miller, 2011). The JPMorgan settlement produces the largest deterrence-signed coefficient (−1.94 events/security-day, $t = -7.26$). The Citigroup coefficient is essentially zero; Taglich is positive, consistent with heightened detection rather than immediate deterrence.

Event	Date	$\widehat{\beta}_e$ (events/security-day)	t
SEC v. Citigroup	Jan 2017	0.01	0.02
SEC v. Taglich Brothers	Oct 2019	1.24	4.48
DOJ/CFTC v. JPMorgan	Sep 2020	−1.94	−7.26
SEC v. Singh	Sep 2022	−0.77	−3.46

$\widehat{\Lambda}^{\text{live}}/\widehat{\Lambda}^{\text{pre}}$ is 0.53, with $t = -4.23$ for the test against one. In the third quartile, the ratio is 0.58, with $t = -3.21$. In the bottom two quartiles, the drop is absent or reversed. Thus, the depth-inflation channel appears precisely where the trigger score says the display-and-trigger mechanism is strongest. Panel B pools across quartiles and adds lagged SPY returns as additional instruments. The Hansen overidentification tests do not reject, with a p-value of 0.09 at one lag and 0.18 at two lags, and the pooled estimates also show a lower live-window price impact than the pre-window benchmark.

Several features of the design require caution, and they are why we treat this evidence as corroborating rather than core. The identifying assumption is an exclusion restriction. SPY moves at this horizon are assumed to affect the individual-stock midpoint only through that stock's own signed flow. This is plausible enough to be informative at sub-second horizons, but it is not beyond dispute. Market-maker repricing algorithms condition on index futures and ETF quotes directly, which would let SPY returns enter the stock midpoint through a channel other than the stock's own flow and would violate the exclusion restriction. Cross-asset price discovery is weak at this frequency, but it is not zero.

The instrument is not weak. The first-stage F-statistics are mechanically very large, from 3.3×10^3 to 4.3×10^8, because the SPY return is a strong proxy for the systematic component of order flow. Instrument strength is therefore not the binding concern. The exclusion restriction is.

The sample is a 235-day subset of the 1,632-day panel. These are the days for which the SPY NBBO was extracted at 250 millisecond resolution. The subset is defined by data-extraction coverage rather than by any episode-level outcome, so it is not a selected sample, but it does limit the precision of the quartile cells. Two of those cells should not be read as point estimates. The Q_1 pre-window estimate of 39.98 has a standard error of 27.00 and is uninformative on its own. The Q_2 live-to-pre ratio of 1.50 is not significantly different from one, with $t = 1.47$. Neither quartile is where the model predicts the depth-inflation effect. The prediction is for the high-trigger quartiles, and the ratio test is significant only in Q_3 and Q_4.

Finally, the ratio test in Panel A computes the standard error of $\widehat{\Lambda}^{\text{live}}/\widehat{\Lambda}^{\text{pre}}$ by the delta method under approximate independence of the two window estimates. Because the pre- and live-window regressions are estimated on the same episodes, this independence is imperfect and the reported t-statistics are approximate. A joint stacked estimation of the two windows would deliver an exact covariance for the ratio. An identification that does not rely on the SPY exclusion restriction, such as a Hasbrouck-style vector-autoregression price-impact estimate or a within-episode comparison using only liquidity-trader-classified flow, would provide a more robust check. We leave both to future revision and, for this reason, read the live-versus-pre drop in the high-trigger quartiles as corroborating the depth-inflation channel rather than as establishing it.

Table G.5: **Within-stock panel: spoofing frequency and lagged cancel-to-trade ratio.** This table reports within-stock panel regressions of monthly phantom-depth frequency on the lagged log cancel-to-trade ratio. The dependent variable is events per trading day for security s in month m. Within-stock variation in the cancel-to-trade ratio is dominated by non-manipulative cancellation components. Column (1) uses HC1 robust standard errors; columns (2)–(3) cluster standard errors by symbol following Bertrand, Duflo, and Mullainathan (2004) since stock fixed effects absorb cross-sectional means and within-stock residuals are serially correlated across months.

	(1) Pooled OLS	(2) Stock FE	(3) Stock + Month FE
$\log(\text{CancelRatio}_{s,m-1})$	0.320^{***}	-2.181^{***}	-1.739^{***}
	(0.047)	(0.242)	(0.292)
Stock FE	No	Yes	Yes
Month FE	No	No	Yes
N security-months	36,688	36,688	36,688
N securities	494	494	494

Table G.6: **Trigger-weighted cross-section of phantom-depth prevalence.** This table re-estimates the cross-sectional regression of Section 7.1 (Table 11) with dependent variables that weight episodes by the cross-fitted trigger score. The dependent variable is a per-security frequency regressed on the log cancel-to-trade ratio, the hidden-order rate, and controls for spread, volatility, log volume, and market-capitalization rank, with HC1 robust standard errors. The raw eligible count reproduces the published regression. The top-quartile-trigger count is the number of episodes whose trigger score $\widehat{T}_i^{(-k)}$ falls in the top quartile within symbol-year. The sum of trigger scores is the per-security total of $\widehat{T}_i^{(-k)}$. Each cell reports the hidden-order coefficient and its t-statistic, for the baseline specification, the specification with an ETP dummy, and the specification excluding ETPs. The hidden-order rate keeps a large negative and significant loading under trigger-weighting. The log cancel-to-trade ratio is statistically insignificant in every cell, with t between -1.71 and $+0.83$.

	Hidden-order rate coefficient (t)		
Dependent variable	Baseline	With ETP dummy	ETPs excluded
Raw eligible count	−34,258 (−4.20)	−46,002 (−5.35)	−45,527 (−4.95)
Top-quartile-trigger count	−905 (−4.78)	−1,178 (−5.87)	−1,145 (−5.35)
Sum of trigger scores	−1,441 (−4.08)	−1,952 (−5.35)	−1,930 (−4.95)

Table G.7: **Trigger-weighted within-stock visibility panel.** This table re-estimates the within-stock visibility panel of Section 7.1 with dependent variables that weight episodes by the cross-fitted trigger score. Each row is a separate regression of a log monthly count on the hidden-order rate and the cancel-to-trade ratio, with security and month fixed effects and standard errors clustered by symbol, the specification that produces the hidden-order coefficients reported in the text. The raw count uses all 90.6 million detected episodes. The scored count uses the 527,019 episodes for which the trigger score is defined. The top-quartile-trigger count is the number of scored episodes whose trigger score $\widehat{T}_i^{(-k)}$ falls in the top quartile within symbol-year. The sum of trigger scores is the monthly total of $\widehat{T}_i^{(-k)}$. The hidden-order rate keeps a negative and significant loading across all four measures. The cancel-to-trade coefficient is null on the full detected set and weakly negative on the scored and trigger-weighted measures, and never takes the positive sign predicted by the cancellation-feasibility channel; its shift tracks the move to the smaller scored sample rather than the weighting. Coefficients are in the native units of each regressor and are not directly comparable in magnitude, since the hidden-order rate lies in $[0, 1]$ while the cancel-to-trade ratio ranges into the hundreds.

	Hidden-order rate		Cancel-to-trade ratio	
Dependent variable, $\log(1+y)$	Coef.	t	Coef.	t
Raw count, all detected episodes	−2.77	−8.8	0.000	+0.5
Scored count, trigger-score sample	−2.44	−8.5	−0.004	−4.6
Top-quartile-trigger count	−1.28	−6.4	−0.004	−9.3
Sum of trigger scores	−1.96	−7.9	−0.003	−5.3

Table G.8: **Trigger-weighted bunching of per-stock break dates.** This table repeats the direction-differential bunching of Table 12 with the per-stock sup-Wald break estimated on a trigger-weighted series, the daily count of episodes whose cross-fitted trigger score $\widehat{T}_i^{(-k)}$ falls in the top quartile within symbol-year. The break procedure is otherwise identical, a linear-trend sup-Wald with 15% trimming and a Hansen p-value, with significance at $p < 0.05$. The downward cohort contains 173 securities and the upward cohort 189. Shares are the fraction of break dates within a ±60-day window of each anchor. The uniform-random-anchor null share is 0.072. The down-minus-up differential p-value is from a 2,000-draw permutation. Both the downward bunching at the anchors and the down-minus-up differential survive trigger-weighting.

Anchor	Down share	Up share	Null share	Differential p down − up
CAT Phase 2a (22 Jun 2020)	0.156	0.074	0.072	0.002
CAT Phase 2b (20 Jul 2020)	0.156	0.058	0.072	< 0.001
JPMorgan (29 Sep 2020)	0.156	0.026	0.072	< 0.001
CAT Phase 2c (26 Apr 2021)	0.110	0.021	0.072	< 0.001

Table G.9: **Cancel-to-reprice vs. cancel-to-withdraw discrimination.** This table reports the interaction between the cross-fitted trigger score and the post-collapse reprice signal, from a regression of the live-window signed dislocation on the trigger score, the reprice signal, their interaction, controls (pre-event spread, volatility, log volume, log duration), and symbol fixed effects, with standard errors clustered by symbol, on all 1,632 trading days (416,504 episodes). The reprice signal is measured over a one, two, or five second window after collapse, re-extracted from the NBBO. The price axis is the spike-side best-price improvement relative to the pre-spike level (the repricing signature); the depth axis is whether spike-side touch depth reposts near the touch. A negative interaction means the trigger-dislocation slope is weaker in the reprice class, the direction opposite to the cancel-and-reprice alternative. The price axis is negative and significant across windows; the depth axis is null.

Window	reprice-price interaction (t)	reprice-depth interaction (t)
1 second	−0.00030 (−5.42)	−0.00002 (−0.54)
2 seconds	−0.00027 (−5.04)	−0.00004 (−1.26)
5 seconds	−0.00023 (−4.71)	−0.00007 (−2.42)

Table G.10: **Regression discontinuity at the $z = 3$ threshold.** This table reports the regression-discontinuity test for a behavioral break at the $z = 3$ detection threshold. The specification is a local linear RD in the $z \in [1, 5]$ bandwidth. 85.5M collapsed spikes. No sharp discontinuity. Reversal intensity varies smoothly with z, indicating that $K = 3$ is a useful classification threshold but not a behavioral break.

	Coefficient	t
Discontinuity $\widehat{\tau}$ at $z = 3$	−0.24	−1.39

Table G.11: **Per-stock change in phantom-depth frequency across the March 2021 break.** This table reports the per-stock cross-section of the phantom-depth-frequency change across the aggregate March 2021 break. For each of the 494 S&P 500 securities, average phantom-depth events per day before and after March 2021 (pre = 2017-01 through 2021-02, 1,043 trading days; post = 2021-03 through 2023-06, 589 trading days). The break is broad but not uniform across the cross-section.

	Count	Percent
Panel A. Direction of change		
Securities with $\Delta < 0$ (frequency fell)	378	76.5%
Securities with $\Delta > 0$ (frequency rose)	116	23.5%
Panel B. Distribution of Δ (events/security-day)	Value	Percent change
Median	−3.51	−31.4%
Mean	−3.42	−17.5%
p_{10}	−11.96	—
p_{90}	+4.43	—
Panel C. Largest drops	Δ	Percent change
IJR	−27.02	−77.9%
SMH	−26.68	−88.7%
KBE	−24.28	−57.8%
TQQQ	−22.01	−61.7%
FB	−21.13	−66.1%
IVV	−21.06	−88.5%
IWD	−18.91	−68.1%
DIA	−18.38	−85.6%
VOO	−18.22	−84.9%
XBI	−20.89	−64.4%
Panel D. Largest increases	Δ	Percent change
AMC	+32.00	+178.8%
MARA	+27.31	+210.3%
SOXL	+21.99	+381.4%
ZOM	+22.03	+208.9%
CVNA	+19.49	+358.7%
ARKK	+15.53	+178.2%
RIOT	+15.36	+89.0%
GLD	+12.53	+76.8%
TSLA	+12.06	+38.1%
MVIS	+20.18	+203.9%

Internet Appendix

Gaussian-mixture inference and the linear pricing rule

This Appendix explains the market makers' inference problem created by order cancellations, which is a standard inference problem generated by a latent mixture state. Conditional on whether cancellation succeeds, the market maker faces a linear Gaussian signal-extraction problem. Unconditionally, the cancellation state is unobserved, so the posterior is a mixture of Gaussian posteriors, and the exact conditional mean is generally nonlinear.

Work in normalized variables and let

$$\widehat{v} \sim N(0,1), \qquad u \sim N(0,1), \qquad \varepsilon \sim N(0,1),$$

with independent primitive shocks. The manipulator's signal is

$$\widehat{\xi} = \widehat{v} + \sqrt{\tau}\,\varepsilon, \qquad \mathbb{E}[\widehat{\xi}^2] = 1 + \tau.$$

Let $Z = s$ denote successful cancellation, which occurs with probability q, and let $Z = u$ denote unsuccessful cancellation, which occurs with probability $1 - q$. Conditional on the two states, the observed order flow is

$$Y_s = \widehat{\beta}_N \widehat{v} + u, \tag{IA.1}$$

and

$$\begin{aligned} Y_u &= \widehat{\beta}_N \widehat{v} + \widehat{\beta}_S \widehat{\xi} + (1+\gamma_S)u \\ &= (\widehat{\beta}_N + \widehat{\beta}_S)\widehat{v} + \widehat{\beta}_S\sqrt{\tau}\,\varepsilon + (1+\gamma_S)u. \end{aligned} \tag{IA.2}$$

Thus the unconditional order flow is the two-component Gaussian mixture

$$Y = \begin{cases} Y_s, & \text{with probability } q, \\ Y_u, & \text{with probability } 1-q. \end{cases}$$

Conditional on the mixture component, the inference problem is linear Gaussian. Without observing the component, the posterior is no longer Gaussian and the exact conditional mean need not be linear.

Define

$$A_s \equiv \mathrm{Var}(Y_s) = \widehat{\beta}_N^2 + 1, \qquad C_s \equiv \mathrm{Cov}(\widehat{v}, Y_s) = \widehat{\beta}_N, \qquad \lambda_s \equiv \frac{C_s}{A_s}, \tag{IA.3}$$

and

$$\begin{aligned} &A_u \equiv \mathrm{Var}(Y_u) = \widehat{\beta}_N^2 + 2\widehat{\beta}_N\widehat{\beta}_S + (1+\tau)\widehat{\beta}_S^2 + (1+\gamma_S)^2, \\ &C_u \equiv \mathrm{Cov}(\widehat{v}, Y_u) = \widehat{\beta}_N + \widehat{\beta}_S, \qquad \lambda_u \equiv \frac{C_u}{A_u}. \end{aligned} \tag{IA.4}$$

For a fixed state $z \in \{s, u\}$, the pair $(\widehat{v}, Y_z)$ is jointly Gaussian. The componentwise posterior mean is therefore linear:

$$\mathbb{E}[\widehat{v} \mid Y_z = y, Z = z] = \lambda_z y. \tag{IA.5}$$

The market makers do not observe Z. Let f_s and f_u denote centered normal densities with

variances A_s and A_u. Bayes' rule gives the posterior weight

$$\omega_s(y) \equiv \Pr(Z = s \mid Y = y) = \frac{q f_s(y)}{q f_s(y) + (1-q) f_u(y)}, \qquad \omega_u(y) = 1 - \omega_s(y). \tag{IA.6}$$

The exact Bayesian posterior mean is therefore

$$m(y) \equiv \mathbb{E}[\widehat{v} \mid Y = y] = [\omega_s(y)\lambda_s + \omega_u(y)\lambda_u]\, y. \tag{IA.7}$$

Equation (IA.7) is the key object. The posterior mean is a weighted average of two linear conditional means. It is generally nonlinear because the posterior weights depend on y. It becomes linear only in special cases, such as $q \in \{0, 1\}$, $A_s = A_u$, or $\lambda_s = \lambda_u$.

IA.1 The global linear projection

The linear pricing rule used in the model corresponds to the scalar linear minimum mean-square-error estimator. It is the best linear projection of $\widehat{v}$ on Y:

$$\ell(y) = \widehat{\Lambda} y, \qquad \widehat{\Lambda} = \frac{\mathrm{Cov}(\widehat{v}, Y)}{\mathrm{Var}(Y)} = \frac{q C_s + (1-q) C_u}{q A_s + (1-q) A_u}. \tag{IA.8}$$

Equivalently, $\widehat{\Lambda}$ satisfies the orthogonality condition

$$\mathbb{E}\Big[\Big(\widehat{v} - \widehat{\Lambda} Y\Big)\, Y\Big] = 0, \tag{IA.9}$$

and solves

$$\widehat{\Lambda} = \arg\min_{\lambda \in \mathbb{R}} \mathbb{E}\big[(\widehat{v} - \lambda Y)^2\big]. \tag{IA.10}$$

This object is standard in signal processing. It requires only first and second moments and remains well-defined for any finite-variance distribution. Under a joint Gaussian distribution, the linear projection coincides with the full conditional mean. Under a Gaussian mixture, it is instead the best estimator within the linear class.

There are two equivalent ways to interpret (IA.8). First, it is the global LMMSE estimator of $\widehat{v}$ given Y. Second, it is the conditional mean that would obtain if the two-component mixture distribution of $(\widehat{v}, Y)$ were replaced by a single Gaussian distribution with the same mean and covariance matrix. In this sense, the linear pricing rule collapses the mixture into one moment-matched Gaussian component and then applies the usual Gaussian conditional-mean formula.

This approximation is common because it is simple and robust. Gaussian-sum filters keep track of several Gaussian components. A global LMMSE rule is the coarsest possible reduction of that idea. It keeps only the first two moments of the mixture and uses the implied linear update.

Exact MMSE and global LMMSE under Gaussian mixtures. The signal-processing literature distinguishes sharply between the exact MMSE estimator and the LMMSE estimator under mixture distributions. With a Gaussian mixture prior, Gaussian mixture noise, or a latent switching state, the exact MMSE estimator is not a single global linear rule. It is a posterior mixture of componentwise linear rules. In the scalar two-state environment above, this is exactly equation (IA.7).

In vector notation, a Gaussian-mixture model gives a posterior mean of the form

$$\mathbb{E}[X \mid Y = y] = \sum_{k=1}^{K} \pi_k(y)\,(a_k + B_k y)\,, \tag{IA.11}$$

where $a_k + B_k y$ is the conditional mean under component k, and $\pi_k(y)$ is the posterior responsibility of that component. The estimator is affine within each component but globally nonlinear because the weights $\pi_k(y)$ change with the observation.

The closest signal-processing analogue is Flåm et al. (2012). They study a linear observation model in which the signal or noise has Gaussian mixture statistics, derive the MMSE estimator in closed form, and compare it with the LMMSE estimator. They emphasize a useful bounding logic. The LMMSE mean-square error is an upper bound on the optimal MMSE mean-square error, while a genie-aided estimator that observes the active mixture component gives a lower bound. In their numerical analysis, the LMMSE can be close to the MMSE in some parameter regions, but the difference is not structurally zero. It depends on the signal-to-noise ratio and on the separation between mixture components.

The same logic applies here. The cancellation outcome is the latent mixture component. Conditional on that outcome, the market maker's inference problem is linear Gaussian. Without observing it, the exact Bayesian price is a posterior-weighted average of two linear prices. The linear pricing rule is the global LMMSE approximation to this nonlinear Bayesian price.

The linear restriction also has a behavioral interpretation. The linear-projection slope minimizes $\mathbb{E}[(\widehat{v} - \widehat{\Lambda}\widehat{y})^2]$, so restricting market makers to linear pricing is equivalent to asking them to use the best quadratic-loss forecast of value from order flow. Under this rule, large pricing errors receive especially high weight. The manipulator's bluff can therefore affect the pricing rule by creating the possibility of rare but extreme order-flow realizations, namely the uncancelled display.

The exact Bayesian price behaves differently. Since the cancellation state is latent, the posterior is a Gaussian mixture. For extreme order-flow realizations, the posterior increasingly assigns weight to the component in which the display failed to cancel, so the conditional mean flattens relative to the global linear projection. This nonlinearity is captured by the mixture weights $\pi_k(y)$ in (IA.11). The residual η_0^2 measured below quantifies the resulting gap between the exact Bayesian price and the linear price.

This gap is negligible in the moderate parameter region used for the numerical equilibria, but it can become economically meaningful in aggressive bluffing regions. For this reason, we report η_0^2 as a diagnostic and treat the small-$a\tau$ limit as an illustration of the mechanism rather than as the empirically relevant region. The within-episode trigger results do not rely on this linear-pricing approximation. They require only that the pricing rule be locally sticky inside the episode, not that it be linear or that $\widehat{\Lambda}^*$ be close to zero.

IA.2 The nonlinear residual

Define the residual between the exact Bayesian price and the global linear price by

$$e(y) \equiv m(y) - \widehat{\Lambda} y.$$

The natural diagnostic is its squared L_2 norm:

$$\eta_0^2 \equiv \mathbb{E}\left[\left(m(Y) - \widehat{\Lambda} Y\right)^2\right] = \mathbb{E}[e(Y)^2]. \tag{IA.12}$$

Using the mixture density, this can be written as a one-dimensional integral

$$\eta_0^2 = \int_{\mathbb{R}} \left[\omega_s(y)\lambda_s + \{1-\omega_s(y)\}\lambda_u - \widehat{\Lambda}\right]^2 y^2 \left[q f_s(y) + (1-q) f_u(y)\right] dy. \tag{IA.13}$$

This diagnostic has a precise decision-theoretic interpretation. The exact posterior mean $m(Y) = \mathbb{E}[\widehat{v} \mid Y]$ is the L_2 projection of $\widehat{v}$ onto all square-integrable functions of Y. The linear rule $\widehat{\Lambda} Y$ is the L_2 projection onto the one-dimensional linear subspace spanned by Y. The Pythagorean identity gives

$$\mathbb{E}\Big[(\widehat{v} - \widehat{\Lambda} Y)^2\Big] = \mathbb{E}\big[(\widehat{v} - m(Y))^2\big] + \eta_0^2. \tag{IA.14}$$

Thus η_0^2 is exactly the incremental mean-square error from imposing a global linear pricing rule.

For the economics of the spoofing model, the same object also bounds the payoff relevance of the approximation. For any square-integrable demand x, Cauchy-Schwarz gives

$$\left|\mathbb{E}\Big[x\{m(Y) - \widehat{\Lambda} Y\}\Big]\right| \le \|x\|_2 \, \eta_0. \tag{IA.15}$$

For a linear demand $x = \beta\widehat{\xi} + \gamma u$,

$$\|x\|_2^2 = (1+\tau)\beta^2 + \gamma^2. \tag{IA.16}$$

The same residual that measures the statistical cost of linear pricing also bounds the contribution of nonlinear pricing errors to expected trading gains for fixed linear strategies.

Two normalized versions are useful:

$$R_V^2 \equiv \frac{\eta_0^2}{\mathrm{Var}(\widehat{v})} = \eta_0^2, \qquad R_m^2 \equiv \frac{\eta_0^2}{\mathbb{E}[m(Y)^2]}. \tag{IA.17}$$

The first normalization measures the nonlinear residual relative to the variance of fundamentals. The second measures it relative to the variance of the exact Bayesian price.

IA.3 A numerical illustration

Table IA.1 reports the nonlinear residual for several parameter choices. The first two rows use coefficients obtained by optimizing the linear-pricing objective at the indicated values of a. The remaining rows are stress tests. They show how the approximation behaves when the spoofer's signal loading or liquidity loading is forced to be large.

Around the moderate equilibrium region, the nonlinear residual is extremely small. In the first row, the nonlinear correction is about 0.0035% of the variance of the exact Bayesian price. In the second row, it is smaller still. By contrast, the stress tests show that the approximation can become economically meaningful when the mixture components are forced far apart. When the spoofer's signal loading is set to $\widehat{\beta}_S = 3$, the residual rises to about 12% of Bayesian price variation.

The numerical exercise therefore supports the linear approximation in the moderate equilibrium region. In regions with very aggressive spoofing intensities, the residual should be recomputed. For first-order-condition calculations, the stronger diagnostic is the derivative of the nonlinear correction, because local slope differences rather than only level differences affect optimality.

Table IA.1: **Nonlinear-pricing residual under Gaussian mixture inference.** The table compares the exact Bayesian posterior mean $m(Y) = \mathbb{E}[\widehat{v} \mid Y]$ with the global linear projection $\widehat{\Lambda}Y$. The reported ratio $\eta_0^2/\mathbb{E}[m(Y)^2]$ measures the incremental mean-square error from imposing a global linear price as a fraction of the variance of the exact Bayesian price.

Case	q	τ	a	$\widehat{\beta}_S$	γ_S	$\widehat{\beta}_N$	$\widehat{\Lambda}$	$\eta_0^2/\mathbb{E}[m(Y)^2]$
Linear-equilibrium coefficients, moderate risk aversion	0.8	0.5	0.1	0.3663	−0.4051	0.9157	0.5250	3.45×10^{-5}
Linear-equilibrium coefficients, higher risk aversion	0.8	0.5	0.5	0.2816	−0.3774	0.9200	0.5274	5.18×10^{-7}
Moderate stress test	0.8	0.5	–	1.0000	−0.4000	0.9686	0.4679	1.72×10^{-3}
Aggressive spoofing-loading stress test	0.8	0.5	–	3.0000	−0.4000	1.5171	0.2752	1.20×10^{-1}
Large liquidity-loading stress test	0.8	0.5	–	1.0000	2.0000	1.5941	0.2951	9.76×10^{-2}

IA.4 Relation to the signal-processing literature

The LMMSE estimator is a workhorse of signal processing. Its appeal is that it uses only first and second moments, is the best linear estimator for any finite-variance distribution, and satisfies the orthogonality principle. Gaussianity is needed for the LMMSE estimator to coincide with the full conditional mean. It is not needed for optimality within the linear class. Using a linear projection under a non-Gaussian distribution is therefore standard, provided that the accuracy loss is checked.

Gaussian-sum filtering gives the closest analogy. The central idea in Alspach and Sorenson (1972) is to approximate a non-Gaussian density by a finite sum of Gaussian components and update each component separately. Each component admits a Kalman-type linear update, but the mixture weights are revised after the observation. The posterior remains a Gaussian mixture, and the posterior mean is a weighted average of component posterior means.

Our structure is exactly this two-component case. Successful and unsuccessful cancellation are the two Gaussian components. Conditional on each component, the price is linear. The full Bayesian price is a weighted average of those two linear prices, with weights that depend on order flow. The global linear projection is an additional reduction that replaces the mixture posterior with a single linear rule.

The same idea appears in later Gaussian-mixture filtering and inverse-problem work. Gaussian-sum filters are used precisely because a single Gaussian update can miss non-Gaussian features such as multimodality, skewness, or latent-state ambiguity. Modern adaptive Gaussian-sum filters manage the tradeoff between accuracy and complexity by splitting, merging, or pruning mixture components (Terejanu et al., 2011; Pishdad and Labeau, 2014; Zanetti and Tuggle, 2018). In image inverse problems, Yu, Sapiro, and Mallat (2012) use Gaussian mixtures to obtain estimators that are locally linear within each component but globally nonlinear through component selection or posterior weighting.

The conclusion for the spoofing model is simple. Linear pricing is not the exact Bayesian price in the latent-cancellation environment. It is the global LMMSE approximation to that price. This approximation is standard, transparent, and easy to diagnose. The residual η_0^2 measures the incremental statistical cost of imposing linearity and bounds the payoff relevance of the nonlinear pricing correction for fixed linear demands.

References

Alspach, D. L. and H. W. Sorenson (1972): "Nonlinear Bayesian Estimation Using Gaussian Sum Approximations," *IEEE Transactions on Automatic Control*, 17(4), 439–448.

Flåm, J. T., S. Chatterjee, K. Kansanen, and T. Ekman (2012): "On MMSE Estimation: A Linear Model under Gaussian Mixture Statistics," *IEEE Transactions on Signal Processing*, 60(7), 3840–3845.

Flåm, J. T., S. Chatterjee, K. Kansanen, and T. Ekman (2011): "Minimum Mean Square Error Estimation under Gaussian Mixture Statistics," arXiv:1108.3410.

Kay, S. M. (1993): *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall.

Poor, H. V. (1994): *An Introduction to Signal Detection and Estimation*, 2nd ed., Springer.

Terejanu, G., P. Singla, T. Singh, and P. Scott (2011): "Adaptive Gaussian Sum Filter for Nonlinear Bayesian Estimation," *IEEE Transactions on Automatic Control*, 56(9), 2151–2156.

Pishdad, L. and F. Labeau (2014): "Approximate MMSE Estimator for Linear Dynamic Systems with Gaussian Mixture Noise," arXiv:1404.3638.

Yu, G., G. Sapiro, and S. Mallat (2012): "Solving Inverse Problems with Piecewise Linear Estimators: From Gaussian Mixture Models to Structured Sparsity," *IEEE Transactions on Image Processing*, 21(5), 2481–2499.

Kundu, A., S. Chatterjee, A. S. Murthy, and T. V. Sreenivas (2008): "Subspace Based Speech Enhancement Using Gaussian Mixture Model," *Interspeech*.

Zanetti, R. and K. Tuggle (2018): "A Novel Gaussian Mixture Approximation for Nonlinear Estimation," *Proceedings of the International Conference on Information Fusion*.

Fesl, B. and W. Utschick (2024): "Linear and Nonlinear MMSE Estimation in One-Bit Quantized Systems under a Gaussian Mixture Prior," arXiv:2407.01305.