Queue Priority as Adverse-Selection Exposure*

ARYAN AYYAR

June 2026

Abstract

Price-time priority is usually treated as an execution advantage. We show that FIFO priority also allocates adverse-selection exposure: moving one rank forward adds exactly one marginal execution state, so priority has the sign of that state's expected surplus. A toxicity threshold and a shield theorem characterize when depth ahead insures a passive order. In an entry-and-placement equilibrium with endogenous informed order size, displayed queues with toxic front priority survive when reposting costs exceed the surrendered block premium. Cancellation is an option that truncates, but cannot reverse, a state-wise toxic margin. Priority rules allocate exposure, not only speed.

*Aryan Ayyar is at the Manipal Academy of Higher Education (ayyar.aryan@manipal.edu). The author thanks seminar participants and colleagues for comments and has no conflicts of interest to disclose.

Introduction

Modern electronic markets run on queues. A trader who wants immediacy sends a marketable order and consumes resting liquidity; a trader who is willing to wait posts a limit order and joins the book. Because many limit orders can rest at the same price, every exchange needs a rule for deciding which same-price order trades first. The standard rule is price-time priority: better-priced orders execute first, and orders at the same price execute in first-in-first-out (FIFO) order. Queue position under this rule is a scarce and valuable resource. In liquid, tick-constrained securities, queues at the best quotes are long, priority cannot be bought by improving the price, and a position near the front of the queue can be worth a meaningful fraction of the spread (Moallemi and Yuan, 2016; Yao and Ye, 2018). Market makers, execution algorithms, and high-frequency traders spend resources—speed, monitoring, message traffic—to obtain queue position early and to protect it. The conventional view behind that expenditure is that earlier rank is better: a front-rank order trades sooner and more often, so time priority is an execution advantage and depth ahead of an order is simply delay.

We ask when this intuition fails. Moving forward in a FIFO queue raises the probability of execution, but it also changes *which* executions the order receives. A front-rank order receives the earliest same-price fills. If those fills are disproportionately informed, stale-quote-seeking, or otherwise adverse, then priority buys fill probability by increasing adverse-selection exposure, and the queue ahead of an order is not delay but protection. The question is therefore a risk tradeoff: priority is valuable when the marginal fill is benign and costly when the marginal fill is toxic. This paper characterizes that tradeoff—when each case obtains, what it implies for cancellation and queue stability, who stands at a toxic front in equilibrium, and what it means for the design of priority rules.

The analysis is organized around one observation. Compare two otherwise identical sell orders at the same ask, one at FIFO rank k and one at rank $k+1$, exchanging only their positions in line while the book and the order-flow environment are held fixed. The two positions pay identically except in exactly one event: the event that cumulative marketable buy volume stops precisely at

k units, reaching the earlier order but not the later one. One step of FIFO priority is therefore a claim on a single marginal execution state, and its value is the probability of that state times the expected surplus of the fill received in it. Fill probability and rank value can accordingly move in opposite directions: a rank can be filled often and still be worth less than the rank behind it, because the additional fills it receives arrive precisely in the states in which the quote is stale. The first part of the paper turns this margin into a sign characterization. In a primitive informed/noise representation, priority at a rank margin is negative exactly when a rank-specific adverse-selection intensity exceeds the prior odds of benign flow, and a passive order's optimal queue position is governed by block averages of the same intensity: an interior rank is optimal when every block of ranks ahead is on average too toxic to want and every block behind is not toxic enough to surrender. Queue depth ahead of such an order is insurance. It filters out the marginal execution states with negative expected surplus before they reach the order.

Which ranks are toxic is a statement about where informed marketable volume stops, not about the unconditional price impact of large trades. When information is exploited through large aggressive orders, informed volume reaches deep ranks, adverse-selection intensity rises with rank, and the front of the queue is the safe place to stand. When information is exploited through stale-quote sniping or small clips that hit the top of the book before quotes adjust, adverse-selection intensity is front-loaded, and a passive order can rationally prefer to stand behind a shield of other orders. Mixed regimes generate interior optima. These order-flow technologies are observable in principle, and they reverse the sign of the paper's predictions in a way that disciplines empirical work.

A valuation lens, however, raises an equilibrium objection: if the front of the queue is toxic and everyone knows it, who stands there, and why does the displayed queue not unravel backward? The second part of the paper answers this question with an entry-and-placement equilibrium in which informed order size, displayed depth, and queue positions are jointly determined. An informed trader with a binary signal chooses a clip size subject to increasing marginal costs of taking deeper units, generating an endogenous distribution over how deep informed volume reaches. Passive

traders enter the queue subject to a display cost and may cancel and repost subject to a reposting friction. Two structural results make the equilibrium tractable. First, a truncation-invariance lemma: with separable informed costs, the value of a given rank does not depend on the depth displayed behind it, because deeper displayed depth changes informed demand only through a capacity cap that never binds for reaching that rank. Entry is therefore disciplined rank by rank, the set of viable depths is closed downward, and the equilibrium displayed depth is the first rank whose standalone value cannot cover the display cost—a rank-level analogue of Glosten's marginal-unit condition with an endogenous informed size distribution. Second, a feasibility observation: under FIFO, a canceled order re-enters at the back of the queue, so the only priority block a trader can surrender is the entire block between her rank and the back. Interior toxic blocks therefore persist even with free cancellation whenever the cumulative block to the back is positive, and queues whose front-to-back block is negative remain stable exactly when reposting frictions exceed the surrendered block premium. The model produces open sets of primitives in which a five-deep queue is entry-stable while the first two priority margins are strictly negative. The same conditions sort traders: ranks followed by toxic blocks must be occupied by traders with high reposting frictions or private execution motives, so toxic priority changes the composition of the queue, not only its value profile. The equilibrium also delivers a sharp cross-sectional comparative static: a shift of informed technology toward sweeps thins the book but detoxifies the front, while a shift toward sniping deepens the book but poisons the front.

The third part of the paper confronts the cancellation option directly. A resting order is not a hold-to-horizon claim; its owner can monitor the book and cancel when toxicity signals arrive, and this option is a large part of queue-position value in the dynamic queue literature. We formalize cancellation as an optimal stopping problem and prove a sandwich theorem: the option-adjusted value of one step of priority is bounded between two truncated versions of the same marginal-state claim, truncated at the optimal cancellation times of the two adjacent ranks. The marginal-state architecture therefore survives optimal stopping intact, and it yields a sign-preservation result: if the conditional expected markout of the marginal fill is nonpositive in every state in which the marginal

exposure is open, then no cancellation policy can make that step of priority valuable. The option truncates exposure to the toxic margin; it cannot change its sign. This is the precise sense in which monitoring and cancellation attenuate, rather than overturn, the paper's static characterization.

The fourth part places the same priority condition in a dynamic information state. A Kyle/Back filter supplies the market's posterior mean and variance at the moment the queue decision is priced, and a marked order-arrival process determines which ranks are reached next. A rank margin is toxic exactly when the normalized information content of its marginal fill, scaled by remaining value uncertainty, exceeds the quote cushion. Toxicity is therefore a moving boundary: a rank that is dangerous early in price discovery can become safe after order flow has revealed information, and the rank profile of fill informativeness is generated by informed clip-size choice rather than assumed—small informed clips produce front-loaded informativeness, sweeps push it deeper.

The final part draws out design and measurement implications. A matching rule allocates the marginal execution states at a price: FIFO concentrates them by time rank, while pro-rata spreads the same states across displayed size, so each pro-rata unit holds the average FIFO rank value. The choice between FIFO and pro-rata is therefore a choice about who bears adverse-selection exposure, not only about who trades first—in sniping regimes pro-rata insures the front of the queue, while in sweep regimes it dilutes the front's valuable claims. On the measurement side, displayed rank is an error-ridden proxy for the rank at which an order actually absorbs execution risk. Hidden and reserve quantity, refresh rules, and reconstruction error create an effective-rank wedge, and the value of that wedge is the negative of the cumulative priority premium over the skipped ranks. A tomography procedure that infers the wedge from message-level evidence therefore measures not only position but exposure, and the sign of the wedge's value—helpful in sniping states, harmful in sweep states—is itself a falsifiable diagnostic of the order-flow technology.

The paper's relation to existing work is developed in Section X, but three boundaries belong up front. The observation that limit orders are exposed to adverse selection and picking-off risk is classical (Copeland and Galai, 1983; Glosten, 1994; Foucault, 1999), and practitioners have long understood that front-of-queue fills can be toxic in sniping regimes. Our contribution is not

the exposure observation but its formalization at the rank margin: the within-price decomposition under which adjacent FIFO ranks can carry opposite-signed marginal values, the threshold and shield characterizations built on it, and the equilibrium and stopping-time layers that discipline it. Glosten's marginal condition prices the marginal unit of depth across price levels through a tail expectation; the object here is the stopping event at a rank within a single price, which is what makes priority itself—rather than depth—the priced object. Queue-position valuation models (Moallemi and Yuan, 2016; Lehalle and Mounjid, 2017) compute the value of rank in rich dynamic environments in which adverse selection on fills is present, but the sign of the priority margin, its reversal across order-flow regimes, and its consequences for queue stability and matching-rule design are not their object. Work on priority rules and queue rationing (Field and Large, 2008; Yao and Ye, 2018; Budish et al., 2015) studies how matching rules allocate rents and races; the exposure-sharing result here isolates a complementary allocation—the same marginal states shared differently across the same liquidity.

The paper proceeds as follows. Section I sets up the static rank-comparison experiment. Section II derives the marginal priority identity. Section III characterizes toxic priority and the shield. Section IV maps order-flow technologies into toxicity profiles. Section V develops the entry-and-placement equilibrium. Section VI treats cancellation as an optimal stopping problem. Section VII embeds the priority condition in a Kyle/Back information state. Section VIII develops the market-design and effective-rank implications. Section IX translates the theory into an empirical design, Section X discusses the literature, and Section XI concludes. Proofs are in Appendix B.

I The Model

Consider a limit order book on a discrete price grid. At each price, orders queue on one of two sides: asks, which sell to incoming marketable buy volume, and bids, which buy from incoming marketable sell volume. Within a price, orders are matched by first-in-first-out priority. The two sides are symmetric after reversing signs, so we work with the ask side throughout. Fix the best ask

a at a public state ω, and let the feasible effective ranks in the ask queue be $\mathscr{K} = \{1, \ldots, K\}$.

A passive trader evaluates a small unit sell order, measured in queue-lot units, at rank $k \in \mathscr{K}$, against the outside option of not supplying liquidity at this ask. Effective rank is the position at which the order actually absorbs execution risk under the matching rule; it can be generated by arrival time, queue state, cancellation and reposting, hidden or reserve priority, or reconstruction error. For the static comparison, the rank is the counterfactual object being valued rather than a fully modeled placement action; Section V endogenizes placement. After the rank is fixed, cumulative marketable buy volume Y and the execution-relevant value v are realized. The variable Y is integer-valued, measured in queue-lot units over the order's exposure horizon. The value v is the asset value at the payoff horizon used to evaluate the fill: a reduced-form terminal value in the static model, the terminal Kyle value in Section VII, and a post-fill markout proxy in the empirical design of Section IX. A rank-k ask is filled if and only if $Y \geq k$. Conditional on execution, the order sells an asset worth v at price a, so its execution surplus is $a - v$.

A passive ask order supplies immediacy by agreeing to sell at the posted ask when marketable buy volume arrives. Its compensation is the quote cushion $a - v$: the price it receives less the execution-relevant value of what it gives up. The joint law of (Y, v) is the reduced-form adverse-selection environment. Marketable buy volume may be more likely, or larger, precisely when v is high relative to a. A rank is therefore not only a fill probability; it is exposure to the value distribution of the executions that reach that rank. All probabilities and expectations are conditional on ω, which includes the displayed book, public information, the exposure horizon, and the matching environment; the conditioning is suppressed.

Assumption 1 (Static rank-comparison experiment)**.** *Conditional on ω, the ask price a, and the matching rule, the counterfactual comparison across feasible effective ranks holds fixed the conditional joint law of (Y, v). The comparison is a swap experiment: the trader's unit is exchanged with an adjacent, otherwise identical unit, so total displayed depth, the depth profile, and the matching environment are unchanged. It is not an insertion, deletion, or queue-depth change. Under anonymity, order flow cannot condition on which identical unit occupies which same-price rank.*

The random variable Y is integer-valued in queue-lot units, and each marginal event used in an adjacent-rank comparison has positive probability.

The swap interpretation matters because insertion, deletion, or displayed-depth changes may alter informed order-size choice and hence the law of (Y, v). Section V exhibits exactly this dependence: there, the informed trader's demand responds to displayed depth, and the rank comparison deliberately holds that depth fixed to price the marginal exposure created by FIFO priority alone.

Definition 1 (Value of a queue rank)**.** With the queue environment and payoff metric fixed, define the rank-k gross execution payoff

$$X_k \equiv (a - v)\mathbf{1}_{\{Y \geq k\}}, \tag{1}$$

and the rank value $V_k \equiv \mathbb{E}[X_k]$. For $k = 1, \ldots, K-1$, define the one-rank payoff increment $Z_k \equiv X_k - X_{k+1} = (a - v)\mathbf{1}_{\{Y=k\}}$ and the marginal priority value, or one-rank priority premium,

$$\Delta_k \equiv V_k - V_{k+1} = \mathbb{E}[Z_k]. \tag{2}$$

The object is not fill probability alone, because the payoff $a - v$ need not be positive in every execution state. Inventory, risk-bearing, waiting costs, and cancellation costs are omitted here so that V_k measures the pure execution-surplus value of rank; Section VI restores the cancellation option, and Appendix A records a mean-variance extension. A positive Δ_k means that moving one step forward is valuable. A negative Δ_k means that the order prefers the shield provided by the rank ahead.

II The Priority Margin

Consider two otherwise identical sell limit orders posted at the same ask price a, one at FIFO rank k and one at rank $k+1$. The earlier order executes whenever marketable buy volume reaches at least k units; the later order executes only if that volume reaches at least $k+1$ units. If fewer than k units arrive, neither order trades. If at least $k+1$ units arrive, both trade. The rank move therefore

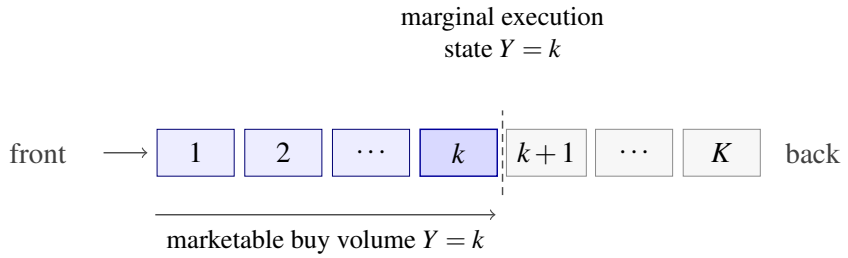

Figure 1: FIFO priority at one ask price. Moving from rank $k+1$ to rank k changes payoff only on the marginal execution state $Y = k$. The sign of priority is therefore the sign of the expected execution surplus conditional on that marginal state.

matters only at the boundary where buy volume is just large enough to reach the earlier order and not the later one. One step of FIFO priority is a claim on the markout of this single extra marginal fill. Figure 1 depicts the comparison.

Lemma 1 (Marginal priority identity)**.** *For every adjacent pair of ranks $k, k+1$ with $\mathbb{P}(Y = k) > 0$,*

$$\Delta_k = \mathbb{E}\left[(a - v)\mathbf{1}_{\{Y=k\}}\right] = \mathbb{P}(Y = k)\,(a - \mathbb{E}[v \mid Y = k])\,. \tag{3}$$

Therefore the marginal priority value is negative if and only if $\mathbb{E}[v \mid Y = k] > a$.

Remark 1. The lemma says why fill probability and rank value can move in opposite directions. Moving from rank $k+1$ to rank k raises fill probability by $\mathbb{P}(Y = k)$, strictly when this event has positive probability. The term $a - \mathbb{E}[v \mid Y = k]$ is the expected post-fill markout on that incremental fill. A high-fill rank is valuable when the additional fill is benign, and costly when the additional fill is informed or stale-quote-seeking. Queue depth ahead is therefore valuable exactly when it absorbs marginal execution states with negative expected markout.

III Toxic Priority and the Shield

Lemma 1 locates the only state priced by a one-rank move: the marginal event $Y = k$. It does not by itself say whether the priority premium is positive, negative, monotone, or front-loaded across ranks. Those properties come from the joint distribution of marketable volume and post-fill value.

This section supplies a primitive informed/noise representation that turns the distributional content of Δ_k into a rank-specific adverse-selection threshold, and then aggregates the threshold over blocks of ranks to characterize optimal queue position.

III.A A Toxicity Threshold

Suppose the marginal execution environment is generated by one of two latent order-flow regimes. In the informed regime, marketable buy volume is associated with high asset values relative to the ask. In the noise regime, execution reflects liquidity demand and can compensate the passive seller for supplying immediacy. Let $\pi \in (0,1)$ be the prior probability of the informed regime, and write $I = 1$ for informed flow and $I = 0$ for noise flow. The regime is not a signal observed by the passive order before choosing rank; it is a latent state that governs the marginal-fill distribution. For each rank margin k, define

$$p_I(k) = \mathbb{P}(Y = k \mid I = 1), \qquad p_N(k) = \mathbb{P}(Y = k \mid I = 0), \tag{4}$$

and the corresponding execution-state means

$$m_I(k) = \mathbb{E}[v \mid Y = k, I = 1], \qquad m_N(k) = \mathbb{E}[v \mid Y = k, I = 0]. \tag{5}$$

These probabilities describe where informed and noise flow stop in the queue; the means describe the value of the asset conditional on that marginal execution. For the transparent two-regime threshold, assume $p_I(k) > 0$, $p_N(k) > 0$, and $m_N(k) < a < m_I(k)$. This sign normalization is stronger than the logic requires: the threshold form below applies whenever the net marginal surplus at rank k is decreasing in the informed-flow weight and crosses zero. The normalization makes that crossing transparent by writing the gain and loss terms as positive quantities.

The sign of priority is a comparison between a weighted gain and a weighted loss. In the noise regime, the marginal event occurs with probability $p_N(k)$ and pays the passive seller $a - m_N(k)$. In the informed regime, the same event occurs with probability $p_I(k)$ and costs the seller $m_I(k) - a$.

The priority premium is therefore

$$\Delta_k(\pi) = (1-\pi)p_N(k)(a - m_N(k)) - \pi p_I(k)(m_I(k) - a). \tag{6}$$

A noise-driven marginal fill has expected gross benefit $b_k \equiv p_N(k)(a - m_N(k)) > 0$, the benchmark compensation for supplying liquidity at that rank margin. Informed flow changes both the probability of reaching the same margin and the severity of the loss conditional on reaching it. The rank-specific adverse-selection intensity is

$$A_k \equiv \frac{p_I(k)}{p_N(k)} \frac{m_I(k) - a}{a - m_N(k)}. \tag{7}$$

It is large when informed flow is disproportionately likely to stop at margin k, or when the informed markout loss is large relative to the noise-state gain. The last ingredient is the prior noise-to-informed odds, $\phi(\pi) \equiv (1-\pi)/\pi$. Priority is valuable when these benign-flow odds are large enough to offset the rank's adverse-selection intensity, and toxic when A_k exceeds them.

Proposition 1 (Toxicity threshold)**.** *The marginal priority value at rank margin k is*

$$\Delta_k(\pi) = \pi b_k \left[\phi(\pi) - A_k\right]. \tag{8}$$

Hence priority at margin k is negative if and only if $A_k > \phi(\pi)$. Equivalently, with $\bar{\pi}_k \equiv (1 + A_k)^{-1}$, $\Delta_k(\pi) < 0$ if and only if $\pi > \bar{\pi}_k$.

Equivalently, the posterior probability of the informed regime on the marginal event is

$$\mathbb{P}(I = 1 \mid Y = k) = \frac{\pi p_I(k)}{\pi p_I(k) + (1-\pi)p_N(k)},$$

and priority is toxic when the posterior-weighted informed loss exceeds the posterior-weighted noise gain:

$$\mathbb{P}(I = 1 \mid Y = k)(m_I(k) - a) > \mathbb{P}(I = 0 \mid Y = k)(a - m_N(k)).$$

The prior-odds threshold writes the same comparison before conditioning on $Y = k$: the likelihood-

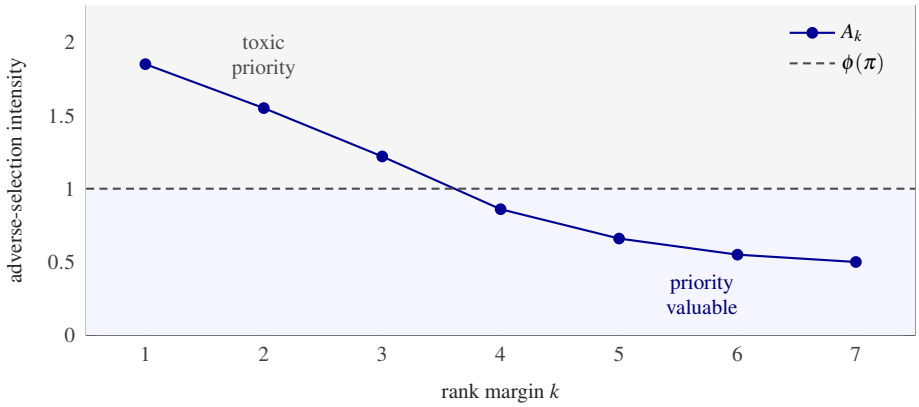

Figure 2: Toxicity threshold for one-rank priority. The horizontal line is the prior noise-to-informed odds $\phi(\pi)$. A rank margin is toxic when its adverse-selection intensity A_k lies above this line. The figure shows a schematic front-loaded example; Section IV shows how different order-flow technologies generate different A_k profiles.

ratio update $p_I(k)/p_N(k)$ and the relative loss severity are already absorbed into A_k. The ordering of toxic ranks is therefore the ordering of A_k. When A_k rises with k, deeper executions carry greater adverse-selection intensity. When A_k falls with k, the front of the queue is the most exposed part of the book.

III.B Depth as a Shield

The threshold in Proposition 1 prices one step of priority. A trader choosing a queue position prices a block of such steps. Moving from rank k^* to a better rank $l < k^*$ adds the marginal execution states $l, \ldots, k^* - 1$; moving to a worse rank $r > k^*$ gives up the marginal execution states $k^*, \ldots, r-1$. Queue depth is insurance when the priority block ahead is sufficiently adverse: the trader prefers to let earlier orders absorb those marginal executions rather than obtain the additional fill probability herself.

Definition 2 (Block toxicity)**.** For any nonempty finite block S of adjacent rank margins, let $B_S \equiv \sum_{j \in S} b_j$. If $B_S > 0$, define the b-weighted average adverse-selection intensity

$$\overline{A}_S \equiv \frac{\sum_{j \in S} b_j A_j}{B_S}.$$

The next result is the rank-choice version of the toxicity threshold; its proof is block arithmetic, since moving forward adds the intervening priority premia while moving backward gives them up. An interior rank is optimal when every feasible move forward would add a toxic priority block, while every feasible move backward would surrender a non-toxic block. The condition $V_{k^*} \geq 0$ is part of the characterization: it is what makes the rank preferred to the outside option.

Theorem 1 (Adverse-selection shield)**.** *Suppose the informed/noise mixture representation holds on all margins in the finite feasible rank set $\mathscr{K} \subset \{1,2,\ldots\}$, with $p_I(j), p_N(j) > 0$, $m_N(j) < a < m_I(j)$, and hence b_j, A_j well defined. Let $k^* \in \mathscr{K}$ have feasible ranks on both sides, and suppose the outside option has value normalized to zero with $V_{k^*} \geq 0$. Then k^* is a global maximizer over $\mathscr{K} \cup \{\text{outside}\}$ if and only if, for every feasible $l \in \mathscr{K}$ with $l < k^*$ and every feasible $r \in \mathscr{K}$ with $r > k^*$,*

$$\overline{A}_{\{l,\ldots,k^*-1\}} \geq \phi(\pi) \qquad \text{and} \qquad \overline{A}_{\{k^*,\ldots,r-1\}} \leq \phi(\pi).$$

Strict inequalities, together with $V_{k^*} > 0$, imply uniqueness over ranks and strict preference over the outside option.

Corollary 1 (Single crossing)**.** *If A_k is strictly decreasing in rank and there exists an interior feasible rank k^* such that $A_{k^*-1} > \phi(\pi) > A_{k^*}$, then k^* is the unique interior maximizer among feasible ranks. If, in addition, $V_{k^*} > 0$, the same rank is strictly preferred to the outside option.*

The queue ahead of k^* is therefore not merely delay. It is a filter. In a front-loaded adverse-selection regime, the orders ahead of k^* absorb marginal executions with negative expected surplus, while rank k^* retains exposure only once that toxic block has been cleared.

IV Order-Flow Regimes

Order-flow primitives discipline the cross-rank profile of adverse-selection intensity. The shield result requires front ranks to be more exposed to adverse marginal fills than the ranks behind them. This is not an assumption about the book alone; it is an assumption about where informed

marketable flow stops relative to ordinary liquidity demand. Throughout this section, "front-loaded" and "deeper" refer to stopping probabilities for the marginal event $Y = k$, not to the cumulative probability $\mathbb{P}(Y \geq k)$. The question is where marketable volume stops, because that stopping event is the extra state created by one step of priority.

Noise demand is naturally centered on ordinary trade sizes. A liquidity buyer who wants immediacy may consume the best ask, or several displayed lots, but her order size is not chosen to exploit a particular stale quote. Informed flow can have a different size profile. When information is used through a large aggressive order, informed volume reaches deeper ranks and $p_I(k)/p_N(k)$ tends to rise with k. When information is used through stale-quote sniping, latency races, or small clips designed to hit the front before quotes adjust, informed volume is concentrated near the front and $p_I(k)/p_N(k)$ tends to fall with k. A market can also contain both forms: some informed trades sweep the book, while others pick off the first displayed quantity.

The adverse-selection intensity (7) combines this stopping-rank likelihood ratio with the conditional loss on execution. Holding the loss-to-gain term fixed, the shape of A_k is the shape of the informed-to-noise stopping ratio; more generally, the same conclusion holds when the conditional loss varies with rank. The shield theorem is therefore not driven by an arbitrary choice of A_k. It is driven by order-size technologies that determine which ranks informed and noise traders are likely to reach.

These primitives give three economically distinct priority regimes. Under informed sweeps, adverse-selection intensity rises with rank; the front of the queue is attractive because the most adverse fills are deep fills. Under sniping, adverse-selection intensity is front-loaded; the first ranks absorb stale-quote executions, and a passive order may prefer to wait behind them. Under mixed informed flow, early ranks can be toxic, middle ranks benign, and deep ranks toxic again. The resulting queue-value profile can have an interior maximum even though all orders quote the same price. Figure 3 illustrates the three regimes. Section V shows how an informed trader's order-size problem generates these profiles endogenously.

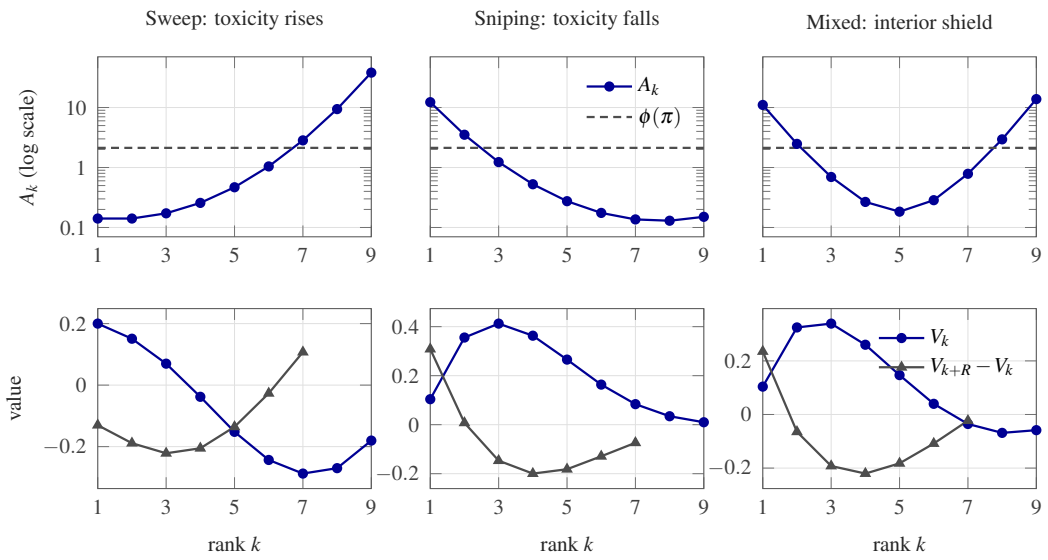

Figure 3: Primitive priority regimes generated with $\pi = 0.32$ and effective-rank wedge $R = 2$. The top row compares adverse-selection intensity A_k (log scale) with the toxicity threshold $\phi(\pi) = 2.125$; margins above the dashed line are toxic. In the bottom row, blue circles plot queue value V_k and gray triangles plot the skipped-block value $V_{k+R} - V_k$. Noise flow has ordinary middle-depth clips; informed sweep flow is concentrated in large clips; informed sniping flow is concentrated in immediate clips; and the mixed regime combines the two.

V Equilibrium Depth and Placement

The results so far price queue positions within a fixed environment. They invite an equilibrium objection: if the front of the queue is toxic and everyone can see the state, who stands there, and why does the displayed queue not unravel backward as each front order cancels and reposts behind the others? This section answers the objection with an entry-and-placement equilibrium in which informed order size, displayed depth, and the occupied ranks are jointly determined. The model generalizes the two-unit example that motivated Assumption 1 to a book of arbitrary depth.

V.A Informed Order-Size Choice

Let the asset value be binary, $v \in \{L, H\}$ with $L < a < H$ and $\pi = \mathbb{P}(v = H)$. Noise marketable buy demand is U, an integer random variable with probability mass function $u \mapsto \mathbb{P}(U = u)$ on $\{0, 1, \ldots, \bar{U}\}$. An informed trader observes v. If $v = L$ she does not buy. If $v = H$ she chooses a marketable clip of x queue-lot units, subject to displayed depth q at the ask, $x \leq q$. Taking deeper

units is increasingly costly: the jth unit of the clip carries incremental cost κg_j, where $g_1 = 0$, g_j is nondecreasing in j with $g_j > 0$ for $j \geq 2$, and the cost scale $\kappa \geq 0$ is drawn from a distribution G and observed by the informed trader. The schedule $\{g_j\}$ stands for detection risk, price impact beyond the displayed quote, or capital limits; its slope is the informed trader's order-size technology. Informed profit from a clip of size $x \leq q$ is $x(H-a) - \kappa \sum_{j \leq x} g_j$.

Lemma 2 (Cutoff clip demand)**.** *The informed trader's optimal clip is* $x^*(\kappa, q) = \min\{q, \hat{x}(\kappa)\}$, *where*

$$\hat{x}(\kappa) \equiv \max\{x \geq 1 : \kappa g_x \leq H - a\}$$

is the unconstrained clip, set to $+\infty$ *when the defining set is unbounded, in which case the constrained clip is q. The unconstrained clip has tail probabilities* $\eta_x \equiv \mathbb{P}(\hat{x} \geq x) = G((H-a)/g_x)$ *for* $x \geq 2$, *with* $\eta_1 = 1$ *and* η_x *nonincreasing in x.*

The tails $\{\eta_x\}$ are the model's sufficient statistic for informed order-flow technology. A sniping technology has steep costs and fast-decaying tails: informed flow mostly takes the first unit. A sweep technology has flat costs and heavy tails: informed flow reaches deep into the book. In the two-unit special case ($\bar{U} \leq 2$, clips capped at two, $g_2 = 1$), the single number $\eta_2 = G(H-a)$ is the sweep probability, and the model reduces to the example referenced in Assumption 1.

V.B Truncation Invariance

Total marketable buy volume at the ask is $Y = U + x^*(\kappa, q)\mathbf{1}_{\{v=H\}}$, with U, κ, and v independent. A displayed book of depth q gives the rank-k unit, $k \leq q$, the value $V_k(q) = \mathbb{E}[(a-v)\mathbf{1}_{\{Y \geq k\}}]$. Displayed depth enters informed demand through the constraint $x \leq q$, so in principle every rank value depends on the whole book. The next lemma shows that this dependence vanishes for occupied ranks.

Lemma 3 (Truncation invariance)**.** *For every* $1 \leq k \leq q$, *the event* $\{\min\{q, \hat{x}\} + U \geq k\}$ *coincides*

with $\{\hat{x}+U \geq k\}$. *Hence* $V_k(q)$ *does not depend on* q*: for all* $q \geq k$*,*

$$V_k(q) = V_k \equiv (1-\pi)(a-L)\,\mathbb{P}(U \geq k) - \pi(H-a)\,\mathbb{P}(\hat{x}+U \geq k). \tag{9}$$

The economics is that the depth constraint binds only at the cap: a clip constrained by displayed depth q was going to take at least $q \geq k$ units anyway, so whether deeper depth lets it take more is irrelevant to whether volume reaches rank k. Displayed depth behind an order therefore does not change the order's own exposure, and entry can be analyzed rank by rank. Two remarks bound the scope of this result. First, it is a property of separable clip costs; if informed costs or signals depend directly on the shape of the book, depth behind an order can change its exposure, and the swap experiment of Assumption 1 is then exactly the counterfactual that removes this channel from the rank comparison. Second, invariance is a statement about occupied ranks, not about the informed trader: her realized clip, and hence the deepest rank reached, does respond to displayed depth.

Equation (9) also delivers the informed/noise mixture of Section III from primitives: $p_N(k) = \mathbb{P}(U=k)$, $p_I(k) = \mathbb{P}(\hat{x}+U=k)$, $m_N(k) = L$, $m_I(k) = H$, so that

$$\Delta_k = (1-\pi)(a-L)\,\mathbb{P}(U=k) - \pi(H-a)\,\mathbb{P}(\hat{x}+U=k), \qquad A_k = \frac{\mathbb{P}(\hat{x}+U=k)}{\mathbb{P}(U=k)} \cdot \frac{H-a}{a-L}, \tag{10}$$

whenever $\mathbb{P}(U=k) > 0$; the identification of the stopping event $\{Y=k\}$ with $\{\hat{x}+U=k\}$ holds for $k < q$, and at every rank when the depth cap almost surely never binds. The cross-rank toxicity profile of Section IV is now generated, not assumed: fast-decaying clip tails concentrate $\mathbb{P}(\hat{x}+U=k)$ at low k and front-load A_k; heavy tails push it deeper.

V.C Entry-Stable Depth

Passive liquidity is supplied competitively by traders who pay a display cost $\zeta > 0$ to stand in the book; ζ collects message, monitoring, and capital costs and can be arbitrarily small. A depth $q \geq 0$ is *viable* if every occupied rank covers the display cost, $V_k \geq \zeta$ for all $k \leq q$. Entry exhausts viable ranks.

Definition 3 (Entry-stable depth)**.** A displayed depth q^* is entry-stable if it is viable and no deeper book is viable: $V_k \geq \zeta$ for all $k \leq q^*$, and for every $q > q^*$ some rank $k \leq q$ has $V_k < \zeta$.

Proposition 2 (Existence and uniqueness of entry-stable depth)**.** *The set of viable depths is* $\{0, 1, \ldots, q^*\}$ *with*

$$q^* = \min\{k \geq 0 : V_{k+1} < \zeta\}, \tag{11}$$

which exists, is unique, and is finite. Equilibrium displayed depth is therefore the first rank whose standalone value cannot cover the display cost.

The condition is a rank-level analogue of the marginal-unit condition in Glosten (1994), with two differences. The priced event is the stopping event at a rank within a single price rather than a tail event across prices, and the informed size distribution that drives it is endogenous to the order-size technology. Because V_k need not be monotone in k, the binding rank is whichever rank fails first, and the book can end even though deeper ranks would have been viable in isolation; with truncation invariance, however, ranks are positions rather than identities, so an unviable rank simply terminates the queue.

Proposition 3 (Depth and the order-size technology)**.** *(i) V_k is strictly decreasing in π at every rank reached with positive probability, so the entry-stable depth q^* is nonincreasing in π. (ii) A first-order stochastic dominance increase in the unconstrained clip $\hat{x}$ weakly lowers V_k for every $k \geq 2$ and leaves V_1 unchanged, so q^* is nonincreasing in sweep mass. (iii) The front margin satisfies*

$$\Delta_1 = (1-\pi)(a-L)\,\mathbb{P}(U=1) - \pi(H-a)\,(1-\eta_2)\,\mathbb{P}(U=0),$$

which is strictly increasing in the sweep tail η_2 whenever $\mathbb{P}(U=0) > 0$.

Parts (ii) and (iii) move in opposite directions, and the tension is the model's sharpest cross-sectional prediction: a shift of informed technology toward sweeps thins the book but detoxifies the front, while a shift toward sniping deepens the book but poisons the front. Books that are deep because informed flow rarely reaches their interior are exactly the books whose front ranks bear

concentrated stale-quote risk.

V.D Repost Stability and Equilibrium

Placement discipline comes from the deviations FIFO actually allows. Time priority cannot be bought, so a trader cannot insert herself at an interior rank: the only same-price repositioning available to an occupied rank is to cancel and re-enter at the back of the displayed queue. The feasible one-shot deviations from rank $k < q$ are therefore exit, with outside value zero, and back-of-queue reposting at an incremental cost $c_k \geq 0$ that collects message costs, latency, implementation risk, and the chance that the book state changes before the order is restored.

Definition 4 (Repost-stable queue)**.** A displayed queue of depth q with repost costs $\{c_k\}$ is repost-stable if every occupied rank $k \leq q$ prefers staying to exiting, $V_k \geq 0$, and every occupied rank $k < q$ prefers staying to back-of-queue reposting, $V_k \geq V_q - c_k$.

Definition 5 (Entry-and-placement equilibrium)**.** An entry-and-placement equilibrium is an informed demand rule and a displayed depth q^* such that (i) the informed clip solves the order-size problem of Lemma 2 at every depth; (ii) q^* is entry-stable; and (iii) the queue of depth q^* is repost-stable.

Theorem 2 (Equilibrium with toxic front priority)**.** *Define the surrendered-block threshold*

$$C_k^* \equiv \max\left\{0,\, V_{q^*} - V_k\right\} = \max\left\{0,\, -\textstyle\sum_{j=k}^{q^*-1} \Delta_j\right\}, \qquad k < q^*.$$

(i) An entry-and-placement equilibrium exists if and only if $c_k \geq C_k^*$ for every $k < q^*$, where q^* is the unique entry-stable depth. (ii) If $V_k \geq V_{q^*}$ for every $k < q^*$, the equilibrium exists for all repost costs, including $c_k = 0$. (iii) There is an open set of primitives for which the equilibrium queue has depth $q^* \geq 2$ and strictly negative front priority, $\Delta_1 < 0$.

Part (ii) isolates a stabilizing feature of the matching rule itself. Under FIFO, a trader cannot surrender only a toxic local block; she must surrender the entire block between her rank and the

Rank k	1	2	3	4	5
Rank value V_k	0.624	0.723	0.725	0.720	0.712
Priority premium Δ_k	-0.099	-0.002	0.005	0.008	–
Adverse-selection intensity A_k	28.8	4.40	2.70	2.03	0.03
Repost threshold C_k^*	0.088	0	0	0	–

Table 1: An entry-and-placement equilibrium with toxic front priority. Primitives: $a = 10$, $L = 8$, $H = 12$, $\pi = 0.2$ (so $\phi(\pi) = 4$), display cost $\zeta = 0.01$; noise demand $\mathbb{P}(U = 0) = 0.36$, $\mathbb{P}(U = u) = 0.01$ for $u \in \{1,2,3,4\}$, $\mathbb{P}(U = 5) = 0.60$; clip tails $\eta = (1, 0.20, 0.10, 0.05, 0.02)$ with $\eta_x = 0$ for $x \geq 6$. The entry-stable depth is $q^* = 5$ ($V_6 = -0.24 < \zeta$). The first two margins are toxic, $A_1, A_2 > \phi(\pi)$, yet all five ranks are viable; rank 3 is the interior optimum, consistent with Theorem 1. Only the front rank requires a positive reposting friction, $c_1 \geq C_1^* = 0.088$.

back. An interior toxic margin therefore triggers no reposting whenever the cumulative block to the back remains positive, and displayed queues can carry locally negative priority premia even when reposting is free. Repost costs are needed only where the whole remaining block is negative, and the threshold C_k^* prices exactly that block. In the two-unit special case, back reposting is literally the swap experiment of Assumption 1: the front and back units exchange positions at unchanged displayed depth, so the stability analysis respects the counterfactual under which the rank values were derived.

Table 1 reports an equilibrium that exhibits the full mechanism. Noise demand is bimodal—either no liquidity demand arrives or a parent order of five lots does—while informed flow mostly snipes the first unit. Marginal fills that stop in the interior of the queue are then disproportionately informed, the first two priority margins are toxic, and the best position in the queue is third in line. The queue nevertheless displays five units: ranks two through four are protected by FIFO feasibility, since their blocks to the back are positive and $C_k^* = 0$, and the front rank is held by any trader whose reposting friction exceeds $C_1^* = 0.088$, about four percent of the noise-state gain $a - L$. The equilibrium thus reconciles, at arbitrary depth, the three objects the two-unit example could only exhibit pairwise: positive displayed depth, negative front priority, and no unraveling. Figure 4 plots the equilibrium's rank-value and intensity profiles.

Corollary 2 (Sorting by reposting frictions)**.** *Suppose traders differ in their effective cost c_k^i of canceling and reposting from rank k to the back; this cost may include message and latency costs,*

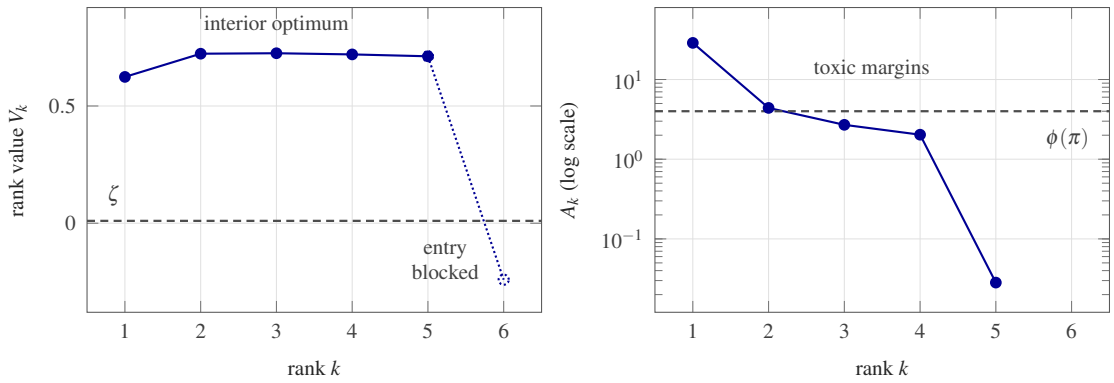

Figure 4: The entry-and-placement equilibrium of Table 1. The left panel plots rank values V_k; the dashed line is the display cost $\zeta = 0.01$. All five occupied ranks are viable, rank 3 is the interior optimum, and the queue ends at $q^* = 5$ because $V_6 < \zeta$ (open marker). The right panel plots adverse-selection intensity A_k on a log scale against the odds threshold $\phi(\pi) = 4$: the first two margins are toxic while deeper margins are not, the front-loaded profile of Corollary 1.

monitoring ability, inventory urgency, or a private execution motive that makes remaining exposed at the ask valuable. In any equilibrium, a trader i occupying rank $k < q^*$ must satisfy $c_k^i \geq C_k^*$. Ranks followed by negative cumulative blocks to the back require traders with higher effective reposting frictions: low-friction, high-monitoring types sort away from those ranks, while high-friction or high-execution-need types can rationally occupy them.

Toxic priority therefore changes not only the value of rank but the composition of the displayed queue. The corollary is a participation-constraint characterization, not a matching theorem: it identifies which trader types can occupy which ranks in any stable configuration without modeling the assignment mechanism. It also reframes a familiar empirical object. A trader who cancels and reposts deeper in the same queue has not lost interest in the price; she is buying insurance against the priority block she surrendered while keeping the price exposure. Section VI prices that option formally.

VI Cancellation as an Optimal Stopping Problem

The rank values of Sections I through V price hold-to-horizon claims. A resting order is not such a claim: its owner can monitor the book and cancel when toxicity signals arrive, and this option

is a central source of queue-position value in dynamic queue models (Moallemi and Yuan, 2016; Lehalle and Mounjid, 2017). The natural question is whether the cancellation option overturns the static characterization—whether a toxic margin can become valuable once the trader is allowed to escape it. This section shows that, for margins that are toxic state by state, it cannot. The marginal-state architecture survives optimal stopping: the option-adjusted value of one step of priority is a truncated version of the same marginal claim, and no cancellation policy can change the sign of a margin that is toxic state by state.

VI.A Option-Adjusted Rank Values

Work in discrete time $t \in \{0,1,\dots,T\}$ on a filtered probability space with filtration $(\mathscr{F}_t)_{t=0}^{T}$ generated by order flow and public book information. Cumulative marketable buy volume is a nondecreasing, integer-valued, adapted process C_t with $C_0 = 0$, and the execution-relevant value v is an integrable $\mathscr{F}_T$-measurable random variable. The fill time of rank k is

$$F_k \equiv \inf\{t \geq 1 : C_t \geq k\},$$

with $F_k \leq F_{k+1}$ because C_t is nondecreasing. Write $M_t \equiv \mathbb{E}[a - v \mid \mathscr{F}_t]$ for the conditional expected markout process. A cancellation policy is a stopping time τ with values in $\{0,1,\dots,T\}$: the order participates in matching through τ and is withdrawn afterward, so a rank-k order operated under policy τ pays $(a-v)\mathbf{1}_{\{F_k \leq \tau\}}$. The policy $\tau \equiv T$ never cancels and recovers the static value, $V_k = \mathbb{E}[(a-v)\mathbf{1}_{\{F_k \leq T\}}]$, with $Y = C_T$ the exposure-horizon volume of the static sections.

Definition 6 (Option-adjusted rank value)**.** The option-adjusted value of rank k is

$$\widehat{V}_k \equiv \sup_{\tau \in \mathscr{T}} \mathbb{E}\left[(a-v)\mathbf{1}_{\{F_k \leq \tau\}}\right],$$

where $\mathscr{T}$ is the set of stopping times with values in $\{0,\dots,T\}$. The option-adjusted priority premium is $\widehat{\Delta}_k \equiv \widehat{V}_k - \widehat{V}_{k+1}$.

Proposition 4 (Existence and exercise boundary)**.** *The supremum is attained.* $\widehat{V}_k \geq \max\{V_k, 0\}$*, and an optimal policy is the first time the conditional value of remaining exposure is nonpositive:*

$$\tau_k^* = \inf\left\{t \geq 0 : w_k(t) \leq 0\right\} \wedge T,$$

where $w_k(t)$ *is the* $\mathscr{F}_t$*-conditional value of keeping the order alive from* t *onward under optimal future cancellation, computed by backward induction; on* $\{F_k \leq t\}$ *the order has filled and the policy is irrelevant. If cancellation instead means reposting to the back of a queue of depth* q *at cost* c*, the withdrawal payoff* 0 *is replaced on the unfilled event by an* $\mathscr{F}_t$*-measurable floor, the option-adjusted value of the back position at the stopped state net of cost, and at a terminal or once-and-for-all exercise date the policy reduces to the one-shot rule: repost if and only if* $\sum_{j=k}^{q-1} \Delta_j < -c$*, the surrendered-block condition of Theorem 2.*

The proposition formalizes cancellation as the exercise of an American option to shed a priority block, and nests the one-shot reposting comparisons used in the equilibrium of Section V as its degenerate case. The substantive question is what the option does to the priority margin itself.

VI.B The Sandwich Theorem

For any stopping time τ, define the truncated priority premium

$$\Delta_k(\tau) \equiv \mathbb{E}\left[(a-v)\mathbf{1}_{\{F_k \leq \tau < F_{k+1}\}}\right]. \tag{12}$$

The event $\{F_k \leq \tau < F_{k+1}\}$ says that, by the stopped time, marketable volume has reached rank k but not rank $k+1$: the marginal execution state of Lemma 1, evaluated over the exposure window that the policy τ keeps open. In particular $\Delta_k(T) = \Delta_k$, the static premium.

Theorem 3 (Sandwich theorem for option-adjusted priority)**.** *Let* τ_k^* *and* τ_{k+1}^* *be optimal cancella-*

tion policies for ranks k and $k+1$. Then

$$\Delta_k(\tau^*_{k+1}) \;\le\; \widehat{\Delta}_k \;\le\; \Delta_k(\tau^*_k).$$

One step of option-adjusted priority is therefore still a claim on the single marginal execution state; the option only changes the window over which that state is collected. The bounds are tight in the degenerate cases: if neither rank ever cancels, both sides equal Δ_k.

Definition 7 (State-wise toxic margin)**.** The rank-k margin is state-wise toxic if $M_t \le 0$ almost surely on the event $\{F_k \le t < F_{k+1}\}$ for every $t \le T$: whenever the marginal exposure is open, the conditional expected markout is nonpositive. It is state-wise benign under the reverse inequality.

Corollary 3 (Sign preservation)**.** *If the rank-k margin is state-wise toxic, then $\Delta_k(\tau) \le 0$ for every stopping time τ, and hence $\widehat{\Delta}_k \le 0$. Symmetrically, a state-wise benign margin has $\widehat{\Delta}_k \ge 0$.*

The corollary is the precise sense in which cancellation attenuates rather than reverses toxicity. The option truncates exposure to the marginal state and raises the value of every rank, and in computed examples it compresses the magnitude of the priority premium; but it cannot make a state-wise toxic margin worth holding, because every window over which the marginal claim can be collected has nonpositive conditional value. Definition 7 is the dynamic strengthening of the static condition $\mathbb{E}[v \mid Y = k] > a$: the static condition signs the marginal claim on average over the full horizon, while the state-wise condition signs it at every stopping opportunity. When the margin is toxic on average but benign in some states, optimal cancellation harvests the benign states, and the sandwich bounds—rather than the sign result—describe what survives.

The result also bears on the equilibrium of Section V, whose stability comparisons used hold-to-horizon values. When the relevant blocks are state-wise signed, option-adjusted values preserve the sign of every surrendered block, so every zero-cost comparison in that equilibrium survives the option; comparisons against strictly positive thresholds, such as a repost cost lying between the static and option-adjusted block premia, depend on magnitudes that the option does change. The sign structure of the equilibrium is not an artifact of suppressing the option.

VII Belief-State Valuation

The static model prices priority from the conditional surplus of a marginal fill. This section asks how that sign moves as the public information state changes through price discovery. A Kyle/Back filter supplies the market's posterior mean and variance at the instant the queue decision is priced; a separate marked queue process determines which FIFO ranks are reached next. A rank is toxic when the information conveyed by the marginal execution exceeds the current quote cushion. The posterior is taken as the public information state at the valuation instant: the boundary below is a state-contingent pricing condition for queue priority, with the marked process used to price rank exposure rather than to solve a complete filtering equilibrium.

The Kyle/Back component supplies a one-dimensional public information state. Let terminal value $v \sim N(m_0, \Sigma_0)$ be revealed at horizon T, observed by an informed trader. Market makers' posterior mean and variance, $P_t = \mathbb{E}[v \mid \mathscr{F}_t^K]$ and $\Sigma_t = \mathrm{Var}(v \mid \mathscr{F}_t^K)$, are generated by a continuous Kyle signal

$$dY_t^K = \beta_t (v - P_t)\, dt + \sigma_u\, dB_t. \tag{13}$$

When the linear trading-intensity process β_t is deterministic, the Kalman-Bucy filter gives $dP_t = \lambda_t\, dY_t^K$ with $\lambda_t = \Sigma_t \beta_t / \sigma_u^2$, and

$$\Sigma_t = \frac{\Sigma_0}{1 + \Sigma_0 \mathscr{I}_t}, \qquad \mathscr{I}_t \equiv \int_0^t \frac{\beta_s^2}{\sigma_u^2}\, ds. \tag{14}$$

The scalar $\mathscr{I}_t$ is Kyle information time: how much private information order flow has incorporated into price. It is distinct from the static adverse-selection intensity A_k.

Queue rank moves through a separate marked order-arrival process. Let S_t collect the payoff-relevant public state for a passive ask order: the displayed book, the ask quote a_t, the posterior (P_t, Σ_t), the trader's effective rank, and the mark intensities that govern queue movement. Let $N(dt, d\mu)$ count buy market orders of queue-lot size $\mu \in \{1, 2, \ldots\}$, with conditional mark intensity $\nu_t(d\mu \mid S_t, v)$ and posterior intensity $\nu_\mu(S_t) = \mathbb{E}[\nu_t(\{\mu\} \mid S_t, v) \mid S_t]$. The event $Y = k$ of the static

sections is represented locally by the execution-size mark $\mu = k$.

Definition 8 (Rank-specific marginal informativeness)**.** For a mark $\mu = k$ with positive posterior arrival probability at state S_t, and $\Sigma_t > 0$, define

$$\chi_k(S_t) \equiv \frac{\mathbb{E}[v - P_t \mid \mu = k, S_t]}{\sqrt{\Sigma_t}}, \tag{15}$$

so that $\mathbb{E}[v \mid \mu = k, S_t] = P_t + \chi_k(S_t)\sqrt{\Sigma_t}$. The object is defined only on marks with positive posterior probability.

The scalar $\chi_k(S_t)$ is the normalized information content of the marginal execution that reaches rank k. It can be generated from primitives by informed clip-size choice rather than assumed rank by rank. Suppose the mark-k intensity is the sum of noise and informed components,

$$\nu_t(\{k\} \mid S_t, v) = \lambda_N(S_t) p_N(k \mid S_t) + \lambda_I(S_t) p_I(k \mid S_t) \exp\{\theta(S_t)(v - P_t)\}, \tag{16}$$

where $p_N(\cdot \mid S_t)$ and $p_I(\cdot \mid S_t)$ are the mark-size distributions of noise demand and informed clip choice and $\theta(S_t) > 0$ is the value sensitivity of informed arrivals—the dynamic analogue of the clip-size distributions that Lemma 2 generates from order-size costs.

Proposition 5 (Clip-size foundation for mark informativeness)**.** *Under* (16)*, the local exponential coefficient of the mark-k intensity at $v = P_t$ is*

$$\rho_k(S_t) = \theta(S_t) \frac{\lambda_I(S_t) p_I(k \mid S_t)}{\lambda_N(S_t) p_N(k \mid S_t) + \lambda_I(S_t) p_I(k \mid S_t)},$$

and under the locally exponential tilt $\nu_t(\{k\} \mid S_t, v) = \nu_k^0(S_t) \exp\{\rho_k(S_t)(v - P_t)\}$ the Gaussian posterior gives $\mathbb{E}[v - P_t \mid \mu = k, S_t] = \rho_k(S_t)\Sigma_t$ and hence $\chi_k(S_t) = \rho_k(S_t)\sqrt{\Sigma_t}$. For fixed value sensitivity, the cross-rank informativeness profile is determined by the informed-to-noise clip-size likelihood ratio $p_I(k \mid S_t)/p_N(k \mid S_t)$.

Front-loaded sniping corresponds to high ρ_k near the front; informed sweeps push ρ_k deeper,

exactly as in (10).

Proposition 6 (Information-state toxicity boundary)**.** *At state S_t, the instantaneous flow value of the margin-k execution event is $\delta_k(t,S_t) = \nu_k(S_t)\big(a_t - \mathbb{E}[v \mid \mu = k, S_t]\big)$. Margin-k priority therefore has negative instantaneous value if and only if*

$$\chi_k(S_t)\sqrt{\Sigma_t} > a_t - P_t, \tag{17}$$

equivalently $\Sigma_t > \big((a_t - P_t)/\chi_k(S_t)\big)^2$ when $\chi_k(S_t) > 0$. Under the exponential tilt with fixed cushion $a_t - P_t = \bar{c} > 0$ and fixed $\rho_k > 0$, the margin is toxic if and only if Kyle information time satisfies $\mathscr{I}_t < \rho_k/\bar{c} - 1/\Sigma_0$; such a crossing exists only if $\rho_k\Sigma_0 > \bar{c}$.

The boundary compares the quote cushion $a_t - P_t$ with the adverse-selection component of the marginal fill, $\chi_k(S_t)\sqrt{\Sigma_t}$: priority is toxic when the information content of the fill, scaled by remaining value uncertainty, exceeds the cushion. The boundary traces toxicity in the state variables $(\mathscr{I}_t, \bar{c})$, not along a realized sample path: in a coherent environment with informative marks, P_t, Σ_t, and a_t would all move at marks, and the fixed-cushion crossing is a comparative static in information time. Figure 5 plots the same comparative static holding the normalized informativeness χ_k fixed over a short window; under the primitive-consistent normalization, $\chi_k = \rho_k\sqrt{\Sigma_t}$ declines mechanically with information time, so the fixed-χ_k diagram is a short-window approximation rather than an additional restriction.

The boundary converts the static primitive regimes into moving priority regions. When $\chi_1 > \chi_2 > \cdots > \chi_K$, the toxic-priority set is a front block,

$$\mathscr{B}_t = \left\{k : \chi_k > \frac{a_t - P_t}{\sqrt{\Sigma_t}}\right\},$$

which weakly shrinks as information is incorporated and Σ_t falls: in a front-loaded sniping regime, the insurance value of depth ahead declines with Kyle information time. In a sweep regime the toxic block is rear-loaded, and in a mixed regime the toxic set can be nonmonotone, giving a dynamic

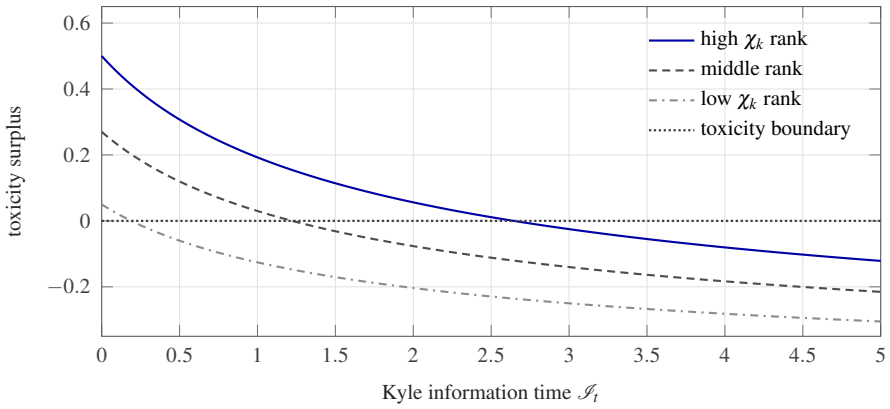

Figure 5: Schematic Kyle/Back toxicity boundary under fixed quote cushion $\bar{c} = 0.55$, $\Sigma_0 = 1$, and fixed normalized rank informativeness $\chi_k \in \{1.05, 0.82, 0.60\}$. The plotted object is $\chi_k/\sqrt{1+\mathscr{I}_t} - \bar{c}$. Posterior uncertainty decays with Kyle information time, so toxic ranks can cross into benign territory once the quote cushion exceeds the information content of the marginal fill.

route to interior queue positions. The dynamic extension does not change the object being priced; it changes the state variables that determine its sign. A full dynamic placement model would solve jointly for stopping rules, mark intensities, and quote adjustment; the boundary here is the input that such a model, and the empirical design of Section IX, both price.

VIII Market Design and Measurement

The block premium $D_S = \sum_{j \in S} \Delta_j$ links rank value to two further objects: the design of the priority rule itself, and the measurement of the rank that actually bears exposure.

VIII.A Priority-Rule Exposure Sharing

A matching rule does not only allocate speed; it allocates the marginal execution states at a price. FIFO assigns those states sequentially by time rank. A pro-rata rule spreads the same states across same-price displayed size. The comparison here is an exposure-accounting comparison at a fixed order-flow law; endogenous responses to the matching rule, including passive oversizing, changes in time-priority races, and informed clip-size choice, are outside the result.

Consider a same-price book with q equal unit orders at the ask. Under FIFO, rank k has

value V_k. Under an equal-size pro-rata benchmark, each unit receives the fraction $1/q$ of every marketable buy unit that executes against the first q units, so a pro-rata unit's value is $V^{PR}(q) \equiv \mathbb{E}[(a-v)\min\{Y,q\}/q]$.

Proposition 7 (Priority-rule exposure sharing)**.** *In the equal-size book, pro-rata priority gives each unit the average FIFO rank value:*

$$V^{PR}(q) = \frac{1}{q}\sum_{j=1}^{q} V_j.$$

Each marginal execution state $Y = j$, $j \leq q$, is borne by one FIFO rank but by all q pro-rata units in equal shares, and

$$V_1 - V^{PR}(q) = \frac{1}{q}\sum_{j=1}^{q-1}(q-j)\Delta_j, \qquad V^{PR}(q) - V_q = \frac{1}{q}\sum_{j=1}^{q-1} j\Delta_j.$$

Aggregate exposure at the price is identical, $qV^{PR}(q) = \sum_{j=1}^{q} V_j$; the difference is cross-sectional allocation. In a sniping regime, where front margins are toxic and the weighted sums above are negative, the front FIFO rank is worth less than a pro-rata share: pro-rata insures the front of the queue by spreading the toxic early states across all same-price liquidity. In a sweep regime, where front margins are benign, the front FIFO rank is worth more than a pro-rata share: pro-rata dilutes the front's valuable early claims and transfers them toward deeper liquidity. The choice of priority rule is therefore a choice about who bears adverse-selection exposure, not only about who trades first. FIFO creates rank-specific exposure and thereby makes reposting, queue-position races, and the toxic-front sorting of Corollary 2 economically meaningful; pro-rata pools the same exposure and mutes all three. This complements work in which matching rules allocate rents and race incentives (Field and Large, 2008; Budish et al., 2015): holding the race fixed, the rule still decides whose fills are the toxic fills.

VIII.B Effective Rank and Queue Tomography

Displayed rank is a measurement of queue position, not necessarily the rank at which an order absorbs execution risk. Suppose a trader appears second in the displayed ask queue; two displayed units trade, but her order does not fill. The missing queue movement reveals that displayed rank was not the rank governing execution. The relevant object is effective rank: the rank at which the order behaves after hidden and reserve priority, refresh rules, routing, and reconstruction error are incorporated (Bessembinder et al., 2009; Buti and Rindi, 2013; Boulatov and George, 2013; Yueshen, 2025).

Let $R \in \{0,1,2,\ldots\}$ denote the effective-rank wedge ahead of the visible order, measured in queue-lot units, so a visible rank k has effective rank $k+R$. If the wedge is known, the visible order has surrendered the priority block $k,\ldots,k+R-1$, and its value is governed by the same skipped-block premium as cancellation:

$$V_{k+R} - V_k > 0 \quad \Longleftrightarrow \quad \sum_{j=k}^{k+R-1} \Delta_j < 0, \tag{18}$$

which in the informed/noise mixture is $\overline{A}_{\{k,\ldots,k+R-1\}} > \phi(\pi)$. Queue tomography is the measurement step that maps message-level evidence—executions, queue advancement, displayed-depth depletion, replenishment, cancellations, and no-fill events—into a posterior over the wedge. Let $\mathscr{Q}$ denote the tomography signal. The signal induces a posterior over R and may also update the surplus of the marginal states reached at each effective rank, since a message sequence that reveals a wedge may also reveal whether executions are sniping clips, ordinary liquidity demand, or sweeps. For each posterior state define the conditional margins and rank values

$$\Delta_j(\mathscr{Q},h) \equiv \mathbb{E}\left[(a-v)\mathbf{1}_{\{Y=j\}} \mid \mathscr{Q}, R=h\right], \qquad V_r(\mathscr{Q},h) \equiv \sum_{j=r}^{\infty} \Delta_j(\mathscr{Q},h).$$

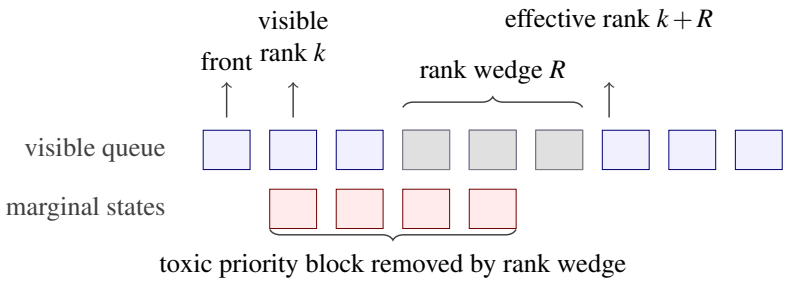

Figure 6: Effective-rank reconstruction. The trader observes displayed rank k, but an effective-rank wedge R shifts the order to effective rank $k+R$. Tomography uses message-level discrepancies to infer this gap. The value of the wedge is the negative of the cumulative priority premium over the skipped marginal states.

Definition 9 (Tomography-adjusted rank value)**.** The tomography-adjusted value of visible rank k is

$$W_k(\mathscr{Q}) \equiv \sum_h \mathbb{P}(R=h \mid \mathscr{Q})\, V_{k+h}(\mathscr{Q},h).$$

Proposition 8 (Tomography-adjusted priority)**.** *The marginal value of one visible step of priority is*

$$W_k(\mathscr{Q}) - W_{k+1}(\mathscr{Q}) = \sum_h \mathbb{P}(R=h \mid \mathscr{Q})\, \Delta_{k+h}(\mathscr{Q},h),$$

and the expected value of the wedge at visible rank k, relative to the no-wedge benchmark under the same posterior states, is

$$\sum_h \mathbb{P}(R=h \mid \mathscr{Q}) \big[V_{k+h}(\mathscr{Q},h) - V_k(\mathscr{Q},h)\big] = -\sum_h \mathbb{P}(R=h \mid \mathscr{Q}) \sum_{j=k}^{k+h-1} \Delta_j(\mathscr{Q},h).$$

A tomography-implied rank wedge is therefore valuable precisely when the posterior expected cumulative priority premium over the skipped interval is negative, and the decision margin it informs is the trader's repost-or-stay choice given the posterior over R. The sign of the tomographic shield is diagnostic of the order-flow technology: in a sniping regime, a wedge improves passive post-fill returns by placing other quantity in front of toxic immediate fills; in a sweep regime, the same wedge mainly delays execution until larger informed orders arrive and can reverse sign. Section IX flags this sign reversal as the design's sharpest test.

Estimated versions of these block decisions inherit a simple robustness property. If each margin is estimated with error at most ε, then for any finite block S the estimated and true block premia differ by at most $|S|\varepsilon$, so the toxic-or-insurance classification of a block agrees with the truth whenever $|D_S| > |S|\varepsilon$. Classification is fragile only inside an error tube around the switching surface $D_S = 0$; away from it, the discrete decisions that the theory maps to behavior are stable to estimation error.

IX Empirical Design

The empirical content is joint rather than univariate. The model's core predictions are that marginal fills have negative markouts exactly when the estimated priority premium is negative; that cancellation and reposting rise when the surrendered block premium is negative; that effective-rank wedges help only when they skip toxic marginal states; and, from the equilibrium of Section V, that informed order-size technology moves displayed depth and front-rank toxicity in opposite directions across securities and venues. These signs should reverse across sniping and sweep regimes and move with the information-state boundary of Section VII.

Before measurement, the model's three appearances of "the marginal fill" must be reconciled. The static sections use Y, cumulative marketable volume over the order's exposure horizon, and the marginal event $Y = k$. The dynamic section uses μ_t, the size mark of an individual queue-moving arrival. The stopping section uses the fill times F_k of a cumulative volume process. Empirically we adopt the per-arrival object: the observed mark $\mu_t = k$ is the event-study analogue of the static marginal event, with queue positions updated between arrivals, and the static Y is recovered by cumulating marks over a fixed exposure window. The empirical objects are then built in layers. The first layer is conditional fill quality: for an ask-side order at price a_t, the state-contingent marginal markout

$$M_k(s) = \mathbb{E}\left[a_t - v_{t+h} \mid \mu_t = k, S_t = s\right], \tag{19}$$

where v_{t+h} is a post-fill fundamental proxy such as a future midpoint or microprice. The second

layer is the event probability $q_k(s) = \mathbb{P}(\mu_t = k \mid S_t = s)$, giving the unnormalized marginal premium $\Delta_k(s) = q_k(s)M_k(s)$. The third layer aggregates margins into blocks for cancellation, reposting, and effective-rank shields, with the effective-rank target

$$\mathbb{E}\left[(a_t - v_{t+h})\mathbf{1}_{\{\mu_t = k+R_t\}} \mid \mathscr{Q}_t, S_t = s\right] \tag{20}$$

replacing the visible-rank version when hidden liquidity or reconstruction error is material.

The dynamic extension adds a state variable for when the signs should change. The empirical analogue of the toxicity boundary is

$$\widehat{\chi}_k(S_t)\widehat{\Sigma}_t^{1/2} > a_t - \widehat{P}_t, \tag{21}$$

where $a_t - \widehat{P}_t$ is the quote cushion, $\widehat{\Sigma}_t$ proxies remaining value uncertainty, and $\widehat{\chi}_k$ is the normalized informativeness of rank-k marginal fills, estimated from the rank-specific response of future midpoints or microprices to marginal executions. To avoid a mechanical test, the informativeness estimate must be formed separately from the markout used to test the premium: by sample splitting, by pre-period calibration, or by estimating $\widehat{\chi}_k$ from short-horizon price responses while testing Δ_k on a distinct performance horizon. The prediction is not that front priority is bad; it is that front priority is bad when the estimated boundary is crossed, and the sign need not decay mechanically if the cushion adjusts with information risk.

Cancellation and reposting are competing-risk outcomes rather than separate primitives: conditional on a live passive order, the next state is fill, repost deeper, disappearance, or continued exposure. The model predicts that reposting away from the front rises when the estimated surrendered block $\sum_{j=k}^{r-1}\widehat{\Delta}_j(s)$ is negative net of costs, with the FIFO-feasibility restriction of Theorem 2: the relevant block runs to the back of the queue, so interior toxic margins alone should not trigger reposting. High fill probability should predict worse post-fill performance only in states where the estimated marginal premium is negative. A tomography-implied wedge should improve passive performance when the posterior skipped block is negative and lose value, or reverse sign, in sweep

Theory object	Empirical counterpart	Use in test
Visible rank k	Displayed queue position reconstructed from add, cancel, and execute messages	Baseline rank measure before hidden-liquidity correction
Effective rank $k + R_t$	Displayed rank plus tomography-implied rank wedge	Measures the rank that prices priority exposure
Marginal fill event	Market-order mark that just reaches the order's effective rank	Identifies $\Delta_k(s)$ through marginal markouts
Post-fill value	Side-adjusted future midpoint or microprice markout	Tests whether marginal priority has negative surplus
Boundary components	$\widehat{\chi}_k$ from rank-specific price responses, $\widehat{\Sigma}_t$ from value-uncertainty proxies, $a_t - \widehat{P}_t$ from quote cushion	Tests when the priority sign should switch
Rank-wedge shield	Tomography-implied skipped priority block	Tests whether the wedge helps only by skipping toxic margins
Order-size technology	Clip-size distribution of aggressive flow by venue-security	Tests the depth-toxicity tradeoff of Proposition 3

Table 2: Mapping theory objects to empirical measurements.

states.

Identification requires care because rank, depth, informed trading, and cancellation are jointly determined. Orders alive at rank k when a trade arrives are survivors of an endogenous cancellation process, and front ranks are disproportionately occupied by fast traders, so rank-specific markouts confound rank with trader identity; the sorting result of Corollary 2 makes this confound a prediction of the model, which is why state-contingent and within-trader designs are needed. The tests require message-level data with order identifiers, exchange-provided queue position, or enough order-event detail to reconstruct effective rank; trade-and-quote data are not sufficient for the rank-wedge and reposting tests. The cleanest designs use short-horizon event windows, effective-rank reconstruction, controls for book state and order-flow imbalance, and variation that moves priority exposure without directly changing adverse-selection risk: FIFO versus pro-rata venue differences, tick-size changes that lengthen or shorten same-price queues, speed bumps and latency floors, exchange outages, throttle events, and rule changes that alter effective rank while leaving fundamentals comparable. Without such variation, estimates should be read as conditional measurement of the mechanism

Finding	Why it weakens the mechanism
Front priority is always bad after conditioning on state	The model predicts state-contingent, not unconditional, priority toxicity
Effective-rank wedges are always beneficial	The shield should hurt or lose value in sweep regimes
Cancellation responds to volatility or imbalance but not to estimated priority blocks	The cancellation channel should be tied to surrendered marginal priority value
Reposting triggered by interior toxic margins with positive blocks to the back	FIFO feasibility implies only the full back block matters
Fill markouts do not switch around the information-state boundary	The dynamic extension predicts sign changes with information risk relative to the quote cushion
Placebo ranks with similar fill probabilities but low $\widehat{\chi}_k$ show the same bad markouts	The mechanism is about informative marginal fills, not fill probability alone
Sweep-dominated markets have deeper books and more toxic fronts	The equilibrium predicts the opposite pairing

Table 3: Falsification tests for the priority-exposure mechanism.

rather than structural identification.

The strongest empirical content is joint. The same pre-fill proxy for toxic marginal states should predict poor post-fill value on high-fill orders, greater willingness to surrender front priority, and positive shield value from skipped blocks; the same clip-size technology measure should predict deep books with toxic fronts in sniping markets and shallow books with benign fronts in sweep markets. Evidence that only one margin moves is suggestive but incomplete, and several findings would weaken the mechanism outright. Table 3 collects them.

X Related Literature

This paper connects six strands of market-microstructure research. The first is adverse selection in limit order books. Copeland and Galai (1983) model the limit order as a free option to informed traders, and Glosten (1994) prices the marginal unit of depth across price levels through the tail expectation of value given total executed volume; Sandås (2001) tests that marginal-depth logic. Lemma 1 is a rank-level analogue of Glosten's condition, not a claim that adverse selection in

limit-order fills is new. The difference is that the comparison is within a single quoted price and prices the stopping event rather than the tail: adjacent FIFO ranks can carry opposite-signed marginal values because each rank adds a different marginal execution state. That within-price decomposition generates the shield, equilibrium, stopping, priority-rule, and effective-rank results, none of which appear in a price-level marginal-depth condition. The entry condition of Proposition 2 is the corresponding rank-level free-entry discipline, with the informed size distribution endogenous to an order-size technology.

The second strand is queue-position valuation and queue-reactive modeling. Moallemi and Yuan (2016) and Lehalle and Mounjid (2017) compute the value of queue position in rich dynamic environments where fill probabilities, queue dynamics, and cancellation options interact, and queue-reactive models make order-flow intensities depend on the book state (Cont et al., 2010; Huang et al., 2015). This paper changes the priced object rather than the environment: one step of priority is valuable or costly according to the surplus of the single marginal state it adds, and Theorem 3 shows that this object survives the cancellation option that is central to those models. The queue-reactive estimates are natural empirical discipline for the stopping probabilities $p_I(k)$ and $p_N(k)$.

The third strand studies informed trading, picking-off risk, latency, and price discovery (Kyle, 1985; Back, 1992; Foucault, 1999; Foucault et al., 2003; Hoffmann, 2014; Menkveld and Zoican, 2017; Budish et al., 2015). Flow-toxicity measures and the empirical sniping literature show that some executions are bad news for liquidity suppliers and that latency races concentrate them (Easley et al., 2011; Foucault et al., 2017; Aquilina et al., 2022). The belief-state boundary of Section VII expresses this in the Kyle/Back state: priority is toxic when rank-specific fill informativeness, scaled by remaining uncertainty, exceeds the quote cushion, with the rank profile of informativeness generated by informed clip-size choice.

The fourth strand is order submission, cancellation, and fleeting liquidity in dynamic limit order markets (Parlour, 1998; Foucault et al., 2005; Goettler et al., 2005; Rosu, 2009, 2020; Hollifield et al., 2004; Hasbrouck and Saar, 2009, 2013; Baruch and Glosten, 2013). In these models cancellation reflects changing valuations, monitoring, or strategic quote dynamics. Here cancellation is tied

to a specific object, the cumulative value of the priority block surrendered by reposting, and the equilibrium of Section V adds the matching-rule feasibility constraint: under FIFO only the block to the back can be surrendered, which is itself a stabilizing force. Fleeting orders and flickering quotes thus admit a complementary reading as exercises of the priority-shedding option.

The fifth strand is market design through priority rules, tick sizes, and fees. Pro-rata matching changes incentives to supply and cancel same-price depth (Field and Large, 2008); tick sizes determine how time priority rations liquidity-provision rents (Yao and Ye, 2018); and fee structures shift liquidity-supply incentives at given quotes (Colliard and Foucault, 2012; Malinova and Park, 2015; Battalio et al., 2016). Proposition 7 isolates a complementary allocation: holding the same-price aggregate exposure fixed, FIFO concentrates marginal states by time rank while pro-rata shares them across displayed size. Maker rebates and access fees shift the effective execution surplus $a - v$ in all of the paper's formulas and can therefore move a priority margin across the zero threshold; a complete fee-design treatment is left for future work.

The final strand concerns hidden liquidity and queue reconstruction (Bessembinder et al., 2009; Buti and Rindi, 2013; Boulatov and George, 2013; Yueshen, 2025). The tomography results give reconstruction an economic target: once an effective-rank wedge shifts a displayed order from rank k to $k+R$, its value is governed by the skipped priority block, so measuring the wedge measures exposure, and the sign of the wedge's value is diagnostic of the order-flow technology.

XI Conclusion

FIFO priority is usually described as a queueing advantage. This paper shows that it is better understood as exposure to marginal execution states. A trader closer to the front receives extra fills, but those fills need not be good fills: one step of priority is a claim on exactly one marginal execution state, and when that state is adverse, priority is costly and the queue ahead is insurance. The threshold and shield results characterize when each case obtains. The entry-and-placement equilibrium shows that toxic front priority is not self-erasing: displayed depth is determined

rank by rank, FIFO's own feasibility constraint protects interior toxic blocks, reposting frictions price the front, and the queue's composition sorts on those frictions. The stopping-time results show that the characterization is not an artifact of suppressing the cancellation option: the option truncates exposure to the marginal state but cannot reverse a state-wise toxic margin. In a Kyle/Back information state the same condition becomes a moving boundary, crossed as price discovery raises the quote cushion relative to fill informativeness.

The model does not predict that front priority is always bad or that rank wedges are always good. It predicts that FIFO priority allocates adverse-selection exposure, that the sign of that exposure switches with the informed order-size technology, and that depth and front-rank toxicity move together in a specific way across markets. A venue that cares about the quality of passive fill outcomes, and not only execution speed, should treat its priority rule as an adverse-selection allocation mechanism: FIFO and pro-rata are not merely different queueing protocols but different insurance arrangements for same-price liquidity suppliers.

References

Matteo Aquilina, Eric Budish, and Peter O'Neill. Quantifying the high-frequency trading "arms race". *The Quarterly Journal of Economics*, 137(1):493–564, 2022. doi: 10.1093/qje/qjab032.

Kerry Back. Insider trading in continuous time. *The Review of Financial Studies*, 5(3):387–409, 1992.

Shmuel Baruch and Lawrence R. Glosten. Fleeting orders. *Columbia Business School Research Paper*, 2013. doi: 10.2139/ssrn.2278457. No. 13-43, available at SSRN 2278457.

Robert Battalio, Shane A. Corwin, and Robert Jennings. Can brokers have it all? on the relation between make-take fees and limit order execution quality. *The Journal of Finance*, 71(5): 2193–2238, 2016. doi: 10.1111/jofi.12422.

Hendrik Bessembinder, Marios Panayides, and Kumar Venkataraman. Hidden liquidity: An analysis

of order exposure strategies in electronic stock markets. *Journal of Financial Economics*, 94(3): 361–383, 2009. doi: 10.1016/j.jfineco.2009.02.001.

Alex Boulatov and Thomas J. George. Hidden and displayed liquidity in securities markets with informed liquidity providers. *The Review of Financial Studies*, 26(8):2096–2137, 2013. doi: 10.1093/rfs/hhs123.

Eric Budish, Peter Cramton, and John Shim. The high-frequency trading arms race: Frequent batch auctions as a market design response. *The Quarterly Journal of Economics*, 130(4):1547–1621, 2015.

Sabrina Buti and Barbara Rindi. Undisclosed orders and optimal submission strategies in a limit order market. *Journal of Financial Economics*, 109(3):797–812, 2013. doi: 10.1016/j.jfineco.2013.04.002.

Jean-Edouard Colliard and Thierry Foucault. Trading fees and efficiency in limit order markets. *The Review of Financial Studies*, 25(11):3389–3421, 2012. doi: 10.1093/rfs/hhs089.

Rama Cont, Sasha Stoikov, and Rishi Talreja. A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563, 2010.

Thomas E. Copeland and Dan Galai. Information effects on the bid-ask spread. *The Journal of Finance*, 38(5):1457–1469, 1983. doi: 10.1111/j.1540-6261.1983.tb03834.x.

David Easley, Marcos M. López de Prado, and Maureen O'Hara. The microstructure of the "flash crash": Flow toxicity, liquidity crashes, and the probability of informed trading. *The Journal of Portfolio Management*, 37(2):118–128, 2011. doi: 10.3905/jpm.2011.37.2.118.

Jonathan Field and Jeremy Large. Pro-rata matching and one-tick futures markets. CFS Working Paper Series 2008/40, Center for Financial Studies, 2008.

Thierry Foucault. Order flow composition and trading costs in a dynamic limit order market. *The Journal of Finance*, 54(1):99–134, 1999.

Thierry Foucault, Ailsa Röell, and Patrik Sandås. Market making with costly monitoring: An analysis of the soes controversy. *The Review of Financial Studies*, 16(2):345–384, 2003.

Thierry Foucault, Ohad Kadan, and Eugene Kandel. Limit order book as a market for liquidity. *The Review of Financial Studies*, 18(4):1171–1217, 2005.

Thierry Foucault, Roman Kozhan, and Wing Wah Tham. Toxic arbitrage. *The Review of Financial Studies*, 30(4):1053–1094, 2017. doi: 10.1093/rfs/hhw103.

Lawrence R. Glosten. Is the electronic open limit order book inevitable? *The Journal of Finance*, 49(4):1127–1161, 1994.

Ronald L. Goettler, Christine A. Parlour, and Uday Rajan. Equilibrium in a dynamic limit order market. *Journal of Finance*, 60(5):2149–2192, 2005.

Joel Hasbrouck and Gideon Saar. Technology and liquidity provision: The blurring of traditional definitions. *Journal of Financial Markets*, 12(2):143–172, 2009. doi: 10.1016/j.finmar.2008.06.002.

Joel Hasbrouck and Gideon Saar. Low-latency trading. *Journal of Financial Markets*, 16(4):646–679, 2013. doi: 10.1016/j.finmar.2013.05.003.

Peter Hoffmann. A dynamic limit order market with fast and slow traders. *Journal of Financial Economics*, 113(1):156–169, 2014.

Burton Hollifield, Robert A. Miller, and Patrik Sandås. Empirical analysis of limit order markets. *The Review of Economic Studies*, 71(4):1027–1063, 2004.

Weibing Huang, Charles-Albert Lehalle, and Mathieu Rosenbaum. Simulating and analyzing order book data: The queue-reactive model. *Journal of the American Statistical Association*, 110(509):107–122, 2015. doi: 10.1080/01621459.2014.982278.

Albert S. Kyle. Continuous auctions and insider trading. *Econometrica*, 53(6):1315–1335, 1985.

Charles-Albert Lehalle and Othmane Mounjid. Limit order strategic placement with adverse selection risk and the role of latency. *Market Microstructure and Liquidity*, 3(1):1750009, 2017.

Katya Malinova and Andreas Park. Subsidizing liquidity: The impact of make/take fees on market quality. *The Journal of Finance*, 70(2):509–536, 2015. doi: 10.1111/jofi.12230.

Albert J. Menkveld and Marius A. Zoican. Need for speed? exchange latency and liquidity. *The Review of Financial Studies*, 30(4):1188–1228, 2017.

Ciamac C. Moallemi and Kai Yuan. A model for queue position valuation in a limit order book. *Columbia Business School Research Paper*, 2016. doi: 10.2139/ssrn.2996221. No. 17-70, available at SSRN 2996221.

Christine A. Parlour. Price dynamics in limit order markets. *The Review of Financial Studies*, 11(4): 789–816, 1998.

Ioanid Rosu. A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11): 4601–4641, 2009.

Ioanid Rosu. Liquidity and information in limit order markets. *Journal of Financial and Quantitative Analysis*, 55(6):1792–1839, 2020.

Patrik Sandås. Adverse selection and competitive market making: Empirical evidence from a limit order market. *The Review of Financial Studies*, 14(3):705–734, 2001.

Chen Yao and Mao Ye. Why trading speed matters: A tale of queue rationing under price controls. *The Review of Financial Studies*, 31(6):2157–2183, 2018. doi: 10.1093/rfs/hhy002.

Bart Zhou Yueshen. Queuing uncertainty of limit orders. *Management Science*, 72(6):4760–4779, 2025. doi: 10.1287/mnsc.2023.03371.

A Risk-Adjusted Priority

The main results rank queue positions by expected surplus. A risk-averse liquidity supplier may also penalize the variance of execution surplus. With $X_k = (a-v)\mathbf{1}_{\{Y \geq k\}}$ and mean-variance objective $J_k(\gamma) = \mathbb{E}[X_k] - \frac{\gamma}{2}\mathrm{Var}(X_k)$ for risk aversion $\gamma \geq 0$:

Proposition 9 (Mean-variance priority)**.** *With* $Z_k = (a-v)\mathbf{1}_{\{Y=k\}}$,

$$J_k(\gamma) - J_{k+1}(\gamma) = \Delta_k - \frac{\gamma}{2}\left[\mathrm{Var}(Z_k) - 2\Delta_k V_{k+1}\right].$$

The adjustment has a closed form because adjacent FIFO payoffs are supported on disjoint execution events: $Z_k X_{k+1} = 0$ state by state, so $\mathrm{Cov}(Z_k, X_{k+1}) = -\Delta_k V_{k+1}$ exactly, and no primitive covariance object is introduced. When $\Delta_k < 0$ and $V_{k+1} > 0$, the term $-2\Delta_k V_{k+1}$ is positive: toxic priority simultaneously lowers the mean and raises the variance, so mean-variance preferences strictly reinforce the shield motive rather than merely qualifying it. The efficient queue rank need not be the front rank: a trader in second or third position can be close enough to capture benign flow yet far enough back that other orders absorb toxic immediacy.

B Proofs

Proof of Lemma 1. By definition, $V_k - V_{k+1} = \mathbb{E}[(a-v)(\mathbf{1}_{\{Y \geq k\}} - \mathbf{1}_{\{Y \geq k+1\}})]$. The difference of indicators is $\mathbf{1}_{\{Y=k\}}$. Conditioning on the event $Y = k$ gives the second equality, and the sign statement follows because $\mathbb{P}(Y = k) > 0$. □

Proof of Proposition 1. Conditioning the marginal priority identity on the order-flow regime gives (6). Using the definitions of b_k, A_k, and $\phi(\pi)$, $\Delta_k(\pi) = (1-\pi)b_k - \pi b_k A_k = \pi b_k[\phi(\pi) - A_k]$, and $\pi b_k > 0$ gives the threshold. The posterior form follows from Bayes' rule on the event $Y = k$. □

Proof of Theorem 1. For a feasible move forward to $l < k^*$, telescoping gives

$$V_l - V_{k^*} = \sum_{j=l}^{k^*-1} \Delta_j(\pi) = \pi \sum_{j=l}^{k^*-1} b_j \big[\phi(\pi) - A_j\big],$$

so $V_l \le V_{k^*}$ if and only if $\sum_{j=l}^{k^*-1} b_j A_j \ge \phi(\pi) \sum_{j=l}^{k^*-1} b_j$, the first block-average condition. Moving backward to $r > k^*$ surrenders the block beginning at k^*: $V_{k^*} - V_r = \sum_{j=k^*}^{r-1} \Delta_j(\pi)$, so $V_{k^*} \ge V_r$ if and only if the second condition holds. These inequalities against all feasible ranks, together with $V_{k^*} \ge 0$ against the outside option, are necessary and sufficient for global optimality. Strict inequalities and $V_{k^*} > 0$ rule out ties. □

Proof of Corollary 1. Each block $\{l, \dots, k^* - 1\}$ consists of margins with $A_j \ge A_{k^*-1} > \phi(\pi)$, so its weighted average exceeds $\phi(\pi)$; each block $\{k^*, \dots, r-1\}$ consists of margins with $A_j \le A_{k^*} < \phi(\pi)$, so its weighted average is below $\phi(\pi)$. Theorem 1 applies, with strict inequalities. □

Proof of Lemma 2. Informed profit from a clip of $x \le q$ units is $\Pi(x) = x(H-a) - \kappa \sum_{j \le x} g_j$, with increment $\Pi(x) - \Pi(x-1) = (H-a) - \kappa g_x$, nonincreasing in x because g is nondecreasing and $\kappa \ge 0$. The objective is therefore discretely concave, and the largest unconstrained maximizer is $\hat{x}(\kappa) = \max\{x \ge 1 : \kappa g_x \le H - a\}$, well defined because $g_1 = 0$ and $H > a$. Concavity implies that the constrained optimum is the truncation $x^*(\kappa, q) = \min\{q, \hat{x}(\kappa)\}$. For $x \ge 2$, $\hat{x} \ge x$ if and only if $\kappa \le (H-a)/g_x$, an event of probability $G((H-a)/g_x)$, nonincreasing in x because g is nondecreasing. □

Proof of Lemma 3. Fix $1 \le k \le q$. Since $\min\{q, \hat{x}\} \ge k - U$ holds if and only if both $q \ge k - U$ and $\hat{x} \ge k - U$, and $q \ge k \ge k - U$ holds always because $U \ge 0$, the events $\{\min\{q, \hat{x}\} + U \ge k\}$ and $\{\hat{x} + U \ge k\}$ coincide. In the low state $Y = U$ and the unit earns $a - L > 0$ on $\{U \ge k\}$; in the high state $Y = \min\{q, \hat{x}\} + U$ and the unit earns $a - H < 0$ on the event above. Taking expectations gives (9), which does not involve q. □

Proof of Proposition 2. By Lemma 3, depth q is viable if and only if $V_k \geq \zeta$ for all $k \leq q$, a condition that is a subset of the corresponding condition at $q+1$; the viable set is therefore closed downward and equals $\{0,\ldots,q^*\}$ with $q^* = \min\{k \geq 0 : V_{k+1} < \zeta\}$, provided this minimum exists. It does: $V_k \leq (1-\pi)(a-L)\,\mathbb{P}(U \geq k) \to 0$ as $k \to \infty$, so $V_k < \zeta$ for all large k. Depth 0 is viable vacuously, so q^* exists, is unique, and is finite. □

Proof of Proposition 3. (i) From (9), $\partial V_k/\partial \pi = -(a-L)\,\mathbb{P}(U \geq k) - (H-a)\,\mathbb{P}(\hat{x}+U \geq k) < 0$ whenever either probability is positive, which holds for every $k \leq q^*$ since $V_{q^*} \geq \zeta > 0$ requires $\mathbb{P}(U \geq q^*) > 0$. Each viability constraint tightens, so the viable set shrinks and q^* is nonincreasing. (ii) A first-order stochastic dominance increase in $\hat{x}$ weakly raises $\mathbb{P}(\hat{x} \geq k-u)$ for every u, hence weakly raises $\mathbb{P}(\hat{x}+U \geq k)$ and weakly lowers V_k; for $k=1$, $\mathbb{P}(\hat{x}+U \geq 1) = 1$ before and after, so V_1 is unchanged. (iii) $\mathbb{P}(\hat{x}+U=1) = \mathbb{P}(\hat{x}=1)\,\mathbb{P}(U=0) = (1-\eta_2)\,\mathbb{P}(U=0)$ because $\hat{x} \geq 1$ always, so $\Delta_1 = (1-\pi)(a-L)\,\mathbb{P}(U=1) - \pi(H-a)(1-\eta_2)\,\mathbb{P}(U=0)$, which is strictly increasing in η_2 whenever $\mathbb{P}(U=0) > 0$. □

Proof of Theorem 2. (i) Lemma 2 gives informed optimality at every depth, and Proposition 2 gives the unique entry-stable q^*. Exit deviations are unprofitable because $V_k \geq \zeta > 0$ for $k \leq q^*$. Under FIFO, the unique same-price repositioning deviation from rank $k < q^*$ is back-of-queue reposting, with deviation gain $V_{q^*} - V_k - c_k$; by truncation invariance the values are unchanged by the reordering of occupied units. All such deviations are unprofitable if and only if $c_k \geq V_{q^*} - V_k$ for every k with $V_{q^*} > V_k$, which is $c_k \geq C_k^*$ for all $k < q^*$. (ii) If $V_k \geq V_{q^*}$ for all $k < q^*$, then $C_k^* = 0$ and the constraint never binds. (iii) The primitives of Table 1 satisfy $V_k > \zeta$ for $k \leq 5$, $V_6 < \zeta$, $\Delta_1 < 0$, and $C_1^* > 0$ with strict inequalities throughout; every V_k is a polynomial in (π, a, L, H), the noise probability masses, and the clip tails, so all strict inequalities persist on an open neighborhood of these primitives, and any repost costs with $c_k > C_k^*$ remain admissible on that neighborhood. □

Proof of Corollary 2. Apply Theorem 2(i) to trader i's friction at rank k: stability requires $c_k^i \geq V_{q^*} - V_k$, and only ranks with $V_{q^*} > V_k$ impose a positive requirement. □

Proof of Proposition 4. For any stopping time τ, iterated expectations give $\mathbb{E}[(a-v)\mathbf{1}_{\{F_k\le\tau\}}] = \mathbb{E}[\mathbf{1}_{\{F_k\le\tau\}}M_\tau] = \mathbb{E}[\xi_\tau]$, where $\xi_t \equiv \mathbf{1}_{\{F_k\le t\}}M_t$ is adapted and integrable, using $\{F_k \le \tau\} \in \mathscr{F}_\tau$ and $M_\tau = \mathbb{E}[a-v \mid \mathscr{F}_\tau]$ by optional sampling at the bounded time τ. The problem $\sup_\tau \mathbb{E}[\xi_\tau]$ is a finite-horizon optimal stopping problem; its Snell envelope $\Lambda_t = \max\{\xi_t, \mathbb{E}[\Lambda_{t+1} \mid \mathscr{F}_t]\}$, $\Lambda_T = \xi_T$, attains the supremum at $\tau^* = \inf\{t : \Lambda_t = \xi_t\}$. Choosing $\tau \equiv T$ gives $\widehat{V}_k \ge V_k$, and $\tau \equiv 0$ gives $\widehat{V}_k \ge 0$ because $F_k \ge 1$. On $\{F_k \le t\}$ the reward is $\xi_t = M_t$ and $\mathbb{E}[\Lambda_{t+1} \mid \mathscr{F}_t] \ge \mathbb{E}[\xi_{t+1} \mid \mathscr{F}_t] = \mathbb{E}[M_{t+1} \mid \mathscr{F}_t] = M_t = \xi_t$, so continuation is weakly optimal after the fill and the policy is payoff-relevant only on $\{F_k > t\}$, where $\xi_t = 0$ and stopping is optimal exactly when the conditional continuation value $w_k(t) = \mathbb{E}[\Lambda_{t+1} \mid \mathscr{F}_t]$ is nonpositive. Replacing the withdrawal payoff 0 on the unfilled event with an $\mathscr{F}_t$-measurable floor (the option-adjusted value of the back position net of cost) leaves the Snell structure unchanged. At a single terminal exercise date the comparison is between V_k and $V_q - c$, which is $\sum_{j=k}^{q-1}\Delta_j < -c$. □

Proof of Theorem 3. Because C_t is nondecreasing, $F_k \le F_{k+1}$, so for every stopping time τ,

$$\mathbf{1}_{\{F_k\le\tau\}} - \mathbf{1}_{\{F_{k+1}\le\tau\}} = \mathbf{1}_{\{F_k\le\tau<F_{k+1}\}},$$

and hence $\mathbb{E}[(a-v)\mathbf{1}_{\{F_k\le\tau\}}] = \mathbb{E}[(a-v)\mathbf{1}_{\{F_{k+1}\le\tau\}}] + \Delta_k(\tau)$. Evaluating at $\tau = \tau_k^*$,

$$\widehat{V}_k = \mathbb{E}[(a-v)\mathbf{1}_{\{F_{k+1}\le\tau_k^*\}}] + \Delta_k(\tau_k^*) \le \widehat{V}_{k+1} + \Delta_k(\tau_k^*),$$

because τ_k^* is feasible for rank $k+1$. Evaluating at $\tau = \tau_{k+1}^*$,

$$\widehat{V}_k \ge \mathbb{E}[(a-v)\mathbf{1}_{\{F_k\le\tau_{k+1}^*\}}] = \widehat{V}_{k+1} + \Delta_k(\tau_{k+1}^*),$$

because τ_{k+1}^* is feasible for rank k. Subtracting $\widehat{V}_{k+1}$ gives the two bounds. □

Proof of Corollary 3. Fix any stopping time τ and let $A = \{F_k \le \tau < F_{k+1}\}$. For each t, $A \cap \{\tau = t\} = \{F_k \le t < F_{k+1}\} \cap \{\tau = t\} \in \mathscr{F}_t$, so $A \in \mathscr{F}_\tau$, and by iterated expectations and optional sampling

$$\Delta_k(\tau) = \mathbb{E}\left[\mathbf{1}_A \, \mathbb{E}[a - v \mid \mathscr{F}_\tau]\right] = \mathbb{E}[\mathbf{1}_A M_\tau].$$

On $A \cap \{\tau = t\}$ the state-wise toxicity condition gives $M_\tau = M_t \le 0$, so $\Delta_k(\tau) \le 0$. Applying Theorem 3, $\widehat{\Delta}_k \le \Delta_k(\tau_k^*) \le 0$. The benign case is symmetric, using the lower bound. □

Proof of Proposition 5. Write $\alpha_k = \lambda_N(S_t) p_N(k \mid S_t)$ and $\beta_k = \lambda_I(S_t) p_I(k \mid S_t)$, so the mark intensity is $\alpha_k + \beta_k \exp\{\theta(v - P_t)\}$. The local exponential coefficient at $v = P_t$ is

$$\frac{\partial}{\partial v} \log\left(\alpha_k + \beta_k e^{\theta(v - P_t)}\right)\Big|_{v = P_t} = \theta \frac{\beta_k}{\alpha_k + \beta_k},$$

the stated $\rho_k(S_t)$. Under the exponential tilt, observing the mark reweights the Gaussian posterior density of $v - P_t$ by $e^{\rho_k (v - P_t)}$, and the Gaussian exponential-tilt identity gives $\mathbb{E}[v - P_t \mid \mu = k, S_t] = \rho_k(S_t)\Sigma_t$, hence $\chi_k = \rho_k \sqrt{\Sigma_t}$. Since $\beta_k/(\alpha_k + \beta_k)$ is increasing in β_k/α_k, the profile is governed by $p_I(k \mid S_t)/p_N(k \mid S_t)$. □

Proof of Proposition 6. Over $[t, t + dt]$ the probability of the mark $\mu = k$ is $\nu_k(S_t)dt + o(dt)$, with expected surplus $a_t - \mathbb{E}[v \mid \mu = k, S_t]$ conditional on the mark; dividing by dt gives the flow value δ_k, which is negative if and only if $\mathbb{E}[v \mid \mu = k, S_t] > a_t$. Substituting $\mathbb{E}[v \mid \mu = k, S_t] = P_t + \chi_k(S_t)\sqrt{\Sigma_t}$ gives (17). Under the exponential tilt, $\chi_k \sqrt{\Sigma_t} = \rho_k \Sigma_t$, and substituting $\Sigma_t = \Sigma_0/(1 + \Sigma_0 \mathscr{I}_t)$ into $\rho_k \Sigma_t > \bar{c}$ gives $\mathscr{I}_t < \rho_k/\bar{c} - 1/\Sigma_0$, which admits a solution $\mathscr{I}_t \ge 0$ only if $\rho_k \Sigma_0 > \bar{c}$. □

Proof of Proposition 7. Under equal-size pro-rata, a unit receives the fraction $1/q$ of each of the first q executed units, so its payoff is $(a - v)\min\{Y, q\}/q$. Since $\min\{Y, q\} = \sum_{j=1}^{q} \mathbf{1}_{\{Y \ge j\}}$, taking expectations gives $V^{PR}(q) = q^{-1} \sum_{j=1}^{q} V_j$. Substituting $V_j = V_1 - \sum_{\ell=1}^{j-1} \Delta_\ell$ gives the front-rank comparison; substituting $V_j = V_q + \sum_{\ell=j}^{q-1} \Delta_\ell$ gives the back-rank comparison. □

Proof of Proposition 8. For each wedge realization h, the adjacent-rank identity gives $V_{k+h}(\mathscr{Q},h) - V_{k+1+h}(\mathscr{Q},h) = \Delta_{k+h}(\mathscr{Q},h)$; averaging over the posterior of R given $\mathscr{Q}$ gives the first display. The second follows from telescoping $V_{k+h}(\mathscr{Q},h) - V_k(\mathscr{Q},h) = -\sum_{j=k}^{k+h-1} \Delta_j(\mathscr{Q},h)$ and averaging. □

Proof of Proposition 9. Since $X_k = Z_k + X_{k+1}$, the means give $\mathbb{E}[X_k] - \mathbb{E}[X_{k+1}] = \Delta_k$ and the variances give $\mathrm{Var}(X_k) - \mathrm{Var}(X_{k+1}) = \mathrm{Var}(Z_k) + 2\,\mathrm{Cov}(Z_k, X_{k+1})$. The supports of Z_k and X_{k+1} are disjoint: Z_k is nonzero only on $\{Y = k\}$ and X_{k+1} only on $\{Y \geq k+1\}$, so $Z_k X_{k+1} = 0$ state by state and $\mathrm{Cov}(Z_k, X_{k+1}) = -\mathbb{E}[Z_k]\,\mathbb{E}[X_{k+1}] = -\Delta_k V_{k+1}$. Substituting into $J_k(\gamma) - J_{k+1}(\gamma)$ proves the formula. □