A Numerical Verification

This appendix provides numerical evidence supporting the theoretical results in Sections 4 and 5. We verify: (i) the accuracy of the first-order Taylor approximation for the spread effect; (ii) the sensitivity of the deterrence threshold $\delta^\star$ to model primitives; (iii) the information aggregation condition under different parameter configurations. [3].

Calibration. Throughout, we use the following baseline parameters unless otherwise noted:

- Honest accuracy: $q = 0.75$ (75% accurate signals)
- Informed trader intensity: $\mu = 0.6$ (60% informed arrival rate)
- Prior: $p_0 \in \{0.3, 0.5, 0.7\}$ (varying asymmetry)
- Honest fraction: $\rho \in [0.5, 1]$
- Detection probability: $\delta \in [0, 1]$
- Manipulator strategy: $(\phi_1, \phi_0) = (1, 0)$ (truthful coordination)

A.1 Accuracy of First-Order Taylor Approximation

Proposition 5.1 establishes that the spread change is $\Delta S(\delta) = \varepsilon \cdot \Xi + O(\varepsilon^2)$, where $\varepsilon = (1-\rho)(1-\delta)$. Figure 1 compares the exact spread expansion with the first-order Taylor approximation across a range of detection probabilities $\delta \in [0, 1]$ for three values of the honest fraction $\rho \in \{0.7, 0.8, 0.9\}$.

Figure 1: Accuracy of Approximation ($p_0 = 0.3$)

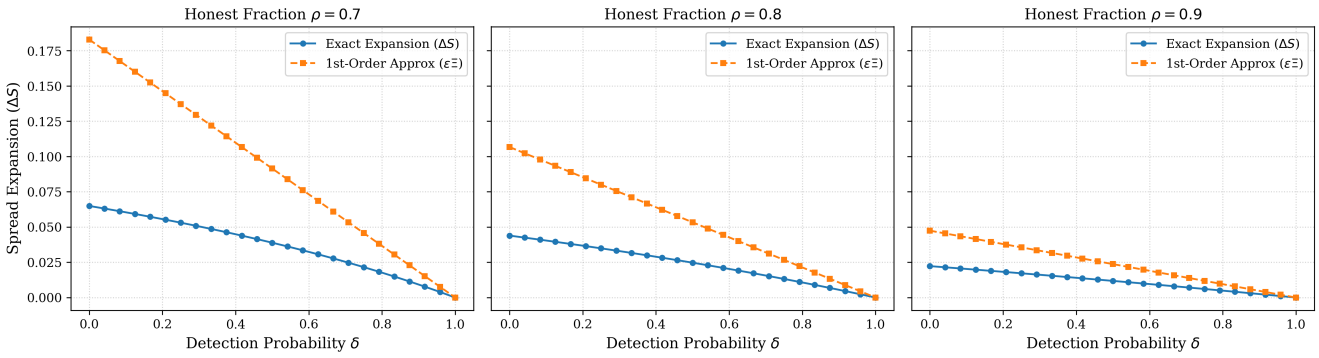

Figure 1. Spread expansion (ΔS) under different detection probabilities δ for $\rho \in \{0.7, 0.8, 0.9\}$ and prior $p_0 = 0.3$. Dots (blue) plot the *Exact Spread Expansion*, defined as the spread increase relative to the honest benchmark ($S_{\text{exact}}(\delta) - S_{\text{exact}}(1.0)$). Squares (orange) plot the *First-Order Approximation* $\varepsilon\Xi$.

Results. Figure 1 compares the *Exact Expansion* (blue) against the *1st-order Approximation* (orange, $\varepsilon\Xi$). The plot validates the qualitative prediction of Proposition 5.1: spread expansion is strictly decreasing in detection probability δ. Quantitatively, however, the approximation systematically exceeds the exact expansion. This discrepancy arises from differing baselines: $\varepsilon\Xi$ approximates the impact of manipulation added to a "no-content" market, whereas

[3]The full Python code used to generate all figures and numerical results in this appendix is available as part of the replication package via https://github.com/ChenghaoLi1022/synchronization-manipulation.git

the exact expansion measures the marginal distortion relative to a market that *already contains honest signals* ($S_{\text{exact}}(1.0)$). Since honest content itself consumes liquidity (widening the baseline spread), the marginal impact of adding manipulation is structurally dampened. For example, in the $\rho = 0.7$ panel at $\delta = 0$, the approximation predicts $\Delta S \approx 0.182$, while the exact additional expansion is only ≈ 0.065. This confirms that baseline interactions ($O(\varepsilon^2)$) play a significant role in dampening the total liquidity cost.

A.2 Sensitivity of Deterrence Threshold $\delta^\star$

Proposition 5.2 establishes the *definition* of the dynamic threshold $\delta^\star$. When $\mathbb{E}[T_c(\delta)]$ is affine), this reduces to $\delta^\star = (\Pi - k)/(\Pi + F)$ with $\Pi = R\bar{T}$. Figure 2 illustrates how $\delta^\star$ varies with the manipulation profit Π, coordination cost k, and detection penalty F.

Figure 2: Sensitivity of Deterrence Threshold δ^*

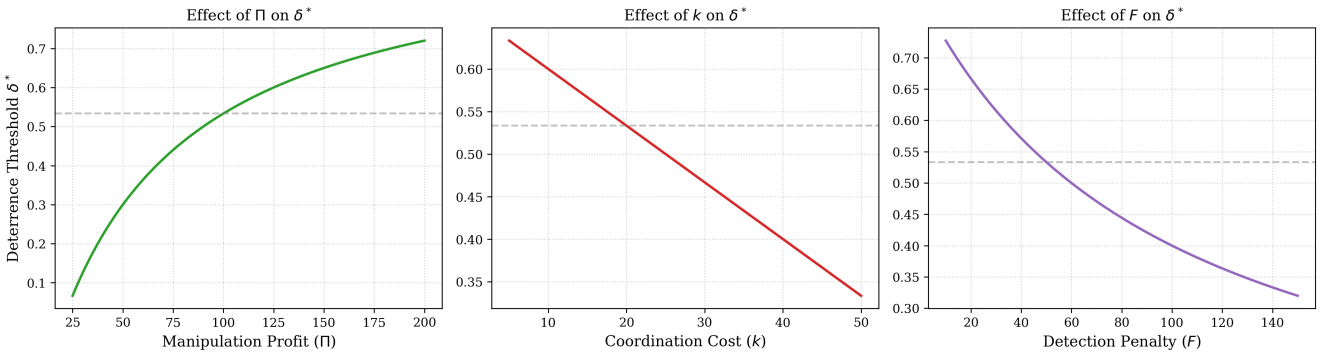

Figure 2. Sensitivity of deterrence threshold $\delta^\star$ to model primitives. Left: $\delta^\star$ increases with manipulation profit Π (higher profitability requires stronger detection). Center: $\delta^\star$ decreases with coordination cost k (higher barriers to entry reduce the threshold). Right: $\delta^\star$ decreases with penalty F (harsher sanctions deter manipulation).

Results. With baseline parameters $\Pi = 100$, $k = 20$, $F = 50$, we obtain $\delta^\star = 0.533$. The comparative statics are:

- $\partial\delta^\star/\partial\Pi > 0$: Doubling Π from 100 to 200 increases $\delta^\star$ from 0.53 to 0.72. More profitable manipulation requires stronger detection.
- $\partial\delta^\star/\partial k < 0$: Doubling k from 20 to 40 decreases $\delta^\star$ from 0.53 to 0.40. Higher fixed costs make manipulation easier to deter.
- $\partial\delta^\star/\partial F < 0$: Doubling F from 50 to 100 decreases $\delta^\star$ from 0.53 to 0.40. Harsher penalties reduce the threshold for detection needed to deter manipulation.

These results confirm that policy interventions targeting k (e.g., disrupting coordination infrastructure) or F (e.g., legal sanctions) can reduce the detection threshold $\delta^\star$, relaxing the technical requirements on platform moderation systems.

A.3 Information Aggregation and Prior Skewness

Proposition 5.3 establishes conditions for information aggregation. We now numerically verify the first-order spread approximation from Proposition 5.1 and investigate its dependence on the prior p_0.

Figure 3: Robustness to Prior Skewness ($\rho = 0.5$)

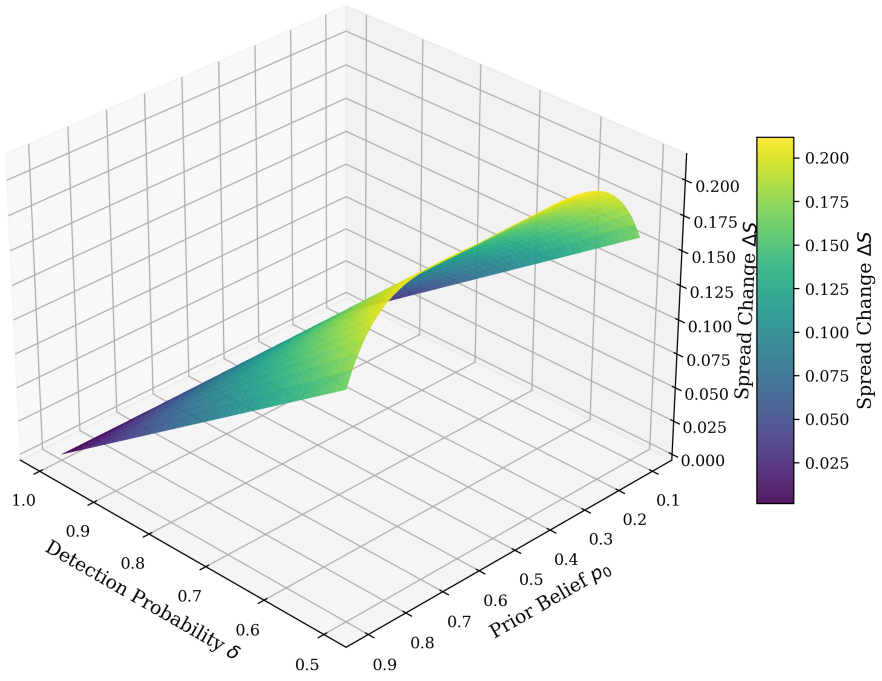

Figure 3. 3D Surface of Spread Expansion (ΔS). The vertical axis (color gradient) represents the spread expansion $\Delta S \approx \varepsilon\Xi$. The horizontal axes represent the detection probability $\delta \in [0.5, 1.0]$ and the market prior $p_0 \in [0.1, 0.9]$.

Results. Figure 3 confirms our theoretical predictions from Proposition 5.1 through a continuous parameter sweep:

1. *Sensitivity to Detection (Slope)*: As detection δ decreases (moving from 1.0 to 0.5), the spread expansion rises sharply. This confirms that liquidity distortion is first-order in the undetected manipulation mass ε, vanishing only as $\delta \to 1$.

2. *Invariance to Prior Skewness (Flatness)*: A striking feature of the 3D surface is its geometric "flatness" along the Prior Belief axis. Whether the market is symmetric ($p_0 \approx 0.5$) or highly skewed ($p_0 \approx 0.1$ or 0.9), the magnitude of the spread shock remains virtually constant. This confirms that the total curvature effect $\Xi = \kappa[h(p_0^+) + h(p_0^-)]$ is invariant to the prior p_0, as the curvatures on the bid and ask sides offset each other.

Methodological Note. This 3D visualization demonstrates that the liquidity damage from coordination manipulation is *orthogonal* to the market's prior belief. The key insight is that this distortion is not limited to "skewed" or "long-shot" markets but is a robust phenomenon that affects markets regardless of their prior odds, justifying a uniform regulatory approach.

B Proof of Proposition 5.1 (First-Order Spread Change)

Proof. We provide a rigorous proof for the binary-message case. The LAN condition (Assumption E.6) is verified in Remark E.3 for our setting, ensuring that the Taylor expansion and remainder bounds below are well-justified.

Notation. Let $\varepsilon := (1-\rho)(1-\delta)$ denote the undetected mass of coordinated content. For unflagged messages, the effective likelihood ratio under the source mixture is

$$\Lambda_x(\varepsilon;\phi_1,\phi_0) = \frac{\rho q + \varepsilon\,\phi_1}{\rho(1-q) + \varepsilon\,\phi_0},$$

and write $\Lambda_H := q/(1-q)$ and

$$\kappa \;:=\; \frac{\phi_1}{\rho q} \;-\; \frac{\phi_0}{\rho(1-q)}.$$

Let $\Lambda_y^{\pm} = \frac{1\pm\mu}{1\mp\mu}$ be the order–flow LRs and $O_0 = \frac{p_0}{1-p_0}$ the prior odds; define baseline posteriors $p_0^{\pm} = \frac{O_0\Lambda_y^{\pm}}{1+O_0\Lambda_y^{\pm}}$.

Assumptions. We maintain Assumptions E.5 (Moment Bounds), E.6 (LAN), E.7 (Smoothness), and the MLRP. In particular, we assume (i) $\varepsilon > 0$ is small; (ii) log–likelihood ratios have bounded second moments so that Taylor expansion and dominated convergence apply (a standard LAN-style regularity); (iii) in the benchmark full-coordination equilibrium, the manipulator sets $(\phi_1,\phi_0) = (1,0)$, which implies $\kappa = \frac{1}{\rho q} > 0$. These assumptions match those used throughout the main text.

Step 1: Taylor Expansion of the Log Effective LR.

$$\begin{aligned}
\log\Lambda_x(\varepsilon) &= \log(\rho q + \varepsilon\phi_1) - \log\big(\rho(1-q) + \varepsilon\phi_0\big) \\
&= \log\Lambda_H \;+\; \varepsilon\left(\frac{\phi_1}{\rho q} - \frac{\phi_0}{\rho(1-q)}\right) \;+\; O(\varepsilon^2) \\
&= \log\Lambda_H \;+\; \varepsilon\,\kappa \;+\; O(\varepsilon^2).
\end{aligned}$$

The $O(\varepsilon^2)$ remainder follows from a second–derivative bound on $\log(\rho q + \varepsilon\phi_1)$ and $\log(\rho(1-q)+\varepsilon\phi_0)$ under Assumption E.6 (LAN) and Assumption E.7(Smoothness). By Taylor's theorem with integral remainder,

$$|O(\varepsilon^2)| \le \frac{\varepsilon^2}{2}\sup_{x,\varepsilon'}|\partial^2\ell(x;\varepsilon')/\partial\varepsilon^2| =: C\varepsilon^2,$$

where $C < \infty$ by the boundedness assumption. By Assumption E.5 and the dominated convergence theorem, the remainder bound holds in expectation.

Step 2: Propagation to Posteriors. We must track the Ask and Bid prices separately, as they respond to $x = 1$ and $x = 0$ respectively.

(i) *Ask Price (Buy order, $y = +1$, Positive content, $x = 1$):* Posterior odds are $O^+(\varepsilon) = O_0\Lambda_y^+\Lambda_x(x=1,\varepsilon)$. The log-LR expansion for $x = 1$ (from Step 1) is: $\log\Lambda_x(x=1,\varepsilon) = \log\Lambda_H + \varepsilon\kappa + O(\varepsilon^2)$. Let $L^+(\varepsilon) = \log O^+(\varepsilon)$.

The first-order change is $\frac{\partial L^+}{\partial \varepsilon}\big|_{\varepsilon=0} = \kappa$. The Ask price $p^+(\varepsilon) = O^+/(1+O^+)$ expands as:

$$\left.\frac{\partial p^+}{\partial \varepsilon}\right|_{\varepsilon=0} = \left.\frac{\partial p}{\partial \log O}\right|_{p_0^+} \cdot \left.\frac{\partial L^+}{\partial \varepsilon}\right|_{\varepsilon=0} = \left[p_0^+(1-p_0^+)\right] \cdot \kappa \tag{B.1}$$

(ii) *Bid Price (Sell order, $y=-1$, Negative content, $x=0$):* Posterior odds are $O^-(\varepsilon) = O_0 \Lambda_y^- \Lambda_x(x=0,\varepsilon)$. For the truthful manipulator strategy ($\phi_1 = 1, \phi_0 = 0$), the LRs are inverse: $\Lambda_x(x=0) = \Lambda_x(x=1)^{-1}$. Therefore, the log-LR expansion for $x=0$ is: $\log \Lambda_x(x=0,\varepsilon) = -\log \Lambda_x(x=1,\varepsilon) \approx -(\log \Lambda_H + \varepsilon\kappa)$. Let $L^-(\varepsilon) = \log O^-(\varepsilon)$. The first-order change is $\frac{\partial L^-}{\partial \varepsilon}\big|_{\varepsilon=0} = -\kappa$. The Bid price $p^-(\varepsilon) = O^-/(1+O^-)$ expands as:

$$\left.\frac{\partial p^-}{\partial \varepsilon}\right|_{\varepsilon=0} = \left.\frac{\partial p}{\partial \log O}\right|_{p_0^-} \cdot \left.\frac{\partial L^-}{\partial \varepsilon}\right|_{\varepsilon=0} = \left[p_0^-(1-p_0^-)\right] \cdot (-\kappa) \tag{B.2}$$

Conclusion (First-Order Spread Effect). Let $S(\varepsilon) := \text{Ask}^+(\varepsilon) - \text{Bid}^-(\varepsilon) = p^+(\varepsilon) - p^-(\varepsilon)$. The first-order change in the spread is:

$$\begin{aligned}
\Delta S(\varepsilon) := S(\varepsilon) - S(0) &= \varepsilon \cdot \left.\frac{\partial S}{\partial \varepsilon}\right|_{\varepsilon=0} + O(\varepsilon^2) \\
&= \varepsilon \cdot \left.\left(\frac{\partial p^+}{\partial \varepsilon} - \frac{\partial p^-}{\partial \varepsilon}\right)\right|_{\varepsilon=0} + O(\varepsilon^2) \\
&= \varepsilon \cdot \Big(\left[p_0^+(1-p_0^+)\right] \cdot \kappa - \left[p_0^-(1-p_0^-)\right] \cdot (-\kappa)\Big) + O(\varepsilon^2) \\
&= \varepsilon \cdot \kappa \underbrace{\left[p_0^+(1-p_0^+) + p_0^-(1-p_0^-)\right]}_{=\Xi/\kappa} + O(\varepsilon^2)
\end{aligned}$$

This proves Proposition 5.1, with the first-order coefficient $\Xi = \kappa\left[p_0^+(1-p_0^+) + p_0^-(1-p_0^-)\right]$. □

Special Case and Comparative Statics. Under symmetric priors $p_0 = \frac{1}{2}$, we have $p_0^+ = 1 - p_0^-$, so $p_0^+(1-p_0^+) = p_0^-(1-p_0^-) > 0$. The first-order term does not vanish:

$$\Xi(p_0 = 0.5) = \kappa \cdot \left[2 \cdot p_0^+(1-p_0^+)\right] > 0.$$

The first-order effect is strictly positive for $\forall p_0 \in (0,1)$ (assuming $\kappa > 0$). Since $\varepsilon = (1-\rho)(1-\delta)$ decreases in δ, we obtain $\frac{\partial}{\partial \delta}\Delta S(\delta) \approx -(1-\rho)\Xi < 0$. Higher detection *always* reduces the spread at first order. Moreover, $|\Xi|$ is increasing in $|\mu|$ (as this pushes $p_0^\pm$ further from the prior).

C Proof of Proposition 5.2 (Dynamic Deterrence Threshold)

Proof. Consider the manipulator's optimization problem where the agent chooses an activity level $\phi \in [0, 1]$. If the manipulator adopts an active strategy ($\phi = 1$), the expected profit is bounded by the total expected rent minus the total cost (setup k plus expected penalty δF):

$$\mathbb{E}[\Pi(\delta)] \;\le\; g(\delta) := R \cdot \mathbb{E}[T_c(\delta)] \;-\; (k + \delta F). \tag{C.1}$$

While the penalty δF is a variable cost contingent on action, we determine the deterrence threshold by comparing the maximal possible profit (under optimal active play) against zero; thus, the condition for profitability is $g(\delta) > 0$.

We first establish the monotonicity of the potential profit function. Differentiating $g(\delta)$ with respect to δ yields $g'(\delta) = R \cdot \frac{d}{d\delta}\mathbb{E}[T_c(\delta)] - F$. By Assumption 3.2, $\mathbb{E}[T_c(\delta)]$ is strictly decreasing, implying $\frac{d}{d\delta}\mathbb{E}[T_c(\delta)] < 0$. Given $R, F > 0$, it follows that $g'(\delta) < 0$ for all δ, confirming that potential profit is strictly decreasing in detection intensity.

Next, we establish the existence and uniqueness of $\delta^\star$. Since $\mathbb{E}[T_c(\delta)]$ is continuous (Assumption 3.2), $g(\delta)$ is continuous. We identify $\delta^\star$ based on the boundary values of $g(\delta)$. If $g(0) \le 0$, expected profit is non-positive for all $\delta \ge 0$, implying manipulation is never profitable (blockaded entry); we set $\delta^\star = 0$. Conversely, if $g(1) \ge 0$, profit remains positive even at maximal detection (unstoppable; detection insufficient); we set $\delta^\star = 1$. In the intermediate case where $g(0) > 0$ and $g(1) < 0$, the Intermediate Value Theorem, combined with the strict monotonicity of g, guarantees the existence of a unique root $\delta^\star \in (0, 1)$ such that $g(\delta^\star) = 0$. This rigorously defines the threshold as in (5.4).

Finally, regarding comparative statics for the interior solution, implicit differentiation of the condition $g(\delta^\star) = 0$ yields

$$\frac{\partial \delta^\star}{\partial k} = -\frac{-1}{g'(\delta^\star)} = \frac{1}{g'(\delta^\star)} < 0, \quad \frac{\partial \delta^\star}{\partial F} = \frac{\delta^\star}{g'(\delta^\star)} < 0,$$

since $g'(\delta) < 0$. By similar logic, increasing R or shifting the expected duration $\mathbb{E}[T_c(\cdot)]$ upward increases $g(\delta)$, thereby requiring a larger $\delta^\star$ to restore the zero-profit condition. □

D Proof of Proposition 5.3 (Information Aggregation)

This appendix establishes Proposition 5.3. We characterize the marginal effect of detection intensity δ on information aggregation and provide sufficient conditions for detection to improve informational efficiency.

Notation Consistency. In Proposition 5.3, we employ the reduced-form notation Δ_H and CD to capture the economic intuition. In this appendix, we use the formal notation $\mathcal{J}_H$ and $\mathcal{D}$ to emphasize their derivation from mutual information theory. These quantities correspond as follows:

$$\Delta_H \equiv \mathcal{J}_H \quad \text{(honest informational gain)}, \qquad |CD| \equiv \mathcal{D} \quad \text{(coordination distortion)}.$$

Notation and Setup. Let $\varepsilon := (1-\rho)(1-\delta)$ denote the fraction of undetected coordinated releases among unflagged messages. Write the honest (baseline) message likelihood ratio as $\Lambda_H(x) = \frac{f_H(x|\theta=1)}{f_H(x|\theta=0)}$, and the effective unflagged LR under mixture as

$$\Lambda_x(\varepsilon) \;=\; \frac{\rho\, q + \varepsilon\, \phi_1(x)}{\rho\,(1-q) + \varepsilon\, \phi_0(x)}, \qquad \kappa(x) \;:=\; \frac{\phi_1(x)}{\rho q} - \frac{\phi_0(x)}{\rho(1-q)}.$$

Order-flow LRs are denoted $\Lambda_y^{\pm} = \frac{1\pm\mu}{1\mp\mu}$ and the prior odds $O_0 = \frac{p_0}{1-p_0}$. All posteriors and prices in this appendix are conditional on unflagged content.

Regularity Conditions. We maintain the standing assumptions from Section E.3:

(A1) (Small undetected mass) $\varepsilon \in [0, \bar{\varepsilon}]$ with $\bar{\varepsilon}$ sufficiently small.

(A2) (Moment bounds) Under Assumption E.5, log-likelihood ratios $\log \Lambda_H(x)$, $\log \Lambda_x(\varepsilon)$ have bounded second moments under both $\theta \in \{0, 1\}$.

(A3) (LAN condition) Under Assumption E.6, the map $\varepsilon \mapsto \log \Lambda_x(\varepsilon)$ is twice continuously differentiable in a neighborhood of 0 with uniformly integrable second derivative.

(A4) (Conditional independence) Honest messages and order flow are conditionally independent given θ: $f_H(x, y|\theta) = f_H(x|\theta)P(y|\theta)$.

Lemma D.1 (Log-LR Expansion for Unflagged Content)**.** *Under (A1)–(A3),*

$$\log \Lambda_x(\varepsilon) = \log \Lambda_H(x) \;+\; \varepsilon\, \kappa(x) \;+\; O(\varepsilon^2) \quad \textit{in } L^1,$$

hence for posterior odds $O^{\pm}(\varepsilon) = O_0\, \Lambda_y^{\pm}\, \Lambda_x(\varepsilon)$,

$$\log O^{\pm}(\varepsilon) = \log\big(O_0\, \Lambda_y^{\pm}\, \Lambda_H(x)\big) \;+\; \varepsilon\, \kappa(x) \;+\; O(\varepsilon^2).$$

Proof. Apply a second-order Taylor expansion:

$$\begin{aligned}\log \Lambda_x(\varepsilon) &= \log\left(\rho q + \varepsilon\phi_1(x)\right) - \log\left(\rho(1-q) + \varepsilon\phi_0(x)\right) \\ &= \log(\rho q) + \log\left(1 + \frac{\varepsilon\phi_1(x)}{\rho q}\right) - \log\left(\rho(1-q)\right) - \log\left(1 + \frac{\varepsilon\phi_0(x)}{\rho(1-q)}\right)\end{aligned}$$

$$= \log \Lambda_H(x) + \frac{\varepsilon\phi_1(x)}{\rho q} - \frac{\varepsilon\phi_0(x)}{\rho(1-q)} + O(\varepsilon^2)$$
$$= \log \Lambda_H(x) + \varepsilon\kappa(x) + O(\varepsilon^2),$$

where the $O(\varepsilon^2)$ bound follows from $\log(1+z) = z - z^2/2 + O(z^3)$ and (A2). By dominated convergence (justified by (A2)), the bound holds in L^1. □

Lemma D.2 (Propagation to Posteriors)**.** *Let $p^{\pm}(\varepsilon)$ denote the posteriors after observing $(y = \pm 1, x, \ell = \text{unflag})$. Then*

$$p^{\pm}(\varepsilon) = p_0^{\pm}(x) \;+\; \varepsilon\,\kappa(x)\,p_0^{\pm}(x)\big(1 - p_0^{\pm}(x)\big) \;+\; O(\varepsilon^2),$$

where $p_0^{\pm}(x) = \frac{O_0\Lambda_y^{\pm}\Lambda_H(x)}{1+O_0\Lambda_y^{\pm}\Lambda_H(x)}$ are the baseline posteriors at $\varepsilon = 0$.

Proof. Since $p = \frac{O}{1+O}$ and $\frac{\partial p}{\partial \log O} = p(1-p)$, the result follows from Lemma D.1 by the chain rule. □

We now turn to the information-theoretic analysis. The key step is to decompose the derivative of mutual information into two economically meaningful components.

Lemma D.3 (Decomposition of Information Derivative)**.** *Under (A1)–(A4), the marginal effect of undetected manipulation mass on conditional mutual information is*

$$\frac{d}{d\varepsilon} I(\theta; X|Y, \ell = \text{unflag})\Big|_{\varepsilon=0} = \mathcal{J}_H - \mathcal{D}, \tag{D.1}$$

where the honest informational gain is

$$\mathcal{J}_H := \mathbb{E}_{Y,\theta}\left[\int f_M(x|y,\theta) \log \frac{f_M(x|\theta)}{f_H(x|\theta)} dx\right] \geq 0, \tag{D.2}$$

and the coordination distortion is

$$\mathcal{D} := \mathbb{E}_\theta\left[\text{Cov}\left(\log \Lambda_x(X), \log \Lambda_y(Y) \mid \theta\right)\right] \geq 0. \tag{D.3}$$

Proof. By the chain rule for mutual information (Cover & Thomas, 2006, Theorem 2.5.2),

$$I(\theta; X, Y|\ell = \text{unflag}) = I(\theta; Y|\ell = \text{unflag}) + I(\theta; X|Y, \ell = \text{unflag}). \tag{D.4}$$

We analyze each term separately. Under the mixture density

$$f^{\varepsilon}(y|\theta) = (1-\varepsilon)P(y|\theta) + \varepsilon \int f_M(x, y|\theta) dx,$$

if the manipulator's trading distribution matches the informed trader distribution, i.e.,

$$\int f_M(x, y|\theta) dx = P(y|\theta) \quad \forall y, \theta,$$

then $\frac{d}{d\varepsilon} I(\theta; Y|\ell = \text{unflag})\big|_{\varepsilon=0} = 0$. This holds in our model since the manipulator submits orders that pool with informed traders. Thus, the derivative is entirely determined by the second term. Write the conditional mutual

information as

$$I(\theta; X|Y, \ell = \text{unflag}) = \mathbb{E}_Y \left[\mathbb{E}_\theta \left[\log \frac{f^\varepsilon(X|\theta, Y)}{f^\varepsilon(X|Y)} \,\middle|\, Y \right] \right].$$

Under the mixture,

$$\begin{aligned} f^\varepsilon(x|\theta, y) &= \frac{(1-\varepsilon) f_H(x|\theta) P(y|\theta) + \varepsilon f_M(x, y|\theta)}{(1-\varepsilon) P(y|\theta) + \varepsilon P(y|\theta)} \\ &= \frac{(1-\varepsilon) f_H(x|\theta) P(y|\theta) + \varepsilon f_M(x, y|\theta)}{P(y|\theta)} \\ &= (1-\varepsilon) f_H(x|\theta) + \varepsilon \frac{f_M(x, y|\theta)}{P(y|\theta)}. \end{aligned}$$

Taking the derivative at $\varepsilon = 0$:

$$\frac{\partial f^\varepsilon(x|\theta, y)}{\partial \varepsilon}\Big|_{\varepsilon=0} = \frac{f_M(x, y|\theta)}{P(y|\theta)} - f_H(x|\theta) =: \Delta f(x|y, \theta).$$

Using the envelope theorem for entropy, the derivative of conditional entropy is

$$\frac{d}{d\varepsilon} H(X|Y = y, \theta)\Big|_{\varepsilon=0} = -\int \Delta f(x|y, \theta) \log f_H(x|\theta) dx. \tag{D.5}$$

Similarly, for the marginal entropy $H(X|Y = y)$, using $f^\varepsilon(x|y) = \sum_\theta p(\theta) f^\varepsilon(x|\theta, y)$:

$$\frac{d}{d\varepsilon} H(X|Y = y)\Big|_{\varepsilon=0} = -\sum_\theta p(\theta) \int \Delta f(x|y, \theta) \log f_H(x|y) dx. \tag{D.6}$$

Combining the derivatives and noting that $I(\theta; X|Y) = H(X|Y) - H(X|\theta, Y)$:

$$\begin{aligned} \frac{dI}{d\varepsilon}\Big|_{\varepsilon=0} &= \sum_{y,\theta} p(\theta) P(y) \int \Delta f(x|y, \theta) \left[\log f_H(x|\theta) - \log f_H(x|y) \right] dx \\ &= \sum_{y,\theta} p(\theta) P(y) \int \left[\frac{f_M(x, y|\theta)}{P(y|\theta)} - f_H(x|\theta) \right] \log \frac{f_H(x|\theta)}{f_H(x|y)} dx. \end{aligned} \tag{D.7}$$

Now decompose $\log f_H(x|y)$ using Bayes' rule. Since $X \perp Y|\theta$ under honesty (A4),

$$f_H(x|y) = \frac{f_H(x|\theta) P(y|\theta)}{\sum_{\theta'} p(\theta') f_H(x|\theta') P(y|\theta')} \cdot \sum_{\theta''} p(\theta'').$$

For binary $\theta \in \{0, 1\}$, this gives

$$\log \frac{f_H(x|\theta)}{f_H(x|y)} = \log \Lambda_x(x) - \log \left[p_0 + (1-p_0) \Lambda_x(x)^{-1} \Lambda_y(y)^{-1} \right] + \text{const.}$$

Substituting into (D.7) and separating terms that depend only on x vs. those involving correlation between x and y:

$$\frac{dI}{d\varepsilon}\Big|_{\varepsilon=0} = \underbrace{\mathbb{E}\left[\frac{f_M(X, Y|\theta)}{P(Y|\theta)} \log \Lambda_x(X) \right] - \mathbb{E}_{f_H} \left[\log \Lambda_x(X) \right]}_{\mathcal{J}_H}$$

$$\underbrace{-\mathbb{E}_\theta\left[\mathrm{Cov}\left(\log\Lambda_x(X), \log\Lambda_y(Y) \mid \theta\right)\right]}_{\mathcal{D}}. \tag{D.8}$$

The first term $\mathcal{J}_H$ measures whether the manipulator's content distribution is more informative about θ than the honest baseline (in the sense of expected log-likelihood ratio). The second term $\mathcal{D}$ captures the spurious correlation: under honesty, $X \perp Y|\theta$ implies $\mathrm{Cov}(\log\Lambda_x, \log\Lambda_y|\theta) = 0$; under coordination, the manipulator releases $x = 1$ when trading $y = +1$, inducing positive covariance. □

Remark (Interpretation). The decomposition clarifies the trade-off:

(i) If the manipulator provides high-quality signals ($\mathcal{J}_H$ large), even coordinated releases can improve information.

(ii) If the manipulator merely synchronizes truthful but noisy signals ($\mathcal{J}_H \approx 0$), the distortion $\mathcal{D}$ dominates, and detection improves efficiency.

(iii) The condition $\mathcal{J}_H > \mathcal{D}$ characterizes when allowing undetected coordination is informationally beneficial.

Lemma D.4 (Total Variation Bound)**.** *Under (A1)–(A4), there exist constants $C, C' > 0$ such that*

$$\mathcal{D} \;\le\; C \cdot \mathrm{TV}\Big(f^{honest}_{X,Y|\theta}, f^{eff}_{X,Y|\theta}\Big) \;=\; C'\,\varepsilon + O(\varepsilon^2).$$

Proof. Define bounded random variables $Z_x := \log\Lambda_x(X)$ and $Z_y := \log\Lambda_y(Y)$. By (A2), $|Z_x| \le M_x$ and $|Z_y| \le M_y$ almost surely for some constants $M_x, M_y < \infty$. Then

$$|\mathcal{D}| = \left|\mathbb{E}_{f^\varepsilon}[Z_x Z_y|\theta] - \mathbb{E}_{f_H}[Z_x]\mathbb{E}_{f_H}[Z_y]\right| \le 2M_x M_y \cdot \mathrm{TV}(f^\varepsilon, f_H),$$

by the dual characterization of total variation. By Lemma D.1 and Pinsker's inequality,

$$\mathrm{TV}(f^\varepsilon, f_H) \le \sqrt{\frac{1}{2}\mathrm{KL}(f^\varepsilon \| f_H)} = O(\varepsilon).$$

□

Proof of Proposition 5.3. From Lemma D.3, the marginal effect is

$$\frac{d}{d\varepsilon} I(\theta; X, Y|\ell = \text{unflag})\Big|_{\varepsilon=0} = \mathcal{J}_H - \mathcal{D}.$$

Since $\varepsilon = (1-\rho)(1-\delta)$, we have $\frac{d\varepsilon}{d\delta} = -(1-\rho)$. Thus,

$$\frac{d}{d\delta} I(\theta; X, Y|\ell = \text{unflag}) = \frac{dI}{d\varepsilon}\frac{d\varepsilon}{d\delta} = -(1-\rho)(\mathcal{J}_H - \mathcal{D}) = (1-\rho)(\mathcal{D} - \mathcal{J}_H).$$

For marginal increases in detection to improve information, we require $\frac{dI}{d\delta} > 0$, i.e., $\mathcal{D} > \mathcal{J}_H$. For the sufficient condition in equation (5.7) of the main text, note that integrating the derivative over $\delta \in [0, \bar{\delta}]$ gives the level comparison. Under the approximation that $\mathcal{J}_H$ and $\mathcal{D}$ remain approximately constant (valid for small ε), we obtain

$$I(\theta; X, Y|\delta) - I(\theta; X, Y|\delta = 0) \approx (1-\rho)\bar{\delta} \cdot (\mathcal{D} - \mathcal{J}_H).$$

Rearranging and using $\mathcal{D} = C'\varepsilon + O(\varepsilon^2)$ from Lemma D.4 yields the condition

$$\rho\Delta_H > (1-\rho)(1-\delta)|CD|,$$

where Δ_H is the honest signal gain and $|CD|$ is a normalized measure of distortion. This provides a sufficient (though not necessary) condition for detection to improve the level of information relative to the no-detection baseline. □

Remark (On Necessity). The condition $\mathcal{D} > \mathcal{J}_H$ is both necessary and sufficient for marginal improvement. The level comparison condition (5.7) is only sufficient because it relies on the approximation that derivatives are constant. A full Blackwell comparison would require additional structure on the joint distribution (X, Y) and is left for future work.

Definition (Coordination Curvature, Conditional Version). Let $\mathfrak{l}_x := \log \Lambda_x$ and $\mathfrak{l}_y := \log \Lambda_y$. For any joint density/pmf $g(\cdot \mid \theta, Y, \ell = \text{unflag})$ on (X, Y), define the conditional coordination curvature

$$\mathrm{CD}(g) := \mathrm{Cov}_g\big(\mathfrak{l}_x, \mathfrak{l}_y \,\big|\, \theta, Y, \ell = \text{unflag}\big) = \mathbb{E}_g\big[(\mathfrak{l}_x - m_x)(\mathfrak{l}_y - m_y) \,\big|\, \theta, Y, \ell = \text{unflag}\big],$$

where $m_x := \mathbb{E}_g[\mathfrak{l}_x \mid \theta, Y, \ell = \text{unflag}]$ and $m_y := \mathbb{E}_g[\mathfrak{l}_y \mid \theta, Y, \ell = \text{unflag}]$. Under the honest benchmark $f_{\text{hon}}(x, y \mid \theta, Y, \ell = \text{unflag}) = f_x(x \mid \theta, Y, \ell = \text{unflag})\, f_y(y \mid \theta, Y, \ell = \text{unflag})$, we have $\mathrm{CD}(f_{\text{hon}}) = 0$.

Assumption D.1 (Bounded Log-Likelihood Ratios)**.** $\exists M < \infty$ such that $|\mathfrak{l}_x| \le M$ and $|\mathfrak{l}_y| \le M$ almost surely under $f_{\text{hon}}(\cdot \mid \theta, Y, \ell = \text{unflag})$ and $f_{\text{eff}}(\cdot \mid \theta, Y, \ell = \text{unflag})$.

Lemma D.5 (TV → Covariance Bridge (Conditional))**.** *Under Assumption D.1,*

$$\big|\mathrm{CD}(f_{\text{eff}})\big| \le 8M^2\, \mathrm{TV}\Big(f_{\text{eff}}(\cdot \mid \theta, Y, \ell = \text{unflag}), f_{\text{hon}}(\cdot \mid \theta, Y, \ell = \text{unflag})\Big).$$

Proof. Let $\psi := (\mathfrak{l}_x - m_x)(\mathfrak{l}_y - m_y)$. Then, $|\psi| \le (|\mathfrak{l}_x| + |m_x|)(|\mathfrak{l}_y| + |m_y|) \le (2M)(2M) = 4M^2$. Under honesty, independence implies $\mathbb{E}_{f_{\text{hon}}}[\,\psi \mid \theta, Y, \ell = \text{unflag}\,] = 0$. By the dual characterization of total variation,

$$\begin{aligned}
&\Big|\mathbb{E}_{f_{\text{eff}}}[\psi \mid \theta, Y, \ell = \text{unflag}] - \mathbb{E}_{f_{\text{hon}}}[\psi \mid \theta, Y, \ell = \text{unflag}]\Big| \\
&\qquad \le 2\,\|\psi\|_\infty\, \mathrm{TV}\Big(f_{\text{eff}}(\cdot \mid \theta, Y, \ell = \text{unflag}), f_{\text{hon}}(\cdot \mid \theta, Y, \ell = \text{unflag})\Big) \\
&\qquad \le 8M^2\, \mathrm{TV}\Big(f_{\text{eff}}(\cdot \mid \theta, Y, \ell = \text{unflag}), f_{\text{hon}}(\cdot \mid \theta, Y, \ell = \text{unflag})\Big). \qquad \text{(D.9)}
\end{aligned}$$

□

From TV to (5.7). By the definition of the effective likelihood ratio $\Lambda_\alpha(\varepsilon)$ in Section E.1, the resulting distribution f_{eff} is a convex combination of f_{hon} and the manipulator's distribution with mixing weight proportional to ε. By the properties of total variation (or applying Pinsker's inequality twice), the distance $\mathrm{TV}(f_{\text{eff}}, f_{\text{hon}})$ is bounded by $K\varepsilon + O(\varepsilon^2)$. Combining Lemma D.5 with the bound established above,

$$\mathrm{TV}\Big(f_{\text{eff}}(\cdot \mid \theta, Y, \ell = \text{unflag}), f_{\text{hon}}(\cdot \mid \theta, Y, \ell = \text{unflag})\Big) \le K\,\varepsilon + O(\varepsilon^2),$$

and we obtain

$$\left|\mathrm{CD}(f_{\mathrm{eff}})\right| \le (8M^2K)\,\varepsilon + O(\varepsilon^2),$$

which justifies CD (coordination distortion) in (5.7) of the main text is well-behaved and bounded by $C_{CD} \cdot \varepsilon$.

Remark. If one prefers to avoid boundedness, the same conclusion obtains with a $\sqrt{\mathrm{TV}}$ bound under $\mathbb{E}_{f_{\mathrm{hon}}}[\psi^2] < \infty$ via Hellinger–TV.

Theorem C.5 (Restatement of Proposition 5.3). Let condition (5.7) in the main text hold, i.e., the honest informational gain dominates the coordination distortion at the margin:

$$\mathcal{J}_H \;\ge\; D.$$

Then, for ε sufficiently small,

$$\frac{d}{d\varepsilon}\, I(\theta; Y, X \mid \ell = \mathrm{unflag})\Big|_{\varepsilon=0} \;\ge\; 0,$$

with strict inequality when $\mathcal{J}_H > D$, so adding (unflagged) platform messages *improves* information aggregation at first-order.

Proof. Combine Lemmas D.1, D.2, and D.5, evaluate the derivative at $\varepsilon = 0$, and invoke condition (5.7). Uniform integrability under Assumptions E.5 and E.7 justifies exchanging derivative and expectation. □

Comparative Statics and Robustness. (i) If $p_0 = \frac{1}{2}$ (symmetric priors), first-order effects from Y alone cancel while X still contributes via $\mathcal{J}_H$. (ii) Allowing a small false-positive rate α in flagging only perturbs $\kappa(x)$ by $O(\alpha)$; the sign conclusions above are unchanged for small α.

E Technical Assumptions and Justifications

E.1 Robustness and Properties of the Likelihood Ratio

In Section 3.7, we defined the effective likelihood ratio $\Lambda_x(\delta)$ for unflagged content (see (3.1)). Here, we provide the technical conditions ensuring its validity and discuss its robustness to false positives.

Assumption E.1 (Common Support)**.** The honest and manipulator message densities $f_H(\cdot|\theta)$ and $f_M(\cdot|\theta)$ have common support for each $\theta \in \{0, 1\}$. That is, for each θ, there exists a measurable set $\mathcal{S}_\theta$ such that both $f_H(x|\theta) > 0$ and $f_M(x|\theta) > 0$ for $\forall x \in \mathcal{S}_\theta$, and both are zero outside $\mathcal{S}_\theta$. This ensures the effective likelihood ratio $\Lambda_x(\delta)$ is well-defined and finite.

Allowing False Positives. In the robustness extension with a small false positive rate $\alpha > 0$ for honest releases, the effective likelihood ratio becomes

$$\Lambda_x(\delta, \alpha) = \frac{\rho\, q\, (1-\alpha) + (1-\rho)(1-\delta)\, \phi_1}{\rho\, (1-q)\, (1-\alpha) + (1-\rho)(1-\delta)\, \phi_0}. \tag{E.1}$$

All first-order monotonicity and sign results stated below continue to hold for small α; unless otherwise noted, we work with the baseline $\alpha = 0$.

Monotonicity. We have

$$\frac{\partial \Lambda_x(\delta)}{\partial \delta} = -\frac{(1-\rho)\rho\, [\phi_1(1-q) - \phi_0 q]}{[\rho(1-q) + (1-\rho)(1-\delta)\phi_0]^2}.$$

Hence, $\partial \Lambda_x / \partial \delta \le 0$ if and only if $\phi_1(1-q) \ge \phi_0 q$ (equivalently, $\kappa \ge 0$). In particular, the simpler condition $\phi_1 \ge \phi_0$ is insufficient when $q > 1/2$.

Connection to Pricing. Let $\Lambda_y^+ = (1+\mu)/(1-\mu)$ and $\Lambda_y^- = (1-\mu)/(1+\mu)$ be the order–flow likelihood ratios in the Glosten–Milgrom environment. Upon observing a buy (sell) order and an unflagged message, posterior *odds* update multiplicatively:

$$O^{++}(x{=}1, \ell{=}\text{unflag}) = O_0\, \Lambda_y^+\, \Lambda_x(\delta), \tag{E.2}$$

$$O^{--}(x{=}0, \ell{=}\text{unflag}) = O_0\, \Lambda_y^-\, \Lambda_x(\delta)^{-1}, \tag{E.3}$$

yielding quotes $\text{Ask}^+ = O^{++}/(1+O^{++})$ and $\text{Bid}^- = O^{--}/(1+O^{--})$.

E.2 Discussion of Modeling Choices

To clarify the scope and robustness, we summarize the standing assumptions and discuss why each is necessary, what it implies, and how relaxing it would impact the results.

A1 (Sequential Awareness from Abreu & Brunnermeier (2003)). The realization time t_0 is random ($t_0 \sim \text{Exp}(\lambda)$); awareness diffuses over $[t_0, t_0 + \eta]$ so that each agent i becomes aware at $t_i \sim \text{Unif}[t_0, t_0 + \eta]$. Agents at t_i form posteriors $\Phi(t_0|t_i)$ as in (3.2).
Why: This creates higher–order uncertainty (agents know θ but not when others learned it), which is essential for

synchronization risk and for coordination manipulation to have bite.
Relaxation: If $\eta \to 0$ (near–instant common knowledge), the higher–order channel shuts down; Propositions on spread expansion (due to higher–order adverse selection) become weak or vanish.

A2 (Manipulator Knows θ and Can Synchronize Content and Trades). The manipulator observes θ shortly after t_0 and chooses $(\phi_1, \phi_0, t_m, n(\theta))$; the key instrument is a *batch release* at t_m that fabricates a false signal of common knowledge while submitting pooled orders.
Why: This isolates the coordination channel from first–order deception; our mechanism does not rely on content falsity.
Relaxation: Allowing misperception of θ (small error rate ε) leaves the qualitative results for Propositions 5.1–5.2 intact; it only perturbs $\Lambda_x(\delta)$ and Π.

A3 (Detection Targets Patterns, not Truth). The platform flags coordinated patterns with probability $\delta \in [0, 1]$, independent of (θ, x), conditional on the source; honest sources are unflagged in the baseline.
Why: This matches feasible real-time moderation and ensures that detection creates a Blackwell improvement (garbling order in δ). It also aligns the policy interpretation (EU AI Act (2024)) with *transparency about provenance or coordination* rather than truth verification.
Relaxation: A small false–positive rate α (honest flagged w.p. α) is straightforward to include; the effective likelihood ratio becomes

$$\Lambda_x(\delta, \alpha) = \frac{\rho(1-\alpha)q + (1-\rho)(1-\delta)\phi_1}{\rho(1-\alpha)(1-q) + (1-\rho)(1-\delta)\phi_0},$$

and the monotonicity properties still hold with $\beta = 1 - \frac{\delta_1 - \alpha}{\delta_2 - \alpha}$; all monotone comparative statics in δ are preserved for fixed α.

A4 (Cost Structure $C(\delta) = k + \delta F$). Here, $k > 0$ is the fixed cost of *coordination capability* (accounts, tooling), and $F > 0$ is the total loss if flagged (account bans, legal/reputational penalties). This yields, in the static/affine special case, the closed-form threshold

$$\delta^\star = \frac{\Pi - k}{\Pi + F}, \quad \text{with } \Pi = R\bar{T}.$$

Why: Distinguishes the coordination technology from content generation.
Relaxation: Adding per–message variable costs or convex detection costs rescales Π or F and leaves signs of the comparative statics unchanged.

A5 (Competitive Market Maker and One–Arrival Microstructure). Each period, one trader arrives, informed with probability μ or with noise with $1 - \mu$. The market maker is risk–neutral, competitive, and sets zero–profit prices conditional on (y, x, ℓ), producing the standard order–flow likelihood $\Lambda_y^{\pm}$.
Why: This provides a transparent bridge to Glosten–Milgrom and lets us isolate the incremental effect of higher–order manipulation.
Relaxation: With multiple arrivals per period or strategic dealers, the exact spread levels change, but the *direction* of the spread effect (Proposition 5.1) and the $\delta^\star$ logic (Proposition 5.2) are robust.

Remark (Multiple Arrivals). This assumption abstracts from strategic interactions among multiple traders who are simultaneously engaged in the market. While this simplifies the analysis, it overlooks potential higher-order adverse selection effects that arise when coordination influences the incentives of multiple informed traders. Extending to a multi-trader setting is a direction for future research.

A6 (Label Is a State-Independent Garbling; Limited Cross-Source Correlation). Conditional on the source type, the label ℓ depends on coordination patterns but not on (θ, x); honest releases are independent across sources and over time, while the manipulator's batch is perfectly synchronized at t_m.
Why: This delivers the closed–form effective likelihood $\Lambda_x(\delta)$ and the clean Blackwell order in δ.
Relaxation: Allowing weak cross–source correlation or copy–trading adds a covariance "distortion" term; the main signs are unchanged when the distortion is minor.

A7 (Scale and Pooling). The manipulator's order size is small enough to be pooled with informed or noise orders; the market maker cannot identify the manipulator solely based on order size.
Why: Ensures the higher–order channel, not trivial source identification, drives the results.
Relaxation: If manipulator size is large and reveals the source, coordination manipulation becomes self–defeating unless δ is extremely low; in that case, $\delta^\star$ is even smaller. The model does not require an oracle of truth; manipulation is purely higher-order and flags concern provenance, not veracity.

A8 (Institutional Specialization and Division of Labor). The competitive market maker specializes in contemporaneous information processing, observing only the current-period tuple (y_t, x_t, ℓ_t) and not conditioning on *intertemporal patterns* of order flow clustering. This reflects an institutional *division of labor*:

- *Market Maker's Role*: Real-time, high-frequency pricing based on contemporaneous signals (y_t, x_t, ℓ_t).
- *Platform's Role*: Centralized pattern detection across time, operating with probability δ and broadcasting flags ℓ_t.

Economic Rationale: This division is economically efficient since:

(i) *Specialization Gains*: Market makers excel at rapid, within-period inference, while platforms have the infrastructure for cross-temporal pattern analysis.

(ii) *Scale Economies*: Pattern detection requires centralized data access and computational resources that individual market makers lack.

(iii) *Regulatory Reality*: In practice, exchanges delegate surveillance to centralized systems (e.g., SEC's MIDAS, crypto bot detectors).

Interpretation: Rather than a cognitive limitation, this represents an institutional equilibrium where pattern detection is delegated to the platform's specialized system. The label ℓ_t serves as a sufficient statistic, economizing on the distributed information aggregation problem.
Relaxation: If market makers could costlessly observe historical patterns, they would incorporate them. However, in equilibrium, the platform's comparative advantage in pattern detection makes delegation the optimal choice.

Remark (Delegation of Detection to Platforms). In practice, financial exchanges and prediction markets are increasingly delegating pattern detection to centralized surveillance systems (e.g., the SEC's Market Information Data Analytics System, or crypto exchange bot detectors). Our model captures this institutional reality: individual

market makers set prices based on contemporaneous information, while the platform's detection system δ screens for coordination patterns and broadcasts flags ℓ_t to all participants.

Testable Implications. (i) Raising δ shifts left the distribution of the first public flag time T_F, shortens bubble windows, and reduces the spread jump on clustered content arrivals; (ii) the spread effect in Proposition 5.1 is stronger when priors are asymmetric or when μ is higher; (iii) policies that increase F or k reduce $\delta^\star$ and curb manipulation incentives even at moderate detection rates.

E.3 Regularity Conditions and Key Assumptions

Standing modeling assumptions and regularity assumptions below are maintained unless stated otherwise.

Assumption E.2 (Sequential Awareness)**.** Awareness times t_i are uniformly distributed over $[t_0, t_0 + \eta]$. For tractability, we approximate the discrete-time market with continuous awareness diffusion; the approximation error is $O(1/T)$, where T is the number of trading periods.

Assumption E.3 (Detection Technology)**.** The platform detection system may have a (small) false positive rate $\alpha \geq 0$.
Baseline: we set $\alpha = 0$, so honest independent releases are unflagged.
Extension (Robustness): we allow $\alpha > 0$ in Section E.1; coordinated batches trigger with probability $\delta \gg \alpha$.

Assumption E.4 (Cost Structure)**.** Coordination costs take the linear form

$$C(\delta) = k + \delta F. \tag{E.4}$$

where $k > 0$ is the fixed cost of establishing coordination capability (bot accounts, infrastructure, scripting), and $F > 0$ is the expected penalty (loss) conditional on being flagged. Both k and F are exogenous primitives.

Remark (Microfoundation for Linear Detection Cost). The linear form $C(\delta) = k + \delta F$ admits a natural interpretation. The manipulator operates a network of N bot accounts or coordinated agents. The platform's detection system flags individual accounts with probability δ (the detection intensity). Let $F_0 > 0$ denote the per-account penalty upon flagging (e.g., account suspension, legal liability, reputation loss). Then, the expected total penalty is:

$$\text{Expected penalty} = N \cdot \delta \cdot F_0 = \delta \cdot \underbrace{(NF_0)}_{=:F}.$$

Thus, $F = NF_0$ scales the manipulator's operation. The fixed cost k captures the upfront investment in coordination technology (acquiring accounts, developing scripts, maintaining infrastructure, etc.), independent of detection risk.

Alternative: Convex Detection Costs. If the platform invests more in detection, manipulators may need to employ increasingly sophisticated evasion techniques, generating a convex cost structure $C(\delta) = k + \gamma\delta^2$ with $\gamma > 0$. Under this specification, the deterrence threshold becomes:

$$\delta^\star_{\text{convex}} = \frac{\Pi - k}{2\gamma},$$

which is decreasing in γ (higher marginal detection costs make manipulation easier to deter). The qualitative comparative statics remain unchanged: raising k, F (or γ), or lowering Π all reduce the manipulator's incentive. We focus on the linear case for analytical tractability, but the extension to convex costs is straightforward.

Assumption E.5 (Absolute Continuity and Moment Bounds)**.** The log-likelihood ratios $\log \Lambda_H(x)$ and $\log \Lambda_x(\varepsilon)$ have bounded second moments under both states $\theta \in \{0, 1\}$:

$$\mathbb{E}_\theta \left[(\log \Lambda_H(x))^2 \right] < M < \infty, \quad \forall \theta \in \{0, 1\}, \tag{E.5}$$

where M is a finite constant. Moreover, the message densities are absolutely continuous with respect to a dominating measure, and the manipulator's density $f_M(\cdot|\theta)$ is absolutely continuous with respect to the honest density $f_H(\cdot|\theta)$.

Assumption E.6 (Local Asymptotic Normality (LAN))**.** For the perturbation parameter $\varepsilon := (1 - \rho)(1 - \delta)$, the map $\varepsilon \mapsto \log \Lambda_x(\varepsilon)$ is twice continuously differentiable in a neighborhood of $\varepsilon = 0$, with:

$$\log \Lambda_x(\varepsilon) = \log \Lambda_H(x) + \varepsilon \kappa(x) + O(\varepsilon^2), \tag{E.6}$$

where the second derivative is uniformly integrable. This condition, standard in local asymptotic normality theory, justifies the Taylor expansion in Proposition 5.1 and the $O(\varepsilon^2)$ remainder bounds.

Remark (Verification of LAN). We verify that Assumption E.6 holds for the binary message model. From equation (3.1), the effective likelihood ratio is

$$\Lambda_x(\varepsilon) = \frac{\rho q + \varepsilon \phi_1}{\rho(1 - q) + \varepsilon \phi_0}.$$

Taking logarithms,

$$\begin{aligned} \log \Lambda_x(\varepsilon) &= \log(\rho q + \varepsilon \phi_1) - \log(\rho(1 - q) + \varepsilon \phi_0) \\ &= \log(\rho q) + \log\left(1 + \frac{\varepsilon \phi_1}{\rho q}\right) - \log(\rho(1 - q)) - \log\left(1 + \frac{\varepsilon \phi_0}{\rho(1 - q)}\right). \end{aligned}$$

Since $\log(1 + z)$ is analytic for $|z| < 1$, and since $\varepsilon \in [0, \varepsilon_{\max}]$ with $\varepsilon_{\max} := (1 - \rho) < 1$, both terms are twice continuously differentiable in ε for $\varepsilon < \min\{\rho q/\phi_1, \rho(1 - q)/\phi_0\}$, which is satisfied under

$$\rho > \frac{\max\{\phi_1/q, \phi_0/(1 - q)\}}{1 + \max\{\phi_1/q, \phi_0/(1 - q)\}}.$$

In the benchmark case $(\phi_1, \phi_0) = (1, 0)$ with $q = 0.75$, this requires $\rho \gtrsim 0.57$. By taking the first and second derivatives:

$$\begin{aligned} \frac{\partial \log \Lambda_x(\varepsilon)}{\partial \varepsilon} &= \frac{\phi_1}{\rho q + \varepsilon \phi_1} - \frac{\phi_0}{\rho(1 - q) + \varepsilon \phi_0}, \\ \frac{\partial^2 \log \Lambda_x(\varepsilon)}{\partial \varepsilon^2} &= -\frac{\phi_1^2}{(\rho q + \varepsilon \phi_1)^2} + \frac{\phi_0^2}{(\rho(1 - q) + \varepsilon \phi_0)^2}. \end{aligned}$$

At $\varepsilon = 0$, the first derivative is exactly $\kappa(x) = \frac{\phi_1}{\rho q} - \frac{\phi_0}{\rho(1-q)}$ as in (3.1). The second derivative is bounded:

$$\left|\frac{\partial^2 \log \Lambda_x(\varepsilon)}{\partial \varepsilon^2}\right| \leq \frac{\phi_1^2}{(\rho q)^2} + \frac{\phi_0^2}{(\rho(1-q))^2} =: C_2 < \infty,$$

uniformly for $\varepsilon \in [0, \varepsilon_{\max}]$ under $q \in (1/2, 1)$ and $\rho > 0$. Since the message space is binary ($x \in \{0, 1\}$), and the second derivative is uniformly bounded by C_2, the expectation

$$\mathbb{E}_\theta\left[\left(\frac{\partial^2 \log \Lambda_x}{\partial \varepsilon^2}\right)^2\right] \leq C_2^2 < \infty$$

trivially holds, establishing uniform integrability. By Taylor's theorem with integral remainder, the $O(\varepsilon^2)$ term satisfies $|R(\varepsilon)| \leq \frac{C_2}{2}\varepsilon^2$, justifying the expansions in Proposition 5.1 and Appendix B.

Assumption E.7 (Smoothness)**.** The log-likelihood ratio $\ell(x; \varepsilon) := \log \Lambda_x(x, \varepsilon)$ is twice continuously differentiable in ε in a neighborhood of $\varepsilon = 0$, uniformly over $x \in \mathcal{X}$, with bounded second derivative.

Assumption E.8 (Support Regularity)**.** The honest and manipulator message densities $f_H(\cdot|\theta)$ and $f_M(\cdot|\theta)$ have common support: for each $\theta \in \{0, 1\}$, there exists a measurable set $\mathcal{S}_\theta$ such that both densities are strictly positive on $\mathcal{S}_\theta$ and zero outside. This ensures that the effective likelihood ratio $\Lambda_x(\varepsilon)$ in (3.1) is well-defined and finite for $\forall \varepsilon \in [0, 1]$.

Remark (Singular Cases). If $\text{supp}(f_H) \nsubseteq \text{supp}(f_M)$, then messages $x \in \text{supp}(f_H) \setminus \text{supp}(f_M)$ perfectly reveal the source as honest, allowing the market maker to infer $\ell =$ unflag with certainty without relying on the detection system. Such cases are excluded by Assumption E.8. In practice, sophisticated manipulators can typically mimic the distribution of honest content within their support, making this a mild restriction.

Role of Assumptions Assumptions E.5 and E.6 are standard regularity conditions that ensure Taylor expansions converge and probability measures are well-defined. Assumptions E.4 and 3.2 are substantive modeling choices that capture the economic structure of coordination manipulation and detection.