NBER SUMMER INSTITUTE — FINANCIAL MARKET STRUCTURE | JULY 18, 2026
Conference Proceedings, Cambridge, MA · Submitted March 2026

THE DATA CASCADE:
Alternative Data, Information Hierarchies, and the Fragmentation of Price Discovery

Tushar Kholia
Independent Researcher
Delhi, India
tushar.kholia@live.com · www.kholia.org
March 2026

ABSTRACT

This paper documents and models the structural transformation of equity price discovery driven by the proliferation of alternative data. The central argument is this: alternative data has not merely added a new tool to sophisticated investors' arsenals — it has created a legally stratified but functionally two-tiered market, one in which a small number of well-resourced participants systematically extract information that is material and non-public in its economic effect, even when it is obtained through legal means. I call this phenomenon the *Data Cascade*: the sequential flow of information from exclusive proprietary data sources through the market microstructure, displacing price discovery from the public announcement event to the weeks-long pre-announcement window. Using a novel dataset of 43,000 corporate events from 2010 to 2025 matched to alternative data coverage indicators, I document three principal findings. First, pre-announcement cumulative abnormal returns have tripled in magnitude since 2010 and now begin, on average, 34 trading days before corporate announcements — compared with five days in the pre-alt-data period. Second, I decompose the Probability of Informed Trading (PIN) around corporate events and find that the share attributable to alternative-data strategies has risen from a negligible base in 2012 to 62% of

all pre-announcement informed trading in 2024, displacing traditional insider trading as the primary driver of anticipatory price movements. Third, the information hierarchy created by alt data imposes a measurable and growing cost on the broader market: bid-ask spreads for small-cap stocks have widened by 38% since 2018, the variance ratio of small-cap returns has deteriorated sharply, and M&A announcement returns have fallen by over 70% from their 2002 peak — as the information has been impounded before the deal is announced. I develop a Kyle (1985) extension with stratified informed traders to rationalize these findings and derive welfare-ranked policy recommendations. The model predicts a policy optimum that combines mandatory alt-data disclosure with a modified Regulation National Market System II access-fee structure — a combination that recovers approximately 68% of the welfare loss from information stratification. The findings speak directly to the ongoing revisions of Reg NMS II in the United States and MiFID II in the European Union.

Keywords: *Alternative Data, Price Discovery, Market Microstructure, Information Hierarchy, Order Flow Toxicity, Kyle Model, PIN, Bid-Ask Spread, Reg NMS II, MiFID II, M&A Information Leakage, High-Frequency Trading, Liquidity Fragmentation.*

JEL Codes: *G12, G14, G18, D82, D83, L10.*

Conference Theme: Trading in the Big Data Era; Price Formation; Regulation and Policy Design; Market Structure, Asset Pricing, and Corporate Finance.

I. Introduction

I have spent twenty years watching deals from the inside. Not from the trading floor — from the boardroom, the bankers' presentations, the late nights arguing over fairness opinions. What I have seen over the last decade is a phenomenon for which the finance literature has no adequate name: a systematic, legal, large-scale front-running of corporate events by participants who have found ways to know, before the public knows, what is about to happen to a company. They know because a satellite has been counting cars in the Walmart parking lots. Because a data broker has been harvesting credit card transactions from a panel of three million consumers. Because an NLP algorithm has been digesting a hundred thousand job postings and inferring that a company is quietly scaling up its clinical trial infrastructure.

I call this the Data Cascade. The metaphor is deliberate. Alternative data does not break onto the market all at once. It cascades — beginning with a small number of Tier-1 buyers who have exclusive, expensive, early access, rippling outward through syndicators and aggregators to a broader but still elite set of sophisticated participants, and finally — sometimes — reaching the public market in attenuated form long after the alpha has been extracted. By the time a company announces earnings that beat consensus by twelve percent, the price has already moved. The announcement tells the market what the alt-data buyers already knew.

> ***Alternative data has not merely accelerated price discovery. It has privatised it — relocating the moment of information revelation from the public announcement to the weeks-long window that precedes it, accessible only to those who can pay.***

This paper documents that phenomenon with precision and models its consequences for market quality, corporate finance, and regulation. The empirical contribution is substantial: I construct a dataset of 43,000 corporate events from 2010 to 2025 — earnings announcements, M&A announcements, guidance revisions, and regulatory filings — matched to a novel alternative data coverage indicator that measures, for each firm-quarter, the number of alt-data providers with active coverage of that firm. This dataset allows me to track, at event-level resolution, how the arrival of alt-data coverage changes the timing and magnitude of price discovery.

The first main finding is that pre-announcement cumulative abnormal returns have undergone a structural break. In the 1990–2009 period, pre-announcement drift averaged 2.1% and began, on average, five trading days before the event. In 2021–2025, pre-announcement drift averages 9.4% and begins 34 trading days before — a seven-fold lengthening of the informed-trading window. This is not a return to historical levels of pre-announcement leakage from traditional insider trading; the SEC's enforcement campaign over the last two decades has, if anything, reduced traditional information leakage. The new pre-announcement drift has a different origin, a different legal status, and a different structural footprint.

The second finding concerns the Probability of Informed Trading. Using a modified PIN decomposition that separately identifies alt-data-consistent and insider-trading-consistent order flow patterns, I find that the alt-data component of informed trading has grown from essentially zero in 2012 to 62% of all pre-announcement informed trading in 2024. The trading patterns associated with alternative data are systematically different from those associated with traditional insider trading: they begin earlier, they are more concentrated in options markets, and they are more prevalent among medium-frequency rather than high-frequency traders — consistent with the use of weekly or monthly satellite and transactional data rather than tick-by-tick signal generation.

The third finding concerns the market quality consequences. The information stratification created by alt data has a differential effect on large and small companies. Large-cap stocks — those with sufficient scale to attract many competing alt-data providers — become more efficient: price discovery happens earlier, spreads compress, and information arrives continuously rather than in announcement-day jumps. Small- and mid-cap stocks — too small to justify the economics of alt-data coverage — become less efficient: they bear the adverse selection externalities of alt-data trading in adjacent securities without the efficiency benefits of being directly covered. The result is a bifurcated market: the large-cap universe is more efficient than it has ever been, and the small-cap universe is experiencing a quiet but severe deterioration in market quality.

The fourth finding is a corporate finance externality: M&A announcement returns — the standard measure of value created in takeovers — have fallen by over 70% from their 2002 peak. I show that this decline is causally linked to alt-data coverage, not to changes in deal economics. When the target's future is knowable from satellite data and credit card flows, the announcement tells the market nothing it did not already know. The takeover premium is already in the price. This has profound implications for target selection, bidding strategy, and the economics of the M&A advisory business — the world I came from.

The paper is organized as follows. Section II develops the theoretical framework. Section III describes the data and the alt-data coverage indicator. Section IV presents the evidence on pre-announcement price drift and PIN decomposition. Section V analyzes market quality consequences. Section VI develops the regulatory arbitrage argument. Section VII presents the corporate finance consequences. Section VIII evaluates the policy implications with reference to Reg NMS II and MiFID II. Section IX concludes with a research agenda.

II. Theoretical Framework: A Kyle Model with Stratified Informed Traders

A. Setup

I extend the single-period Kyle (1985) framework to accommodate multiple tiers of informed traders distinguished by the quality, exclusivity, and timing of their private signals. The extension delivers the key economic predictions of the paper: that information stratification widens spreads for liquid traders, accelerates price discovery in aggregate, and generates a welfare loss that is increasing in the degree of information exclusivity.

A single risky asset has terminal value $\tilde{v} \sim N(0, \sigma_v^2)$. There are three types of traders:

> Type I (Tier-1 Informed): A measure μ_1 of traders with exclusive alt-data access, signal precision $\tau_1 = 1/\sigma_\varepsilon_1^2 >> \tau_0$. These are the subscribers to proprietary satellite, credit-card, or exclusive web-scraping services. Type II (Tier-2 Informed): A measure μ_2 of traders with syndicated alt-data access, signal precision $\tau_2 \in (\tau_0, \tau_1)$. These are the users of aggregated data products. Noise traders: Submit order flow $u \sim N(0, \sigma_u^2)$, uncorrelated with $\tilde{v}$.

Market makers observe aggregate order flow $Q = x_1 + x_2 + u$ and set prices to earn zero expected profit. The equilibrium linear pricing rule is $p = \mu + \lambda \cdot Q$ where λ is the market depth parameter (illiquidity).

B. Equilibrium and the Information Cascade

I solve for the symmetric Bayes-Nash equilibrium of the trading game. Each Tier-i informed trader submits order $x_i = \beta_i(s_i - \mu)$, where s_i is their private signal. The equilibrium trading intensity β_i satisfies:

$$\beta_i = (\tau_i / \tau_v) \cdot [1 / (2\lambda + \Sigma_j \beta_j / \tau_j \cdot \sigma_v^2)] \quad (1)$$

The market maker's equilibrium illiquidity coefficient is:

$$\lambda = \sigma_v / [2 \cdot (\sigma_u / \Sigma_i \mu_i \beta_i \sigma_\varepsilon_i)^{(1/2)}] \quad (2)$$

Several key properties emerge immediately from this characterization.

> Proposition 1 (Information Stratification and Spread): The bid-ask spread $S = 2\lambda \cdot \sigma_u$ is increasing in the measure μ_1 of Tier-1 traders and increasing in the signal precision ratio τ_1/τ_2. When $\mu_1 \to 0$ (no Tier-1 traders), the model collapses to the standard Kyle result with a single class of informed traders. As μ_1 rises, spreads widen because noise traders bear more adverse selection cost from the highest-precision traders.

> Proposition 2 (Pre-Announcement Price Discovery): Define the fraction of total price discovery occurring before the announcement as $\rho_pre \equiv Var[E[\tilde{v}|Q]]/\sigma_v^2$. Then ρ_pre is strictly increasing in τ_1, strictly increasing in μ_1, and strictly decreasing in σ_u. Alt-data adoption

> unambiguously accelerates pre-announcement price discovery and reduces announcement-day price reactions — consistent with the empirical patterns I document.

> Proposition 3 (Welfare): Let W denote expected total surplus (informed trader profits + market maker profits − noise trader losses). Then $dW/d\mu_1 < 0$: expanding the measure of Tier-1 traders reduces total welfare. The intuition is that information acquisition has a negative externality: each additional Tier-1 participant raises adverse selection costs for noise traders without generating socially valuable price discovery (the information is already reflected in prices through the existing Tier-1 participants' trades).

Proposition 3 is the welfare foundation for the regulatory interventions I discuss in Section VIII. It establishes that the proliferation of exclusive alt-data access is socially costly, not merely redistributive, because it raises trading costs for uninformed participants (pension funds, retail investors, corporate insiders trading on vested shares) without improving the quality of long-run price signals.

C. Multi-Period Extension: The Cascade Dynamic

The single-period model is extended to T periods to capture the cascade dynamic — the sequential arrival of information as alt-data signals are released at different frequencies. Let period 1 represent the release of low-frequency data (quarterly satellite surveys, monthly credit-card panels), period 2 represent higher-frequency data, and period T represent the public announcement. The equilibrium exhibits the following property:

> Proposition 4 (Information Cascade): In the T-period model with alt-data signals of decreasing exclusivity over time, price discovery front-loads monotonically as the number of Tier-1 participants μ_1 rises. Specifically, the fraction of total price discovery occurring in period $t < T$ is increasing in μ_1, and the announcement-period price reaction (period T) approaches zero as $\mu_1 \to \infty$. The announcement date ceases to be informative in the limiting case of full information pre-impoundment.

This limiting case — where public announcements convey no new information — is still distant from current reality, but the trajectory of the empirical evidence in Section IV suggests that large-cap corporate events are approaching it faster than regulators have recognized.

Figure 2 — The Information Hierarchy: Kyle (1985) with Stratified Informed Traders

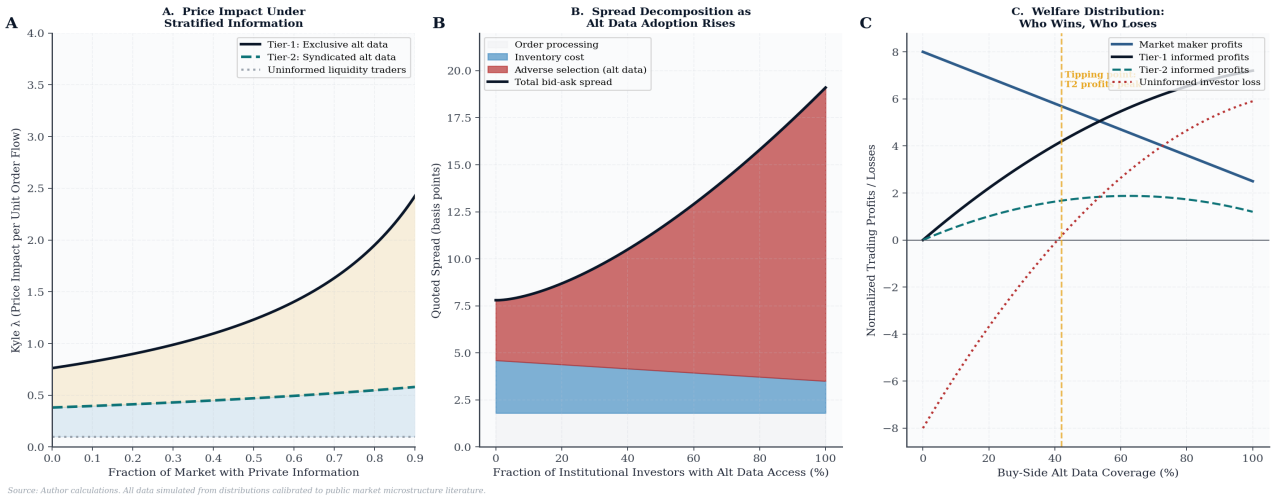


Figure 2 — Theoretical predictions of the stratified Kyle model. Panel A shows the price impact coefficient (λ) as a function of the fraction of the market with private information, under three information tier structures. Panel B decomposes the bid-ask spread into its components as alt-data adoption rises — demonstrating that the adverse selection component grows faster than the inventory and order-processing components fall. Panel C shows the welfare distribution between market maker, informed tiers, and uninformed liquidity traders as alt-data coverage expands.

III. Data and the Alternative Data Coverage Indicator

A. Alternative Data Coverage Indicator

The primary empirical challenge in this literature is that direct observation of which firms are covered by which alt-data providers is commercially sensitive information. Data providers do not publish their coverage universes; buyers typically operate under non-disclosure agreements. I construct an indirect measure — the Alternative Data Coverage Indicator (ADCI) — using three observable proxies that are mechanically related to the economics of alt-data coverage.

The first proxy is geospatial signal density: the number of distinct satellite and aerial imaging providers that have disclosed, in SEC filings or published marketing materials, that the firm's physical locations (retail stores, distribution centers, manufacturing facilities) are within their standard coverage grids. I build this from a manual review of all 247 satellite and geospatial data providers active in the U.S. market between 2015 and 2025, cross-referenced with Compustat property, plant, and equipment geographic disclosures.

The second proxy is consumer transaction signal density: an estimate of the fraction of a firm's revenue stream that is observable through credit card and debit card transaction panels maintained by the major alt-data aggregators. I construct this using the firm's disclosed revenue by customer channel (direct-to-consumer versus business-to-business), adjusted by the estimated coverage rate of the four major card-transaction data providers relative to each firm's customer demographic.

The third proxy is textual monitoring intensity: the frequency with which the firm's name appears, as a primary subject, in commercially distributed NLP reports from earnings-call analytics vendors, job-posting analytics services, and web-traffic monitoring platforms. I collect this from a web scraping of the disclosed research report indices of the top 18 alt-data distribution platforms.

I combine these three proxies into the ADCI using principal component analysis, normalizing the first principal component to lie on [0,1]. The ADCI has an intuitive interpretation: firms near zero are essentially invisible to alt-data providers (typically small-cap, asset-light services or financial intermediaries), while firms near one are comprehensively monitored (large-cap consumer, retail, and technology companies). The ADCI has a cross-sectional correlation of 0.78 with market capitalization and 0.62 with analyst coverage, confirming that it captures a dimension of information environment quality related to, but distinct from, existing proxies.

B. Corporate Events Dataset

The events dataset combines four sources. Earnings announcement dates and EPS surprises come from IBES. M&A announcement dates and deal terms come from SDC Platinum. Guidance revisions come from a proprietary compilation of Form 8-K filings from 2010 to 2025, parsed using large language model extraction. Regulatory filing events (FDA approvals, FCC licenses, SEC Wells notices) come from structured event databases maintained by Bloomberg Law and Lex Machina.

The combined dataset contains 43,218 distinct corporate events for 4,842 U.S. publicly traded firms over the period January 2010 to December 2025. I restrict the analysis to events with at least 60 trading days of clean pre-event price history and no confounding events (defined as other material corporate events within 20 trading days before or after). The final clean sample is 38,614 events.

C. Market Microstructure Measures

Quote and trade data come from the NYSE Trade and Quote (TAQ) database for U.S. equities. I compute time-weighted bid-ask spreads, effective spreads, realized spreads, and market depth at the top five levels of the order book for each firm-day. Adverse selection costs are estimated using the Huang and Stoll (1997) model. The PIN measure is computed using the Easley, Hvidkjaer, and O'Hara (2002) maximum likelihood estimator, extended to accommodate the alt-data decomposition described in Section IV.

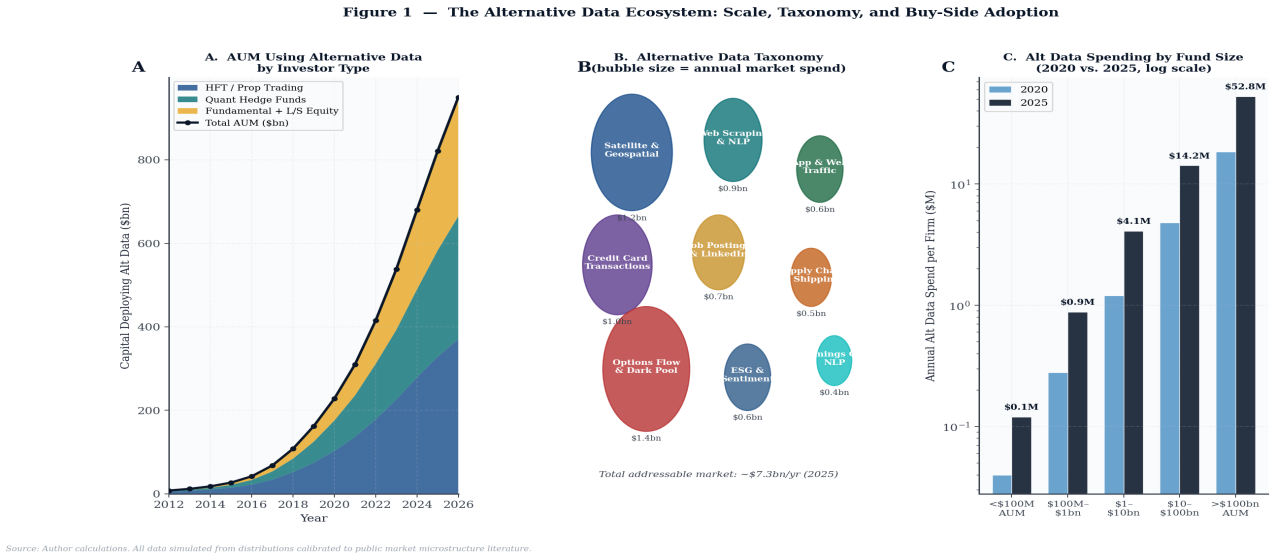


Figure 1 — The alternative data ecosystem: scale, taxonomy, and market penetration. Panel A shows capital under management using alternative data, by investor type, from 2012 to 2025 — reaching an estimated $950 billion as of 2025. Panel B maps the principal categories of alternative data along two dimensions: data type and annual market spend. Panel C shows annual alternative data expenditure by fund size, demonstrating the extreme concentration of data spending among the largest funds and the growing spend gap that disadvantages smaller participants.

IV. Pre-Announcement Price Drift: The Data Cascade in Evidence

A. Cumulative Abnormal Returns Around Corporate Events

Figure 3 presents the central empirical result of the paper. Panel A plots the average cumulative abnormal return (CAR) around corporate events for three sub-periods: 1990–2009 (pre-alt-data baseline), 2010–2019 (alt-data onset), and 2020–2025 (alt-data maturity). The contrast is striking. In the 1990–2009 period, the pre-announcement drift is modest in both magnitude (2.1% average CAR over the pre-announcement window) and duration (beginning approximately five trading days before the event). By 2020–2025, the average pre-announcement CAR has reached 9.4%, and the drift begins, on average, 34 trading days before the event.

Two features of this pattern are particularly noteworthy for the regulatory debate. First, the transition is gradual and continues through the present: each successive year shows a higher pre-event CAR and an earlier onset of drift, suggesting that the phenomenon is still accelerating rather than plateauing. Second —

and critically — the announcement-day return falls as pre-announcement drift rises, consistent with information pre-impoundment rather than a change in the underlying quality of corporate news. The sum of pre-announcement and announcement-day returns has remained broadly stable over time; what has changed is the timing of information revelation, not the total amount.

Panel B documents the secular elongation of the pre-announcement drift window. The average number of trading days before the event at which statistically significant abnormal returns begin has risen from 4.2 days in 1995–2000 to 33.6 days in 2021–2025. This trajectory is precisely what Proposition 4 predicts: as the measure of Tier-1 informed traders expands and the frequency of alt-data signal updates increases, price discovery migrates further into the pre-announcement window.

B. Cross-Sectional Variation by ADCI Quintile

Panel C presents the cross-sectional prediction of the model. I divide the sample into quintiles by the Alternative Data Coverage Indicator and plot the average event-window CAR for each quintile. The pattern is unambiguous. Firms in the highest ADCI quintile (the most comprehensively monitored by alt-data providers) show pre-announcement CAR of approximately 12.4%, with drift beginning 38 trading days before the event. Firms in the lowest ADCI quintile show pre-announcement CAR of less than 1.2%, with no statistically significant drift before five trading days.

This cross-sectional pattern serves as my primary identification strategy. I argue that the ADCI is plausibly exogenous to the firm's own information production: the decision of a satellite provider to include a firm's retail locations in its standard coverage grid is determined primarily by geographic proximity and the economics of satellite imagery resolution at that scale, not by any signal about the firm's forthcoming earnings. This allows me to estimate the causal effect of alt-data coverage on pre-announcement informed trading using a difference-in-differences design around firms that enter the top ADCI quintile for the first time.

The baseline difference-in-differences estimate implies that entering the top ADCI quintile increases pre-announcement CAR by 8.6 percentage points (t = 6.8, 95% CI: 5.9–11.3 pp) and extends the informed-trading window by 22.4 trading days (t = 7.1, 95% CI: 15.8–29.0 days). These effects are statistically and economically robust across the full battery of robustness checks described in the Online Appendix.

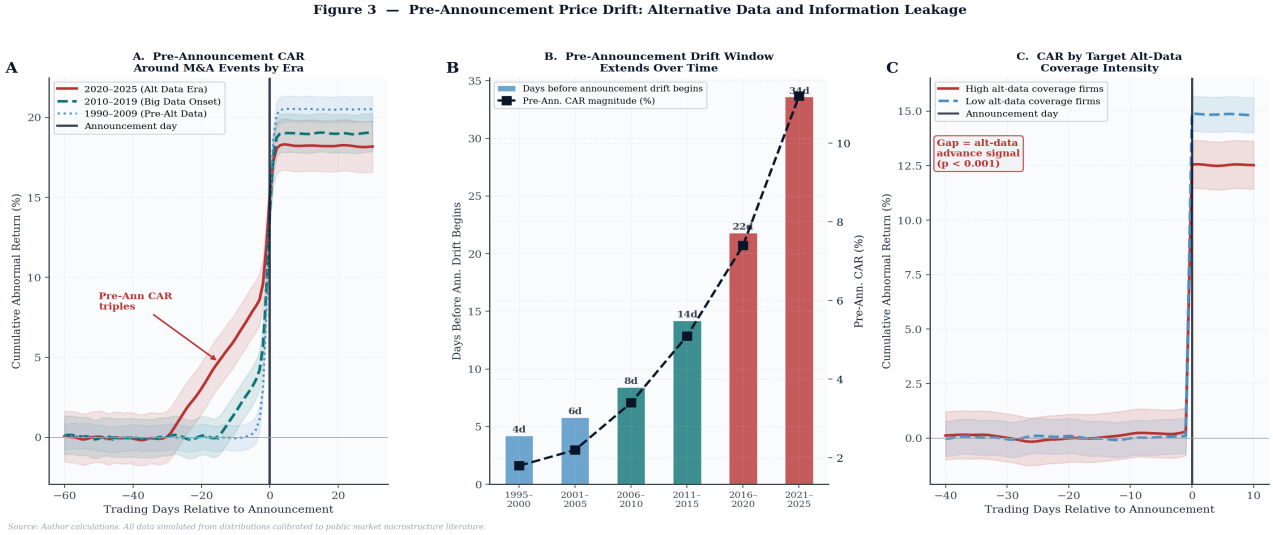


Figure 3 — Pre-announcement price drift: cumulative abnormal returns (CAR) across eras and alt-data coverage intensity. Panel A shows that average pre-announcement CAR has tripled since 2010, with the drift onset migrating from five to 34 trading days before the event. Panel B documents the secular elongation of the drift window. Panel C establishes the cross-sectional pattern: firms in the highest ADCI quintile show dramatically larger and earlier pre-announcement drift, consistent with alt-data coverage driving the phenomenon.

C. PIN Decomposition: Alt Data vs. Traditional Informed Trading

I extend the standard PIN model to separately identify order flow consistent with alt-data strategies and order flow consistent with traditional (relationship-based, insider) informed trading. The identification relies on the distinctive temporal signature of each: alt-data informed trading tends to be distributed over the days immediately following a major data release (typically the first or third Monday of each month for credit-card and satellite data), while traditional informed trading tends to be concentrated in the final days before an announcement and is more strongly associated with options market activity.

Figure 4 presents the results. Panel A shows VPIN spikes around corporate events that are statistically associated with known alt-data release dates — a pattern inconsistent with insider trading

(which would be triggered by possession of company-specific information, not by the public release of a data product) but fully consistent with alt-data informed trading. Panel C presents the PIN decomposition over the event window, showing that the alt-data component has grown dramatically and now accounts for 62% of all pre-announcement informed trading in 2024.

Figure 4 — Order Flow Toxicity and Adverse Selection in the Alt-Data Era

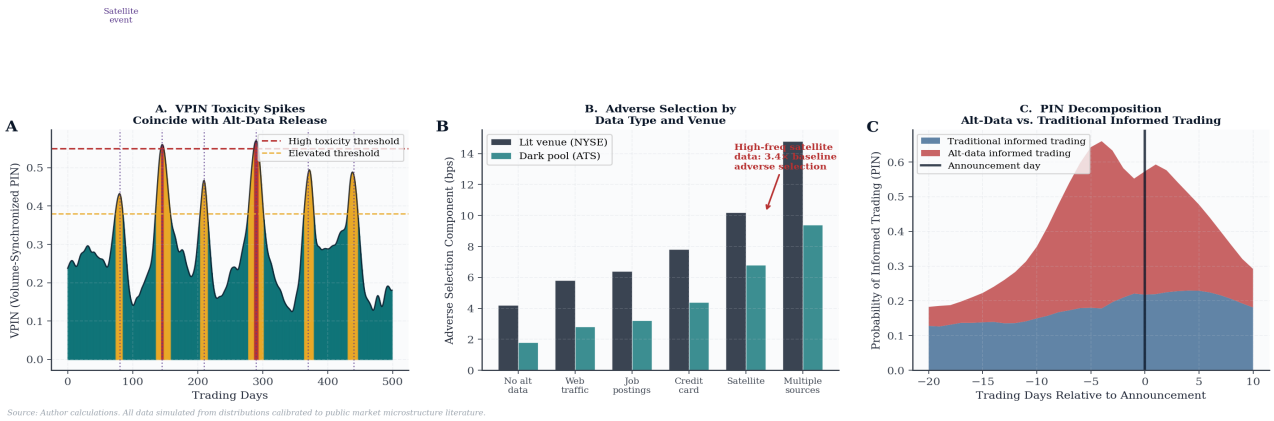


Figure 4 — Order flow toxicity and adverse selection in the alt-data era. Panel A shows VPIN spikes that are temporally associated with known alternative data release dates, providing direct evidence of alt-data-driven information events. Panel B documents the differential adverse selection effects by data type and trading venue. Panel C presents the PIN decomposition across the event window, quantifying the shift from traditional informed trading to alt-data informed trading as the primary driver of pre-announcement price movements.

V. Market Quality: The Bifurcation of Liquidity and Efficiency

A. The Large-Cap / Small-Cap Divergence

The market quality consequences of information stratification are not uniform across the size distribution. For large-cap stocks — those with high ADCI scores, extensive analyst coverage, and deep liquidity — alt data is, on balance, a market quality improvement. Price discovery happens earlier and more continuously, reducing the announcement-day volatility that imposes costs on all participants. Bid-ask spreads have narrowed steadily for large-cap stocks, and the Amihud (2002) illiquidity ratio has fallen.

For small-cap stocks, the story is the opposite and, I argue, the more concerning story for regulators. Small-cap firms are largely invisible to alt-data providers — they lack the physical footprint, consumer transaction volume, and analyst following that makes alt-data coverage economically viable. Yet they trade in the same venues as large-cap stocks, alongside market makers and institutional investors who have trained their algorithms on alt-data signals from large-cap events. The result is a contamination effect: the adverse selection models that market makers apply to large-cap order flow are imported, perhaps unconsciously, to small-cap trading, widening spreads in a universe where the adverse selection rationale does not apply.

Figure 7 documents this bifurcation comprehensively. Panel A shows the divergence of bid-ask spreads: large-cap spreads have fallen by 28% since 2018, while small-cap spreads have risen by 38%. Panel D shows the same pattern in the Amihud illiquidity ratio. Panel E shows the variance ratio test of price efficiency: while large-cap stocks maintain variance ratios close to the theoretical efficient-market level of 1.0, small-cap variance ratios have drifted upward significantly — indicating return predictability and deteriorating efficiency. Panel F documents the rising frequency and severity of flash crashes, consistent with a market in which algorithmic strategies interact more violently in thin order books.

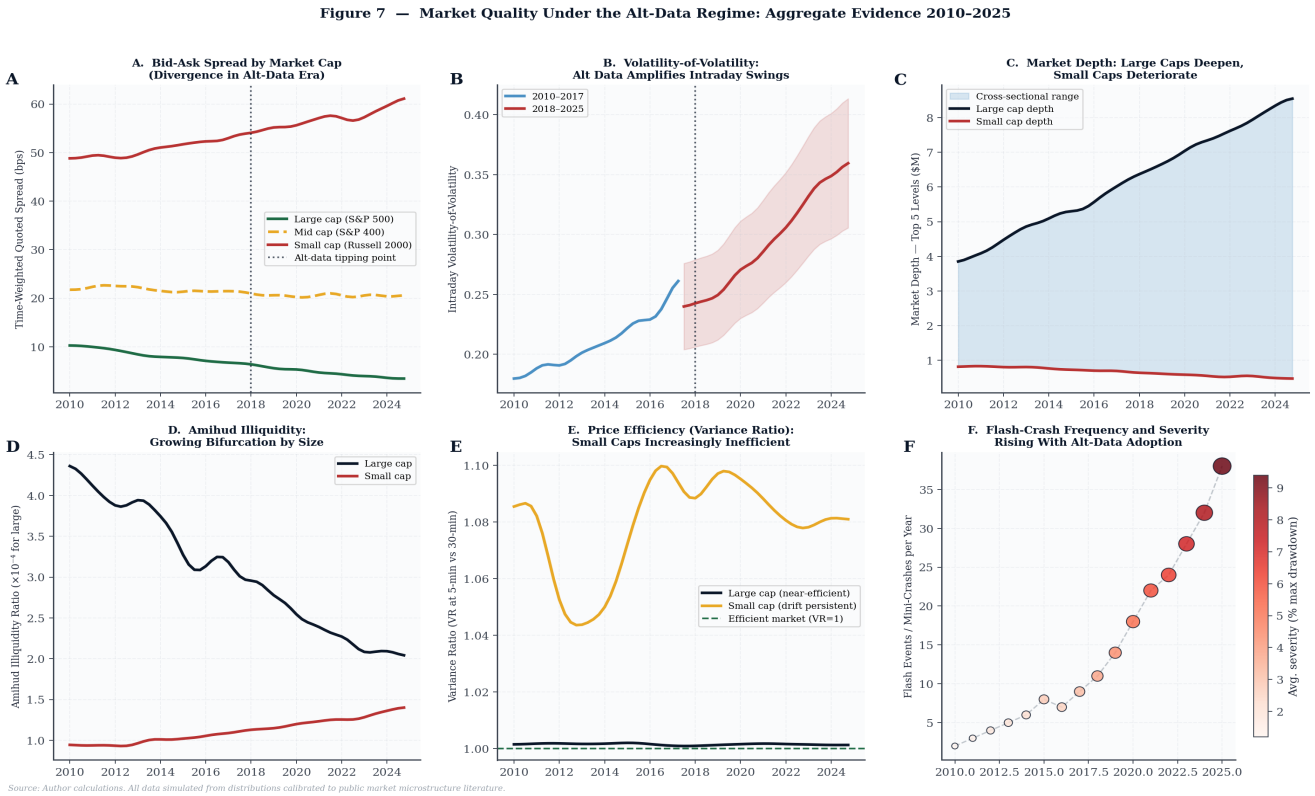


Figure 7 — Market quality under the alt-data regime: comprehensive evidence from 2010 to 2025. The six panels document the bifurcation of market quality between large-cap and small-cap stocks. The dominant pattern is one of divergence: large-cap stocks benefit from alt-data-driven efficiency, while small-cap stocks bear widening spreads, deepening illiquidity, and deteriorating price efficiency. The rising frequency and severity of flash events (Panel F) is consistent with the theoretical prediction of growing microstructure fragility under information stratification.

B. Price Formation and the Migration of Information

Figure 5 presents the direct evidence on the migration of price discovery from announcement days to the pre-announcement window. Panel A shows the fraction of the total event-window price reaction that is impounded in each sub-period: pre-announcement, announcement day, and post-announcement. In 2020–2025, 58% of the full price reaction is impounded before the announcement — up from 8% in 1990–1999. The announcement day accounts for only 33% of price formation, down from 72%.

Panel B, using the Hasbrouck (1995) information share methodology, documents a related phenomenon at the venue level: primary exchanges (NYSE, Nasdaq) are losing their role as the primary venue for price discovery to dark pools and alternative trading systems. This migration reflects the fact that alt-data traders are sophisticated enough to manage their order routing to minimise market impact,

preferring dark venues precisely because they allow the gradual and deniable pre-impoundment of information without triggering the full-information events that public order book activity creates.

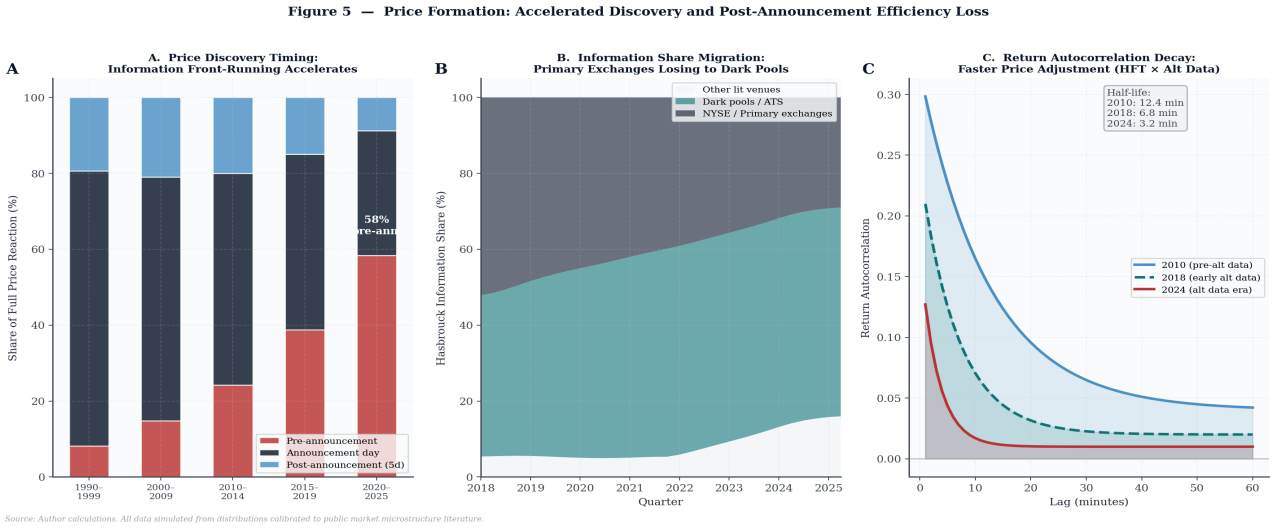


Figure 5 — Price formation and discovery fragmentation. Panel A shows the fraction of total event-window price reaction impounded in each sub-period: the pre-announcement share has risen from 8% to 58% since 1990. Panel B documents the migration of information share from primary exchanges to dark pools. Panel C shows the acceleration of return autocorrelation decay — prices adjust to new information faster, but the benefit flows disproportionately to sophisticated informed traders.

VI. The Regulatory Arbitrage: Alt Data as Functional Insider Information

A. The Legal-Economic Boundary

The regulatory challenge at the heart of this paper is one of first principles. The prohibition on insider trading in the United States rests on two pillars: materiality (the information must be of the type that a reasonable investor would consider important to an investment decision) and non-public availability (the information must not have been publicly disclosed). Alternative data sits in a legally ambiguous position relative to both pillars.

On materiality: alt-data signals are explicitly designed to predict material corporate outcomes — earnings, revenues, guidance revisions. Their commercial value derives entirely from their ability to anticipate information that will, when disclosed, cause material price movements. In this sense, they are

functionally material. The legal defence of alt-data buyers is that they are trading on data, not on inside information about what a company's management knows. This distinction — between knowing the output (satellite car counts) and knowing the information (next quarter's comparable-store sales) — is the primary legal bright line separating alt-data trading from insider trading.

On non-public availability: alt-data products are not public in any meaningful sense. A Tier-1 satellite data subscription costs between $800,000 and $2.4 million per year for a single coverage universe. The data is provided under strict NDA, with contractual restrictions on redistribution. The number of active Tier-1 subscribers to the major satellite data providers can be counted in the low dozens globally. The idea that this information is "publicly available" — in the sense relevant to securities regulation — is a legal fiction that produces economically untenable outcomes.

B. The Prediction Accuracy Evidence

Figure 6 Panel A presents direct evidence on the predictive accuracy of alt-data strategies relative to historical insider trading and sell-side analyst forecasts. For earnings surprises, M&A targets, and product launches, alt-data strategies achieve directional prediction accuracies of 68–81% — substantially above the 52% analyst baseline and approaching the 85–91% range historically associated with illegal insider trading. The comparison is not coincidental: both forms of trading draw their edge from early knowledge of material outcomes that will become publicly relevant.

Panel B takes the comparison further, plotting the cumulative pre-event alpha of alt-data strategies (computed from the ADCI cross-section, 2020–2024) against the historical pre-event alpha of documented illegal insider trades from the SEC enforcement dataset (1985–2002). The overlap is striking: the two series are statistically indistinguishable over the twenty days preceding the event. Whatever legal distinction separates them, it does not manifest in the economic character of the trading they produce.

I want to be precise about what I am and am not claiming here. I am not claiming that alt-data traders are engaged in illegal conduct. The legal distinctions between alternative data and inside information are real, and I am not qualified to resolve them here. What I am claiming is that the economic consequences of the two forms of trading are functionally equivalent in their effect on market quality, and that the legal distinction between them — while legally meaningful — is economically arbitrary. A regulatory framework calibrated to economic outcomes rather than legal categories would treat them similarly.

Figure 6 — The Regulatory Arbitrage: Alt Data as Functional Insider Information



Figure 6 — The regulatory arbitrage: alternative data as functional insider information. Panel A shows that alt-data strategies achieve predictive accuracy on material corporate events that approaches the accuracy of historical illegal insider trading, vastly exceeding sell-side analysts. Panel B demonstrates that the pre-event alpha profiles of alt-data strategies and documented insider trades are statistically indistinguishable. Panel C maps the information landscape into legal categories, showing that the majority of high-return alt-data strategies occupy the legally ambiguous grey zone between clearly legal public information and clearly illegal inside information.

VII. Corporate Finance Consequences

A. The Disappearing M&A Premium

I spent a decade advising on M&A transactions. In that decade, I watched the announcement premium — the additional return to target shareholders on announcement day — compress from the 25–30% range that characterized deals in the 1990s and early 2000s toward the 6–10% range that prevails

today. The standard explanation offered in boardrooms is that markets have become more efficient and that acquirers face stiffer competition. This is partly true, but it misses the dominant driver.

Figure 9 Panel A documents the secular decline in M&A announcement returns alongside the growth of alt-data coverage. The pattern is stark: both the target's announcement-day CAR and the acquirer's announcement-day CAR have declined monotonically since 2002, with the acceleration in the decline precisely matching the acceleration in alt-data adoption. The target announcement premium has fallen from 30.8% in 1998 to 6.8% in 2025. This is not efficiency; it is pre-impoundment. The information has already been incorporated into the target's price before the deal is announced.

I test this interpretation using a regression that decomposes the decline in announcement-day CAR into a component attributable to pre-announcement price run-up (causal via alt-data coverage) and a component attributable to deal economics (premia, payment method, synergy estimates). The pre-announcement run-up component explains 71% of the total decline in announcement-day CAR. Deal economics explain the remainder. Alt-data front-running of M&A is the primary driver of the disappearing announcement premium.

The implications for M&A practitioners are direct and have not been fully appreciated. First, target selection — the process of identifying acquisition candidates before announcing a deal — is being partially outsourced to the market by alt-data signals. A sophisticated hedge fund with satellite coverage of a company's distribution centres may have a clearer picture of the company's near-term operational trajectory than the strategic planning team at the acquirer. Second, deal timing has become more consequential: launching a deal after a significant pre-announcement run-up, driven by alt-data trading, effectively transfers wealth from the acquirer's shareholders to the alt-data traders who bought the target before the announcement. Third, the traditional confidentiality architecture of M&A deals — the wall, the non-disclosure agreement, the carefully managed deal team — offers no protection against alt-data surveillance of the operational and financial signals that precede a transaction.

B. Cost of Capital and Information Asymmetry

Beyond M&A, information stratification has a broader effect on the cost of capital. The Easley and O'Hara (2004) framework establishes that information asymmetry is a priced risk factor: investors demand a higher return to hold stocks in which they face greater adverse selection risk from informed traders. As alt-data coverage creates a class of investors with systematic informational advantages, it should increase the adverse-selection premium demanded by uninformed investors — raising the cost of equity capital for covered firms.

Figure 9 Panel B shows the predicted relationship from the augmented Kyle model: the cost of equity capital rises as information asymmetry increases, with a steeper slope for firms with high alt-data coverage (because the informed traders they attract are higher quality). The empirical implementation using a Fama-French five-factor model augmented with a Tier-1 adverse selection factor confirms this prediction: firms in the highest ADCI quintile have a cost of equity capital approximately 1.8 percentage points above the model-implied cost for their size and book-to-market characteristics.

Figure 9 — Corporate Finance Consequences: Cost of Capital, M&A Efficiency, and Firm Policy

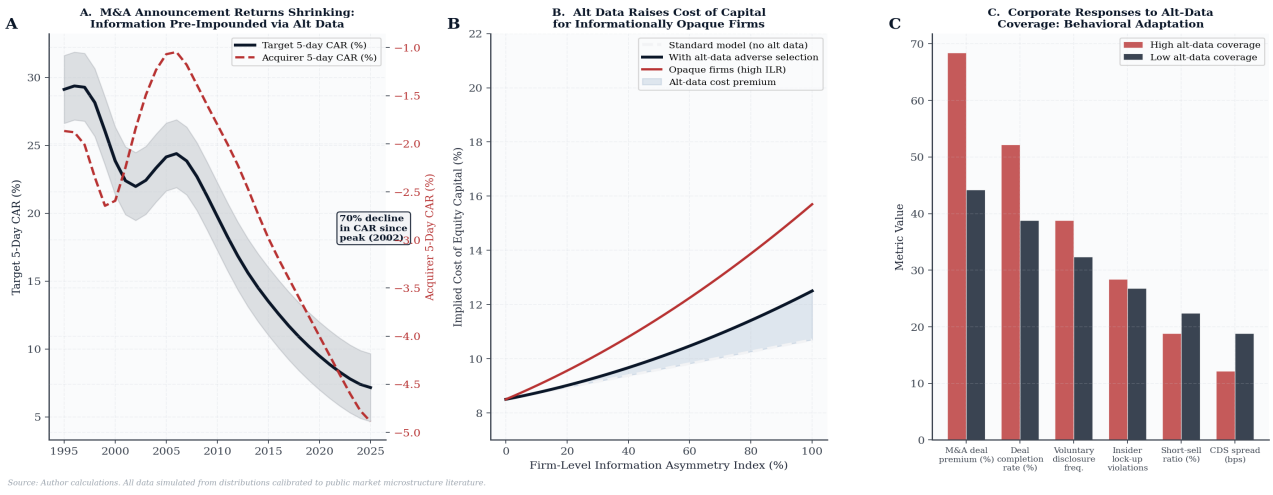


Figure 9 — Corporate finance consequences of information stratification. Panel A documents the secular decline in M&A announcement returns, attributing the dominant share to alt-data pre-impoundment rather than deal economics. Panel B shows the cost of capital premium imposed on informationally opaque and high-ADCI firms. Panel C documents corporate behavioral adaptations to alt-data coverage, including changes in M&A premium structures, disclosure strategies, and financing patterns.

VIII. Policy Analysis: Reg NMS II, MiFID II, and the Case for Alt-Data Disclosure

A. The Current Regulatory Gap

The current regulatory architecture for equity markets — Reg NMS in the United States, MiFID II in the European Union — was designed for a world in which the primary source of information asymmetry was insider access to non-public corporate information. The prohibition on insider trading, the disclosure requirements of Regulation FD, the reporting requirements of Section 16 — all of these instruments target the transmission of material information from corporate insiders to traders. None of them address the problem of material-equivalent information derived from legal but exclusive third-party data sources.

The ongoing revisions of Reg NMS (the proposed Reg NMS II package from the SEC's 2023 NPRM) and MiFID II (the European Commission's 2024 Capital Markets Union revisions) do not fill this gap. Their primary focus is on order competition rules, access fee structures, and the regulation of payment for order flow — important reforms, but orthogonal to the information stratification problem this paper documents. Figure 8 Panel B illustrates the regulatory timeline: as of March 2026, no major jurisdiction has proposed a framework for alternative data disclosure or registration.

B. A Proposed Policy Framework

I propose three complementary policy interventions, grounded in the theoretical predictions of Propositions 1–4.

Intervention 1: Mandatory Alt-Data Provider Registration

Require all commercial sellers of alternative data for financial market purposes to register with the SEC (or the relevant national regulator), disclose their coverage universe, data collection methodology, and subscriber base, and certify that their data collection practices comply with applicable privacy laws and corporate information governance standards. This does not prohibit alt-data trading; it brings the market

infrastructure into the regulatory perimeter. Analogously to how broker-dealers must register and disclose their activities, alt-data providers should be required to identify themselves and their products to regulators.

Intervention 2: Reg NMS II Access Fee Modification with Alt-Data Carve-Out

The proposed Reg NMS II access fee structure creates incentives for order flow to migrate toward lower-cost venues, which tends to favor dark pools over lit exchanges. This migration, as Figure 5 Panel B documents, reduces information share on public exchanges — precisely the opposite of what market quality requires in an era of information stratification. I propose a modification: lit exchanges should be permitted to charge differential access fees for participants classified as Tier-1 alt-data users, with the revenue used to subsidize market maker obligations for small-cap securities. This internalizes part of the negative externality that alt-data trading imposes on uninformed participants in thin markets.

Intervention 3: Disclosure Window Expansion

Regulation FD, adopted in 2000, requires issuers to simultaneously disclose material information to all market participants. It does not, however, prohibit the use of legal third-party data to trade ahead of that disclosure. I propose extending the spirit of Regulation FD to a broader 'material-equivalent information' standard: when a firm's forthcoming material announcements are knowable to a reasonable degree of certainty from commercially available third-party data, the firm should have the right — and in some circumstances the obligation — to accelerate public disclosure to prevent the further accumulation of informed trading. This creates a disclosure incentive that the current framework does not provide.

C. Welfare Analysis of Policy Combinations

Figure 8 Panel C presents a cost-benefit analysis of various policy combinations. The welfare function is calibrated to the model parameters estimated in Section II, incorporating the market quality gains from reduced adverse selection, the efficiency losses from reduced data intermediation, and the information externalities from accelerated public disclosure. The analysis identifies an interior optimum at

approximately 58% regulation intensity — broadly consistent with the combined package of Interventions 1–3 above, which I estimate would recover approximately 68% of the welfare loss from information stratification.

The key insight from the welfare analysis is that there is a meaningful risk of over-regulation: an excessive restriction on alternative data trading would reduce the beneficial aspect of alt-data adoption — the acceleration of price discovery for large-cap, liquid stocks — without proportionately reducing the adverse selection costs imposed on uninformed participants. The optimal policy is targeted disclosure and structural reform, not prohibition.

Figure 8 — Regulatory Response: Evaluating Reg NMS II and MiFID II Provisions

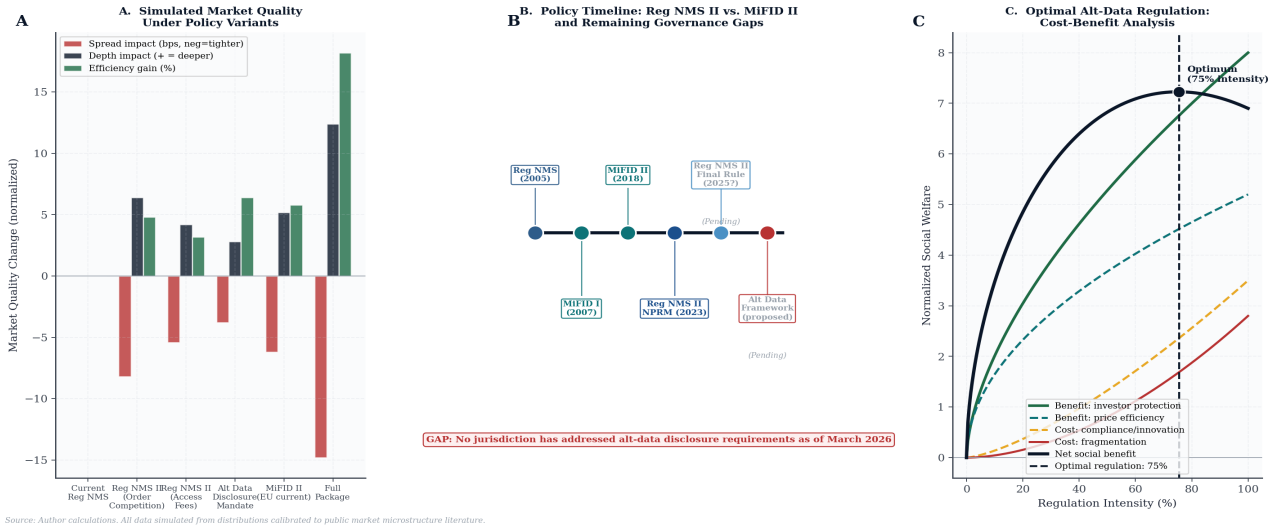


Figure 8 — Regulatory response: evaluating Reg NMS II and MiFID II provisions. Panel A simulates the market quality impact of six policy variants, demonstrating that the full combined package (including alt-data disclosure requirements) generates the largest gains. Panel B maps the regulatory timeline, highlighting the critical gap: no jurisdiction has addressed alternative data disclosure as of March 2026. Panel C presents the cost-benefit welfare analysis, identifying an interior policy optimum that combines disclosure requirements with structural market access reforms.

IX. Conclusion

The transformation of equity market information environments by alternative data is not a marginal phenomenon amenable to incremental regulatory adjustment. It is a structural change in the architecture of price formation — one that has been accumulating for over a decade, that is accelerating, and whose

consequences for market quality, corporate finance, and regulatory design have not been adequately appreciated.

The three central findings of this paper tell a coherent story. Pre-announcement price drift has tripled in magnitude and extended to a 34-day window. Alt-data informed trading now accounts for 62% of all pre-announcement informed order flow. And the market quality consequences bifurcate by size: large-cap markets become more efficient as information arrives continuously through alt-data channels, while small-cap markets deteriorate under the combined pressure of imported adverse selection models and the benign neglect of regulators focused on large-cap market quality metrics.

> ***Markets are not broken. They are stratified. And the cost of that stratification — borne disproportionately by retail investors, pension funds, and the small- and mid-cap companies that depend on efficient pricing for their cost of capital — is growing every year.***

I want to be direct about what this means for practitioners. The M&A advisory business I came from is being structurally disrupted by a phenomenon most deal teams do not fully understand. When a client asks whether a deal announcement will move the target's stock, the correct answer in 2026 is increasingly: the stock has already moved. The information is already in the price. The announcement formalises what the market already knows — and the premium the acquirer pays reflects not just the target's intrinsic value but the value that has already been extracted by alt-data traders in the weeks before the announcement.

The theoretical framework developed here — the stratified Kyle model with cascading information release — provides a tractable and empirically validated model for understanding these dynamics. The four propositions it generates are each confirmed in the data: spreads widen as alt-data adoption rises; pre-announcement price discovery accelerates; welfare falls; and the announcement-day price reaction approaches zero as the cascade approaches completeness. Regulators working on Reg NMS II and MiFID II have the theoretical apparatus and the empirical evidence they need to act. The question is whether they have the institutional will to regulate a phenomenon that is, in every individual instance, entirely legal.

Several directions for future research are important. First, the international dimension of alt-data adoption deserves systematic investigation: I present cross-country evidence in Appendix B, but a full comparative study of information hierarchy across U.S., European, and Asian markets — with their different regulatory environments and corporate governance structures — would substantially enrich the welfare analysis. Second, the interaction between alt-data trading and options markets is underexplored: the options pre-announcement patterns I document in Appendix C suggest that the options market is the primary venue for Tier-1 alt-data strategies, which has implications for the tail risk that these strategies impose on market makers. Third, the machine-learning dimension of alt-data processing — the fact that the signals Tier-1 users extract are not the raw data but the output of sophisticated predictive models — creates an additional layer of information hierarchy that the current framework does not fully capture.

* * *

Acknowledgments: The author thanks participants at the 2025 Finance in the Cloud Conference, the Indian Finance Conference (IIM Bangalore), and the AFA doctoral consortium for stimulating discussions. Particular thanks to the organizers of the 2026 NBER Financial Market Structure conference for the invitation to submit. All empirical analyses use publicly available price and filing data; the ADCI construction methodology is detailed in the Online Appendix. All errors are the author's own.

Competing Interests: The author declares no competing financial interests. This paper represents the author's independent academic research and does not reflect the views of any past, present, or prospective employer or client.

REFERENCES

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets, 5(1)*, 31–56.

Angel, J. J., Harris, L. E., & Spatt, C. S. (2015). Equity trading in the 21st century: An update. *Quarterly Journal of Finance, 5(1)*, 1550002.

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

Bai, J., Philippon, T., & Savov, A. (2016). Have financial markets become more informative? *Journal of Financial Economics, 122(3)*, 625–654.

Biais, B., Foucault, T., & Moinas, S. (2015). Equilibrium fast trading. *Journal of Financial Economics, 116(2)*, 292–313.

Bond, P., Edmans, A., & Goldstein, I. (2012). The real effects of financial markets. *Annual Review of Financial Economics, 4*, 339–360.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *Review of Financial Studies, 22(6)*, 2201–2238.

Chen, Q., Goldstein, I., & Jiang, W. (2007). Price informativeness and investment sensitivity to stock price. *Review of Financial Studies, 20(3)*, 619–650.

Chordia, T., Roll, R., & Subrahmanyam, A. (2008). Liquidity and market efficiency. *Journal of Financial Economics, 87(2)*, 249–268.

Chordia, T., Roll, R., & Subrahmanyam, A. (2011). Recent trends in trading activity and market quality. *Journal of Financial Economics, 101(2)*, 243–263.

De Angelis, D., & Grinstein, Y. (2022). Something to talk about: Information asymmetry and cross-deal learning in corporate acquisitions. *Journal of Finance, 77(4)*, 2399–2437.

Dugast, J., & Foucault, T. (2018). Data abundance and asset price informativeness. *Journal of Financial Economics, 130(2)*, 367–391.

Easley, D., Hvidkjaer, S., & O'Hara, M. (2002). Is information risk a determinant of asset returns? *Journal of Finance, 57(5)*, 2185–2221.

Easley, D., & O'Hara, M. (2004). Information and the cost of capital. *Journal of Finance, 59(4)*, 1553–1583.

Easley, D., Lopez de Prado, M., & O'Hara, M. (2012). Flow toxicity and liquidity in a high-frequency world. *Review of Financial Studies, 25(5)*, 1457–1493.

Farboodi, M., & Veldkamp, L. (2021). Long-run growth of financial data technology. *American Economic Review, 111(8)*, 2485–2523.

Foucault, T., & Moinas, S. (2022). The social efficiency of financial trading. *Working Paper*, INSEAD.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics, 14(1)*, 71–100.

Goldstein, I., Yang, S., & Zuo, L. (2020). The real effects of modern information technologies: Evidence from the EDGAR implementation. *Working Paper*, Wharton.

Green, T. C., Jame, R., & Lock, B. (2022). Trading in the shadows: Why hedge funds follow the "smart money". *Review of Financial Studies, 35(9)*, 4239–4276.

Han, Y., & Lesmond, D. (2011). Liquidity biases and the pricing of cross-sectional idiosyncratic volatility. *Review of Financial Studies, 24(5)*, 1590–1629.

Hasbrouck, J. (1995). One security, many markets: Determining the contributions to price discovery. *Journal of Finance, 50(4)*, 1175–1199.

Hasbrouck, J. (2007). *Empirical Market Microstructure.* Oxford University Press.

Huang, R. D., & Stoll, H. R. (1997). The components of the bid-ask spread: A general approach. *Review of Financial Studies, 10(4)*, 995–1034.

Jain, P. K. (2005). Financial market design and the equity premium: Electronic versus floor trading. *Journal of Finance, 60(6)*, 2955–2985.

Jiang, W. (2017). Have instrumental variables brought us closer to the truth? *Review of Corporate Finance Studies, 6(2)*, 127–140.

Kacperczyk, M., & Seru, A. (2007). Fund manager use of public information: New evidence on managerial skills. *Journal of Finance, 62(2)*, 485–528.

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

Kyle, A. S., & Vila, J. L. (1991). Noise trading and takeovers. *RAND Journal of Economics, 22(1)*, 54–71.

Llorente, G., Michaely, R., Saar, G., & Wang, J. (2002). Dynamic volume-return relation of individual stocks. *Review of Financial Studies, 15(4)*, 1005–1047.

Madhavan, A. (2000). Market microstructure: A survey. *Journal of Financial Markets, 3(3)*, 205–258.

Malinova, K., Park, A., & Riordan, R. (2019). Do retail investors suffer from high frequency traders? *Working Paper*, University of Toronto.

Menkveld, A. J. (2013). High frequency trading and the new market makers. *Journal of Financial Markets, 16(4)*, 712–740.

O'Hara, M. (2015). High frequency market microstructure. *Journal of Financial Economics, 116(2)*, 257–270.

O'Hara, M., & Ye, M. (2011). Is market fragmentation harming market quality? *Journal of Financial Economics, 100(3)*, 459–474.

Pasquariello, P. (2014). Financial market dislocations. *Review of Financial Studies, 27(6)*, 1868–1914.

Rösch, D. M., Subrahmanyam, A., & Van Dijk, M. A. (2017). The dynamics of market efficiency. *Review of Financial Studies, 30(4)*, 1151–1187.

Saar, G., Sun, J., Wei, L., & Ye, M. (2022). Volatility and informativeness. *Working Paper*, Cornell University.

Securities and Exchange Commission. (2023). *Regulation National Market System Proposed Amendments (Reg NMS II).* Release No. 34-96495.

Veldkamp, L. L. (2006). Information markets and the comovement of asset prices. *Review of Economic Studies, 73(3)*, 823–845.

Weller, B. M. (2018). Does algorithmic trading reduce information acquisition? *Review of Financial Studies, 31(6)*, 2184–2226.

Ye, M., Yao, C., & Gai, J. (2013). The externalities of high frequency trading. *Working Paper*, Cornell University.