Master in International Finance

RESEARCH PAPER

Academic Year 2025–2026

Detecting Informed Trading Before Resolution

An Ex-Ante Screen for Insider Activity on Polymarket

Viktor Bock & Julius Guenthoer

Under the supervision of Prof. Daniel Schmidt

President of the Jury: Prof. Olivier Darmouni

Submitted on 1 June 2026

☒ PUBLIC REPORT ☐ CONFIDENTIAL REPORT

Abstract

Insider trading on prediction markets is real, large, and now criminally prosecuted, yet existing detection methods work only after a market resolves. This thesis builds the first ex-ante detector of informed trading on Polymarket: a buy-anchored screen that identifies positioning ahead of an others-driven price jump, using only information available before resolution. A five-rung hypothesis ladder validates the flag. In documented insider cases, the detector surfaces the same wallets about a week before any existing methods. Like all established approaches, it does not prove that flagged traders acted on inside information; its advance is timing, not evidentiary certainty.

Keywords: prediction markets, Polymarket, insider trading, informed trading, forensic finance, ex-ante detection.

JEL Classification: C58, D82, G14, G18

Acknowledgements

The authors thank Prof. Daniel Schmidt for supervision throughout the research and Prof. Olivier Darmouni for serving as President of the Jury. We are also grateful to Prof. Joshua Della Vedova, who kindly sent us the exact classification code and prompt configuration underlying his study, enabling us to replicate his results and build on his methodology. The empirical analysis rests on the public Polymarket on-chain dataset compiled by Z. Wang et al. (2026), which we gratefully acknowledge.[1]

Viktor Bock

HEC Paris
1 Rue de la Libération
78350 Jouy-en-Josas
France

viktor.bock@hec.edu
+49 176 26355618

Julius Guenthoer

HEC Paris
1 Rue de la Libération
78350 Jouy-en-Josas
France

julius-maximilian.guenthoer@hec.edu
+49 1514 6179267

[1] In preparing this paper, the authors used Anthropic's Claude as a writing and analytical aid, including for language refinement, review of analysis code, and discussion of alternative statistical approaches. The authors retained full decision-making authority over research design, methodology, and the interpretation of results, verified all AI-assisted output, and take responsibility for the content of this paper.

Contents

List of tables

List of figures

Introduction

In December 2025 a Polymarket account called AlphaRaccoon turned roughly $2.7 million of bets on Google's annual "Year in Search" rankings into about $1.2 million in winnings, a run so improbable that rival traders concluded on Discord and X that its owner had to be a Google insider. They were right. In May 2026 the U.S. Department of Justice charged Michele Spagnuolo, a Google software engineer, with commodities fraud, wire fraud, and money laundering, alleging that he had used confidential internal search data, accessed through a tool marked "Google Confidential", to bet on figures such as the singer D4vd and Pope Leo XIV before the rankings were public (United States Department of Justice, 2026a). "Unlike the counterparties to his trades," the complaint stated, "Spagnuolo knew the outcome of these wagers before the trading public did."

Spagnuolo's was neither an isolated case nor the first. A month earlier the Department of Justice had charged U.S. Special Forces soldier Van Dyke with using classified information about the operation to capture Nicolás Maduro to win some $400,000 on Polymarket (United States Department of Justice, 2026b), and over the preceding year investigators and journalists had documented suspected insider trading around the February 2026 US–Israel strike on Iran, the 2025 Nobel Peace Prize, and Taylor Swift's engagement (Eckert and Fouche, 2026; Mitts and Ofir, 2026). These episodes accompanied the rise of Polymarket and its rival Kalshi to mainstream venues on which billions are now wagered on elections, government actions, and corporate events (Tavva, 2026). They have also drawn a response: the first criminal prosecutions of prediction-market insider trading, Congressional scrutiny, a Senate vote to bar its own members from these markets, and rulemaking by the Commodity Futures Trading Commission (Commodity Futures Trading Commission, 2026). Insider trading on prediction markets is, in short, real, large, and now a prosecutorial and regulatory concern.

Prediction markets are unusually exposed to it. A binary contract pays one dollar or nothing, so an agent who knows an event's outcome holds an almost risk-free position rather than the noisy edge typical of equity markets; outcomes resolve on discrete, datable moments around which contract value shifts instantly; and on a decentralised venue like Polymarket anyone can trade pseudonymously from a freshly created wallet, without the disclosure or identity requirements that govern regulated markets. The very features that make these markets efficient aggregators of information also make them efficient vehicles for trading on information that is not yet public.

Yet every case above was caught the same way: after the fact, by traditional investigation. Van Dyke was found through nondisclosure agreements and fund transfers, Spagnuolo in part through other traders' suspicion. The academic detection literature that has grown up alongside these markets shares the same limitation. Each published method, whether it scores realised bet size and profitability (Mitts and Ofir, 2026), account-level fingerprints (Liu, 2026), or excess forecasting accuracy (Della Vedova, 2026a), keys on a market's resolved outcome or its post-resolution fund flows, and so cannot flag a wallet until its market has settled, often weeks or months after the trade. Surveillance is therefore forensic and not preventive: it reconstructs who likely traded on private information once the money has been made, rather than flagging the act while there is still time to respond.

This thesis asks whether that order can be reversed: whether informed trading on prediction markets can be detected ex ante, from the anticipatory act and before a market resolves, and what such a detector identifies. The choice of *informed* over *insider* trading is deliberate: trading on private information leaves

an observable footprint in prices, but whether that information was material and non-public, the legal test for insider trading, turns on a question of intent that market data alone cannot settle. It builds, to our knowledge, the first ex-ante, act-level detector of informed trading on a prediction market. Its advance is one of timing, not coverage. It does reach traders the existing methods miss, but the point is that it flags them earlier, on the buy itself, not on the outcome the buy anticipates. Even the very wallet an ex-post method would eventually flag, this detector surfaces before the market resolves, a flag determinable about a day after the trade instead of months after settlement. That lead is the difference between preventing misconduct and reconstructing it.

The detector adapts a paradigm from equity-market event studies, where informed trading leaves its footprint as accumulation ahead of a discrete price jump, to the binary-contract price path. It is buy-anchored: it begins from a single purchase and asks whether that purchase preceded a material, others-driven price move that the buyer neither caused nor followed, applying a set of six gates, conditions that must all hold, evaluated only on information available before the market resolves and never on the final settlement value. It is offered as one transparent configuration, a proof of concept that such detection is feasible, not a tuned or optimal instrument.

We validate the detector through a five-rung hypothesis ladder, each rung removing a benign explanation for the flag, and then characterise what it surfaces. The flag proves to be a real timing signal rather than a volume artefact; it converges with three independent ex-post detectors while reaching traders they miss; its edge is specific to the anticipatory act rather than to the trader; and it concentrates, robustly in direction, in markets where non-public information can plausibly exist. The population it surfaces is led by sophisticated, active accounts trading consequential,information-rich events. Against a set of documented insider cases, including AlphaRaccoon, the detector flags the same wallets a median of about a week before any outcome an ex-post method could act on, and over two weeks earlier on individual markets. Throughout, it is a screen that surfaces candidates for scrutiny and not an adjudicator of intent: pre-resolution price behaviour cannot prove that a flagged trader held material non-public information, only that the trade was consistent with trading on private information.

The thesis proceeds as follows. Chapter 2 describes prediction markets and their regulatory setting. Chapter 3 reviews the theory and evidence on informed trading, the existing detection methods, and the equity-market literature the detector draws on, and develops the hypotheses. Chapter 4 sets out the data and the detector. Chapter 5 reports the validation ladder and the characterisation, and Chapter 6 concludes.

Prediction markets: An overview

2.1 What are prediction markets?

Prediction markets are financial exchanges in which participants trade contracts whose payoffs depend on the outcomes of future events. In their most common form, a prediction market offers a binary contract that pays one dollar if a specified event occurs and zero otherwise. The market price of such a contract can be read as the crowd's aggregate estimate of the probability that the event will happen (Wolfers and Zitzewitz, 2004). If a contract asking "Will Candidate X win the election?" trades at 53 cents, the market

collectively assigns roughly a 53% chance to that outcome.

2.2 A brief history

Prediction markets have a longer history than their recent prominence suggests. The intellectual foundation traces to Hayek (1945), who argued that prices aggregate knowledge dispersed across many individuals, an idea Arrow et al. (2008) extended explicitly to prediction markets as forecasting tools. The Iowa Electronic Markets (IEM), launched in 1988, were the earliest modern implementation: Forsythe et al. (1992) found that IEM contracts on the 1988 U.S. presidential election outperformed opinion polls, a result replicated across subsequent cycles (Berg et al., 2008). Intrade and TradeSports extended the model to a wider range of events during the 2000s. Rothschild and Sethi (2015) documented diverse trading strategies and suggestive evidence of manipulation by a single large trader on Intrade's 2012 presidential market, foreshadowing recurring concerns on modern platforms.

Two developments brought prediction markets into the financial mainstream. In November 2020, the U.S. Commodity Futures Trading Commission (CFTC) licensed Kalshi as the first federally registered designated contract market for event contracts, permitting it to operate without the stake limits that constrained PredictIt and the IEM (Burgi et al., 2025). The same year, Polymarket was launched as a globally accessible, blockchain-settled venue on the Polygon network, allowing trading without any intermediaries.

The scale of growth has been extraordinary[1], driven in particular by the 2024 U.S. presidential cycle, in which Polymarket's markets on Donald Trump's election became a cultural and media reference point and triggered a step-change in volumes. Polymarket processed more than $13 billion in cumulative trading volume by early 2025 and reached $12 billion in monthly volume in January 2026 (Tavva, 2026). Monthly active accounts on Polymarket grew from approximately 1,600 in December 2023 to over 519,000 by December 2025 (Gomez-Cram et al., 2026), making it a significant financial market in its own right.

2.3 Market participants

The rapid growth of prediction markets has attracted a diverse participant base, but trading activity is distributed highly unevenly. In December 2025, the median active Polymarket account traded just $72, while the 99th percentile account traded $74,000, a ratio exceeding 1,000 to one (Gomez-Cram et al., 2026).

The roles participants occupy in the market structure are common across both Polymarket and Kalshi. *Makers* post resting limit orders or quotes that other participants can execute against; they earn rebates and supply the liquidity on which the market depends. *Takers* execute against existing orders or quotes and pay fees in return for immediate execution. Beyond this maker-taker distinction, participants vary widely in sophistication and intent. *Retail traders* form the largest population by headcount but contribute disproportionately small volume per account. Profit distribution reflects this heterogeneity. E. Wang (2026), studying the 273 most-analyzed traders on Polymarket, finds that the top decile of profitable wallets captures 52% of aggregate positive profit ($98 million of $189 million), while the top quintile captures 89%. Overall, only 30% of Polymarket traders earn positive returns at all (Reichenbach and

[1] Daily updated Polymarket activity data is available on https://dune.com/filarm/polymarket-activity

Walther, 2025). *Professional arbitrageurs and market makers* operate continuously across markets, often using automation; E. Wang (2026) reports that 67% of analyzed wallets show at least partial automation, although the most profitable cohort consists of "likely human" traders using fast interfaces. *Whales*, concentrated trader-investors with large positions, dominate volume and price discovery (Abid, 2026; Cheong and Tamayo, 2026), and a subset of whales operate informed strategies that motivate the detection literature that follows. *Informed traders*, defined here as participants who hold material non-public information about contract outcomes, are the population this thesis is centrally concerned with detecting.

2.4 How prediction markets work

Contract types. Prediction market contracts are event derivatives whose payoffs are tied to discrete real-world outcomes. The simplest and most common is the *binary contract*: two mutually exclusive outcome shares, YES and NO, that resolve to either $1 or $0 upon the event's conclusion. A trader who believes an event is more likely than the current price implies buys YES shares; a trader who disagrees buys NO shares or sells YES shares. Since exactly one outcome must occur, the prices of a YES and a NO share in the same market sum to $1 in equilibrium, and any deviation creates a risk-free arbitrage opportunity (Saguillo et al., 2025). *Categorical contracts* extend the same logic to events with more than two outcomes, with each candidate represented by a separate token and only the winning outcome paying $1. Unlike equity shares, which trade against a fixed supply issued by a firm, prediction-market shares are not pre-issued: new YES/NO pairs are *minted* on demand when two traders take opposite sides of a market, and *burnt* when complete pairs are redeemed against collateral, so the outstanding share count expands and contracts with trading activity.. The full token-level mechanics are set out in Appendix A.

Trading infrastructure. Polymarket combines the speed of a centralised exchange with the transparency of blockchain settlement through a two-layer hybrid architecture (Gomez-Cram et al., 2026; Tsang and Yang, 2026). The first layer is an off-chain Central Limit Order Book (CLOB) in which cryptographically signed order messages are matched by price-time priority. The second layer is on-chain settlement on the Polygon blockchain: matched orders are submitted to exchange smart contracts that atomically execute the trade and record it on-chain (Tsang and Yang, 2026). Every settled trade is therefore publicly visible, creating a fully auditable ledger that earlier prediction-market platforms did not offer. This transparency is what makes transaction-level academic research on these markets possible, and distinguishes Polymarket from platforms such as Intrade, where researchers had access only to aggregate price and volume data (Rothschild and Sethi, 2015). The fee structure and a comparison with Kalshi's contrasting CFTC-regulated, identity-verified, and centralized design are deferred to Appendix A.

Outcome resolution. Once an event concludes, the market must determine which outcome occurred so that winning shares can be redeemed and losing shares expire worthless. Polymarket uses the Universal Market Access (UMA) Protocol, a decentralised verification system. After an event concludes, an outcome is proposed by a reporter and submitted to the UMA oracle network, where independent token holders vote on its validity. If no dispute is raised within a specified challenge period, settlement proceeds automatically; otherwise a formal challenge is triggered before final resolution (Gomez-Cram et al., 2026). The decentralised design makes it costly for any single party to manipulate resolution, but it is not immune to capital concentration: documented governance disagreements on disputed multi-million-dollar markets

are discussed in Appendix A.

2.5 Regulatory landscape

Prediction markets sit between two legal categories. An event contract is economically a binary option, which places it under financial-derivatives law, but functionally a fixed-odds wager, which places it under gambling law. Jurisdictions have resolved the duality differently and largely through enforcement rather than legislation. The U.S. Commodity Futures Trading Commission (CFTC) has gradually accepted event contracts as "swaps" under the Commodity Exchange Act, allowing Polymarket to resume U.S. operations through a CFTC-licensed exchange in December 2025 (Commodity Futures Trading Commission, 2025). France, Belgium, Italy, Poland, Romania, and Spain have instead treated the same product as unlicensed gambling and ordered geo-blocks, with the European Markets in Crypto-Assets Regulation (MiCA) leaving classification of event-contingent derivatives to Member-State law (Norton Rose Fulbright, 2026; *Polymarket Blocks French Traders Amid Gambling Inquiry* 2024).

Regulatory development has tracked specific incidents. The French ban followed media coverage of a trader who realised approximately $80 million on the 2024 U.S. presidential election; in the United States, the Southern District of New York brought the first two criminal prosecutions[2] of prediction-market insider trading in 2026 after the Van Dyke and Spagnuolo cases became public (Sorkin et al., 2026; United States Department of Justice, 2026a,b). Detailed case descriptions can be found in Appendix G. The gap between observed and prosecuted conduct is nevertheless striking: Mitts and Ofir (2026) estimate roughly $143 million in aggregate anomalous profits across over 210,000 flagged wallet-market pairs, against only these two indictments. The CFTC's March 2026 Advance Notice of Proposed Rulemaking contemplates surveillance and information-asymmetry rules (Commodity Futures Trading Commission, 2026).

Polymarket has begun to police insider trading itself. Its terms of service prohibit trading on stolen confidential information, on illegal tips, and by any user able to influence the outcome (Polymarket, 2026b). In early 2026 the platform partnered with Palantir Technologies (deploying its Vergence AI engine for real-time anomaly detection) and Chainalysis (extending on-chain forensics into compliance workflows) (*Polymarket Partners With Palantir to Combat Insider Trading in Prediction Markets* 2026; TradingView, 2026), and reports having shared over 315 wallet identifiers with authorities, contributing to the arrests of *Van Dyke* and *Spagnuolo* (Polymarket, 2026a).

Literature review and hypotheses

3.1 Literature review

3.1.1 Why prediction markets are vulnerable to insider trading

The theoretical susceptibility of financial markets to informed trading is well established. Kyle (1985) shows that a strategic insider trading on private information leaves a measurable footprint in prices and

[2]Insider trading on Polymarket is reached through CFTC anti-fraud authority over swaps: CEA § 6(c)(1) and Regulation 180.1 mirror SEC Rule 10b-5, and CEA § 4c(a)(4) prohibits federal employees from trading on government non-public information (CFTC, 2011; Congressional Research Service, 2025; United States Congress, n.d.).

order flow, profiting without fully revealing it, and Glosten and Milgrom (1985) show that the bid-ask spread arises endogenously as the market maker's protection against adverse selection, the risk that the counterparty on any given trade possesses superior information.

Prediction markets inherit these vulnerabilities but add features that make them especially acute. First, information advantages are often *absolute* rather than relative: an insider who knows a geopolitical event will occur within 48 hours effectively holds a binary \$1 or \$0 payoff, a far cleaner edge than the noisy signals typical in equity markets. Second, the event-driven structure of prediction markets creates sharp information arrival points (diplomatic decisions, corporate announcements, military operations) around which contract value can shift instantly. Third, and most importantly for detection, decentralised platforms like Polymarket allow pseudonymous participation. Unlike regulated equity markets with mandatory identity disclosure, traders on Polymarket need only a blockchain wallet, which can be created in seconds at no cost and with no identity verification (Liu, 2026; Tavva, 2026). This pseudonymity enables Sybil attacks, in which a single operator disperses informed trades across many wallets to conceal both identity and the scale of their position.

The legal and regulatory context amplifies this vulnerability. Ashar (2025) compares insider-trading regulation across securities, commodities, and gambling and finds prediction markets deficient on all three; unlike equity markets, they impose no disclosure obligations or trading windows on insiders (Ashar, 2025; Beylin, 2026).

3.1.2 Empirical evidence of anomalous trading behaviour

The theoretical vulnerability described above is not merely hypothetical. A growing body of empirical work has documented suspicious trading behaviour on prediction markets at a scale that has attracted both academic and prosecutorial attention. Tavva (2026) observes that the rapid expansion of users and volume has outpaced the development of compliance infrastructure on decentralised platforms, leaving a widening gap between trading activity and the surveillance resources available to detect misconduct.

The most prominent legal case emerged in April 2026, when the U.S. Department of Justice charged a U.S. Army Special Forces soldier with using classified intelligence about the U.S. operation to capture Nicolás Maduro in Venezuela to earn \$409,000 on Polymarket (United States Department of Justice, 2026b); tellingly, it was detected through traditional investigation (nondisclosure agreements, military photographs, suspicious fund transfers), not market surveillance. Beyond individual cases, Mitts and Ofir (2026) provide the most comprehensive case-study analysis to date, documenting episodes from the February 2026 US–Israel strike on Iran to pre-announcement trading ahead of Taylor Swift's engagement, and estimate roughly \$143 million in aggregate anomalous profit across more than 210,000 wallet–market pairs. Abid (2026) documents the scale of concentration: a single French trader earned roughly \$85 million on the 2024 presidential election.

Della Vedova (2026a) studies 222 million settled Polymarket trades and identifies 6,032 of 450,048 wallets as likely informed, which together extract about \$150 million. Liu (2026) reports 518 markets in which his classifier detects at least one insider; although these flagged wallets constitute only 0.08% of all participants, they account for 7.1% of total winning profits. Importantly, Liu (2026) shows that market makers absorb the loss, earning 18.5 cents less per dollar of volume in insider-affected markets, in line with the Kyle-Glosten adverse selection mechanism.

3.1.3 Detection methods

The empirical findings above have been generated by a range of detection methodologies that differ in their theoretical foundations, data requirements, and the type of informed trading they target. We describe in detail the methods we reimplement and benchmark in this thesis and then summarise the remaining approaches. A parallel commercial ecosystem of on-chain forensics and whale-tracking services (Arkham, Chainalysis) provides related tools used by traders and journalists; we follow the academic literature in treating these as auxiliary infrastructure rather than detection methods in their own right.

Behavioural screening and on-chain analysis

Mitts and Ofir (2026) take a practitioner-focused perspective, developing a composite screening measure that aggregates several behavioural signals into a single suspiciousness score for each wallet–market pair. They use a weighted combination of indicators, including anomalies in bet size (both cross-sectional and within-trader z-scores), realised profitability, the clustering of trades before events, and directional concentration. Traders flagged by this procedure exhibit a win rate of 69.9%, far above what would be expected by chance.

Liu (2026) develops a related but distinct behavioural classifier based on five trader-market fingerprints: directional conviction (one-sided rather than hedged exposure), low VWAP entry price (buying before the price moves), fresh wallet age at first trade (wallet created shortly before the event), concentration on few markets, and large realised profit. The classifier then adds a Sybil-cluster propagation step that traces on-chain USDC fund flows (via Etherscan and Relay.link) to link wallets sharing a common non-exchange withdrawal destination: in the largest documented cluster, 37 of 43 known Iran-strike insider wallets withdrew to one counterparty receiving $2.08 million in aggregate flow, evidence that a single operator can control dozens of nominally independent accounts, a coordination structure that wallet-level screens miss. In a related on-chain vein, Cheong and Tamayo (2026) trace wallets' incoming USDC and find that the largest whales are disproportionately funded entirely from untraceable, non-exchange sources, and that such funding opacity predicts higher accuracy among them, consistent with concealment by informed traders.

We explicitly replicate Mitts and Ofir, 2026's and Liu, 2026's behavioural scans to benchmark our ex-ante detection method. The exact signal constructions, eligibility filters, and aggregation rules for both methods are reproduced in Sections 4.4.1, D.1, 4.4.2, and D.2.

Structural tests: The orthogonality approach

Della Vedova (2026a) develops the most theoretically grounded detection method in the literature. The key insight is that trading performance can be decomposed into *forecasting accuracy* (whether the trader predicts the paying side of a binary contract) and *execution quality* (whether the trader enters at a better-than-average market price). For uninformed traders, including skilled analysts using public information, these dimensions should be statistically *orthogonal*: a trader may predict outcomes well or time entries well, but the two skills need not co-move. For informed traders, by contrast, the same private signal identifies both the correct side and the current price as wrong, generating a positive correlation between accuracy and execution. The flag itself is a wallet-level test of excess directional accuracy against the price-implied benchmark; the positive accuracy-execution co-movement is the population-level signature

that validates the flagged set as informed rather than merely lucky. Della Vedova (2026b) finds that for uninformed traders, forecasting accuracy and execution quality are nearly independent, sharing less than 1% of variance for human traders. The full implementation we replicate is given in Sections 4.4.3 and D.3.

Other approaches

Two further methods shift the unit of analysis from the wallet to the market, and we do not reimplement either. Nechepurenko (2026a)'s Information Leakage Score quantifies, for a resolved binary market, the fraction of total price movement that had already occurred before the relevant news became public (via a Murphy decomposition of the Brier score); a high score is consistent with informed traders acting ahead of the announcement, but the measure is computable for only a small fraction of markets. Zhu et al. (2026) merge low- and high-frequency time-series anomaly detection (gradual position-building and abrupt pre-announcement jumps) into a market-level anomaly score weighted around manually curated events, but provide no implementation or empirical evaluation, so we treat it as a conceptual framework rather than a replicable detector.

3.1.4 Insider trading and price jumps: Lessons from equity markets

The detection methods reviewed above were all developed in the past six months in response to large-scale prediction markets. Their novelty risks obscuring a much older empirical and theoretical literature in equity finance that studies the same core phenomenon: the footprint of informed trading around discrete price moves. This literature provides both the intellectual basis for the thesis's methods and the empirical regularities guiding its design. This section organises the equity-market work into three threads: (1) sudden price jumps as information-arrival events; (2) systematic pre-event run-up as an empirical signature of informed trading; and (3) what does and does not translate from this body of work to prediction markets.

Price jumps as information-arrival events

Modern asset-pricing econometrics distinguishes diffusive from jump components of returns: small continuous moves reflect order-flow noise and gradual updating, while discrete jumps reflect material information. Using high-frequency data, Barndorff-Nielsen and Shephard (2006) and Lee and Mykland (2008) show that identified jumps cluster around information arrivals, government macroeconomic announcements and firm-specific news, rather than occurring randomly. Lee (2012) sharpens this for detection: individual-stock jumps are predictable from scheduled information events, and jump probability is elevated in the window immediately *before* scheduled earnings releases, which she reads as information leakage ahead of public announcements. The corollary that guides this thesis is that a jump preceding a public event is itself a signature of leaked information.

Pre-event run-up and the timing of informed trading

The pre-event window is the primary setting for studying informed trading in equities, and four findings shape our design. Keown and Pinkerton (1981) document significant abnormal-return run-up beginning about twelve trading days before merger announcements, attributed to leakage among the advisers involved. Meulbroek (1992) supplies the clearest validation, linking that run-up to proven illegal trading: in SEC enforcement cases, nearly half of the pre-announcement run-up occurs on identified insider-trading days,

so the run-up is largely driven by insider activity rather than merely correlated with it. Marin and Olivier (2008) show the pattern is asymmetric, insider *purchases* spike in the month before a large positive jump, with the link strongest and the pre-event window shortest for positive moves; this motivates anchoring the detector on a buy and reading the move in that buy's own direction (its one-sided jump gate), the binary-market analogue of purchases ahead of a favourable jump. Augustin et al. (2019), examining option trading before 1,859 takeovers, find abnormal pre-announcement volume concentrated in short-maturity out-of-the-money calls and conclude, after ruling out public-information explanations, that more than half is unlikely to be public-information-based. Two implications follow: informed traders choose the venue that maximises leverage on their edge (equity options there, binary-payoff contracts here), and abnormal pre-event positioning, once non-informational explanations are eliminated, identifies informed trading without observing the traders directly. Our detector is built on this principle.

What translates, and the ex-ante gap

The equity-market literature supplies two transferable ideas. First, the framing of jumps as information-arrival events (Barndorff-Nielsen and Shephard, 2006; Lee and Mykland, 2008) carries over: binary contracts likewise reprice discretely in response to private information about resolution. Second, the regularity that informed trading clusters in the pre-event window (Augustin et al., 2019; Keown and Pinkerton, 1981; Marin and Olivier, 2008; Meulbroek, 1992) makes the immediate pre-jump period the natural window for ex-ante detection. The one feature that does not carry over is infrastructure: the equity run-up literature leans on disclosure and enforcement records (Section 16 filings, SEC cases, Schedule 13D) that prediction markets lack. But the on-chain trade record supplies what that infrastructure cannot, a complete, public history of every position, so the lessons transfer conceptually and the detector is built natively from order flow rather than transplanted from equity estimators.

This defines the contribution of the present thesis. Existing prediction-market detection methods are, without exception, *ex post*: each computes its signal only after a market resolves. The three we replicate key on realised outcomes, the composite screens of Mitts and Ofir (2026) and Liu (2026) on realised size, profitability, and account fingerprints, and Della Vedova (2026a) on realised accuracy. The two methods that gesture toward earlier detection do not close the gap: Nechepurenko (2026b)'s information-leakage score is intended for real-time use but is computable for only 0.7 percent of markets and Zhu et al. (2026) provide no implementation. The toolkit therefore identifies who likely traded on private information only after the fact, leaving platforms and regulators to act forensically rather than preventively. Liu (2026) explicitly names pre-event jump trading as a direction for future research, and to our knowledge no published work has implemented it. This thesis does so: an ex-ante detector that translates the equity-market jump-and-run-up paradigm to prediction markets, combining a fixed-magnitude jump threshold on the binary-contract price path (Lee and Mykland, 2008) with a buy-anchored screen of wallets positioning in the 24-hour window before the jump.

3.2 Hypothesis development

The research question of this paper, namely whether informed trading on prediction markets can be detected ex-ante through pre-event jump anticipation and what such a detector identifies, splits into two parts. The first is a methods-validation question and is the substantive content of this section. The second

is exploratory and is addressed through descriptive characterisation in Chapter 5.

These hypotheses apply the theory of informed trading to a new detection problem. Microstructure models (Easley et al., 1996; Glosten and Milgrom, 1985; Kyle, 1985) characterise the informed agent as positioning ahead of information events when an asymmetry exists; from this we derive the observable conditions an ex-ante flag must satisfy. Because insider trading is defined by an unobservable intent, the use of material non-public information, and no prediction-market dataset carries ground-truth labels (a constraint shared by the benchmark detectors of Mitts and Ofir (2026), Della Vedova (2026a), and Liu (2026)), we cannot test intent directly. Instead we organise the test as a ladder that eliminates benign explanations in turn: a flag may appear because it is (i) a mechanical artefact of trading volume, (ii) statistically unrelated to the informed-trading construct captured by existing ex-post detectors, (iii) a by-product of generic trading skill, (iv) driven by forecasting a gradual price drift rather than reacting to an information event, or (v) forecasting ability whose edge does not depend on private information. Each hypothesis removes one rung; a flag that passes all five is a construct-valid, behaviour-specific, information-driven, and information-dependent signal, the strongest informed-trading case observational prediction-market data permit. The final step, to demonstrated use of material non-public information, requires ground truth on intent and is the epistemic ceiling, taken up in the conclusion (Chapter 6).

3.2.1 H1: The signal is real

Flagged wallets position before jumps at a rate exceeding that produced by their own trading under random timing. The null is that the pre-jump concentration is a mechanical artefact of trading volume. Kyle-style models depict the informed agent positioning non-randomly with respect to information arrival, splitting orders ahead of public revelation to limit price impact (Kyle, 1985); if timing carries information, then *when* a wallet bought, not merely how much it traded, must place it ahead of jumps beyond what activity alone would imply.

3.2.2 H2: The signal is construct-valid

The cohort flagged by the ex-ante detector overlaps the cohorts flagged by the established ex-post detectors far above chance. The null is that the ex-ante flag is statistically independent of the benchmark detectors' flags; we additionally report, as a descriptive matter, the share of flagged wallets that no benchmark identifies (incremental coverage). Convergent validity requires a new instrument to correlate with established measures of the same construct, so above-chance overlap shows the flag surfaces the same traders that detectors keying on realised outcomes independently regard as informed, while the non-overlapping share shows the detector extends the identified population from its distinct, earlier vantage point rather than re-deriving the existing flags.

3.2.3 H3: The edge is behaviour-specific, not generic skill

Within flagged wallets, performance is concentrated in their flagged anticipatory pairs and not in the rest of their trading. The null is that flagged wallets earn a uniform edge across their activity, consistent with generic skill. In the microstructure tradition (Glosten and Milgrom, 1985; Kyle, 1985) the informed agent's edge is event-driven, realised on the trades that exploit a specific information asymmetry rather than as a persistent, general skill, so a within-wallet comparison that holds each trader's overall ability

fixed localises the signal to the act of anticipation rather than to the type of person.

3.2.4 H4: The anticipation is information-driven, not drift-forecasting

Flagged traders distinctively anticipate sudden, discrete repricings: the anticipated moves are more discrete than those preceding comparable unflagged anticipations in the same markets. The null is that they are no more discrete than this baseline, so that discreteness is a property of large moves in general rather than a mark of the flag. The high-frequency jump-detection literature treats a discontinuous price move as the empirical footprint of discrete information arrival (Lee and Mykland, 2008), so a discrete share in excess of the market-matched baseline separates reacting to an information event from extrapolating a forecastable drift; the test is posed comparatively because large moves in these markets may be generically discrete.

3.2.5 H5: The signal is information-dependent

The flag concentrates in markets in which material non-public information plausibly exists, relative to markets resolved by purely public data. The null is that flag intensity is independent of a market's information-richness. Informed-trading theory ties the informed trader's advantage to the existence of an information asymmetry (Glosten and Milgrom, 1985; Kyle, 1985): an edge indifferent to whether exploitable non-public information exists is forecasting skill, whereas an edge conditional on it is informed trading, making H5 the closest observational test of that conditionality and the closest the data come to the insider question itself.

3.2.6 Characterisation questions

The second part of the research question, namely *what* the detector identifies, is exploratory and is treated separately. We address it through descriptive characterisation rather than tests against a null, in order to avoid post-hoc framing of patterns that emerge in the data.

C1 (who?) The composition of the flagged cohort by trader type, classified using the Della Vedova (2026a) wallet taxonomy.

C2 (what?) The composition of the flagged events by market category, economic magnitude, and concentration across the wallet distribution.

C3 (why this population?) Whether the composition reflects genuine informed concentration or the mechanics of the detector's gates, decomposed against the size-floor and exposure-rate channels.

C4 (which documented cases?) Whether the detector flags a set of highly-likely insider wallets, and which trader-type population and behavioural archetype the caught and missed cases fall into.

C5 (implication?) What the composition and recall pattern imply for the detector's role: a screen for consequential informed anticipation, a superset of insider trading in precision but a subset of it in recall, surfacing rather than adjudicating.

3.2.7 From hypotheses to interpretation

If H1–H5 hold, the detector isolates a real, construct-valid, behaviour-specific, information-driven signal concentrated where private information can exist: the strongest informed-trading case that observational

prediction-market data permit. The remaining step, from informed-consistent anticipation to demonstrated use of material non-public information, requires ground truth on intent and lies beyond the paper's claim. The detector identifies informed-consistent anticipation and surfaces a candidate population; it does not determine insider status. Its role, and its relation to the complementary ex-post tools of Mitts and Ofir (2026), Della Vedova (2026a), and Liu (2026), are discussed in Chapter 6.

Data and methodology

4.1 Sample construction

The analysis rests on the raw on-chain record rather than Polymarket's curated API. We use "OrderFilled" events emitted by the two Polymarket exchange contracts on the Polygon blockchain, the CTF Exchange and the NegRisk Exchange, obtained from the public dataset compiled by Z. Wang et al. (2026), which reconstructs the full Polymarket on-chain history from these events.[1] We draw our universe from this record and filter it down as described below. The record preserves, for each fill, the maker and taker addresses, the trade direction, the price, the size, and the token side, so that taker-side activity can be isolated without recourse to any platform-level aggregation. Market metadata and the category tags applied at stage F8 are drawn from the same snapshot, supplemented by event tags retrieved from Polymarket's Gamma API. We work from a frozen snapshot dated 2026-05-04, which fixes the universe and makes the construction reproducible. This universe is the sample on which the MNPI taxonomy (4.2) and the trader-type taxonomy (4.3) are subsequently defined.

One row is a single fill: one matched maker–taker pair recorded by an OrderFilled event. A single market order can match several resting limit orders within one blockchain transaction, so one trading decision can generate several fills. The in-scope record comprises 63,343,761 fills, corresponding to 42,854,686 distinct transactions, a mean of 1.48 fills per transaction, with 18.5% of transactions multi-fill and exactly one taker per transaction. The NegRisk router decomposes a single economic trade across complementary legs, which we collapse to 32,702,089 economic trades. This normalisation preserves the maker and taker roles, so the trader-type taxonomy (4.3) can count intentional taker-side decisions.

The analysis universe is constructed by nine filters, F1–F9, applied to the full snapshot (F0) and reported in Table 1. A *market* is a single binary YES/NO leg, and an *event* is its parent (`event_id`), which may group several markets into one economic question. Each stage acts only on units that pass all previous stages, so the table forms a funnel from the full snapshot to the final sample. The filters reduce the snapshot from 1,019,069 markets across 437,231 events to 26,791 binary markets across 7,168 events, or 2.6% of the original market count.

[Table 1 about here.]

The filtering serves four design goals rather than nine independent criteria. *Determinacy* (F1 and F3) retains only resolved markets with a clean binary YES/NO outcome, so that every market carries a determinate result for the downstream analysis. *Tradeability* (F2 and F4 to F7) restricts to markets created from 2024-02-01 onward that cleared non-trivial liquidity floors and ran long enough for informed

[1] The dataset reproduces the on-chain event log directly; we use it in place of node access so that the universe is fully reproducible from a fixed artefact.

positioning to appear, within parent events liquid enough to carry reliable category tags. The 2024-02-01 cut-off follows Mitts and Ofir (2026) and brackets the documented insider-trading episodes the study examines; it also excludes the platform's single largest market, the 2024 US presidential-election winner, whose volume would otherwise dominate the sample.[2] *Information scope* (F8) restricts the universe to eight subject categories (Politics, Elections, Finance, Business, Economy, Geopolitics, Tech, and Entertainment) and vetoes sports, weather, and single-tweet markets. These categories are retained because their resolution tracks frequent, widely reported developments that attract dense news coverage and the continuous public-information flow sustaining active price discovery; they are selected on subject matter and news intensity, not on any presumption of greater susceptibility to informed trading. That question is decomposed within the retained sample by the MNPI taxonomy (4.2), and the exclusion of sports here is why the taxonomy's Performance class is small by construction. *Structural validity* (F9) drops partial NegRisk bundles, so that every retained multi-leg event is economically complete.

The retained sample carries $20.67 billion in volume and trade timestamps from 2024-02-01 to 2026-05-04. The 63.3 million fills above are executed by 1,696,645 unique wallets across 16,753,150 wallet–market pairs; a wallet is counted whether it appears as maker or taker (1,669,389 as takers and 637,881 as makers, with overlap), and market makers are not excluded. The large majority of markets resolve NO, a YES rate of 23.9% (6,401 of 26,791). Two markets report metadata volume but have no on-chain trades and enter the market-level descriptives only, so all trade-level quantities are computed over the 26,789 traded markets.

Table 2 describes the sample across the eight categories. Volume is concentrated in the political and geopolitical questions, with Geopolitics, Politics, and Elections together accounting for roughly 57% of total volume. Participation follows a related but distinct ordering: Table 3 shows that Business, Politics, and Geopolitics dominate participation by both wallet–market pairs and unique wallets, while Finance is the thinnest category on either measure. Most wallets are narrow in scope: Table 4 shows that 36.1% trade in a single category, with breadth tapering steadily to the 1.3% active across all eight, a dispersion that motivates the market-concentration (HHI) feature used in the trader-type taxonomy (4.3).

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

4.2 MNPI taxonomy

We adopt the market-level MNPI taxonomy of Della Vedova (2026a), replicating his four-category classification procedure on our analysis universe[3]. Each market is classified by the *information-generating process* that determines its outcome, not by topic. The taxonomy orders markets by the structural plausibility that material non-public information (MNPI), namely information that is material to the resolution, not yet public, and concentrated in a small and identifiable set of agents, is held by some party ahead of the public. The ordering is ordinal and ex-ante: each label is assigned from the question text alone, before any trade data is observed.

[2] Created January 2024, so outside the window; at roughly $3.7 billion across its legs it is the platform's largest market and would otherwise dominate every volume-weighted statistic

[3] We are grateful to Professor Della Vedova, who kindly sent us the exact classification code and the verbatim prompt configuration underlying his study. This allows us to reproduce his procedure precisely rather than reconstruct it from the paper; our only deviations, three in total, are documented in Appendix B.

Four categories result, ordered as a monotone descent in the structural plausibility of concentrated MNPI:

Vote. Markets resolved by a sealed ballot or committee decision: elections, party nominations, award-jury votes, confirmation or no-confidence votes, reality-television eliminations. A defined finite set of voters or panelists learns the result before the public, making this the category in which concentrated MNPI is most plausible.

Action. Markets resolved by a strategic decision taken by a few named actors: regulatory approvals, executive orders, central-bank rate decisions, mergers and acquisitions, executive appointments, indictments, treaties. Plausible but narrower than Vote: the deliberating group knows the outcome before announcing it.

Performance. Markets resolved by open competition whose result emerges during play: tournaments, athletic records, head-to-head contests. Structurally implausible: no party controls or learns the outcome before it occurs, and the determining event is public as it unfolds.

Stochastic. Markets resolved by aggregate processes that no single agent controls: cryptocurrency prices, asset and index levels, macroeconomic releases, polling aggregates, weather, viewership figures. Least plausible: the outcome is the product of many dispersed actors, and concentrated MNPI about the resolution is structurally absent.

Vote/Action form the *MNPI-admitting* class; Performance/Stochastic form the *MNPI-resistant* class.

Less plausible, not impossible. The taxonomy grades plausibility and concentration, not strict possibility. Informed trading is not claimed to be impossible in Performance or Stochastic markets. A whale intending to liquidate a large Bitcoin position holds genuinely private, price-moving information; that information, however, concerns their own order flow rather than foreknowledge of the fundamental resolution, and it is dispersed across a large anonymous population rather than held by a small identifiable committee. The same logic applies to an analyst with a superior macroeconomic forecast or an early read on polling crosstabs: the edge is anticipatory or statistical and widely contestable, not discrete insider leakage. The distinction the taxonomy targets is therefore between private information about the fundamental that determines resolution (Vote and Action) and private information about order flow or superior reading of public signals, which can arise anywhere. The taxonomy is an ordinal prior about where concentrated insider trading should be most detectable, not a filter that asserts where it can occur; the empirical tests adjudicate the prior.

4.3 Trader-type taxonomy

To characterise whom the detector flags (see 5.7), we assign each wallet in the analysis universe to a trader type. We adapt the activity-based taxonomy of Della Vedova (2026a): we use his five classes (algorithmic, sophisticated, active retail, casual, one-shot) and their published thresholds, but redefine the activity unit as a distinct taker transaction and add a sixth, residual maker-only class for this dataset. These labels are behavioural, not identity-based. "Algorithmic" denotes a trading frequency and volume strongly suggestive of automation, not a verified bot: on-chain data can neither prove a wallet is automated nor rule out an exceptionally active human. This holds for all classes: the taxonomy groups wallets by observable activity, not by ground-truth identity.

Unit of analysis. We count distinct taker transactions, not order-book fills. A *taker* is the party that crosses the spread and initiates a trade against resting liquidity; one taker decision can match several resting maker orders in a single blockchain transaction, producing multiple fills from one intent. Counting fills would overstate activity for wallets trading against fragmented liquidity and misclassify active traders as algorithmic, so the taker transaction, which corresponds to a trading decision rather than an order-book event, is the cleanest observable proxy for intent. Wallets that appear only as makers have no observable taker activity and form the residual Maker-only class. This unit choice is material: it is what produces the sophisticated-dominated composition of the flagged cohort, and the sensitivity of that composition to the choice of unit is quantified in Appendix C.

Features and classification. Each wallet is described by five lifetime activity features computed over its taker-side history: transaction count, dollar volume, active span, transactions per active day, and a market-concentration index (HHI[4]). It is then assigned to the first class for which it qualifies, scanning from the most to the least active: *Algorithmic* (very high-frequency or high-count activity suggestive of automation), *Sophisticated* (substantial, diversified, sustained activity without that frequency), *Active retail* (regular but smaller-scale trading), *Casual* (a handful of trades), and *One-shot* (a single trade), with the residual *Maker-only* class for wallets that never take. The exact feature definitions and the threshold values defining each class are given in Appendix C.

Population distribution. Table 5 reports the breakdown across the 1,696,645 wallets in the analysis universe.

[Table 5 about here.]

The universe is overwhelmingly low-activity: Casual, Active retail, and One-shot together are approximately 96% of the 1,696,645 wallets, with Sophisticated (2.4%), Maker-only (1.6%), and Algorithmic (0.5%) the small high-activity and passive remainder. The flagged cohort inverts this distribution sharply (see 5.7).

4.4 Replication of existing detection methods

We benchmark the ex-ante detector against three detection methods replicated from the recent prediction-market literature, each keying on a distinct evidentiary primitive: Mitts and Ofir (2026) on the size, profitability, timing, and directionality of realised bets; Liu (2026) on an account-level behavioural fingerprint; and Della Vedova (2026a) on realised forecasting accuracy. That the three rest on different primitives is what makes their agreement, and their disagreement, informative for the complementarity analysis of Chapter 5. This section gives only what those comparisons require: what each method measures, its unit of analysis, and the single adaptation each demands on the Polymarket data. The full signal constructions, eligibility filters, thresholds and weights are collected in Appendix D, which a reader unfamiliar with these methods should consult for the complete specifications.

[4]The Herfindahl-Hirschman Index, originally a measure of market concentration in industrial organisation, sums squared activity shares.

4.4.1 Mitts and Ofir: Large-bet composite score

Mitts and Ofir (2026) screen for informed trading at the (wallet, market) level on the premise that informed traders deviate from the rest of the market along several dimensions at once. Five signals, namely cross-sectional and within-wallet bet size, realised profit, late-buy timing, and directional concentration, are standardised and combined into a single composite suspiciousness score over pairs in which the wallet bought at least $500. The method's own flag is a Layer-2 bet-size gate, firing when either bet-size signal exceeds a within-distribution z of two; the composite then ranks the gated pairs rather than serving as a flag of its own. We depart from the original on the final selection step alone. On our sample the bare gate is dominated by high-volume wallets whose stakes are large for mechanical reasons, so we retain only the top composite-score quartile within the gated set. This keeps the pairs that are most anomalous across all five dimensions and brings the flagged count into line with our own detector, so that the cross-method comparisons reported later reflect which wallets each method selects rather than how many it admits. The benchmark used throughout this thesis is therefore the intersection of the Layer-2 gate and the top composite-score quartile. Additional details on our replication can be found in Appendix D.1.

4.4.2 Liu: Five-flag behavioural fingerprint

Liu (2026) assembles an account-level fingerprint of five trader-market characteristics that, in combination, separate informed traders from the merely successful crowd: high directional conviction, favourable (low-VWAP) entry, a young or single-purpose wallet, focused activity across few markets, and large realised profit. A (wallet, market) pair is flagged when at least four of the five fingerprints fire, evaluated over winning pairs with at least $1,000 of taker-side volume. Liu's full classifier augments this behavioural scan with a Sybil-cluster propagation step that links wallets sharing a common off-chain withdrawal destination; we replicate the behavioural scan alone and set the fund-flow enrichment aside, since it relies on off-chain transaction tracing outside this study's on-chain scope. Because Liu fixes the five signals and the four-of-five aggregation rule but not the threshold at which each signal fires, we operationalise every fingerprint as an extreme-quartile cut, defined in Appendix D.2.

4.4.3 Della Vedova: Excess accuracy and execution quality

Where the other two methods score how a wallet trades, Della Vedova (2026a) asks only whether it was right. A wallet is flagged by a one-sided binomial test of excess directional accuracy against the price-implied benchmark, the accuracy a trader who merely follows the market would obtain, evaluated over wallets with at least ten resolved buy trades. A structural orthogonality test then asks whether accuracy co-moves with execution quality, the prices a wallet obtains relative to the market's volume-weighted average, on the logic that private information delivers a correct side and a favourable entry at once whereas luck delivers neither systematically. The accuracy screen is the flag; the accuracy-execution co-movement is the population-level validation that the flagged accuracy reflects information rather than chance, not a per-wallet flag in its own right. The excess-accuracy statistic, its binomial z-score, and the orthogonality regression are given in Appendix D.3.

4.5 The ex-ante anticipation detector

4.5.1 Conceptual framework

Our detector rests on three premises drawn from the market-microstructure and event-study literatures. *First*, a persistent shift in a market's price encodes the arrival of information. In a binary prediction market the price lives on the probability scale $[0,1]$, so a directional move of fixed magnitude carries a common meaning across markets: it is a change in the implied probability of the event. We therefore threshold an absolute move, 0.15 on the $[0,1]$ scale, rather than a volatility-normalised one, treating a fifteen-point probability move as economically material wherever it occurs. The jump-detection literature motivates reading a discrete, persistent move as information arrival (Barndorff-Nielsen and Shephard, 2006; Lee, 2012; Lee and Mykland, 2008). A volatility-scaled analogue following Lee and Mykland (2008) is left to future work.

Second, agents who hold information position before the move rather than after it. The event-study and insider-trading literatures document pre-event accumulation by informed traders (Keown and Pinkerton, 1981; Meulbroek, 1992), and Kyle (1985) shows why such traders split orders and trade in advance of the price adjustment they anticipate. We read positioning over a 24-hour horizon, the scale of a news cycle; the window is varied over 12 to 48 hours in Appendix E.

Third, this pre-move positioning is recoverable. Because every Polymarket trade is recorded on-chain (4.1), the timing, direction, and size of a wallet's buys ahead of a move can be reconstructed at the (wallet, market) level from public order flow.

Ex ante describes the object of detection, not its deployment. The behaviour the detector recovers, the timing, direction, and size of a wallet's pre-jump buys, was anticipatory at the moment of trade, and it is recovered retrospectively from the historical record using only information generated before the market resolves: the qualifying buy and the price path observed in live trading over the following 24 hours, never the realised settlement value. The detector therefore sits between real-time surveillance and the ex-post benchmarks it complements. It is not real-time: a flag cannot be confirmed at the instant of the trade, since it requires the 24-hour forward window to elapse. But it is far more ex ante than the three benchmarks, which key on the resolved outcome and so cannot flag a wallet until its market has settled, often weeks or months after the trade: here a flag is determinable about a day after the qualifying buy and typically well before resolution.

4.5.2 From order flow to anticipatory buys

Buyer convention. For each trade the buyer is the party taking on directional exposure: the taker when the trade is taker-initiated (`taker_direction` = BUY), and otherwise the resting maker. The gates below are evaluated on the on-chain trade record described in 4.1.

Prices and windows. All prices are expressed on the YES leg, and each buy carries a direction $d_b \in \{+1,-1\}$, with $d_b = +1$ for a YES buy and $d_b = -1$ for a NO buy. With $\mathit{Window} = 24\,\text{h}$, the

forward and backward directional moves around a buy at time t_b are

$$\Delta_{fwd} = d_b \left(p(t_b + Window) - p(t_b) \right), \tag{4.1}$$

$$\Delta_{bwd} = d_b \left(p(t_b) - p(t_b - Window) \right), \tag{4.2}$$

where $p(t)$ is the last trade price at or before t.

The jump. A buy clears the jump gate when $\Delta_{fwd} \geq JumpSize = 0.15$. A move of 0.15 in the buyer's own direction is rare: over a 24-hour window the median directional move is essentially zero. Only 3.8% of buys with an observed forward window see a forward move of 0.15 or more in their favour (Table 6). A qualifying jump therefore sits near the 96th percentile of directional moves, well into the tail that substantive public news can rationalise.

[Table 6 about here.]

Persistence and direction. The move is evaluated at the window endpoint $t_b + Window$, which removes transient intra-window reversals without requiring a microstructure model of the price path. The gate is one-sided: a move against the buy is never flagged.

4.5.3 The six gates

The detector is conjunctive: a (wallet, market) pair either passes all six gates or does not enter the flagged set, and each gate rules out one specific non-anticipatory explanation for the same observable pattern. Gates G1 to G4 are evaluated for each individual buy; a buy that clears them is a *qualifying buy*. Gates G5 and G6 are then evaluated at the (wallet, market) level over the wallet's pooled qualifying buys. The detector assigns no graded score and produces no ranking among flagged pairs; aggregation to the wallet level is handled with the population (5.1). Figure 1 illustrates the gates on a stylised example. A (wallet w, market m) pair is flagged when all of the following hold.

G1 (forward window observed). The full 24 hours after each qualifying buy is present in live trading ($t_b + Window \leq$ the market's last trade), and the prior 24 hours is present ($t_b - Window \geq$ the first trade) for G4. This is the outcome-independence guarantee: the forward move is read from real pre-resolution trading, never from the settlement value.

G2 (jump). $\Delta_{fwd} \geq 0.15$, with the endpoint and one-sided rules above.

G3 (reverse-causality control). The wallet's own volume in the forward window is below 20% of total contemporaneous market volume, so the move is driven by other participants rather than by the wallet's own price impact.

G4 (de-momentum). $\Delta_{bwd} < 0.10$: the buy does not follow an existing trend in its own direction. The gate is one-sided, leaving adverse prior moves unconstrained, since a prior move against the buy is the opposite of the trend-following the gate excludes.

G5 (materiality). The wallet's pooled qualifying buys in the market total at least \$1,000. This anchors to the size gates of Mitts and Ofir (2026) (\$500) and Liu (2026) (\$1,000); it is the conservative choice and the dominant lever in leave-one-out sensitivity.

G6 (conviction). Net-to-gross directional volume exceeds 0.80, defined below.

All qualifying buys by a wallet in a market are pooled, whether they anticipate one move or several, and the materiality floor is applied to the pooled total.

4.5.4 Guards against spurious flags

A flag reflects anticipation rather than a data artefact because four guards, built into the gates, each close off a mechanical alternative (Table 7).

[Table 7 about here.]

The jump gate also confines flagged buys to the mid-range: because a qualifying buy needs a 0.15 directional move on $[0,1]$, a YES-direction buy can qualify only at a price of 0.85 or below and a NO-direction buy only at 0.15 or above, so flagged buys can never be near-boundary bets that merely ride the terminal convergence to a 0 or 1 settlement that binary contracts exhibit near resolution (Bartlett and O'Hara, 2026).

4.5.5 Conviction and the two analysis sets

Conviction (G6). Let t_0 be the wallet's earliest qualifying buy in the market and j_t the first time within 24 hours of t_0 at which the price crosses the buy price by 0.15 in the modal qualifying direction. This run-up window ends when the move is realised at j_t, at or before $t_0 + 24\,\text{h}$, and so is generally shorter than the jump-detection window. Over the interval $[t_0, j_t]$ (the run-up window of Figure 1) every trade the wallet makes, taker or maker and on either side, is signed by whether it increases or decreases exposure in the flagged direction. Conviction is net signed volume divided by gross volume, on $[-1,+1]$; the 0.80 threshold corresponds to roughly 90% of gross run-up volume being directionally aligned. Because the conviction distribution is bimodal, the gate is insensitive to the exact threshold (E.2).

Two analysis sets. Conviction is a directional-purity filter applied after the fact, not a timing condition, so we carry forward two nested sets. The *timing set* applies gates G1–G5 only and is used for the H1 permutation test (5.2), where randomising a post-hoc purity filter would be inappropriate. The *headline set* applies all six gates, G1–G6, and is used for all later hypotheses and for cohort characterisation.

[Figure 1 about here.]

4.5.6 Parameters

Only the materiality floor is anchored to prior literature; the remaining thresholds rest on prediction-market-specific reasoning rather than transplanted equity-market conventions. The values in Table 8 are the detector's single setting; their formal defence, including leave-one-out and range sensitivity, is the subject of the robustness analysis in E.

[Table 8 about here.]

4.5.7 Scope and interpretation

Three properties define what the detector is and is not. *First*, it is a retrospective screen: it reads only pre-resolution information, the qualifying buy and the observed forward price path made outcome-independent by gate G1, and it is not a real-time surveillance tool. *Second*, it is buy-anchored: every element, the jump window, the de-momentum and others-driven controls, the materiality floor, and conviction, is defined

relative to a single qualifying buy. The detector marks the anticipatory act, the timing, direction, and size of the buy, rather than the trader's eventual profit, accuracy, or account history. This is the substantive contrast with the three ex-post benchmarks, which key on realised outcomes. *Third*, it is one reasonable and transparent configuration, offered as a first and suggestive instrument: a proof of concept that act-level, ex-ante detection is feasible. It is not claimed to be tuned, optimal, or definitive. What the detector can and cannot establish follows from these properties. It surfaces anticipation consistent with informed trading: it is a screen, not an adjudicator. Pre-resolution price behaviour cannot separate trading on non-public information from a fast reaction to public information, so the detector flags such anticipation rather than confirming its source, and it makes no claim about how prevalent informed trading is among flagged pairs. The operational limitations of this configuration, including the recall boundary that omits the late hold-to-resolution bet, the dependence on parameter choices, and its reliance on the endpoint magnitude of the move rather than its full shape, together with the refinement roadmap, are taken up in (6).

Results

5.1 Applying our detector: The flagged population

This section applies the detector specified in 4.5 to the in-scope panel described in Chapter 4.1 and sizes the flagged population, a set of *candidate* (wallet, market) events. Table 9 walks the sample down through the gates. Of 63,343,761 in-scope fills across 26,789 markets, 56,977,751 buys are eligible for the forward-window test (G1), of which 2,182,119 are followed by a directional move of at least 0.15 in the buyer's direction within 24 hours (G2). The reverse-causality control (G3: buyer's own forward-window volume below 20% of total contemporaneous volume) removes only about 2.4% of the surviving buys, the cases in which the buyer's own trading rather than the wider market drove the move; that it binds so rarely indicates that self-driven jumps are uncommon among anticipation buys (its effect on the quality of the flagged population is shown in the gate ablation, Appendix E). The de-momentum condition (G4: prior 24-hour directional move below 0.10) removes a further $\approx 27\%$ and is the dominant filter at the buy level. The remaining qualifying buys aggregate to 396,380 (wallet, market) pairs, which the \$1,000 materiality floor (G5), by some distance the most restrictive filter at the pair level, reduces to the *timing set* of 18,078 pairs across 6,658 wallets; the conviction gate (G6) then yields the *headline set* of 10,298 pairs across 4,839 wallets.

[Table 9 about here.]

Within the headline set, the median realised PnL per flagged event is \$1,947 and the win rate is approximately 81%.[1] Because the flag conditions on the realised forward move, this profitability is mechanically inflated by construction and is reported here solely to characterise the population; it is examined substantively under H3 (5.4) and bounded in the discussion (6), and is never treated as evidence of edge in its own right. The conviction gate is threshold-insensitive: the flagged set varies by about 5% across thresholds in $[0.70, 0.90]$ around the 0.80 default. The median pre-jump conviction in the headline set is $+1.00$, with

[1] Realised PnL for a (wallet, market) pair is the wallet's net trading cash flow in that market (USDC received from sales less USDC paid on purchases, across both outcome tokens) plus the settlement value of any net position held at resolution, where each unit of the winning outcome pays \$1 and the losing outcome \$0; computed gross of fees. The win rate is the share of flagged pairs with strictly positive realised PnL.

93% of flagged pairs at exactly +1.00: strictly one-directional accumulation in the buy's direction up to the jump for the typical flagged pair. Sensitivity of the findings to the remaining gate parameters (jump threshold, share cap, materiality floor, and window length) and a full specification-curve analysis are reported in Appendix E.

5.2 H1: The signal is real

H1 (3.2.1) posits that flagged wallets position before jumps at a rate exceeding what their own trading volume would produce under random timing; the null is that the pre-jump concentration is a mechanical artefact of activity. That *when* a wallet buys carries information beyond *whether* it trades is implied by the predictive role of informed order flow in the microstructure tradition (Glosten and Milgrom, 1985; Kyle, 1985).

Test. The test runs on the timing set (gates G1–G5, 5.1); the conviction-passing headline set is not used here, since randomising a post-hoc purity filter would be inappropriate. For each (wallet, market) pair, the wallet's actual trades (count, sizes, directions) and the market's realised price path are held fixed; only the wallet's buy timestamps are redrawn, from the empirical trade-activity time distribution of that same market. The detector is re-run on the permuted timestamps, over 200 permutations, with the test statistic the count of (wallet, market) pairs that still clear all five gates.

Result. Table 10 reports the test. Under the actual timing the detector flags 17,701 pairs,[2] against a permutation-null mean of 6,890 (SD 39.5) with a maximum of 6,995 across the 200 permutations: a factor of 2.6×, and a standardised distance of $z \approx 273$.[3] No permutation reaches within a factor of two of the observed, so the empirical permutation p-value is at its floor, $\leq 1/(200+1) \approx 0.005$.

[Table 10 about here.]

Interpretation. The null is decisively rejected: pre-jump concentration is not a mechanical by-product of trading volume. *When* a wallet buys places it ahead of jumps beyond what its activity alone would produce, and the gap is categorical rather than marginal (roughly 273 standard deviations above the null mean). The genuine signal is, however, more modest than the headline multiple suggests: random timing alone reproduces approximately 38% of the timing-set flags (6,890 of 18,078 pairs, on average), since in markets with frequent jumps some random buys land before some of them; the remaining approximately 62% is the timing signal that does not reduce to volume. Because each pair is permuted within its own market with the realised price path held fixed, the test isolates the timing channel *given* the market in which the wallet was active, and is silent on whether the wallet selected jumpy markets to begin with, a descriptive question taken up in the characterisation (3.2.6, C3).

Boundaries. The test controls for luck given trading volume, not information or prescience: the comparator is the wallet's own reshuffled trades, not a counterfactual involving non-public information.

[2] The recompute count (17,701) lies marginally below the nominal timing-set headline of 18,078 from 5.1 due to edge effects in the share denominator under the permutation harness. Observed and null counts are produced by identical code, so the ratio and standardised distance are unaffected.

[3] The standardised distance is reported as an effect-size descriptor only. We do not convert it to a Gaussian tail probability, since the permutation null is bounded, discrete, and not Gaussian.

Survival of H1 establishes that the flag is a real timing signal, not that the signal is informed.

5.3 H2: The signal is construct-valid

H2 (3.2.2) asks whether the cohort flagged by the ex-ante detector overlaps the cohorts flagged by the established ex-post detectors of Mitts and Ofir (2026), Liu (2026), and Della Vedova (2026a) far above chance. The desired profile has three components, each diagnostic of a different property: *convergent validity* (above-chance wallet-level overlap, the load-bearing claim, establishing that the ex-ante flag tracks the same underlying construct of informed trading); *incremental novelty* (a substantial share of flagged wallets reached by no benchmark, so the detector extends the identified population); and *event-orthogonality* (even for overlapping wallets, the flagged trades differ across detectors, reflecting the ex-ante versus ex-post vantage points).

Test. The headline set (5.1) is compared, at both the wallet and the (wallet, market) pair level, against the three benchmarks of 4.4: the Mitts and Ofir (2026) composite (its Layer-2 bet-size gate restricted to the top quartile by composite score, size-matched to the headline set), the Liu (2026) behavioural fingerprint, and the Della Vedova (2026a) accuracy flag, which is defined only at the wallet level.[4] The wallet universe is $N_W = 1{,}696{,}645$ and the (wallet, market) pair universe is $N_{\text{pairs}} = 16{,}753{,}150$. For each benchmark we report the overlap with the headline set, the expected overlap under independence given the marginal flag rates, and the enrichment ratio (observed-over-expected). Cohen's κ is reported as a conventional agreement summary; because both flags are rare (each well under 1% of the relevant universe) and κ is universe-sensitive at such rates, it is interpreted alongside the enrichment.

Result. Table 11 reports the comparison. At the wallet level, the headline cohort is 147× over-represented among the gated top-quartile Mitts and Ofir (2026) flags ($\kappa = 0.347$), 47× over-represented among the Liu (2026) fingerprint flags ($\kappa = 0.114$), and 35× over-represented among the Della Vedova (2026a) flags ($\kappa = 0.070$). Independence is rejected against each benchmark by a wide margin (one-sided hypergeometric test, $p < 10^{-200}$ in each case).[5] At the (wallet, market) pair level, 10.9% of flagged pairs coincide with a Mitts–Ofir pair and 1.7% with a Liu (2026) pair; κ is near-zero in both cases (0.091 and 0.024). 60.6% of flagged wallets and 84.7% of flagged pairs are not flagged by any of the three.

[Table 11 about here.]

Interpretation. All three components hold. The wallet-level enrichments (147×, 47×, 35×) decisively reject independence: an act-based, ex-ante detector preferentially surfaces the same traders that three outcome-based detectors, built on distinct evidentiary primitives, independently regard as informed, which is the convergent validity the test was designed to establish. Incremental coverage of 61% of wallets and 85% of pairs shows the detector also extends that population to traders the ex-post methods do not reach. And the near-zero pair-level κ alongside the high wallet-level enrichment is the signature of event-orthogonality: the methods flag overlapping traders on largely different events, because the ex-ante detector marks the anticipatory act while the benchmarks key on realised outcomes. That conjunction, overlapping

[4] The three benchmark detectors are reimplemented as described in the methodology chapter (4.4), including the adaptations to the present data and scope; the reimplementations are validated in Appendix D.

[5] The reported p is the one-sided upper-tail hypergeometric probability of observing at least the realised overlap, given the wallet universe N_W and the two marginal flag counts; it is computed directly from the counts in Table 11.

traders but different events, marks a valid-but-novel measurement dimension; pure independence would instead have signalled that the flag captured something unrelated to informed trading.

Boundaries. H2 establishes that the ex-ante flag tracks the construct measured by the published detectors and extends it; it does not claim superiority. The four methods are complementary, each catching a different archetype of informed behaviour: the ex-ante detector marks the anticipatory act, while the benchmarks key on realised size and profitability (Mitts and Ofir, 2026), realised accuracy (Della Vedova, 2026a), and account-level fingerprints (Liu, 2026). None of these methods, the present detector included, has been validated against an adjudicated set of insiders; each is a published heuristic flag whose precision and recall against true insider status are unknown. The substantial wallet-level overlap therefore shows that the four methods identify a similar underlying population of *candidates* from different evidentiary vantage points: it is evidence of construct similarity, not that the population so identified consists of insiders.

5.4 H3: The edge is behaviour-specific

H3 (3.2.3) asks whether the edge of flagged wallets concentrates in their flagged anticipation pairs or pervades their broader trading activity. Generic skill would predict the latter: the same wallets would be edge-bearing across their other markets, and the flag would be incidental to a trader-level trait rather than a property of the act it identifies. The substantive comparison is therefore within-wallet, holding each trader's general ability fixed; the load-bearing statistics are the magnitude of the within-wallet gap and the share of wallets that exhibit it.

Test. For each of the 4,839 wallets in the headline set, we compare outcomes (win rate and realised PnL per (wallet, market) pair) on flagged anticipation pairs against the same wallets' other (wallet, market) pairs, both pooled and within-wallet (paired). The within-wallet comparison restricts to wallets with at least two flagged pairs and at least five non-flagged pairs, yielding 1,354 wallets eligible for the paired test. The flagged-event sample is selected on the realised forward move, so part of the pooled gap is mechanical; the within-wallet design controls for this bias to the extent that the same selection operates on each wallet's non-flagged activity, and the share of wallets exhibiting a positive within-wallet gap is the statistic most robust to the pooled inflation. To isolate the contribution of the behavioural gates from the residual mechanical bias, we additionally report a *matched-jump* comparison: for each paired wallet, the comparison set is restricted to non-flagged pairs in which the wallet also had at least one buy followed by a ≥ 0.15 directional move within 24 hours but failing one or more of the behavioural gates G3–G6. This control holds selection on the realised forward move constant between the flagged and the comparison set.

Result. Table 12 reports the comparison. Pooled across the headline cohort, the anticipation pairs win 80.7% of the time against 65.0% on the same wallets' 759,590 other pairs (a +15.7pp gap), with median realised PnL roughly three orders of magnitude apart (Panel A). Those other pairs already sit above the global universe baseline of 48.5%, so flagged wallets are on average modestly skilled traders; the edge on their anticipation pairs is far larger than that baseline differential. Within-wallet, the mean win rate is 81.9% on anticipation pairs against 59.6% on the rest, a gap of +22.3pp, and 81% of the 1,354 paired wallets win more often on their anticipation pairs (87% earn a higher median PnL there). These bare within-wallet shares are descriptive and remain inflated by selection on the realised move; the formal test

is the selection-controlled matched-jump comparison below. Anticipation pairs are a median 4% (mean 13%) of a flagged wallet's market-pair count, so the behaviour the detector identifies is a small slice of each wallet's activity yet carries a disproportionate share of its realised edge.

[Table 12 about here.]

Matched-jump robustness. The matched-jump comparison defined above holds selection on the realised move constant; 1,303 of the 1,354 paired wallets have at least one matched-jump-other pair (Table 13). The within-wallet gap narrows to +10.6pp in mean win rate (median +14.0pp), positive in 72% of non-tied wallets, decisively above the 50% null (binomial sign test, $p \approx 7 \times 10^{-52}$). The same wallets' non-matched-other pairs win at only 57.9%, so selection on the realised move alone accounts for roughly 13.6pp of within-wallet edge: about half of the bare +22.3pp gap is mechanical selection, and the residual $\approx$ +10.6pp is the edge contributed by the behavioural gates G3–G6 over and above it.[6]

[Table 13 about here.]

Interpretation. The edge concentrates in the flagged behaviour, not in the flagged wallets as such: the within-wallet concentration is inconsistent with a generic person-level trait pervading a wallet's activity. The matched-jump decomposition refines the magnitude rather than the direction: about half of the bare within-wallet gap is mechanical selection on the realised move, and the residual is the part contributed by the behavioural gates G3–G6. Flagged wallets are modestly skilled traders on average, and the residual matched-jump edge on their anticipation pairs is real, statistically decisive, and persistent across individual wallets, though narrower than the bare comparison alone would suggest. The signal therefore localises to the act of anticipation the detector targets in a manner not reducible to mechanical selection on the realised move alone, while that selection is not negligible.

Boundaries. H3 rejects the "generically skilled whale" explanation: the edge is not uniform across a wallet's activity. But the residual does not by itself separate informed trading from contrarian or skilled forecasting; that is the work of H4 (5.5), on whether the anticipated moves are discrete news-like repricings, and H5 (5.6), on whether flags concentrate where private information can exist. Pure market-making is largely screened out by the conviction gate, which requires one-sided directional accumulation, and negligible activity by the materiality floor (4.5); pure luck given volume is ruled out by H1 (5.2).

5.5 H4: The anticipation is information-driven, not drift-forecasting

H3 established that the flagged edge concentrates in the anticipation events rather than across the wallet's broader trading. H4 (3.2.4) asks whether that behaviour responds to discrete information arrival or, instead, to a gradual trend that a skilled forecaster could extrapolate. The discriminator is the *shape* of the anticipated move: a gradual drift is in principle forecastable through trend extrapolation, whereas a sudden, discrete repricing is the empirical footprint of a discrete information event (Lee and Mykland, 2008). Because large moves in these markets may be generically discrete, the hypothesis is posed comparatively: flagged anticipations precede *distinctively* discrete moves, more discrete than those preceding comparable unflagged anticipations in the same markets. The null is that they are no more discrete than this baseline.

[6] This split is approximate accounting, not an exact partition: the components do not sum exactly because the anticipation baseline and the comparison-set composition differ between the bare and matched-jump comparisons.

Test. For each flagged (wallet, market) pair we take its qualifying buy and classify the anticipated move as *discrete* if a ≥ 0.15 directional move in the buy's direction is realised within a $\leq$ 2h sub-window that carries the price across the qualifying threshold and persists to the 24-hour horizon, so that the move which produced the flag is itself sudden rather than a transient intra-window spike that reverts. We compare the discrete share of flagged pairs against a within-market control of unflagged anticipating pairs (same markets, same ≥ 0.15 move over 24 hours, classified by the identical rule), via a two-proportion test. Each pair is classified on its single qualifying buy, so that differences in trading frequency between flagged and control wallets cannot mechanically inflate the discrete share.

Result. Table 14 reports the comparison. Among the flagged pairs, 47.1% have a discrete flagging move, against 47.0% of the 289,922 within-market control pairs: a gap statistically indistinguishable from zero (two-proportion $z = +0.23$, $p \approx 0.82$), so the null is not rejected and roughly half of flagged moves are gradual repricings rather than sudden jumps. A looser definition that counts *any* $\leq$ 2h, ≥ 0.15 sub-window anywhere in the forward window, including transient spikes that revert before the move that produced the flag, raises the flagged share to 81.0% (against an 83.1% control share), but it conflates incidental intra-window volatility with the anticipated repricing; under the strict, flagging-move definition the flagged share matches the baseline. Because flagged and control pairs are drawn from the same markets, accounting for within-market dependence would only widen the interval around the near-zero gap and cannot produce the positive excess the hypothesis predicts, so the null conclusion is conservative to it.

[Table 14 about here.]

Interpretation. Discreteness does not distinguish the flagged behaviour: it is a property of large moves in these markets generally, not a mark of the flag, so we do not rest the informational interpretation on the shape of the move. This does not reopen the drift-forecasting alternative in its most direct form, because trend extrapolation, forecasting the continuation of a prior move, is already excluded structurally by the de-momentum gate: a flagged move is never the continuation of a prior same-direction trend. What H4 adds is the transparent finding that move-shape provides no *further* discrimination beyond that structural control; the residual distinction between responding to information and forecasting a fresh gradual repricing is left to H5 (5.6). Move-shape is, however, informative about flag *quality*: discrete-move flags realise higher median PnL, win more often, and are more novel relative to the benchmark detectors than gradual-move flags,[7] so it tags the cleaner flags even though it does not separate flagged anticipations from unflagged ones.

Boundaries. The null bounds what move-shape can establish; it does not bear on the detector's validity, which rests on H1, H2, H3, and H5. Where the relevant information lives, whether flagging concentrates in markets in which material non-public information can plausibly exist, is taken up by H5 (§5.6).

[7] Discrete versus gradual flags: median PnL \$2,360 vs \$1,646 (Mann–Whitney $p \approx 2 \times 10^{-24}$); win rate 83.2% vs 78.5% ($p \approx 10^{-9}$); novel to all three benchmarks 76.1% vs 72.5% ($p \approx 3 \times 10^{-5}$).

5.6 H5: The signal is information-dependent

H4 established that the moves the flagged wallets anticipate cannot be separated from those of comparable unflagged anticipators by shape alone. H5 (3.2.5) asks the most insider-specific question of the ladder: does the flagging concentrate in markets in which material non-public information can plausibly exist, relative to markets in which it is implausible by the nature of the resolution process? Informed-trading theory ties the informed trader's edge to the existence of an information asymmetry (Kyle, 1985): an edge that is indifferent to whether exploitable private information can exist is forecasting skill; an edge conditional on it is informed trading. H5 is therefore the closest observational test of that conditionality, and the closest the data come to the insider question itself. The concentration it documents is statistically robust; the interpretation of that concentration is bounded, for the reasons set out below.

Test. Every market in the analysis universe is assigned to one of four information-generating categories that adapt the taxonomy of Della Vedova (2026a): two *MNPI-admitting* categories (*vote* and *action*, in which a narrow set of agents learns or decides the outcome before the public) and two *no-MNPI* categories (*stochastic* and *performance*, in which the resolution mechanism makes the existence of material non-public information implausible (not strictly impossible). The specific classifier's construction is reported in Appendix B. The test statistic is the *flag intensity* of each cell, defined as flagged headline pairs per 10,000 universe (wallet, market) pairs of the same type; the per-pair denominator adjusts for the substantial differences in exposure across cells. MNPI-admitting and no-MNPI are compared and a 95% confidence interval on the ratio is reported, bootstrapped over markets as the unit of independence (2,000 resamples).

Result. Table 15 reports the flag intensity in each cell. The *action* category is the hottest at 8.1 flagged pairs per 10,000 universe pairs, followed by *vote* (7.0), *stochastic* (4.7), and *performance* (1.8). Aggregating, MNPI-admitting markets show a flag intensity of 7.6 per 10,000 against 4.6 in no-MNPI markets, a ratio of 1.63× with a market-clustered bootstrap 95% confidence interval of [1.35, 1.94] that excludes one. Dropping the 17.3% of markets whose category is assigned by the large-language-model classifier rather than deterministically leaves the ratio at 1.68× [1.37, 2.03] (Appendix B); the gradient survives the deterministic-only subset, and is in fact slightly stronger on it.

[Table 15 about here.]

Per-jump robustness. The gradient is not an artefact of MNPI-admitting markets being jumpier. Normalising by jumps rather than by opportunity pairs *raises* the ratio to 1.85× ([1.52, 2.24]), because those markets are in fact slightly *less* jump-prone than no-MNPI ones (Table 16). The two normalisations control different confounds in opposite directions, trader density (per pair) and jump frequency (per jump), so the gradient surviving both leaves no single artefact to explain it.

[Table 16 about here.]

Interpretation. Flagging concentrates in MNPI-admitting markets, and the concentration is robust: the ratio's 95% confidence interval excludes one, and the gradient is, if anything, slightly stronger on the deterministic-only subset. It is not confined to small-group decision markets but spans both MNPI-admitting cells, action and vote alike. Two features bound how far the result can be pushed. First, the concentration is partial, a 1.63× tilt rather than a clean confinement: informed-consistent anticipation is

more intense where private information can exist, not absent where it cannot. Second, roughly 36% of flagged pairs fall in stochastic markets, consistent with the detector partly reflecting market-mix and the higher base rate of anticipatable jumps in event-driven categories rather than information-richness alone, the channel decomposed in C3 (5.9).

Boundaries. H5 is the frontier rung, and the claim is calibrated to what a cross-sectional test on a constructed market typology can carry. The concentration itself is robust and significant. Two things bound its interpretation rather than its existence. First, the typology, though adapted from the published taxonomy of Della Vedova (2026a), is validated here only provisionally: an independent LLM pass agrees on roughly 77% of the validation sample (Appendix B) and human blind-coded validation is pending, so the precise magnitude may move even though the direction is stable. Second, and in principle, H5 cannot establish intent: a higher flag rate in MNPI-admitting markets is consistent with information-dependence but not synonymous with the use of material non-public information, the epistemic ceiling that bounds the whole ladder (3.2.5). The claim is therefore strong in direction and significance, and bounded in magnitude and in what it implies about intent: the flag is *information-dependent*, concentrating where exploitable private information can plausibly exist.

From the hypothesis ladder to characterisation

Taken together, the rungs form a cumulative case, qualified by one transparent null. H1 shows the flag is a real timing signal; H2 that it is construct-valid, concentrating on the population the established detectors regard as informed while extending the identified set; and H3 that the edge is behaviour-specific, narrower than the bare within-wallet comparison once forward-move selection is matched but with a real, persistent residual. H4 returns a null: the anticipated moves are no more discrete than comparable unflagged moves in the same markets, roughly half gradual. Move-shape therefore does not carry the informational reading (within the flagged set, the discrete cases are simply the higher-quality ones), and the drift-forecasting alternative is excluded structurally by the de-momentum gate. H5, more tentatively, finds the flagging concentrates in markets where material non-public information can plausibly exist, relative to those where the resolution mechanism makes it implausible. The detector therefore isolates a real, construct-valid, behaviour-specific signal whose incidence concentrates in information-admitting markets: the strongest informed-trading case that observational prediction-market data permit, with the residual step to demonstrated use of material non-public information requiring ground truth on intent (6). Having established what the signal *is*, the remainder of this chapter characterises what it *identifies*: who the flagged traders are, what they flag, and why the flagged population takes the shape it does.

5.7 C1: Who the flagged traders are

Composition. Classified by the trader taxonomy of Della Vedova (2026a) (4.3), the flagged cohort is dominated by sophisticated traders, who make up 61.4% of the 4,839 flagged wallets and generate 49.7% of the 10,298 flagged events (Table 17). Algorithmic accounts are a smaller but disproportionately active tier: 13.7% of wallets but 37.5% of events, roughly 5.9 flagged events per algorithmic wallet against 1.7 per sophisticated one. Together the two top tiers carry 87.2% of flagged events, leaving active retail, casual, one-shot, and a residual maker-only pocket (33 wallets whose flagged buys arrived as passive

limit-order fills) to share the remaining one in eight. Both top tiers are substantial, multi-market operators: the median sophisticated wallet trades 35 markets at a lifetime $117,584 and the median algorithmic wallet 377 markets at $954,245, and algorithmic accounts post both the highest per-wallet event yield and the highest median realised PnL per flagged event.

[Table 17 about here.]

Implications. Two implications follow. First, the detector surfaces the well-resourced, active trader population that informed-trading theory treats as the natural origin of informed positioning, rather than a naïve retail crowd, consistent with the construct validity established in H2 (5.3). Second, the algorithmic tier is ambiguous in a way the sophisticated tier is not: it mixes wallets exploiting genuine non-public information with fast reactors to public information arrival, and observational data cannot separate the two, so the epistemic ceiling here becomes a concrete property of this sub-population.

5.8 C2: What the detector flags

C2 shifts focus from who is flagged to the flagged events themselves: their markets and economic footprint.

Market composition. By topic, the flagged events concentrate in four categories, Business, Geopolitics, Politics, and Elections, which together account for 88% of the set (Table 18). These labels are topical and cut across resolution types, so the category breakdown describes subject matter rather than information-richness; the informative cut is the orthogonal resolution-type axis (Table 19). There, the two MNPI-admitting types, action (38.6% of events) and vote (24.9%), together carry 63.5% of flagged events and $52.2M of realised PnL, against 36.3% and $15.7M in no-MNPI markets. The PnL concentration in MNPI-admitting markets (3.3×) is larger than the event concentration (1.7×), because vote-resolved markets carry the highest median PnL of any type ($2,785); action and vote thus hold 63% of events but 77% of realised value, the same markets that carried the highest flag intensities in H5 (5.6). The one large exception is a stochastic bloc, 36% of events and $15.7M, in markets where non-public information is implausible by the resolution mechanism; it is not concentrated in any one trader type but mirrors the cohort, reinforcing the C1 reading that the flag blends genuine informed positioning with fast reaction to public information.

[Table 18 about here.]

[Table 19 about here.]

Economic footprint and concentration. The flagged events are economically consequential: $68.0M in total realised PnL across the 10,298 events, a median of $1,947 and a right-skewed mean of $6,599. 80.7% are profitable (8,315 winners against 1,959 losers), the median win and loss roughly symmetric ($3,046 versus −$3,114), and the single largest event realised $2.58M against a largest loss of −$1.59M. This profitability is whole-market (the wallet's full position in the market, not the anticipatory buy alone) and selection-inflated, reported to size the flagged events rather than as evidence of edge (5.4). The mass is concentrated in a small tail: the top 1% of flagged wallets (48 accounts) account for 16.1% of events and roughly 29% of gross realised gains, the top 10% (483 wallets) for 46% of events and roughly 56% of

gains.[8] Most flagged wallets are flagged on a single event (median 1, mean 2.13, maximum 120; Gini 0.47 on events per wallet), so the population is a long tail of one-off flags beneath a smaller, repeatedly-flagged core. That core is jointly sophisticated and algorithmic and tilts algorithmic as the flag count rises: of the 1,399 wallets flagged on two or more events, 62.0% are sophisticated and 31.8% algorithmic, but of the 361 flagged on five or more, 59.6% are algorithmic and 39.9% sophisticated, consistent with algorithmic accounts' higher per-wallet yield and, in part, their greater exposure (decomposed in C3, 5.9). This tail structure bears directly on the detector's role as a screen, taken up in C5 (5.11).

5.9 C3: Why this population

The composition documented in C1 (5.7) and C2 (5.8) could reflect a genuine concentration of informed anticipation or the mechanical reach of the detector's gates. We decompose it against the two binding channels: the \$1,000 materiality floor and trading exposure, the latter expressed as flags per 10,000 (wallet, market) opportunity pairs.

Size-floor. The \$1,000 floor is the dominant filter and it bites unevenly (Table 20): it clears active-retail anticipations at roughly an eighth the rate of algorithmic or sophisticated ones, converting a retail-led pre-floor population into the active-trader-led timing set, after which the conviction gate leaves the headline set sophisticated-dominated. This is a property of the gates, not in itself evidence that active traders are intrinsically more informed.

[Table 20 about here.]

Exposure rate. Normalising flags by opportunity separates a per-opportunity edge from pure exposure (Table 21). The active-trader concentration survives: algorithmic and sophisticated accounts anticipate at almost identical per-opportunity rates, so it sits in the active-trader tier as a whole, not in algorithmic accounts specifically, and the algorithmic-to-retail gradient runs about $6\times$ before the floor and $42\times$ after, a genuine tilt the floor amplifies roughly sevenfold. The gradient is genuine across markets too: both MNPI-admitting resolution types sit clearly above publicly-resolved stochastic markets, so the $1.63\times$ MNPI advantage of H5 is carried by action and vote alike.

[Table 21 about here.]

What is genuine and what is mechanical. Two cautions govern these rates. First, because a flag requires a ≥ 0.15 move, a stratum with a higher base rate of such moves can read hotter for that reason alone, so the rates mark *where* the detector fires, not informed-trader density as such; for the resolution-type gradient this caveat is tested and rejected, since normalising by jumps in H5 (5.6) *strengthens* it, while for the category gradient it stands. Second, all three concentrations survive per-opportunity normalisation, so none is a pure exposure artefact: the within-market permutation test of H1 (5.2) shows roughly three-fifths of the active-trader concentration is timing that random spray cannot reproduce, which the floor then amplifies mechanically. Separating the genuine concentration from this gate-mechanical selection is C3's contribution.

[8] Concentration is reported in shares of gross gains rather than of net PnL, since the net total nets winners against losers and is sensitive to a small set of large negative events; gross-gains and event shares are unaffected and are the clean concentration metrics.

5.10 C4: Which documented cases the detector flags

The preceding characterisation describes the flagged population in aggregate; this section turns to five wallets for which informed trading is independently established or alleged, each concerning material non-public information of the kind insider-trading enforcement contemplates (corporate, governmental, or institutional) and identified through criminal proceedings, regulatory investigations, legislative referrals, or investigative journalism (full per-case narratives and sources in Appendix G). The purpose is mechanistic, not validation: five outcome-selected cases, only two of them charged, are far too few and too selected to estimate recall, and we make no recall claim from them. Because each detector keys on a different signal, each reaches a different kind of insider, and on cases where the informed trading is known we can see concretely which archetype each method catches, which it misses, why, and which it reaches earliest. These five are not a census of insider trading in the sample: public identification is rare and driven by conspicuous outcomes, so the sample almost certainly contains further informed trading, of both the archetypes we reach and those we do not, that no public process has surfaced.

[Table 22 about here.]

What the detector flags. The detector flags two of the five wallets. *Magamyman* is flagged on two Israel–Iran military-action markets, and *AlphaRaccoon* on four: three Gemini 3.0 release-date markets and the Google "Year-in-Search" #1-searched-person market that is the subject of the SDNY complaint. For *Magamyman*, the flagged markets are *not* the single bet that made him publicly known (the Supreme-Leader-removal wager): the detector surfaces him as a repeat anticipator across his broader Iran-war activity, the named-case counterpart of the event-orthogonality established in H2 (5.3). For *AlphaRaccoon*, the detector flags the precise market domain of the criminal charge.

Why caught, why missed. The two caught wallets share a behavioural archetype: repeat, multi-market traders whose positioning is observable mid-market, buying at material size more than twenty-four hours before an others-driven price move and holding a directionally pure position into it. Both are sophisticated accounts, within the active-trader tier that the aggregate composition in C1 (5.7) and C3 (5.9) singled out, so the named cases instantiate the aggregate finding rather than sitting beside it. The three missed wallets fail on distinct edges of the recall boundary, each a property of the detector's design or the sample's coverage rather than a failure to detect a visible anticipation. *6741* created an account solely for the Nobel market and bought hours before resolution, at 99% of the market's life: a pure hold-to-resolution bet with no observable forward window, excluded by gate G1 by construction. *van Dyke* positioned days ahead of the Maduro operation, but his principal bet repriced only at the operation itself, outside the twenty-four-hour window of his entry, and his smaller within-window Venezuela positions fell below the materiality floor. *ricosuave666*'s most-lucrative wagers fall in-window but outside the analysis universe: they are ultra-short-fuse markets that resolve within days of creation and are removed by the minimum-duration filter (F5, 4.1), so the detector never saw them; the longer-dated Israel–Iran markets it did trade in the universe generated only one near-miss anticipation, filtered at the materiality floor. The detector therefore reaches the repeat mid-market anticipator and, by design, cedes the late or out-of-window bet (*6741*, *van Dyke*) to the ex-post benchmarks, while *ricosuave666*'s decisive short-fuse bets fall outside the analysis universe altogether and are missed by every method.

[Figure 2 about here.]

Which method catches whom, and why. Table 23 maps the five cases to each method, and the pattern follows from what each measures, a division of labour rather than a ranking. Mitts–Ofir keys on realised bet size and profitability, so it catches *6741*'s concentrated, high-profit Nobel bet; Liu keys on account-level behavioural fingerprints, so it catches *van Dyke*'s freshly-created, single-purpose account; Della Vedova keys on wallet-level accuracy, so it reaches the consistently-correct *AlphaRaccoon* and *Magamyman*; and our detector, keying on the anticipatory act, reaches those same two repeat mid-market anticipators, but ex ante. No method flags more than three of the five and none flags all; together the four cover four of the five, with only *ricosuave666* escaping every method, its decisive short-fuse bets having been removed from the analysis universe by the minimum-duration filter. That the ex-post benchmarks reach more of this particular set (Mitts–Ofir three, against our two) is expected and not evidence of inferiority: the cases are selected on the realised outcomes those methods measure, the very axis on which the ex-ante detector, by construction, does not compete.

[Table 23 about here.]

Detection lead: the decisive margin. The complementarity is not only in *whom* each method flags but in *when*, and here the ex-ante detector holds a categorical advantage. An outcome-conditioned method cannot act before a market resolves; the ex-ante detector fires at the qualifying buy plus the twenty-four-hour confirmation window, mid-market. Taking each market's last trade as the earliest moment an outcome-based method could act, across the six markets on which the two caught wallets are flagged the detector surfaces the wallet a median of seven days (mean eight) before resolution, and *every* flag precedes resolution. The lead ranges from same-window, on near-resolution bets such as *AlphaRaccoon*'s Year-in-Search market, to seventeen days (*Magamyman*'s two flags lead by seventeen and two days, *AlphaRaccoon*'s four by a median of six). On the four markets Mitts–Ofir also flags, the lead reaches up to seventeen days; Figure 2 shows a representative sixteen-day instance, the *same* wallet, on the *same* market, surfaced over two weeks before the realised outcome the ex-post method requires. This temporal margin, not a higher case count, is the detector's distinctive contribution, and unlike the case count it is not an artefact of how the case list was selected.

Read this way, the five cases say little about *how many* insiders the detector catches and much about *which kind* it catches and *when*. It is a screen for the repeat, mid-market informed anticipator that the aggregate composition already singled out: it surfaces such wallets ex ante, a median of about a week before the outcome any ex-post method could act on, and, as with *Magamyman*, sometimes on events beyond their publicised bets, while ceding the late, out-of-window bet to those methods. The documented cases that fit this archetype are few because public identification is rare, not because the archetype is, and the detector flags thousands of wallets sharing it.

5.11 C5: What the detector identifies

Taken together, the characterisation fixes the detector's role, not its verdict. C1 and C2 show that it concentrates on a well-resourced, active trader population (sophisticated and algorithmic accounts) trading economically consequential, information-laden events; C3 that this concentration is a genuine per-opportunity effect, not a pure artefact of exposure, though the materiality floor amplifies it; and C4 that, among the behavioural archetypes of documented insiders, it reaches the repeat mid-market

anticipator and, by design, not the late pre-resolution bet. None of this establishes that any flagged wallet traded on material non-public information: that determination is external, and the ambiguous algorithmic sub-population together with the substantial no-MNPI bloc places a mixed population inside the flag. Two properties follow, and they pull in opposite directions. In precision, the flag is a *superset* of insider trading: it admits genuine informed anticipation alongside fast reaction to public information and the ordinary exposure of highly active accounts, so a flag is a candidate, not a finding. In recall, it is a *subset*: it cannot reach the insider who positions only at resolution or on an out-of-window event, the archetype the ex-post methods are built for, nor one whose decisive market the sample-construction filters exclude. The detector is therefore a *screen*, not an adjudicator: it surfaces consequential, mid-market, anticipatory positioning ex ante, for downstream scrutiny, and does not by itself determine intent. Its distinctive value is what that screen reaches that the alternatives do not: the anticipatory act itself, at the moment of the trade and a median of about a week before any outcome an ex-post method requires, across a population that public enforcement observes only sparsely. The five cases are a floor, not a census, so the screen's natural object is the far larger set of wallets sharing the same archetype that no charge or investigation has surfaced.

Conclusion

6.1 Findings and contribution

This thesis asked whether informed trading on prediction markets can be detected ex ante, from the anticipatory act and before a market resolves, and what such a detector identifies. The answer to the first question is that it can. A buy-anchored screen, importing the jump-and-run-up paradigm of equity event studies to the binary-contract price path, flags informed-consistent positioning ahead of resolution. A five-rung hypothesis ladder then shows the flag to be a real timing signal, construct-valid against three independent ex-post detectors, specific to the anticipatory act and not to generically skilled wallets, and concentrated, robustly in direction, in markets where material non-public information can plausibly exist. The answer to the second question is that the detector identifies a screen, not a verdict: it surfaces a candidate population of consequential, sophisticated-led, mid-market anticipation, flagged before any outcome an ex-post method could act on, to guide downstream scrutiny.

Our approach was conservative. Instead of claiming to catch insiders, we framed detection as the elimination of benign explanations and tested the flag against each. Four rungs held, and one returned a transparent null: the moves flagged wallets anticipate are no more discrete than comparable unflagged moves in the same markets, so the informational reading rests on timing and information-dependence, not on the shape of the move. That null is itself a result worth keeping, since it marks the limit of what move-shape can establish and keeps the argument honest. The information-dependence rung, by contrast, proved more robust than anticipated, surviving the obvious objection that information-rich markets are merely jumpier. The method's effectiveness lies less in any single statistic than in the convergence of independent tests on the same conclusion, reached from a vantage point no prior method occupies.

That vantage point is the contribution. The field had identified the gap but not closed it: every existing prediction-market detection method is ex post, computable only once a market resolves, and Liu (2026) explicitly named pre-event anticipation detection as a direction for future work. This thesis provides

it, to our knowledge the first ex-ante, act-level detector of informed trading on a prediction market. Its decisive advance is temporal. The detector fires mid-market, as little as a day after the qualifying buy, and on the documented cases it reaches, it surfaces the same wallet a median of about a week before the realised outcome the ex-post methods require. Because that lead reflects *when* the detector acts, not which outcomes were realised, it is immune to the outcome-selection that otherwise favours those methods.

6.2 Scope and boundaries

The detector's reach is bounded in two opposite ways. In precision it is a superset of insider trading, mixing genuine informed anticipation with fast reaction to public information and the ordinary exposure of very active accounts. In recall it is a subset, reaching only positioning observable mid-market. The step from informed-consistent anticipation to demonstrated use of material non-public information requires ground truth on intent that no observational dataset carries, and is the epistemic ceiling that bounds this method as it does every other. The detector therefore informs where to look; it does not pronounce on who is culpable, and it is offered in that spirit.

6.3 Directions for future research

What we present is one reasonable, transparent configuration, a proof of concept that ex-ante, act-level detection is feasible, not a tuned or optimal instrument. Several extensions would accordingly sharpen the instrument, though none is required for the present findings to stand. The most valuable would reach into the near-resolution window: the detector's minimum-duration filter and its reliance on an observable forward window exclude the shortest-fuse markets, exactly those in which anticipation of imminent action is most acute, and a design that recovered late, hold-to-resolution positioning would close the largest gap in recall. A volatility-scaled jump threshold, in the manner of Lee and Mykland (2008), would adapt the detector to each market's own variability instead of applying one fixed move. And extending the analysis to Kalshi, the largest regulated, identity-verified prediction market and a venue with a very different microstructure and data regime, would test how far the signature generalises beyond Polymarket.

6.4 Real-world implications

The wider lesson is that ex-ante, act-level surveillance of informed trading is feasible on transparent venues, and that the on-chain transaction record is what makes it so. The complete, public history of every position that equity-market disclosure cannot supply is precisely what lets the anticipatory act be recovered directly from order flow, and it is the shared foundation on which the criminal, regulatory, and journalistic identifications of these cases have all ultimately relied. For practical surveillance the detector complements the existing methods, it does not supersede them: because they rest on different signals and reach different archetypes, a wallet flagged by several is a stronger candidate than one any single method flags, and a toolkit should combine them. The timing strengthens the point: documented episodes are accumulating, the first prosecutions and regulatory rulemaking are under way, and the platforms are moving toward active surveillance. A screen that flags the anticipatory act before a market resolves should help shift that surveillance from forensic reconstruction toward prevention, pointing to where to look before the outcome settles, not confirming long afterward where the money was made.

References

Abid, M. N. (2026). “The Polymarket Paradox: Manipulation, Whale Concentration, and Predictive Accuracy in the World’s Largest Prediction Market”. Working Paper.

Arrow, K. J., R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. O. Ledyard, S. Levmore, R. Litan, P. Milgrom, F. D. Nelson, G. R. Neumann, M. Ottaviani, T. C. Schelling, R. J. Shiller, V. L. Smith, E. Snowberg, C. R. Sunstein, P. C. Tetlock, P. E. Tetlock, H. R. Varian, J. Wolfers, and E. Zitzewitz (2008). “The Promise of Prediction Markets”. In: *Science* 320.5878, pp. 877–878.

Ashar, R. (2025). “Insider Trading in the Era of Prediction Markets”. Harvard Law School Working Paper.

Augustin, P., M. Brenner, and M. G. Subrahmanyam (2019). “Informed Options Trading Prior to Takeover Announcements: Insider Trading?” In: *Management Science* 65.12, pp. 5697–5720. DOI: 10.1287/mnsc.2018.3122.

Barndorff-Nielsen, O. E. and N. Shephard (2006). “Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation”. In: *Journal of Financial Econometrics* 4.1, pp. 1–30. DOI: 10.1093/jjfinec/nbi022.

Bartlett, R. and M. O’Hara (2026). “Adverse Selection in Prediction Markets: Evidence from Kalshi”. Stanford University and Cornell University Working Paper.

Berg, J. E., F. D. Nelson, and T. A. Rietz (2008). “Prediction Market Accuracy in the Long Run”. In: *International Journal of Forecasting* 24.2, pp. 285–300.

Beylin, I. (2026). “Regulation of Prediction Markets in the U.S.” Seton Hall Law School Working Paper.

Burgi, C., W. Deng, and K. Whelan (2025). “Makers and Takers: The Economics of the Kalshi Prediction Market”. University College Dublin Working Paper.

CFTC (2011). *Regulation 180.1, 17 C.F.R. § 180.1: Prohibition on the Use of Manipulative or Deceptive Devices*. Modelled textually on SEC Rule 10b-5.

Cheong, W. C. and A. Tamayo (2026). “Beyond the Wisdom of the Crowd: Concentrated Informed Trading in Earnings Prediction Markets”. London School of Economics Working Paper.

CNN (Mar. 2026). *Iran-related bets on prediction sites scrutinized over ‘death markets’ and possible insider trades*. CNN Politics. Accessed: 2026-06-01. URL: https://www.cnn.com/2026/03/07/politics/iran-war-prediction-markets-polymarket-kalshi.

Commodity Futures Trading Commission (Nov. 2025). *Amended Order of Designation for QCX, LLC (Polymarket/QCEX)*. Permitting Polymarket, through its acquisition of QCEX, to operate as an intermediated trading venue subject to DCM and DCO requirements.

Commodity Futures Trading Commission (Mar. 2026). *Advance Notice of Proposed Rulemaking: Event Contracts and Prediction Markets*. Federal Register.

Congressional Research Service (2025). *Prediction Markets and Insider Trading Law*. Tech. rep. LSB11406. Library of Congress. URL: https://www.congress.gov/crs-product/LSB11406.

Della Vedova, J. (2026a). "Information and Execution: Detecting Informed Trading in Prediction Markets". In: *Available at SSRN 6567238*.

Della Vedova, J. (2026b). "Who Profits from Prediction Markets? Execution, Not Information". University of San Diego Working Paper.

Dubach, P. D. (2026). "The Anatomy of a Decentralized Prediction Market: Microstructure Evidence from the Polymarket Order Book". Working Paper.

Easley, D., N. M. Kiefer, M. O'Hara, and J. B. Paperman (1996). "Liquidity, Information, and Infrequently Traded Stocks". In: *Journal of Finance* 51.4, pp. 1405–1436.

Eckert, V. and G. Fouche (Jan. 2026). *Nobel Committee Says Peace Prize Winner Likely Revealed Early by Digital Spying*. Reuters. URL: https://shorturl.at/Tj5ib.

Forbes (Oct. 10, 2025). *Did the Nobel Peace Prize Expose Insider Trading on Prediction Market Polymarket?* Forbes. URL: https://shorturl.at/ELUem (visited on 06/01/2026).

Forsythe, R., F. Nelson, G. R. Neumann, and J. Wright (1992). "Anatomy of an Experimental Political Stock Market". In: *American Economic Review* 82.5, pp. 1142–1161.

Glosten, L. R. and P. R. Milgrom (1985). "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders". In: *Journal of Financial Economics* 14.1, pp. 71–100.

Gomez-Cram, R., Y. Guo, T. I. Jensen, and H. Kung (2026). "Prediction Market Accuracy: Crowd Wisdom or Informed Minority?" London Business School / Yale University Working Paper.

Hayek, F. A. (1945). "The Use of Knowledge in Society". In: *American Economic Review* 35.4, pp. 519–530.

Keown, A. J. and J. M. Pinkerton (1981). "Merger Announcements and Insider Trading Activity: An Empirical Investigation". In: *Journal of Finance* 36.4, pp. 855–869. DOI: 10.1111/j.1540-6261.1981.tb04888.x.

Kyle, A. S. (1985). "Continuous Auctions and Insider Trading". In: *Econometrica* 53.6, pp. 1315–1335.

Lee, S. S. (2012). "Jumps and Information Flow in Financial Markets". In: *Review of Financial Studies* 25.2, pp. 439–479. DOI: 10.1093/rfs/hhr084.

Lee, S. S. and P. A. Mykland (2008). "Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics". In: *Review of Financial Studies* 21.6, pp. 2535–2563. DOI: 10.1093/rfs/hhm056.

Liu, S. (2026). "Wisdom of the Crowd or Wisdom of the Insider? Insider Trading on Prediction Markets". In: *Insider Trading on Prediction Markets (April 29, 2026)*.

Mann, H. B. and D. R. Whitney (1947). "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1, pp. 50–60.

Marin, J. M. and J. P. Olivier (2008). "The Dog That Did Not Bark: Insider Trading and Crashes". In: *Journal of Finance* 63.5, pp. 2429–2476. DOI: 10.1111/j.1540-6261.2008.01401.x.

Meulbroek, L. K. (1992). "An Empirical Analysis of Illegal Insider Trading". In: *Journal of Finance* 47.5, pp. 1661–1699. DOI: 10.1111/j.1540-6261.1992.tb04679.x.

Mitts, J. and M. Ofir (Mar. 2026). *From Iran to Taylor Swift: Informed Trading in Prediction Markets*. DOI: 10.2139/ssrn.6426778.

Nechepurenko, M. (2026a). "ForesightFlow: Real-Time Detection of Informed Trading in Decentralized Prediction Markets". In.

Nechepurenko, M. (2026b). "Information Leakage at Population Scale: An Evaluation of the Polymarket Insider-Relevant Subpopulation, 2020–2026". Devnull FZCO Working Paper.

Norton Rose Fulbright (2026). *The EU's Approach to Prediction Markets and Event Contracts*. Publications. URL: https://www.nortonrosefulbright.com/en/knowledge/publications/290d594a/the-eus-approach-to-prediction-markets-and-event-contracts.

Polymarket (2026a). *How We Monitor and Enforce: Real-Time Surveillance, On-Chain Transparency, and Disciplinary Action*. Self-reported figures: 315+ wallet details provided to authorities, 90+ referrals to law enforcement, 2 arrests referred by Polymarket. URL: https://integrity.polymarket.com/.

Polymarket (2026b). *Terms of Use: Restricted Trading Activities*. polymarket.com. Sections prohibiting trading on stolen confidential information, illegal tips, and trading by users in a position to influence the underlying outcome. URL: https://polymarket.com/es/tos.

Polymarket Blocks French Traders Amid Gambling Inquiry (Nov. 2024). CoinDesk. URL: https://www.coindesk.com/policy/2024/11/22/polymarket-blocks-french-users-amid-gambling-inquiry.

Polymarket Partners With Palantir to Combat Insider Trading in Prediction Markets (Mar. 2026). Yahoo Finance. Reporting the deployment of Palantir's Vergence AI engine and the partnership with TWG AI. URL: https://finance.yahoo.com/news/polymarket-joins-forces-palantir-bring-223948342.html.

Reichenbach, F. and M. Walther (2025). "Exploring Decentralized Prediction Markets: Accuracy, Skill, and Bias on Polymarket". Technische Universität Berlin Working Paper.

Rothschild, D. M. and R. Sethi (2015). "Trading Strategies and Market Microstructure: Evidence from a Prediction Market". Microsoft Research / Columbia University Working Paper.

Saguillo, O., V. Ghafouri, L. Kiffer, and G. Suarez-Tangil (2025). "Unravelling the Probabilistic Forest: Arbitrage in Prediction Markets". IMDEA Networks Working Paper.

Sorkin, A. R., B. Warner, S. Kessler, M. J. de la Merced, N. Gallogly, B. O'Keefe, and I. Mount (May 2026). *Another Insider Trading Case Hits Prediction Markets*. The New York Times (DealBook). URL: https://www.nytimes.com/2026/05/28/business/dealbook/insider-trading-polymarket.html.

Tavva, G. A. (2026). "Prediction Markets as a Money Laundering Vector: The Coordinated Loss Problem". Third Door Fund Working Paper.

The Guardian (May 6, 2026). *Polymarket Trader Arrested over Israel–Iran War Bets*. The Guardian. URL: https://www.theguardian.com/world/2026/may/06/polymarket-israel-iran-war-arrest (visited on 06/01/2026).

The Washington Post (May 27, 2026). *Google Employee Charged with Insider Trading on Polymarket*. The Washington Post. URL: https://www.washingtonpost.com/technology/2026/05/27/google-employee-charged-with-insider-trading-polymarket/ (visited on 06/01/2026).

TradingView, I. via (2026). *Polymarket Taps Blockchain Data Firm Chainalysis to Curb Insider Trading*. TradingView News. URL: https://www.tradingview.com/news/invezz:9fa825c8a094b:0-polymarket-taps-blockchain-data-firm-chainalysis-to-curb-insider-trading/.

Tsang, K. P. and Z. Yang (2026). "The Anatomy of a Blockchain Prediction Market: Polymarket in the 2024 U.S. Presidential Election". Virginia Tech Working Paper.

United States Congress (n.d.). *Commodity Exchange Act § 6(c)(1), 7 U.S.C. § 9(1)*. Anti-manipulation and anti-fraud provision applicable to swaps; modelled on Section 10(b) of the Securities Exchange Act of 1934.

United States Department of Justice (May 2026a). *Google Employee Charged With Insider Trading*. U.S. Attorney's Office, Southern District of New York Press Release. URL: https://www.justice.gov/usao-sdny/pr/google-employee-charged-insider-trading.

United States Department of Justice (Apr. 2026b). *U.S. Soldier Charged With Using Classified Information To Profit From Prediction Market Bets*. Office of Public Affairs Press Release. URL: https://www.justice.gov/opa/pr/us-soldier-charged-using-classified-information-profit-prediction-market-bets.

Wang, E. (2026). "Smart Money on Polymarket: A Behavioral Anatomy of 273 Top Prediction-Market Traders". Empirical Study.

Wang, Z., L. Chao, Y. Bao, L. Cheng, J. Liao, and Y. Li (2026). *Polymarket Data: Complete Data Infrastructure for Polymarket*. https://huggingface.co/datasets/SII-WANGZJ/Polymarket_data. A comprehensive dataset and toolkit for Polymarket prediction markets.

Wolfers, J. and E. Zitzewitz (2004). "Prediction Markets". In: *Journal of Economic Perspectives* 18.2, pp. 107–126.

Zhu, J., Y. Wu, and C. Insixiengmay (2026). "Towards Detecting Insider Trading in Prediction Markets: Event-Aligned Multi-Scale Anomaly Detection". Tsinghua University Working Paper.

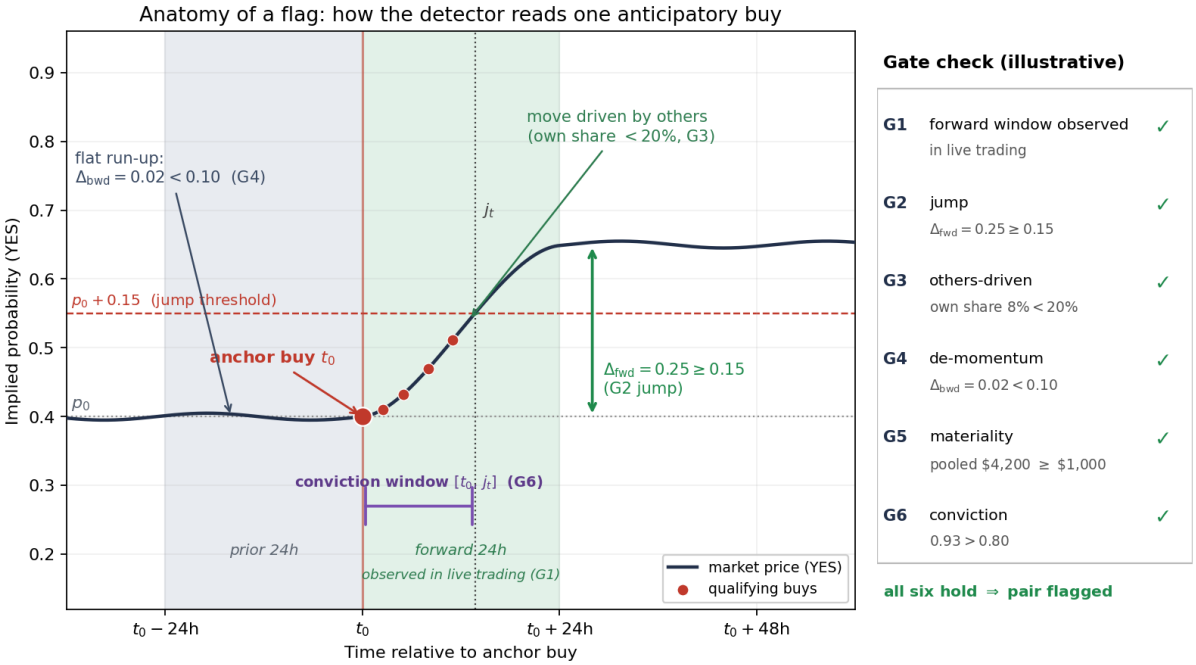

Figure 1. Anatomy of an ex-ante flag. A schematic, idealised illustration of how the detector reads a single anticipatory buy; the price path is stylised and carries no data. A wallet buys at the anchor t_0 while the market is flat (G4), and over the following 24 hours, which are observed in live trading (G1), the price moves at least 0.15 in the buy's direction (G2), driven by other participants rather than by the wallet's own volume (G3). Pooled qualifying buys must clear the materiality floor (G5), and the wallet must accumulate with directional conviction over the run-up $[t_0, j_t]$ until the move is realised (G6). The diagram conveys the mechanics only and is not evidence of the detector's performance.

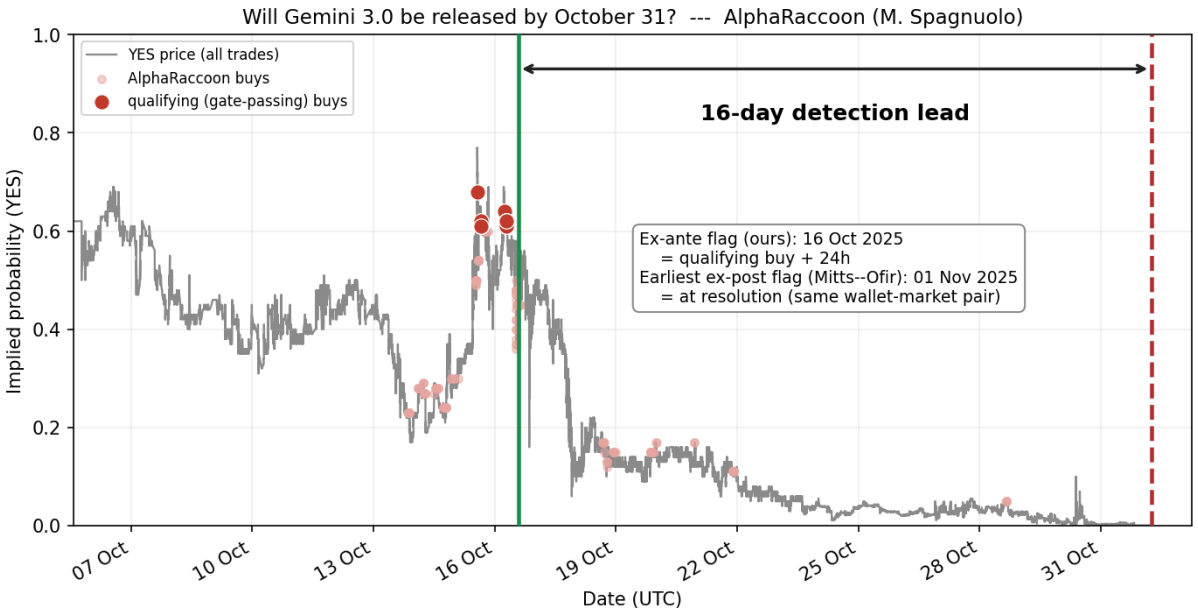

Figure 2. The ex-ante detection lead, illustrated on a single (wallet, market) pair flagged by both the ex-ante detector and Mitts–Ofir: *AlphaRaccoon* (Michele Spagnuolo, SDNY-charged) in "Will Gemini 3.0 be released by October 31?". *AlphaRaccoon* accumulated the NO side; the implied YES probability subsequently fell to zero as Gemini 3.0 did not ship by 31 October, resolving his position in the money. The dark-red points are the qualifying buys that satisfy the detector's per-buy gates (the flag additionally requires the pair-level materiality and conviction gates); the green line marks our ex-ante flag, the earliest qualifying buy plus the twenty-four-hour confirmation window (16 October 2025), and the red dashed line marks the resolution date (1 November 2025), the earliest point at which an outcome-conditioned method such as Mitts–Ofir could flag the same (wallet, market) pair. The sixteen-day gap is the detection lead. The forward move is observed in live trading, not at settlement, so it is a genuine repricing rather than the mechanical resolution.

Table 1. Sample construction funnel: Polymarket universe to analysis sample.

Stage	Filter criterion	Events	Markets	% of F0
F0	All Polymarket markets (snapshot 2026-05-04)	437,231	1,019,069	100.00
F1	Resolved markets, clean binary resolution	376,574	901,284	88.44
F2	Created on/after 2024-02-01	372,034	892,350	87.57
F3	Binary YES/NO markets	51,852	309,855	30.41
F4	Non-zero trading volume	47,367	257,326	25.25
F5	Duration $>$ 7 days	26,577	98,540	9.67
F6	Market volume $\geq$ \$1,000	21,932	75,278	7.39
F7	Parent-event volume $>$ \$10,000	19,100	71,031	6.97
F8	Category bucket and sports/weather/tweet veto	7,483	28,763	2.82
F9	Drop partial negRisk bundles (atomicity rule)	7,168	26,791	2.63

Notes. Filters are applied sequentially: each stage operates on the population that survives all prior stages. *Events* and *Markets* report the surviving counts after the stage's filter; *% of F0* expresses surviving markets as a percentage of the full F0 snapshot. The final analysis sample is the F9 population of 26,791 markets across 7,168 events.
Source. Z. Wang et al. (2026); Polymarket on-chain snapshot, 2026-05-04.

Table 2. Analysis sample by category.

Category	Events	Markets	Volume ($bn)	% Vol	Med. mkt vol. ($)	Med. dur. (days)	YES rate (%)
Geopolitics	1,385	3,359	4.31	20.9	101,834	36.3	23.6
Politics	1,903	5,165	3.94	19.1	63,860	51.8	25.1
Economy	295	1,228	3.55	17.2	40,335	34.4	20.4
Elections	419	2,524	3.46	16.8	56,899	41.7	15.2
Business	1,212	5,157	3.12	15.1	94,314	30.2	27.4
Entertainment	780	4,027	1.23	6.0	25,654	49.2	18.6
Tech	577	2,228	0.83	4.0	51,985	43.2	17.7
Finance	597	3,103	0.22	1.1	10,998	33.9	36.4
Total	7,168	26,791	20.67	100.0	48,472	36.8	23.9

Notes. The table describes the final analysis sample (stage F9), broken down by the eight category buckets. *Events* and *Markets* are counts of unique parent events and binary markets. *Volume* is total dollar trading volume in USD billions; *% Vol* is the category's share of total sample volume. *Med. mkt vol.* is the median market-level dollar volume; *Med. dur.* is the median market duration in days (`end_date − created_at`). *YES rate* is the share of markets resolving YES. Each market is assigned to exactly one bucket, so the columns sum to the total. Categories are sorted by total volume.
Source. Z. Wang et al. (2026); Polymarket on-chain snapshot, 2026-05-04.

Table 3. Trading activity by category.

Category	Wallet–market pairs	Unique wallets
Business	4,500,783	853,687
Politics	3,177,136	845,120
Geopolitics	2,944,242	733,198
Elections	1,724,620	625,601
Tech	1,366,177	449,891
Economy	1,244,492	544,185
Entertainment	1,210,899	394,400
Finance	584,801	175,105
Total	16,753,150	1,696,645

Notes. A wallet–market pair is a unique (wallet, market) combination; a wallet is counted whether it appears as maker or taker. Pairs partition across categories and sum to the total. Unique wallets do not sum to the total: a wallet active in several categories is counted once per category but once overall. Market makers are not excluded.
Source. Z. Wang et al. (2026); F9 analysis sample, snapshot 2026-05-04.

Table 4. Wallet activity breadth: number of categories per wallet.

Categories active in	Unique wallets	% of wallets
1	612,157	36.1
2	349,695	20.6
3	236,478	13.9
4	179,573	10.6
5	135,734	8.0
6	99,034	5.8
7	62,752	3.7
8 (all)	21,222	1.3
Total	1,696,645	100.0

Notes. A wallet is active in a category if it traded at least one market in that category (as maker or taker). Percentages may not sum to 100 due to rounding.
Source. Z. Wang et al. (2026); F9 analysis sample, snapshot 2026-05-04.

Table 5. Trader-type composition of the 1,696,645 wallets in the analysis universe. Shares may not sum to 100% due to rounding.

Trader type	Wallets	Share
Algorithmic	7,903	0.5%
Sophisticated	40,336	2.4%
Active retail	618,899	36.5%
Casual	778,961	45.9%
One-shot	223,290	13.2%
Maker-only	27,256	1.6%
Total	1,696,645	100.0%

Table 6. Distribution of 24-hour forward directional price moves, analysis sample.

Statistic of Δ_{fwd} (YES-equivalent, in the buyer's direction)	Value
Median move	≈ 0.00
90th percentile	≈ 0.05
95th percentile	≈ 0.12
99th percentile	≈ 0.38
Share of buys with $\Delta_{fwd} \geq 0.15$ (*JumpSize*)	3.8%

Notes. Computed over the 56.98 million buys whose full 24-hour forward window falls within their market's trading life, in the analysis sample (4.1). The distribution is a property of the universe and does not depend on the detector's conviction or materiality gates.

Table 7. Anti-artefact guards built into the detector.

Guard	Mechanism	Rules out
Forward window observed in live trading	G1	the jump being the settlement print rather than a real move
Mid-range anchor (≤ 0.85 / ≥ 0.15)	implied by G2	the move being terminal convergence to 0 or 1
Others-driven (own share $< 20\%$)	G3	the buyer's own price impact
De-momentum (prior move < 0.10)	G4	trend continuation or extrapolation

Table 8. Detector parameters.

Parameter	Symbol	Default	Rationale
Jump magnitude	*JumpSize*	0.15	Economic materiality in probability space
Window length	*Window*	24 h	News-cycle horizon; event-study tradition
Materiality floor	*MinSize*	\$1,000	MO \$500 / Liu \$1k anchor; conservative
Reverse-causality cap	*MaxShare*	0.20	Causal control; permissive yet binding
De-momentum bound	*MaxPriorMove*	0.10	Excludes momentum-following
Conviction threshold	*MinConv*	0.80	Clean directional accumulation; bimodal

Table 9. Funnel from the in-scope panel to the flagged population. The *timing set* (gates G1–G5) is the set on which the H1 permutation test is conducted; the *headline set* (gates G1–G6) is used for the remaining hypothesis tests and all descriptive statistics.

Step	Count	Wallets
Universe (in-scope panel)	63,343,761 fills / 26,789 markets	—
G1 (full 24h forward window observed)	56,977,751 candidate buys	—
G2 (forward move ≥ 0.15 in buyer's direction)	2,182,119	—
G3 (buyer's forward-window share < 0.20)	2,129,844	—
G4 (prior 24h directional move < 0.10)	1,546,491 qualifying buys	—
Aggregate qualifying buys to (wallet, market) pairs	396,380 pairs	—
G5 (anticipatory volume $\geq \$1{,}000$) $\rightarrow$ **timing set**	18,078	6,658
G6 (conviction > 0.80) $\rightarrow$ **headline set**	10,298	4,839

Table 10. H1 permutation test on the timing set (gates G1–G5; 18,078 pairs across 6,658 wallets). The null is generated by holding each wallet's trades and the market's price path fixed and redrawing the wallet's buy timestamps from the market's empirical trade-activity time distribution.

Statistic	Value
Observed: pairs clearing G1–G5	17,701
Permutation null: mean (SD), $n_{\text{perm}} = 200$	6,890 (39.5)
Permutation null: maximum across 200 permutations	6,995
Observed / null mean	2.6×
Standardised distance (effect-size descriptor)	$z \approx 273$
Empirical permutation p-value	$\leq 1/201 \approx 0.005$

Table 11. Overlap of the ex-ante headline set (10,298 pairs / 4,839 wallets) with three benchmark ex-post detectors. Panel A reports the load-bearing wallet-level enrichment; Panel B reports the supporting pair-level overlap, available only for the two benchmarks that flag at the trade level (the Della Vedova (2026a) detector flags only at the wallet level). Wallet universe $N_W = 1{,}696{,}645$; pair universe $N_{\text{pairs}} = 16{,}753{,}150$. "MO" is the Mitts and Ofir (2026) Layer-2-gated pairs in the top quartile by composite score; Liu and DV as reimplemented in Appendix D. Hypergeom. p is the one-sided upper-tail probability of the observed wallet overlap under independence.

Panel A. Wallet level (primary: convergent validity)

Benchmark	Their wallets	Overlap	% of ours	Expected	Enrichment	κ	Hypergeom. p
MO (gated, top-q.)	3,423	1,439	29.7%	9.8	147×	0.347	$< 10^{-200}$
Liu	3,643	493	10.2%	10.4	47×	0.114	$< 10^{-200}$
DV	2,683	269	5.6%	7.7	35×	0.070	$< 10^{-200}$

Panel B. Pair / event level (supporting: event orthogonality)

Benchmark	Their pairs	Overlap	% of ours	Jaccard	κ
MO (gated, top-q.)	14,061	1,122	10.9%	0.048	0.091
Liu	4,119	177	1.7%	0.012	0.024

Table 12. H3 comparison of anticipation pairs against the same wallets' other (wallet, market) pairs. Panel A reports the pooled comparison across the headline cohort of 4,839 flagged wallets; Panel B reports the within-wallet (paired) comparison, restricted to wallets with at least two flagged pairs and at least five other pairs ($n = 1{,}354$). Pooled magnitudes are mechanically inflated by the flag's selection on the realised forward move; the within-wallet comparison is the substantive test. The global universe baseline win rate is 48.5%.

Panel A. Pooled across the headline cohort ($n = 4{,}839$ wallets)			
	Anticipation pairs	Other pairs	Gap
n pairs	10,298	759,590	—
Win rate	80.7%	65.0%	+15.7pp
Median PnL	$1,947	$3	—
Mean PnL	$6,599	$23	—

Panel B. Within-wallet (paired; $n = 1{,}354$ wallets with ≥ 2 flagged & ≥ 5 other)			
	Anticipation pairs	Other pairs	Within-wallet gap
Mean win rate	81.9%	59.6%	+22.3pp
Median per-wallet median PnL	$2,265	$8	—
% wallets winning more on anticipation			81%
% wallets with higher median PnL on anticipation			87%

Table 13. H3 matched-jump robustness. Each paired wallet's comparison set is restricted to non-flagged pairs in which the wallet also had at least one buy followed by a ≥ 0.15 directional move within 24 hours but failing one or more of the behavioural gates G3–G6. The eligible sample is the 1,303 of the 1,354 paired wallets that have at least one matched-jump-other pair. The sign test is computed on non-tied wallets. Within-wallet mean win rates on the same wallets' non-matched-other pairs (i.e. non-flagged pairs in which no comparable forward move occurred) are reported for context.

Statistic	Value
Mean win rate on anticipation pairs	82.1%
Mean win rate on matched-jump-other pairs	71.5%
Within-wallet gap (anticipation − matched-jump-other), mean	+10.6pp
Within-wallet gap, median	+14.0pp
Positive / tied / negative wallets (of 1,303)	833 / 145 / 325
% positive among non-tied ($n = 1{,}158$)	72%
Sign test (binomial vs 50%, non-tied)	$p \approx 7 \times 10^{-52}$
Context: mean win rate on non-matched-other pairs	57.9%

Table 14. Discreteness of the anticipated move, flagged pairs versus a within-market control of unflagged anticipating pairs. A move is *discrete* if a ≥ 0.15 directional move is realised within a ≤ 2h sub-window that carries the price across the qualifying threshold and persists to the 24-hour horizon; each pair is classified on its single qualifying buy.

Cohort	*n* pairs	Discrete share
Flagged (headline set)	10,298	47.1%
Within-market control (unflagged anticipating)	289,922	47.0%
Two-proportion test: $z = +0.23$, $p \approx 0.82$ (null not rejected)		

Table 15. Flag intensity by market information-generating category. The numerator of each cell is the count of headline flagged (wallet, market) pairs falling in markets of that category; the denominator is the count of universe (wallet, market) pairs in markets of the same category. Aggregated rows pool the two MNPI-admitting and the two no-MNPI categories. The ratio's 95% confidence interval is bootstrapped over markets (2,000 resamples).

Market category	Flag intensity (per 10,000 universe pairs)
MNPI-admitting	
Action	8.1
Vote	7.0
Combined	7.6
No-MNPI	
Stochastic	4.7
Performance	1.8
Combined	4.6
Ratio (MNPI-admitting / no-MNPI)	1.63× [1.35, 1.94]

Table 16. H5 per-jump normalisation. A jump is a ≥ 0.15 absolute move in a market's daily last-price series, measured identically across market types; flags per 100 jumps removes jump frequency from the comparison. The MNPI-admitting gradient strengthens rather than attenuates, and MNPI-admitting markets are in fact slightly less jump-prone than no-MNPI markets. The two metrics are complementary: per universe pair controls trader density, per jump controls jump frequency.

	per 10k universe pairs	per 100 jumps	jumps per universe pair
MNPI-admitting (vote+action)	7.6	41.2	0.0018
No-MNPI (stochastic+performance)	4.7	22.2	0.0021
Ratio (admitting / no-MNPI)	1.63×	1.85×	0.88×

Table 17. Composition of the flagged cohort by trader type (4,839 wallets / 10,298 events). Types are assigned by an activity heuristic over each wallet's lifetime history (distinct taker transactions, transactions per active day, dollar volume, market-concentration Herfindahl index, activity span); "algorithmic" denotes activity intensity, not a verified automated account. *Orders*, *volume*, and *markets* are wallet-level medians within each type, computed over each flagged wallet's lifetime taker activity in the in-scope panel; *PnL* is the within-class median of realised PnL per flagged anticipation pair. The *maker-only* row captures wallets with no observable taker activity in the panel whose flagged buys arrived passively via resting limit orders.

			Within-type median			
Trader type	Wallets	Events	Orders	Volume	Markets	PnL
Algorithmic	661 (13.7%)	3,864 (37.5%)	2,337	$954,245	377	$2,571
Sophisticated	2,973 (61.4%)	5,114 (49.7%)	123	$117,584	35	$1,780
Active retail	621 (12.8%)	715 (6.9%)	25	$24,044	7	$1,294
Casual	495 (10.2%)	512 (5.0%)	5	$5,915	2	$1,145
One-shot	56 (1.2%)	59 (0.6%)	1	$667	1	$1,959
Maker-only	33 (0.7%)	34 (0.3%)	0	$0	0	$1,338
Cohort	4,839	10,298	—	—	—	—

Table 18. Composition and economics of the 10,298 flagged events by market category (categories assigned by the market-classifier of 4.2). *Median PnL*, *win rate*, and *total PnL* are realised whole-market PnL (settlement-based, gross of fees; 5.1); PnL is mechanically inflated by the detector's conditioning on the realised forward move (5.4) and is reported descriptively.

Category	Events	Median PnL	Win rate	Total PnL
Business	2,801 (27.2%)	$1,576	79%	$9.2M
Geopolitics	2,747 (26.7%)	$1,750	79%	$14.7M
Politics	2,203 (21.4%)	$2,178	83%	$14.1M
Elections	1,333 (12.9%)	$2,978	85%	$22.4M
Tech	384 (3.7%)	$2,308	88%	$2.2M
Entertainment	354 (3.4%)	$2,117	85%	$2.0M
Economy	321 (3.1%)	$2,245	73%	$3.1M
Finance	155 (1.5%)	$1,129	72%	$0.4M
All	10,298	$1,947	80.7%	$68.0M

Table 19. Dissection of the 10,298 flagged events by market resolution type (the MNPI taxonomy of 4.2). Action and vote markets admit material non-public information; stochastic and performance markets resolve on public data, making the existence of material non-public information implausible by the resolution mechanism (not strictly impossible). PnL is realised whole-market PnL (settlement-based, gross of fees; 5.1) and selection-inflated (5.4).

Resolution type	Events	Admits MNPI	Median PnL	Win	Total PnL
Action	3,973 (38.6%)	yes	$1,892	80%	$20.15M
Vote	2,569 (24.9%)	yes	$2,785	84%	$32.06M
Stochastic	3,730 (36.2%)	no	$1,597	79%	$15.67M
Performance	9 (0.1%)	no	$1,102	100%	∼$0
Other	17 (0.2%)	ambiguous	$1,904	76%	∼$0
MNPI-admitting	6,542 (63.5%)				$52.20M
No-MNPI	3,739 (36.3%)				$15.71M

Table 20. Trader-type composition (pair shares) at three stages of the detector funnel, and per-type clearance rate at the $1,000 materiality floor. The clearance rate is each type's share of pre-floor (G1–G4) pairs that satisfy the floor; composition rows are within-stage pair shares. Trader types use the distinct-taker-transaction taxonomy of 4.3.

Stage	Pairs	Algo.	Soph.	Active retail	Casual / smaller
Pre-floor (G1–G4)	396,380	30.2%	24.7%	37.3%	7.9%
Timing set (G1–G5)	18,078	46.0%	42.5%	7.1%	4.5%
Headline set (G1–G6)	10,298	37.5%	49.7%	6.9%	5.9%
$1,000 clearance rate (within type)	—	7.0%	7.9%	0.9%	2.4–3.4%

Table 21. Per-opportunity flag intensity by trader type, market category, and resolution type, per 10,000 opportunities (each stratum's flagged pairs over its own opportunity count; strata partition the 16.75M-pair universe). Panel A reports pre-floor (G1–G4) and headline (G1–G6) rates; Panels B and C report headline rates. Panel C reproduces the H5 intensities (5.6).

Panel A. By trader type

Trader type	Pre-floor	Headline
Algorithmic	891.7	28.81
Sophisticated	514.5	26.91
Active retail	140.2	0.68
Casual	105.5	1.97
One-shot	85.4	2.06
Maker-only	158.2	3.83

Panel B. By market category (headline)

Category	per 10,000
Geopolitics	9.33
Elections	7.73
Politics	6.93
Business	6.22
Entertainment	2.92
Tech	2.81
Finance	2.65
Economy	2.58

Panel C. By resolution type (headline)

Resolution type	per 10,000
Action	8.07
Vote	6.96
Stochastic	4.66
Performance	1.82
MNPI-admitting	7.59
No-MNPI	4.65

Table 22. The five documented-insider cases. Verification tier reflects evidentiary status as of mid-2026; documented profit is the figure in public reporting (Appendix G).

Pseudonym	Domain	Profit	Status
Tier 1: criminally charged			
AlphaRaccoon (M. Spagnuolo)	Google Year-in-Search & products	\$1.2M	Charged, arrested (SDNY)
van Dyke (G. Van Dyke)	Venezuela / Maduro operation	\$0.4M	Indicted, pleaded not guilty
Tier 2: investigated or flagged, no charges			
ricosuave666	Israel–Iran war	\$0.15M	Account deleted; linked in reporting to Israeli case
6741	Nobel Peace Prize	\$0.05M	Nobel Institute investigation
Magamyman	Iran war / geopolitics	\$0.55M	Highlighted by US lawmakers; Israeli police probe

Table 23. Detection of the five cases by the ex-ante detector and three ex-post benchmarks. MO is Mitts–Ofir's gated pairs (their Layer-2 bet-size gate) in the top quartile by composite score; Liu and Della Vedova (DV) as in 5.3 (DV flags at wallet level only).

Wallet	Ours (ex ante)	MO	Liu	DV
AlphaRaccoon (charged)	**Y**	Y	–	Y
van Dyke (charged)	–	–	Y	–
Magamyman	**Y**	Y	Y	Y
6741	–	Y	–	–
ricosuave666	–	–	–	–
Total / 5	2	3	2	2

Further prediction market concepts

This appendix supplements Section 2.4 with operational detail that is not essential to the main analysis but useful for an interested reader. Section A.1 covers the token-level mechanics underlying Polymarket trading volume, Section A.2 describes Polymarket's fee structure, Section A.3 introduces Kalshi as a comparative case that clarifies what Polymarket's pseudonymity and on-chain transparency contribute to the feasibility of insider-trading detection, and Section A.4 documents the governance vulnerabilities that affect outcome resolution on Polymarket.

A.1 Token-level mechanics

On Polymarket, outcome shares are implemented as conditional tokens recorded on the Polygon blockchain. The core guarantee is straightforward: one YES token plus one NO token can always be redeemed for one USDC (a stablecoin pegged to the U.S. dollar) after a market resolves (Dubach, 2026). This redemption guarantee anchors prices and prevents persistent deviations from the no-arbitrage condition discussed in Section 2.4.

New shares are created through a process called *minting*. When two traders take opposite sides, for instance one willing to pay \$0.70 for a YES share and another willing to pay \$0.30 for a NO share, their combined \$1.00 is locked as collateral in a smart contract, and a new pair of YES and NO shares is created and distributed to the respective traders. The inverse operation, *burning* (or "merging"), occurs when a complete set of YES and NO shares is returned to the contract, which destroys the tokens and releases the underlying collateral. A third operation, *converting*, is specific to categorical markets and allows a trader to convert a YES share of one outcome into NO shares of all other outcomes, enabling portfolio rebalancing without full unwinding (Tsang and Yang, 2026). The distinction between trading existing shares and minting new ones matters for measurement: Tsang and Yang (2026) show that naively summing on-chain transaction flows, which mix ordinary trades with minting and burning, can substantially overstate actual trading volume.

A.2 Fee structure

Polymarket uses a maker-taker fee structure. Makers, who post limit orders and provide liquidity to the order book, earn rebates; takers, who execute against existing orders, pay fees that vary by market category, peaking at 1.8% for cryptocurrency contracts as of early 2026 (Gomez-Cram et al., 2026). This incentive structure encourages liquidity provision and is standard in modern exchange design.

A.3 Kalshi as a comparative case

The main text treats Kalshi only briefly, as a comparative case used to motivate the detection challenges specific to Polymarket. The discussion below describes its trading mechanics, regulatory status, outcome-resolution procedure, and the existing empirical literature in greater detail.

Kalshi was the first federally licensed prediction market in the United States, having received a CFTC designation as a designated contract market in November 2020 (Burgi et al., 2025). Unlike Polymarket,

it operates entirely within a centralised infrastructure without a blockchain settlement layer, under the regulatory oversight of the CFTC (Beylin, 2026). Burgi et al. (2025) characterise Kalshi as a quote-driven market in which price quotes range from one cent to 99 cents. Traders participate either as makers, posting quotes that others can accept, or as takers who accept existing quotes. Using transaction-level data on over 300,000 contracts, Burgi et al. (2025) document that Kalshi prices are informative and become increasingly accurate as markets approach resolution, consistent with the theoretical prediction that uncertainty diminishes as new information arrives.

Outcome resolution on Kalshi is also fundamentally different from Polymarket's UMA-based approach. Each contract specifies in advance the resolution source and decision rule, for example, "this contract resolves YES if the Associated Press calls the state for Candidate X by 11:59 p.m. on November 5" (Beylin, 2026). This rules-based approach is simpler and more predictable than oracle voting, but introduces a degree of centralisation risk: participants must trust both the platform and its regulator to apply the stated criteria fairly.

Two structural differences from Polymarket bear directly on the detection problem this thesis addresses. First, Kalshi requires identity verification of all participants under CFTC rules, which both eliminates the pseudonymity that enables Sybil-style identity fragmentation on Polymarket and gives regulators direct access to participant identities should suspicious activity be detected. Second, Kalshi's transaction data are not published publicly on a blockchain, so independent academic verification is more constrained than on Polymarket; the empirical literature on Kalshi consequently relies on private data shares with the platform and cannot exploit wallet-level information. Taken together, these design choices make Kalshi a useful benchmark for assessing what Polymarket's pseudonymity and on-chain transparency contribute, in both directions, to the feasibility of insider-trading detection.

A.4 Governance attacks on the UMA oracle

The decentralised resolution layer described in Section 2.4 is not immune to capital concentration. Abid (2026) documents two oracle resolution disputes in 2025 in which markets totalling more than $250 million settled through governance votes that traders alleged were captured by token whales. The most-discussed instance is the "Will Ukraine agree to Trump's mineral deal before April?" market, in which a single actor controlling approximately 25% of UMA voting power resolved a USD 7 million contract as YES despite no agreement having been publicly reported; the contract's implied probability moved from 9% to 100% over a roughly forty-eight hour window. Although such conduct is plausibly reachable under CEA § 6(c) as manipulation, no enforcement action has been brought to date. The episode illustrates that detection methods focused on pre-event informed trading must be supplemented by separate scrutiny of governance-layer manipulation, a distinction returned to in the empirical chapters.

MNPI-classification: Implementation and validation

This appendix documents the implementation of the MNPI taxonomy introduced in 4.2 and its validation against a hand-coded subsample. The replication package, namely the rule template and the verbatim classification prompt was provided directly by Professor Della Vedova; we reproduce his four-tier

procedure faithfully and disclose three deviations in B.3. Classification uses the market-question text only.

B.1 The four-tier priority cascade

Markets are classified by a priority cascade in which the first tier to resolve a market wins:

Tier 1: Platform category tag. Polymarket's own category field is used only where it is high-precision: the "Elections" category maps to Vote. Other platform categories are not trusted to map one-to-one onto the information-type taxonomy and are passed to the next tier.

Tier 2: Platform event tags. A whitelist of high-precision platform event tags maps to Vote, Performance, or Stochastic. The whitelist is conservative: if whitelisted tags on the same market disagree, Tier 2 abstains and the market passes down.

Tier 3: Deterministic rule corpus. Approximately 329 regular-expression patterns over the question text, grouped by MNPI type (Stochastic, Vote, Action, Performance) and evaluated first-match-wins within a fixed group order. Each rule carries a rationale. Illustrative rules: "Will [team] win [game]?" maps to Performance; "Will [person] win [award or nomination]?" maps to Vote; military strike, attack, or invasion verbs map to Action; price, benchmark, sales, or viewership terms map to Stochastic. The published template ships a representative subset of this corpus; our extensions are documented in B.3.

Tier 4: Language-model fallback. Residual markets unmatched by the deterministic tiers are classified by a large language model using the verbatim prompt plus our addendum (B.3). Operational details: questions are batched into groups of 30, the model returns a JSON array of `{id, category}` pairs, results are checkpointed by `condition_id` so that re-runs only process unresolved markets, and the per-market `mnpi_source` field records the resolving tier.

The output schema for each market is `{condition_id, question, category, mnpi_type, mnpi_source, mnpi_detail}`.

B.2 Decision principle

The single operative question that the rules and hand-coders both apply is: *who, if anyone, could know the resolution before the public?* The classification is by how the outcome is determined, not by what the question is about. Several boundary conventions are used consistently and were also applied during hand-coding:

- Election win is Vote; an approval rating or poll average is Stochastic.
- A regulator or committee approval is Action; the resulting market price is Stochastic.
- A central-bank rate decision, for example an FOMC vote, is Action; the realised rate level is Stochastic.

B.3 Deviations from the published implementation

Three deviations from the published implementation are disclosed for transparency:

1. **Rule-corpus extension.** The published template ships a representative subset of the rule corpus. We extended it toward the full corpus of approximately 329 patterns and added rules for our sample's phrasing, including military-verb conjugations and plurals, benchmark and AI-model scores, sales and sell-out terms, appearance counts, reality-show elimination patterns, and chart-position rules.
2. **Model upgrade.** The published run used Claude Haiku. We used Claude Sonnet 4.6 for the residual after Haiku under-performed on the long tail in our validation: Haiku agreed with the hand-coded sample substantially less often than the deterministic tiers, motivating the upgrade. Sonnet materially improved the residual.
3. **Prompt addendum.** We appended clarifications to the verbatim prompt to sharpen the two boundaries our validation showed the model getting wrong:
   - *Action versus Other.* If a specific identifiable person or small group plausibly knows the outcome first, label Action; reserve "Other" for genuinely ill-posed items.
   - *Performance scope.* Restrict Performance to genuine open competitions; route military actions to Action, polls and ballots to Vote, and benchmark, sales, price, or viewership outcomes to Stochastic.

   The verbatim prompt body provided by Della Vedova is otherwise unchanged.

B.4 Coverage by tier and label distribution

The deterministic tiers (Tiers 1 through 3) settle approximately 82.7% of markets; the language-model fallback (Tier 4) settles the remaining 17.3%. The per-tier breakdown is reported in Table 24, and the final label distribution across the 26,791 markets in the analysis universe in Table 25.

Table 24. Coverage of the four-tier classification cascade across the 26,791 markets in the analysis universe.

Tier	Share of markets settled
Tier 1: Platform category tag	9.4%
Tier 2: Platform event-tag whitelist	37.7%
Tier 3: Deterministic rule corpus	35.6%
Tier 4: Language-model fallback	17.3%
Total	100.0%

Table 25. Final MNPI label distribution across the 26,791 markets in the analysis universe.

MNPI type	Markets	Share
Stochastic	12,987	48.5%
Action	6,850	25.6%
Vote	6,526	24.4%
Performance	328	1.2%
Other	100	0.4%

The small Performance class is a property of the analysis universe rather than a classification failure: the universe excludes sports markets at the upstream filtering stage, so few Performance-type markets survive into the classification step. The "Other" bucket is a tiny residual of genuinely ill-posed questions.

B.5 Validation design and results

Design. A stratified, hand-coded validation sample of 200 markets was drawn, approximately 40 markets per final label, with a fixed random seed for reproducibility. Coding was blind: a workbook presented row identifier, question text, and a dropdown over {vote, action, performance, stochastic, other} together with a rubric sheet; the automated label was held in a separate hidden key not visible to the coder. Each market was hand-coded from the question text alone, then scored against the automated label. Overall agreement, per-category agreement, and a confusion matrix (rows: hand-coded; columns: automated) were computed.

Result. Overall agreement is 81.0%. Per-category agreement is reported in Table 26.

Table 26. Per-category agreement between hand-coded and automated MNPI labels on the 200-market stratified validation sample.

Category	Agreement (hand-coded vs. automated)
Stochastic	97.6%
Action	83.8%
Vote	83.7%
Performance	61.5%
Other	50.0%
Overall	81.0%

Disagreements concentrate at MNPI-type boundaries (the conventions of 4.2) and at the two small classes, Performance and Other, where each individual market is a large share of its stratum and a small number of edge cases moves the per-category percentage sharply. The three classes that carry the analysis (Stochastic, Action, and Vote) agree at 83% to 98%.

The full confusion matrix is reported in Table 27. Two observations are visible in it that the per-category numbers alone do not convey. First, Vote's only meaningful leakage is into Action (7 of 43 hand-coded Vote markets are labelled Action by the automated procedure), which is consistent with the Vote/Action boundary being the closest of the four-way distinctions. Second, the "Other" bucket is rarely chosen by the model: only 3 markets are labelled Other by the automated procedure outside the hand-coded Other stratum, which is the intended effect of the prompt addendum on the Action-versus-Other boundary (B.3).

Table 27. Confusion matrix on the 200-market stratified validation sample. Rows are hand-coded labels; columns are automated labels. The diagonal counts agreements; off-diagonal entries are disagreements. The hand-coded *n* per category differs from the stratified draw target of approximately 40 per automated label, since some markets were recoded by hand.

Hand-coded ↓ / Automated →	Vote	Action	Performance	Stochastic	Other	Hand-coded *n*
Vote	36	7	0	0	0	43
Action	3	62	1	5	3	74
Performance	4	3	16	3	0	26
Stochastic	0	1	0	40	0	41
Other	1	2	1	4	8	16

B.6 Limitations

Four limitations of the classification procedure are worth flagging:

1. Classification uses the question text only.
2. The language-model tier is non-deterministic. Checkpointing and a fixed prompt mitigate this, but exact reproduction of the Tier 4 labels depends on the model snapshot.
3. Hand coding was performed by a single coder; inter-coder reliability was not computed.
4. The Performance and Other strata are small, so per-category agreement on them is noisy.

Trader-type taxonomy: Implementation and sensitivity

This appendix documents how the trader-type taxonomy used throughout the characterisation (5.7) is constructed, and shows that the headline composition is robust to the one implementation choice that affects it materially. It records the provenance of the five-class scheme and what we specified ourselves, the exact computational steps and feature definitions, and the first-match classification rule with its thresholds. It then reports the taxonomy's sensitivity to the activity unit, the choice that most directly shapes the flagged cohort's composition, before closing with the caveats that bound how the labels should be read. Nothing here changes a substantive result; the purpose is to make the classification fully reproducible and to make explicit where the labels are descriptive proxies rather than verified identities.

C.1 Provenance and what we specified

The five classes of Della Vedova (2026a) (algorithmic, sophisticated, active retail, casual, one-shot) and their threshold values are adopted as published in 4.3; we relabel his "Bot" class to "Algorithmic" to signal that the label denotes activity intensity rather than verified automation. The activity unit (the distinct taker transaction) and the sixth, residual Maker-only class are our additions for this dataset; the original published procedure does not prescribe them.

C.2 Exact computational steps

The taxonomy is computed in a single streaming pass over the in-scope trade panel:

1. Read 63,343,761 fills and take the taker wallet on each row.
2. Aggregate per taker wallet: distinct transaction count; volume; first and last transaction timestamps. The span and transactions-per-active-day features are derived from these as specified in the formal definitions below.
3. Compute the HHI: group by $(\text{taker}, \text{market})$, count distinct transactions per market for each wallet, and aggregate to $\sum_m c_{w,m}^2 / \left(\sum_m c_{w,m}\right)^2$ per wallet.
4. Construct the universe as the union of all taker and maker addresses (1,696,645 wallets). Maker-only

wallets, those with $n = 0$, are assigned Maker-only directly.

5. Apply the first-match classification rule of C.5.

C.3 Implementation details

Two implementation details are immaterial to the substantive results but worth recording:

- *Hash for memory.* The `transaction_hash` field is hashed to a 64-bit integer before `n_unique()` is applied, for memory efficiency on the 63-million-row column. Over the approximately 42.9 million distinct transactions in the panel the collision probability under a 64-bit hash is negligible, and the resulting transaction counts are exact for all practical purposes.
- *Volume aggregation unit.* Volume is summed at the fill level rather than the transaction level: a multi-fill order's dollar size is the sum of its constituent fills, so the natural unit for the volume feature is fill-level even though the frequency unit is the deduplicated transaction. Only the count requires deduplication.

C.4 Formal feature definitions

For each wallet w with full in-scope taker history, let T_w denote the set of distinct taker transactions in which w is the taker and $T_{w,m} \subseteq T_w$ the subset in market m. Let F_w denote the set of taker-side fills, $c_{w,m} = |T_{w,m}|$, and $\text{usd}(f)$ the dollar size of fill f. The five features are:

- Transactions: $n_w = |T_w|$.
- Volume: $V_w = \sum_{f \in F_w} \text{usd}(f)$.
- Span: $s_w = (t_{\text{last},w} - t_{\text{first},w})/86{,}400$, in days.
- Transactions per active day: $\text{tpd}_w = n_w / \max(s_w, 1)$.
- Concentration: $\text{HHI}_w = \sum_m \left(c_{w,m}/n_w\right)^2$.

Since every transaction settles in a single market, $\sum_m c_{w,m} = n_w$ for every wallet, so the HHI form above and the step-3 form $\sum_m c_{w,m}^2 / \left(\sum_m c_{w,m}\right)^2$ coincide.

C.5 Classification rule and thresholds

Each wallet is assigned to the first class for which it qualifies, in the order below; the threshold values are those of Della Vedova (2026a), applied to the distinct-taker-transaction unit defined above.

1. *Maker-only*: $n = 0$.
2. *Algorithmic*: more than 50 transactions per active day, or more than 1,000 transactions in total.
3. *Sophisticated*: volume above \$10,000, HHI below 0.5, and span exceeding 30 days, without meeting the Algorithmic thresholds. Substantial, diversified, sustained activity without the high frequency that defines Algorithmic.
4. *Active retail*: 10 to 1,000 transactions, not meeting the Sophisticated criteria.
5. *Casual*: 2 to 9 transactions.
6. *One-shot*: 1 transaction.

C.6 Sensitivity to the activity unit

Because the frequency thresholds in the classification rule are unit-dependent, we report how the flagged-cohort composition responds to the choice of activity unit, confirming that the distinct-taker-transaction unit is the appropriate one for this analysis. Across the whole panel, the 63.3 million fills collapse to 42.9 million distinct taker transactions, a mean of 1.48 fills per transaction, with 18.5% of transactions multi-fill. Table 28 reports the composition of the 4,839-wallet flagged cohort under each unit.

Table 28. Composition of the flagged cohort (4,839 wallets) under a naive fills-based activity count (counting both maker- and taker-side fills) versus the distinct-taker-transaction unit used in the main analysis.

Trader type	Fills (naive)	Taker transactions (headline)
Algorithmic	47.1%	13.7%
Sophisticated	34.8%	61.4%
Active retail	15.7%	12.8%
Casual	2.2%	10.2%
One-shot	0.1%	1.2%
Maker-only	0.0%	0.7%

Of the 4,839 flagged wallets, 2,261 (46.7%) are assigned to a different class under the two units. The dominant reclassification is from Algorithmic under fills to Sophisticated under taker transactions, affecting 1,461 wallets. Of the 2,280 wallets classified as Algorithmic under fills, 71.0% reclassify away under the taker-transaction unit, and 64.1% reclassify to Sophisticated specifically. The mechanism has two components: the taker-transaction unit differs from a naive fill count by (i) collapsing multi-fill orders within a single transaction into one count and (ii) excluding maker-side activity entirely. Both push high-fill wallets out of the Algorithmic class. Under a naive fill count the Algorithmic share of the flagged cohort would be overstated by roughly 3.4× (47% versus 14%), and nearly half the flagged cohort would be classified differently. The taker-transaction unit, which corresponds to a trading decision rather than an order-book event, is what produces the sophisticated-dominated composition of the headline set documented in 5.7.

C.7 Caveats

Four caveats apply to the taxonomy:

1. "Algorithmic" and every other class is a behavioural activity label, not a verified identity. On-chain data cannot confirm automation, market-making, or any other status; the labels are descriptive proxies.
2. The thresholds are the conventional values adopted from Della Vedova (2026a) and are not validated against ground-truth account types.
3. Maker-only is a residual rather than a positive identification of market-making activity: these wallets' decision frequency is unobservable from the trade record, since quote placements are not present at the same granularity as taker fills.
4. Activity is measured over the in-scope panel only; behaviour outside the analysis universe is not counted.

Benchmark detectors

This appendix specifies the three ex-post benchmark detectors against which our screen is validated (5.3, 5.10), so that each can be reproduced exactly and the points at which we depart from the published designs are explicit.

D.1 Mitts and Ofir: Signals, weights, and the flag

Theoretical motivation. Mitts and Ofir (2026) design their detection approach around the premise that informed traders reveal themselves by trading differently from the rest of the market along several dimensions at once: they stake unusually large amounts, both relative to other participants and relative to their own history; they earn disproportionate profits; they buy disproportionately close to the resolution event; and they hold one-sided, directional positions. The method standardises each of these dimensions into a signal and aggregates the five signals into a single composite suspiciousness score for every (wallet, market) pair.

Eligible population and signal construction. The unit of analysis is a (wallet, market) pair. A first filter retains only pairs in which the wallet purchased at least $500 of total volume in that market, removing casual participants whose activity is too small to distinguish from noise. Because the cross-sectional signals are defined relative to a within-market reference distribution, markets with fewer than three eligible wallets are dropped. On this filtered sample the five signals are constructed as follows.

Cross-sectional bet z-score. Within each market, the total USD bought by each eligible wallet is standardised relative to the market mean and standard deviation:

$$z_{\text{bet,cross}} = \frac{\text{total\_USD\_bought} - \mu_{\text{market}}}{\sigma_{\text{market}}}. \tag{D.1}$$

Within-wallet bet z-score. Across all markets in which a wallet holds eligible positions, the wallet's USD bought in each market is standardised relative to the wallet's own mean and standard deviation:

$$z_{\text{bet,within}} = \frac{\text{total\_USD\_bought} - \mu_{\text{wallet}}}{\sigma_{\text{wallet}}}. \tag{D.2}$$

Profit z-score. Within each market, the realised PnL of each eligible wallet is standardised relative to the distribution of PnL across all eligible participants:

$$z_{\text{profit}} = \frac{\text{realised\_PnL} - \mu_{\text{PnL,market}}}{\sigma_{\text{PnL,market}}}. \tag{D.3}$$

Late-buy fraction. The share of a wallet's total USD bought that occurred within the final 48 hours before the market's last recorded trade. The last-trade timestamp is used as the resolution anchor rather than the scheduled `end_date` field, which is frequently inaccurate for NegRisk event markets:

$$\text{late\_buy\_frac} = \frac{\text{USD\_bought\_in\_final\_48h}}{\text{total\_USD\_bought}}. \tag{D.4}$$

Directional concentration. The degree to which a wallet's trades in a market were one-sided. Mitts and Ofir (2026) describe this signal only qualitatively (a value of one denotes a purely one-sided position); we operationalise it as

$$\text{dir\_conc} = \left|\,\text{buy\_share} - 0.5\,\right| \times 2, \qquad \text{buy\_share} = \frac{\text{USD\_bought}}{\text{USD\_bought} + \text{USD\_sold}}. \tag{D.5}$$

Composite score. The five signals are combined into a single composite score using the weights specified in the original paper:

$$25 \cdot z_{\text{bet,cross}} + 20 \cdot z_{\text{bet,within}} + 30 \cdot z_{\text{profit}} + 15 \cdot \text{late\_buy\_frac} + 10 \cdot \text{dir\_conc}. \tag{D.6}$$

The composite is an unbounded linear combination of the raw signal values. Profitability receives the highest weight (30), followed by cross-sectional bet size (25), within-wallet bet size (20), timing (15), and directionality (10).

Flagged population. In Mitts and Ofir (2026)'s original design a pair enters the flagged population if it passes the Layer 2 gate, that is, if $z_{\text{bet,cross}} > 2$ or $z_{\text{bet,within}} > 2$; the composite score then ranks pairs within that gated set rather than serving as a flag criterion of its own. The gate ensures that at least one dimension of bet-size unusualness is statistically meaningful before a pair is considered. We depart from the original on the final selection step. Applied to our sample, the bare gate is dominated by high-volume wallets whose bets are large for mechanical reasons: their baseline stake is simply high, so the gate over-flags whales rather than isolating the most suspicious behaviour. We therefore retain only the top quartile by composite score within the gated set. This restriction keeps the pairs that are most anomalous across the combined bet-size, profitability, timing, and directionality dimensions, and it brings the number of flagged pairs into line with our own detector's flagged count, so that the cross-method comparisons reported later reflect differences in which wallets each method selects rather than how many it admits. The benchmark used throughout this thesis is therefore the intersection of the Layer 2 gate and the top composite-score quartile.

D.2 Liu: Fingerprints, thresholds, and the flag

Theoretical motivation. Liu (2026) approaches detection from a behavioural angle: rather than testing a single statistical signature, the method assembles a fingerprint of five trader characteristics that, in combination, separate informed traders from the winning crowd. An informed trader is expected to bet with high directional conviction, to enter at favourable prices, to operate from a young or single-purpose wallet, to concentrate on few markets, and to realise large profits. Liu's full classifier augments this behavioural scan with a Sybil-cluster propagation step that links wallets sharing a common withdrawal destination. We replicate the behavioural scan alone and set the fund-flow enrichment aside, since it depends on off-chain transaction tracing outside the on-chain scope of this study.

Eligible population. Two conditions define the eligible pool E of (wallet, market) pairs: the wallet won its position in the market (positive realised PnL on the resolved side), and the pair's taker-side volume is at least \$1,000. The winning-side requirement confines the scan to traders who bet correctly, asking

whether some correct bets carry patterns consistent with foreknowledge rather than luck; the volume floor removes negligible positions that could rank extreme by chance. Within each market m we write $W_m \subseteq E$ for the eligible winners, used as the within-market reference whenever $|W_m| \geq 10$.

The five fingerprints. Liu (2026) defines the five signals qualitatively and fixes the aggregation rule stated below, but not the threshold at which each signal fires; we operationalise every flag as a quartile cut. For a metric x assessed over a reference set R, define the extreme-quartile indicators

$$\mathbb{1}^{\uparrow}(x;R) = \mathbb{1}\{x \geq Q_{75}(R)\}, \qquad \mathbb{1}^{\downarrow}(x;R) = \mathbb{1}\{x \leq Q_{25}(R)\},$$

where $Q_{75}(R)$ and $Q_{25}(R)$ are the upper and lower quartiles of the metric over R. The five fingerprints are then:

Directional conviction (f_1). $f_1 = \mathbb{1}^{\uparrow}(Conv; W_m)$, where *Conv* is the direction-concentration metric (Mitts and Ofir (2026), Signal 5): a top-quartile rank against the market's eligible winners.

Favourable entry (f_2). $f_2 = \mathbb{1}^{\downarrow}(VWAP; W_m)$, where *VWAP* is the volume-weighted price paid per winning-side token: a bottom-quartile rank within the market, that is, a cheaper entry.

Young wallet (f_3). $f_3 = \mathbb{1}^{\downarrow}(Age; E)$, where *Age* is the number of days from the wallet's first on-chain activity to its first trade in the market: a bottom-quartile value in the global eligible pool.

Focused activity (f_4). $f_4 = \mathbb{1}^{\downarrow}(Breadth; E)$, where *Breadth* is the count of distinct markets the wallet has ever traded: a bottom-quartile value globally.

High profit (f_5). $f_5 = \mathbb{1}^{\uparrow}(PnL; W_m)$, where *PnL* is realised profit on the pair: a top-quartile rank against the market's eligible winners.

The three within-market fingerprints (f_1, f_2, f_5) rank a pair against the other winners in its market; the two wallet-level fingerprints (f_3, f_4) place the wallet against the global eligible distribution.

Flagged population. A (wallet, market) pair is Liu-flagged when it satisfies at least four of the five fingerprints, $\sum_{k=1}^{5} f_k \geq 4$. A wallet is Liu-flagged when it has at least one Liu-flagged pair.

D.3 Della Vedova: Excess accuracy and the orthogonality test

Theoretical motivation. Where the previous two methods score how a wallet trades, Della Vedova (2026a) asks only whether it was right. Della Vedova sets behaviour aside and measures forecasting accuracy directly, asking whether a wallet's directional bets resolve correctly more often than the market price already implied. The premise is that a genuinely informed trader buys the winning side at a rate no uninformed price-follower could match. The method then adds a structural check: it tests whether this accuracy co-moves with execution quality, the prices a wallet obtains relative to the rest of the market, on the logic that private information delivers a correct side and a favourable entry at once, whereas luck

delivers neither systematically. The accuracy screen is the flag; the orthogonality test is the validation that the accuracy reflects information rather than chance.

Eligible population. A wallet enters the eligible population if it made at least ten resolved buy trades in in-scope markets, enough observations for the binomial test below to have power. Accuracy is assessed on the buy side; selling before resolution does not change a wallet's directional call.

Excess accuracy and the flag. For each buy, a wallet is *correct* if it bought the outcome that resolved. The benchmark probability of being correct on a single trade is the price-implied $\max(p_{yes}, 1-p_{yes})$, the accuracy of a trader who merely follows the market. Excess accuracy aggregates the gap across a wallet's buys:

$$ea = \frac{n_{correct} - \sum \max(p_{yes}, 1-p_{yes})}{n_{buys}}. \tag{D.7}$$

Treating each buy as an independent Bernoulli trial with success probability $\max(p_{yes}, 1-p_{yes})$, the expected correct count is $\sum \max(p_{yes}, 1-p_{yes})$ with variance $\sum \max(p_{yes}, 1-p_{yes})\big(1-\max(p_{yes}, 1-p_{yes})\big)$, giving the one-sided z-score

$$z = \frac{n_{correct} - \mathrm{E}[correct]}{\sqrt{\mathrm{Var}[correct]}}. \tag{D.8}$$

A wallet is flagged (uncorrected) if $z > 2.326$ (one-sided $p < 0.01$) and $ea > 0$; the positive-*ea* requirement prevents significance being driven by a long run of marginally above-benchmark trades.

Execution quality and the orthogonality test. Execution quality measures whether a wallet trades at better prices than the market's volume-weighted average price (VWAP). For a wallet it is the mean over its trades of

$$\varepsilon = (\mathrm{VWAP}_{market} - p_{yes})\, s, \qquad s = +1 \text{ for buys, } -1 \text{ for sells,} \tag{D.9}$$

so a positive ε means buying below and selling above VWAP. The orthogonality regression asks whether accuracy and execution co-move among the accurate wallets:

$$\varepsilon = \beta_0 + \beta_1\, ea + \beta_2\, \mathbf{1}_{ea>0} + \beta_3\, (ea \times \mathbf{1}_{ea>0}) + \beta_4 \ln n_{trades} + \beta_5 \ln vol + u. \tag{D.10}$$

The interaction β_3 is the test statistic: a positive β_3 means wallets with positive-*ea* also obtain better execution, the bundling that private information predicts and that is absent in the uninformed population.

Robustness of the detector specification

The detector is one transparent configuration, not a tuned optimum (4.5). This appendix defends that configuration along three axes: parameter sensitivity (one parameter varied at a time, others at default), gate ablation (each gate disabled), and the empirical distribution of the conviction statistic. Across the twenty-six testable specifications the flagged population varies by roughly an order of magnitude in count, while the H1 timing enrichment remains in a 2.3 to 3.5× band with the permutation p-value at its floor; the central finding is therefore not an artefact of any single parameter.

E.1 Parameter sensitivity and gate ablation

Table 29 reports the timing and headline pair counts and the H1 enrichment under each one-knob-at-a-time perturbation of the detector's six tunable parameters, and under disabling each gate. The permutation p-value is at its floor (< 0.01) in every specification.[1]

[1] H1 enrichment in this table is computed on 100 permutations to keep the specification sweep tractable; the headline H1 result (5.2) is on 200 permutations. The two estimates agree to one decimal place, and the question this table asks is whether the result holds across specifications rather than the precision of any single point.

Table 29. Robustness of the flagged population and the H1 timing enrichment. One parameter is varied at a time, others held at default (default row in bold). Gate-ablation rows disable a single gate, others at default. Timing pairs apply gates G1–G5; headline pairs apply G1–G6. H1 enrichment is the ratio of observed timing-set pairs to the permutation-null mean, computed on 100 permutations.

Specification	Timing pairs	Headline pairs	H1 enrichment
Default	**18,078**	**10,298**	**2.5×**
JumpSize (default 0.15)			
0.10	29,551	18,136	2.4×
0.125	22,843	13,387	2.5×
0.20	11,710	6,300	2.7×
0.25	8,256	4,255	2.9×
0.30	5,882	2,954	3.0×
0.40	3,000	1,367	3.5×
MaxShare (default 0.20)			
0.10	17,217	9,885	2.6×
0.15	17,777	10,159	2.6×
0.30	18,384	10,434	2.5×
0.50	18,604	10,533	2.5×
MinSize (default $1,000)			
$250	52,692	30,659	2.5×
$500	31,993	18,428	2.5×
$2,500	7,716	4,352	2.6×
$5,000	3,654	2,073	2.7×
MaxPriorMove (default 0.10)			
0.05	16,064	9,119	2.6×
0.15	19,417	11,070	2.5×
0.20	20,524	11,753	2.4×
MinConv (default 0.80; timing set fixed, so H1 is unaffected)			
0.70	18,078	10,548	2.5×
0.75	18,078	10,426	2.5×
0.85	18,078	10,152	2.5×
0.90	18,078	10,025	2.5×
Window (default 24h)			
12h	14,889	8,826	2.7×
48h	21,883	11,768	2.4×
(one gate disabled, others at default)			
drop G3 (share cap)	18,743	10,603	2.5×
drop G4 (de-momentum)	24,704	14,404	2.3×
drop G5 (floor)	396,380	—	—
drop G6 (conviction)	18,078	18,058	—

Population size is most sensitive to the materiality floor. Counts move by roughly an order of magnitude when *MinSize* is varied from $250 to $5,000, by a factor of about five across the jump-threshold range $[0.10, 0.40]$, and by little under any other knob. The floor is therefore the dominant lever

on population size, consistent with the funnel reported with the population (5.1), where the floor cuts 396,380 pre-floor pairs to 18,078.

H1 enrichment is stable, and strengthens as the jump bar tightens. The enrichment sits in a 2.3 to 3.5× band across all twenty-six tabulated specifications, with the permutation p-value at its floor in every case. The clearest monotonicity is in the jump threshold: enrichment rises from 2.4× at $JumpSize = 0.10$ to 3.5× at 0.40, consistent with rarer moves being harder to anticipate by chance and the detector becoming more selective at higher thresholds. The other parameter sweeps move enrichment by at most about 0.2× around the default. The headline finding of H1, that timing is not reducible to volume, is therefore not an artefact of any single parameter choice.

Gate ablation localises the active gates. Disabling G3 (the reverse-causality share cap) raises the timing count by under 4% and leaves H1 enrichment essentially unchanged at 2.5×, so the share cap binds rarely; this is consistent with the upstream observation that self-driven jumps are uncommon among anticipation buys (4.5). Disabling G4 (the de-momentum bound) raises the timing count by about a third and softens H1 enrichment to 2.3×, so the gate contributes to selectivity even though most flags pass it under the default. Disabling G5 (the materiality floor) is qualitatively different: the pair count rises by more than an order of magnitude, to 396,380, confirming that the floor is the dominant filter at the pair level. H1 enrichment is not separately tabulated for this specification, since the floor is a size filter that does not interact with the timing signal H1 tests, so the enrichment under the no-floor specification would add no information beyond the headline. Disabling G6 (conviction) reduces the headline count to 18,058, almost exactly the timing set, since conviction is applied after G1–G5 and is empty as a filter once it is no longer enforced; H1 is unaffected by construction, as the timing set on which it is run does not use conviction.

E.2 The conviction gate sits in the bimodality valley

Figure 3 reports the empirical distribution of pre-jump conviction, the net-to-gross signed volume over the run-up window $[t_0, j_t]$ from the wallet's first qualifying buy to the realised jump, across the 18,058 timing-set pairs for which conviction is computed.[2] The distribution is bimodal: 52.8% of pairs sit at exactly $+1.00$, indicating strictly one-directional accumulation in the buy's direction up to the jump, and a softer mass of 21.5% clusters near zero, indicating two-sided or hedged activity. The region between the two modes carries little weight, which is the property that makes the conviction gate insensitive to its threshold.

[2]Twenty of the 18,078 timing-set pairs have no computable conviction, owing to degenerate run-up windows in which the realised jump occurs at or before the first qualifying buy under tied timestamps. They are excluded from this distribution; they remain in Table 29.

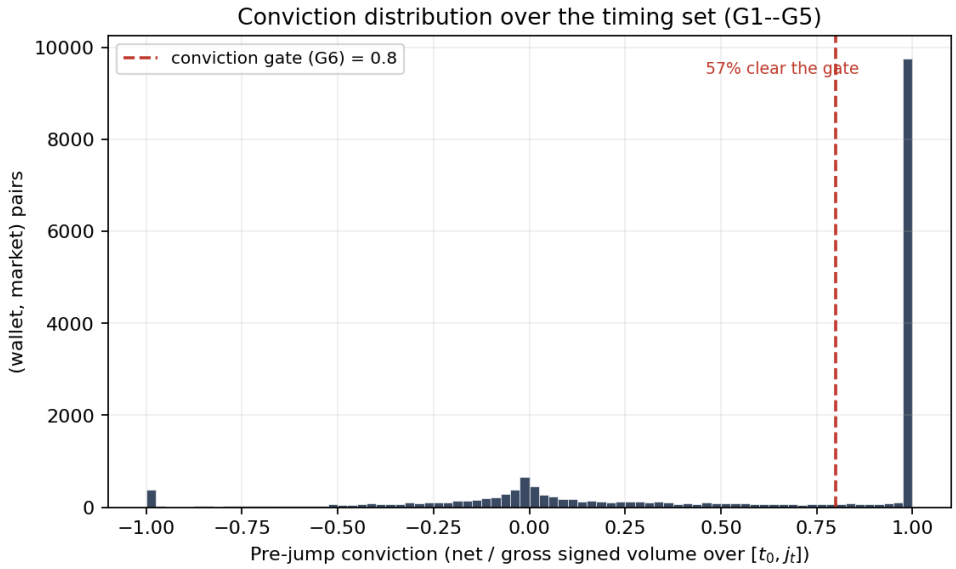

Figure 3. Distribution of pre-jump conviction, the net-to-gross signed volume over the run-up $[t_0, j_t]$, across the 18,058 timing-set pairs with a computed conviction. The dashed line marks the conviction gate $MinConv = 0.80$. The distribution is bimodal: 52.8% of pairs sit at exactly $+1.00$ (strictly one-directional accumulation) and a softer mass concentrates near zero (two-sided activity), with little weight between. The gate at 0.80 therefore falls in the low-density region between the two modes, which is why the flagged set varies by under 5% across thresholds in $[0.70, 0.90]$ (Table 29).

Of the timing set, 57% clear the conviction gate, producing the 10,298 headline pairs; 93% of the headline pairs are at exactly $+1.00$, recovering the median-conviction figure reported with the population (5.1). Because the region between the two modes is sparse, any threshold within it partitions essentially the same set; Table 29 confirms this directly, showing that the flagged set varies by under 5% across thresholds in $[0.70, 0.90]$. The choice of 0.80 therefore sits within an insensitive plateau rather than at a tuned threshold.

Statistical tests and methods

This appendix sets out the statistical procedures used to test the hypotheses of the results chapter, so that the inferential claims in the main text can be read without prior familiarity with the individual tests. Each section defines one method and the setting it serves: the permutation test underlying H1, the overlap measures and hypergeometric test underlying H2, the binomial sign test underlying H3, the two proportion and Mann–Whitney tests underlying H4, and the bootstrap confidence interval underlying H5.

F.1 The permutation test

A permutation test (equivalently, a randomisation test) constructs the distribution of a test statistic under the null hypothesis directly from the data, by repeatedly re-randomising the one feature whose effect is in question while holding every other feature of the data fixed. It assumes no parametric form for that distribution: the null is built by simulation rather than read from a known family such as the normal or the chi-square. The logic rests on exchangeability. If the null hypothesis holds, the feature under test carries no information, so the value actually observed should be no more extreme than a value obtained

by assigning that feature at random. Re-randomising it many times traces out the range of outcomes attributable to chance alone, and the observed statistic is judged by where it falls within that range.

Here the feature under test is the timing of a wallet's purchases, and the question is whether that timing places the wallet ahead of price jumps more often than its trading activity alone would produce. The null hypothesis is that timing is uninformative: a wallet that trades frequently in a volatile market will mechanically buy shortly before some jumps, and under the null the detector's flags are entirely this by-product of volume. To build the null we hold fixed everything about each (wallet, market) pair except the buy timestamps. The wallet's number of trades, their sizes, and their buy and sell directions are preserved, and the market's realised price path is preserved exactly; only the buy timestamps are redrawn, from the empirical distribution of trade times in that same market; that is, from the times at which trading actually occurred there. The full detector (gates G1–G5) is then re-run on the relabelled trades. This is a restricted, within-market permutation: because each pair is reshuffled inside its own market against that market's own price path, the procedure holds market selection and trading volume fixed by construction and isolates the timing channel alone.

The test statistic is the number of (wallet, market) pairs that still clear all five gates after the timestamps are redrawn. Enumerating every possible reassignment is infeasible, so the null distribution is approximated by Monte Carlo: the redraw-and-recount step is repeated $B = 200$ times, giving 200 draws of the statistic under the null. The observed statistic, computed on the true timestamps by identical code, is compared with this set. The empirical permutation p-value is the proportion of null draws at least as extreme as the observed value,

$$p \;=\; \frac{1 + \#\{\, b : T_b \geq T_{\text{obs}} \,\}}{B+1}, \tag{F.1}$$

where T_b is the statistic in permutation b. The observed data count as one further draw from the null (the "$+1$" in numerator and denominator), which keeps the estimate valid and rules out a p-value of exactly zero. When no permutation reaches the observed value, p attains its floor of $1/(B+1)$; with $B = 200$ this is $1/201 \approx 0.005$. The floor is a resolution limit set by the number of permutations, not a feature of the data: a larger B would lower it, but here no permutation comes within a factor of two of the observed count, so further permutations would not alter the conclusion.

Two properties of this null govern how the result is read. First, the statistic is a count and the null is bounded and discrete, so its distribution is not Gaussian. We therefore report the standardised distance between the observed value and the null mean, $(T_{\text{obs}} - \bar{T})/\text{SD}$, as an effect-size descriptor only, and do not convert it to a normal-tail probability, which would be meaningless for a bounded, non-normal null. Second, the test is a control for luck given volume: the comparator is the wallet's own trades reshuffled in time, not a counterfactual involving private information. Rejecting the null therefore establishes that the flag reflects genuine timing rather than mechanical activity; it does not, on its own, establish that the timing is informed. The test is also conditional on the market: holding the price path fixed isolates timing within the markets where the wallet was active, and is silent on whether the wallet selected jump-prone markets to begin with. We used Mitts and Ofir (2026)'s paper as inspiration for our permutation testing.

F.2 Overlap measures and the hypergeometric test

The agreement between two binary flags over a common population is summarised by a 2×2 contingency table. For a universe of N units, let a be the number flagged by both detectors, b by the first only, c by the

second only, and $d = N - a - b - c$ by neither; the two flag counts are $n_A = a + b$ and $n_B = a + c$. The measures below are all functions of these cells.

Expected overlap and enrichment. Under independence, the expected number flagged by both is $E[a] = n_A n_B / N$. The *enrichment ratio* is the observed overlap relative to this expectation, $a/(n_A n_B/N)$; it measures how many times more often the two detectors co-flag a unit than chance would produce. For the joint overlap of three flags the expectation under mutual independence is $E = n_A n_B n_C / N^2$, and the enrichment is again observed over expected.

Hypergeometric test. Enrichment describes magnitude; the hypergeometric test asks whether the observed overlap is too large to be sampling variation under independence. The null is that the two flag sets are unrelated, so that the n_A units flagged by the first detector are, in effect, a random draw without replacement from the N-unit universe, independent of which n_B units the second detector flags. Under this null the overlap follows the hypergeometric distribution,

$$P(a = k) \;=\; \frac{\binom{n_B}{k}\binom{N-n_B}{n_A-k}}{\binom{N}{n_A}}.$$

The alternative is positive association, so the test is one-sided and the p-value is the upper-tail probability of an overlap at least as large as observed, $p = \sum_{k \geq a_{\text{obs}}} P(a = k)$. This is identical to the one-sided Fisher exact test on the 2×2 table. A very small p rejects independence and establishes that the two detectors are positively associated. The test speaks to the presence of association, not its magnitude (which is the enrichment), and not to whether the shared population consists of true insiders.

Cohen's κ. Cohen's κ summarises agreement between two binary classifiers beyond chance,

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{a + d}{N}, \qquad p_e = \frac{n_A n_B + (N - n_A)(N - n_B)}{N^2},$$

where p_o is observed agreement and p_e is the agreement expected under independence given the marginals. It is 1 for perfect agreement, 0 for chance agreement, and negative below chance. One property matters for the interpretation here: when both flags are rare, almost every unit falls in the "neither" cell, so p_o and p_e are both close to one and κ is governed almost entirely by the small flagged cells; κ can therefore remain low even when co-flagging is heavily enriched. At flag rates well below 1% this deflation is substantial, so we read κ as a conventional descriptor alongside the enrichment and the hypergeometric test, not as a stand-alone measure of association.

Jaccard index. For set-overlap comparisons we also report the Jaccard index, $J = a/(a + b + c) = a/(n_A + n_B - a)$, the size of the intersection relative to the union of the two flagged sets. Unlike κ it ignores the shared "neither" cell, so it is not deflated by the large untagged majority; it is a descriptive similarity fraction (0 for disjoint sets, 1 for identical sets), not a test.

F.3 The binomial sign test

The sign test assesses, in a paired comparison, whether one condition exceeds the other more often than chance, using only the *direction* of each pair's difference and not its magnitude. In H3 the unit is the wallet: for each wallet we form the within-wallet gap between its anticipation pairs and its comparison pairs (in win rate, or in median PnL), and classify the wallet as positive, negative, or tied. Ties are discarded, leaving n non-tied wallets, of which k are positive.

Under the null hypothesis the two conditions are exchangeable within a wallet, so a positive and a negative gap are equally likely and k follows a binomial distribution, $k \sim \text{Binomial}(n, 0.5)$. The test statistic is k, and the (one-sided) p-value is the binomial tail probability of at least k positives,

$$p = \sum_{j \geq k} \binom{n}{j} 0.5^n;$$

the two-sided version doubles the smaller tail. A small p rejects the null and establishes that the gap points the same way for a *majority of individual wallets*.

The test is deliberately chosen for its weak assumptions. Because it uses only the sign of each wallet's gap, it is non-parametric and robust to the heavy-tailed, selection-inflated PnL distribution, and it does not require gaps to be comparable in magnitude across wallets. This is exactly the property the within-wallet design needs: it shows the effect is pervasive across traders rather than produced by a few large wallets. The test speaks only to the consistency of direction across wallets; the magnitude of the typical gap is reported alongside as a descriptive complement.

F.4 The two-proportion test

The two-proportion test compares the rate of a binary property between two groups. Let group A have x_1 of n_1 units with the property, at rate $\hat{p}_1 = x_1/n_1$, and group B have $\hat{p}_2 = x_2/n_2$. The null hypothesis is that the two underlying rates are equal, $p_1 = p_2$. Under the null the difference $\hat{p}_1 - \hat{p}_2$ is approximately normal with a variance estimated from the pooled rate $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$, giving the test statistic

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

referred to the standard normal distribution; we report the two-sided p-value, $2\,[1 - \Phi(|z|)]$. The normal approximation is accurate at the sample sizes used here. We use this test for the comparison of discrete-move shares, and for the win-rate and novelty comparisons between discrete- and gradual-move flagged pairs.

The test assumes the observations are independent. When they are clustered, for example when several (wallet, market) pairs come from the same market or the same wallet, the effective sample is smaller than the raw count and the true variance exceeds the formula's estimate, so the nominal p-value is anti-conservative (the reported significance is somewhat overstated). Where a comparison fails to reject, this dependence only widens the interval around the estimated gap and cannot create a difference, so a null conclusion is conservative to it; where a comparison rejects with a large margin, the dependence leaves the qualitative ordering intact even though the exact p-value would be larger under a clustering

adjustment.

F.5 The Mann–Whitney U test

The Mann–Whitney U test (Mann and Whitney, 1947) is a non-parametric test of whether one group's values tend to exceed the other's, using only ranks and assuming no particular distributional form. The null hypothesis is that the two samples are drawn from the same distribution, so that a value drawn at random from one group is equally likely to exceed or fall below a value drawn from the other (no stochastic ordering). The statistic U counts, across all cross-group pairs of observations, how often a value from one group exceeds a value from the other; for large samples it is standardised to a z using its null mean $n_1 n_2/2$ and known variance, from which the two-sided p-value follows.

We use this test for comparisons of realised PnL, where the distribution is heavy-tailed and dominated by a few large outcomes. In that setting rank-based tests are reliable whereas mean-based tests (such as the t-test) are not, since a handful of extreme values can drive a mean comparison; the associated location summary we report is the median. As with the two-proportion test, the procedure assumes independent observations, so clustering within markets or wallets makes the nominal p-value anti-conservative; given the size of the margins in the comparisons reported, the ranking is unaffected.

F.6 The bootstrap confidence interval

The bootstrap estimates the sampling uncertainty of a statistic by resampling the data, when an analytic standard error is unavailable or its assumptions are doubtful. The observed sample is treated as a stand-in for the population: we draw many resamples of the same size with replacement, recompute the statistic on each, and use the spread of the recomputed values to approximate the statistic's sampling distribution. The 95% *percentile* confidence interval is then the 2.5th and 97.5th percentiles of that bootstrap distribution.

The resampling unit must respect the dependence in the data. Here the (wallet, market) pairs within a market are not independent: they share the market's resolution mechanism, its information-generating category, and its realised price path. Resampling individual pairs would therefore understate the uncertainty. We instead use a *cluster bootstrap* that resamples whole markets with replacement, carrying all of each drawn market's pairs, and recomputes the flag-intensity ratio on each of 2,000 resamples. Because markets are the unit of resampling, the effective sample size is the number of markets, and the interval's width reflects between-market variation rather than the much larger pair count.

The statistic is the ratio of flag intensities, MNPI-admitting over no-MNPI. The null of no concentration corresponds to a ratio of one (equal intensity in the two market types). A 95% interval that lies entirely above one therefore rejects that null at the 5% level (two-sided) and indicates higher flag intensity in MNPI-admitting markets; the interval also conveys the magnitude and precision of the effect, not merely its significance. The bootstrap quantifies sampling uncertainty given the category labels; it does not incorporate uncertainty in those labels themselves, which is addressed separately by the robustness check excluding large-language-model-assigned markets and by the pending human validation (Appendix B).

Case studies

This Appendix summarises five case studies of highly likely insider trading occurrences on betting markets. The cases are heterogeneous in the type of nonpublic information at issue, ranging from classified military intelligence and secret committee deliberations to proprietary corporate data and private personal knowledge, but share three recurring on-chain signatures that motivate the empirical strategy of this thesis: recently funded or freshly created wallets, concentrated directional positions in narrowly framed binary markets, and trade timing closely preceding a discrete public news event. The stated wallet level data is based on Mitts and Ofir, 2026's *From Iran to Taylor Swift: Informed Trading in Prediction Markets*, the first systematic empirical study of informed trading on Polymarket and Kalshi.

G.1 The Google "Year in Search" case (AlphaRaccoon)

On 4 December 2025, Google's annual Year in Search report named the singer d4vd as the most trending person globally, an outcome that almost no public commentator had anticipated. A wallet operating under the display name "AlphaRaccoon" (subsequently renamed "0xafEe"; address 0xee50…27ed6) correctly predicted 22 of 23 outcomes across the Polymarket Year-in-Search markets, realising approximately USD 1.15 million in profit in a single day. Its most striking position, a USD 10,647 wager on d4vd at approximately USD 0.05 per share, returned roughly USD 200,000, a return of approximately 1,900%. The same wallet had previously earned approximately USD 150,000 by correctly predicting the exact release date of Google's Gemini 3.0 model. The case extends the typology beyond government secrets to *proprietary corporate* nonpublic information. On 27 May 2026, federal prosecutors in the Southern District of New York identified the wallet's operator as Michele Spagnuolo, a Google information security engineer since 2014. He was charged with commodities fraud, wire fraud, and money laundering for using Google's confidential business information to make over USD 1.2 million on Polymarket, having accessed internal data hours before placing the d4vd wager; his total stake reached approximately USD 2.75 million. The CFTC filed a parallel civil action, and the case is the second known federal criminal prosecution tied to prediction-market trades (The Washington Post, 2026).

G.2 The Maduro case (Van Dyke)

On 3 January 2026, U.S. forces captured Venezuelan President Nicolás Maduro in a closely held nighttime operation. The "Burdensome-Mix" wallet (0x31a5…b8eD9), created only weeks earlier, accumulated 523,879 YES tokens in the "Maduro out by January 31, 2026?" market at an average price of USD 0.074 (an implied 7.4% probability) across five trading sessions between 31 December and 3 January. The largest single accumulation occurred on the evening of 2 January Eastern Time, hours before President Trump's announcement. The position cost approximately USD 38,533 and resolved at approximately USD 523,879, a return of roughly 1,260%. The case prompted the introduction of the Public Integrity in Financial Prediction Markets Act (H.R. 7004) by Representative Ritchie Torres on 9 January 2026 and led to a criminal charge filed against the wallet operator. On 23 April 2026, the Justice Department unsealed an indictment identifying the operator as Gannon Ken Van Dyke, a U.S. Army soldier stationed at Fort Bragg who had participated in the planning of the operation ("Operation Absolute Resolve") and used his

access to classified information to trade. He was charged with unlawful use of confidential government information, theft of nonpublic government information, commodities fraud, wire fraud, and an unlawful monetary transaction, having staked approximately USD 33,034 across roughly thirteen bets and profited approximately USD 409,881; he subsequently attempted to conceal his identity by routing proceeds through a foreign cryptocurrency vault and requesting deletion of his Polymarket account (United States Department of Justice, 2026b).

G.3 The IDF case (ricosuave666)

Between 11 and 13 June 2025, in the days preceding Israel's Operation Rising Lion against Iranian nuclear and military infrastructure, a cluster of jointly funded Polymarket accounts with deliberately vague usernames, one of which operated as "ricosuave666", placed a series of correct directional bets across four markets on Israeli military action and ceasefire timing, realising just over USD 152,000 in profit that month before going dormant. The accounts were reactivated in January 2026 and flagged by Polymarket users speculating on social media, after which positions were cancelled and usernames changed; cumulative profits across June 2025 to January 2026 were approximately USD 156,000. On 12 February 2026, the Tel Aviv District Attorney indicted an IDF reservist and a civilian accomplice on charges including severe security offences, bribery, and obstruction of justice, following a joint investigation by Malmab (the Israeli Director of Security of the Defence Establishment), Shin Bet, and the Israel Police's Lahav 433 cyber unit. The case is, on the public record, the first criminal prosecution worldwide of insider trading on a prediction market. After local journalists successfully appealed to narrow a gag order, Tel Aviv court documents identified the civilian as Omer Ziv, a 30-year-old Tel Aviv resident who worked in online-gambling affiliate marketing; the second defendant, an Israeli Air Force reserve major, remains anonymous on national security grounds. According to the indictment, the major shared confidential information on the timing of military action while Ziv placed the bets using his own funds across the newly created accounts. On 12 June the major was allegedly briefed that the operation would begin before midnight and relayed this to Ziv via WhatsApp, urging larger bets; as the jets took off he allegedly wrote, "It's already happening." The pair allegedly placed further bets tied to subsequent Israeli and US strikes, including operations in Yemen in September. Ziv faces an additional charge of aggravated espionage, which carries a potential life sentence; both defendants have been ordered held in custody until the end of trial (The Guardian, 2026).

G.4 The 2025 Nobel Peace Prize case (6741)

On 10 October 2025 the Norwegian Nobel Committee awarded the Peace Prize to Venezuelan opposition leader María Corina Machado. Beginning approximately eleven hours before the announcement, three accounts ("6741," "dirtycup," and "GayPride") took aggressive directional positions on Machado as the Polymarket implied probability rose from approximately 4% to over 70%. Combined net profits across the three accounts were approximately USD 90,000. On 30 January 2026 the Norwegian Nobel Institute released the findings of its internal investigation, assisted by national security authorities, concluding that a cyber breach of its computer systems, rather than a human leak, was the "most likely" explanation for the early disclosure; its director stated that the "digital domain" remained the main suspect, while acknowledging that investigators could not determine exactly what had happened or who was responsible.

No defendant has been identified. The case is unusual in that the suspicious on-chain activity was itself the principal forensic signal that prompted the institution to confirm a breach (Forbes, 2025).

G.5 The February 2026 Iran Strike case (Magamyman)

On 28 February 2026, the United States and Israel conducted coordinated airstrikes against Iran in which Supreme Leader Ali Khamenei was killed. Traders on online prediction markets wagered over USD 1 billion on aspects of the conflict, with more than USD 194 million staked on Polymarket's offshore site alone on whether Khamenei would be "out as Supreme Leader" by various dates. A pseudonymous account, "Magamyman" (wallet 0x4dfd…f6e4a), made approximately USD 553,000 from Iran bets placed on the eve of war; in one wager it staked USD 32,000 in the early morning of 28 February, hours before the strikes began, correctly predicting strikes that day when Polymarket odds implied only a 17% chance. Blockchain-analytics firm Bubblemaps identified at least half a dozen traders who collectively earned roughly USD 1.2 million betting that the U.S. would strike Iran, most having funded their accounts the same day as their bets, hours before the strikes began. The trades that paid out on Khamenei's death occurred on Polymarket's largely unregulated international platform, beyond the reach of U.S. regulators; the activity drew condemnation from Democratic lawmakers, a proposed bill to bar senior federal officials from trading on prediction markets, and scrutiny of the platform's ties to the president's son. No defendant has been identified (CNN, 2026).

G.6 Synthesis

Taken together, the five cases delineate a typology of nonpublic information traded on prediction markets that spans classified national-security intelligence (the Maduro, IDF/Iran, and Khamenei-strike cases), secret institutional deliberations (the Nobel Peace Prize), and proprietary corporate data (Google Year in Search). Three have produced criminal charges to date, the IDF/Iran, Maduro, and Google cases, while the Nobel and Khamenei-strike cases remain uncharged, with no defendant publicly identified. Across the dataset, Mitts and Ofir estimate approximately USD 143 million in aggregate anomalous profits over the period February 2024 to February 2026 (Mitts and Ofir, 2026). For the purposes of this thesis, the cases illustrate that a recurring set of on-chain signatures, freshly created wallets, narrowly framed binary markets, concentrated directional positions, and tight pre-event timing, appears across very different categories of underlying information, suggesting these features are not specific to any one type of nonpublic information.