Probabilistic Beliefs in Forecasting and Prediction Markets

Aggregation, Inference and Liquidity under Mechanism Constraints

Zane Hassoun

PhD

University of York
Mathematics

February 2026

Abstract

This thesis studies how probabilistic beliefs are aggregated, expressed, and inferred in forecasting platforms and prediction markets. In both settings, observed probabilities reflect not only underlying information but also the mechanisms underlying forecast and trade generation. A central difficulty is that the available data are often incomplete, irregular, or systematically filtered by institutional design. The thesis addresses how meaningful belief dynamics can be recovered under these constraints.

The thesis consists of three self-contained papers. The first develops a Bayesian method for dynamically aggregating human probability forecasts submitted over time. Forecast sequences are modelled as evolving distributions, and a change-point framework is used to identify structural shifts in collective belief. The resulting aggregation weights adapt to bursts of activity and periods of inactivity, providing a flexible alternative to fixed recency-weighting schemes.

The second paper examines belief inference from sparse limit order books in prediction markets. In thin markets, visible orders reflect only a subset of trader beliefs, since moderate beliefs are often unobserved. A simulation-based inversion approach is introduced that combines a parametric belief model with an explicit visibility mechanism, allowing latent belief distributions to be inferred from truncated order-book snapshots while remaining consistent with observed prices.

The third paper studies liquidity in automated market maker based prediction markets. A mechanism-consistent model is developed in which price responsiveness is governed by a composite risk–liquidity parameter reflecting both market design and trader risk tolerance. Window-level estimators are used to characterise how this parameter varies across markets and over time.

Together, the thesis provides a unified statistical perspective on belief dynamics in forecasting and prediction-market environments.

Contents

List of Figures

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

Chapter 3 is based on the paper Kairosis: A Method for Dynamical Probability Forecast Aggregation Informed by Bayesian Change-Point Detection [46], published in the International Journal of Forecasting and co-authored with Niall MacKay and Ben Powell.

Chapter 5 is based on ongoing work co-authored with Ben Powell.

— 1 —

Introduction

Prediction markets and human forecasting platforms offer complementary but distinct windows into how information is aggregated, processed, and expressed as probabilistic beliefs. Forecasting tournaments collect subjective probability forecasts from human participants, typically without identifying individual forecasters or modelling their behaviour explicitly. Prediction markets, by contrast, generate probabilities endogenously through trading activity, with market design playing a central role in determining how information enters prices. Although both environments produce probabilistic signals, they differ substantially in their mechanics. Forecasting tournaments provide rich temporal data but no explicit market mechanism, while automated market makers (AMMs) in prediction markets transform sparse trading flow into probabilistic quotes according to deterministic pricing rules rather than bilateral competition.

Across both environments, researchers face a common set of empirical difficulties. First, the probabilistic signals produced by these systems evolve over time in ways that reflect heterogeneous and irregular information arrival. Methods that treat forecasts or prices as static quantities risk overlooking structural changes in the underlying information set. Second, the data visible to an analyst are rarely a complete representation of the beliefs held by forecasters or traders. Forecast submissions are asynchronous and self-selected, while sparse order books reveal only a filtered subset of trading intentions, particularly in thin markets where moderate beliefs are largely absent from visible orders. Third, the mechanisms that generate probabilities, whether behavioural, statistical, or mechanical, shape the observed data in systematic ways. Human forecasters choose when to submit updates, traders choose whether to reveal beliefs through orders, and AMMs impose mechanical price

sensitivity that is independent of information. These features induce distortions that must be modelled explicitly to enable sensible inference.

This thesis develops statistical and structural methods that address these challenges. The central theme is characterising latent informational structure, namely what people believe, how beliefs evolve over time, and how responsive prices are to trading activity, even when the observed data are incomplete or shaped by the mechanism that produced them. The work is organised as three standalone papers, summarised below.

1.1 DYNAMIC FORECAST AGGREGATION VIA KAIROSIS

The first paper studies the problem of aggregating probability forecasts that are submitted sequentially over time. Forecasting tournaments such as Metaculus collect large numbers of subjective probability assessments from human participants, with forecasters free to update their predictions whenever they choose. In practice, this leads to highly irregular submission patterns. Forecasts often arrive in concentrated bursts following salient news or discussion, while other periods exhibit little or no activity. As a result, the empirical sequence of forecasts typically forms a nonstationary time series whose distribution changes over the lifetime of a question.

Standard aggregation methods do not explicitly account for this temporal structure. Simple averages, medians, or weighted combinations of forecasts implicitly assume that the distribution of crowd belief is stable over time. Under this assumption, older forecasts are treated as equally informative as newer ones, regardless of whether the information environment has changed. In contrast, empirical forecast sequences frequently exhibit distinct phases, with relatively stable periods interrupted by episodes of rapid updating. Aggregation methods that ignore these dynamics risk combining forecasts drawn from different informational regimes.

A common response to this issue is to introduce recency weighting, typically through exponential discounting or rolling windows. These approaches recognise that older information may be less relevant, but they impose a fixed temporal structure that is independent of the data. Exponential discounting applies the same decay rate regardless of whether the belief distribution is stable or has shifted. Rolling windows discard forecasts outside a fixed horizon, even if those forecasts remain informative. Neither approach distinguishes between random noise and a genuine

structural change in the distribution of forecasts. These limitations are particularly apparent when forecast submissions are highly clustered or when bursts of activity follow long periods of inactivity.

We address this problem using a Bayesian change-point framework, which we call 'kairosis', to model the temporal evolution of forecasts. Forecasts are first discretised into a categorical representation, allowing the empirical belief distribution to be compared across time. For each candidate change point, the model evaluates the evidence for a structural break by comparing the forecast distributions before and after the split using a Dirichlet–categorical likelihood. This yields a posterior distribution over possible change points that summarises the degree of support for changes in the underlying belief state.

Aggregation weights are derived from this posterior distribution. Specifically, the weight assigned to each forecast is defined as the cumulative posterior probability that a change point has occurred after the forecast was submitted. Forecasts submitted well before the most likely change point receive relatively low weight, while those submitted after receive higher weight. When the data provide little evidence of a structural break, the resulting weighting function resembles a smooth recency-weighted scheme. When the data support a clear shift, the weights adjust sharply around the inferred change point. In this way, kairosis adapts endogenously to the observed forecast sequence rather than imposing an exogenous time scale.

The empirical analysis applies kairosis to a large collection of forecasting questions from Metaculus. Performance is evaluated relative to static aggregation methods, exponential discounting, and rolling-window approaches. Across these comparisons, kairosis produces aggregate forecasts that are more stable during periods of low informational activity and more responsive following substantive updates. These results illustrate how explicitly modelling changes in the forecast distribution can improve aggregation in environments characterised by irregular participation and heterogeneous information arrival.

1.2 INFERRING LATENT BELIEFS FROM ORDER BOOKS

The second paper studies prediction markets through the lens of the limit order book, with particular emphasis on thin markets in which standing orders are sparse and concentrated away from the midpoint price. In these environments, the visible order

book typically reflects only a small subset of traders' beliefs. Many participants choose not to place limit orders unless their subjective probability differs sufficiently from the prevailing market price. As a result, the midpoint price, often interpreted as a market consensus probability, is consistent with a wide range of underlying belief distributions. Two markets trading at the same price may therefore correspond to very different informational states, including broad agreement, substantial disagreement, or a strategic absence of orders near the centre.

This ambiguity arises because the order book provides a truncated and selectively observed view of trader beliefs. Unlike continuous double auction markets with deep books and frequent order submission, prediction markets often operate with limited liquidity and small order sizes. Traders whose beliefs lie close to the market price may have little incentive to post orders, either because the expected payoff is small or because execution is unlikely. The visible book is therefore biased toward more extreme beliefs, while moderate opinions remain largely unobserved. Any attempt to infer latent belief structure from order book data must account for both the cross-sectional distribution of beliefs and the mechanism that determines whether a belief becomes visible as a standing order, a concern that has been noted previously in the forecasting and decision-theoretic literature (e.g. [51]).

To address this problem, the paper assumes that subjective probabilities across traders are distributed according to a Beta(α, β) distribution, a flexible family on $[0, 1]$ capable of capturing consensus, skewness, and polarisation in beliefs. The inference procedure is based on a forward simulation approach. Candidate belief distributions are generated by drawing beliefs from the Beta distribution, after which a filtering rule is applied to replicate the implicit no-trade region around the midpoint price. Only simulated beliefs that satisfy this visibility condition are retained. The resulting truncated sample is then compared to the observed order book. By adjusting (α, β) to align the filtered simulated data with the empirical book, the method recovers belief distributions that are jointly consistent with both the observed orders and the market price.

This simulation based inversion explicitly models the selective visibility of orders and avoids interpreting the observed book as a direct sample from the underlying belief population. It therefore captures an important strategic feature of thin prediction markets that is missed by direct density estimation methods.

Empirically, the framework is applied to snapshots from Polymarket markets over time, producing trajectories of (α_t, β_t) that summarise the evolution of belief

dispersion and asymmetry. These trajectories allow periods of genuine agreement or disagreement to be distinguished from episodes in which the book appears extreme due solely to sparse order placement. In this way, the analysis provides a principled interpretation of order book data as a filtered reflection of a larger latent belief population, rather than as a complete description of market sentiment.

1.3 STRUCTURAL LIQUIDITY AND RISK IN AUTOMATED MARKET MAKERS

The third paper studies liquidity in prediction markets operated by automated market makers (AMMs), where prices are generated by a deterministic pricing rule rather than through bilateral competition between buyers and sellers. In traditional continuous double-auction markets, liquidity is an emergent property of trading activity: market depth and price impact arise endogenously from the collective behaviour of heterogeneous participants who post bids, asks, and market orders. Classical microstructure measures, including Kyle's λ, Amihud's illiquidity ratio, and variance-based decompositions, are designed for this institutional setting and quantify how order flow translates into prices when liquidity is supplied competitively.

AMM-based prediction markets differ fundamentally. Under mechanisms such as the logarithmic market scoring rule (LMSR), prices are determined by a cost function that maps outstanding quantities into probabilities according to a fixed rule. The sensitivity of prices to order flow is therefore governed mechanically by a liquidity parameter b, rather than by the strategic interaction of multiple liquidity providers. As a result, observed price movements combine two distinct sources of variation. The first reflects changes in trader beliefs driven by information arrival. The second reflects the curvature of the cost function itself, which determines how aggressively prices respond to trades even in the absence of new information. Standard liquidity measures do not disentangle these components and are therefore not directly interpretable in an AMM environment.

To address this, the paper develops a mechanism-consistent framework that links trader behaviour and AMM pricing into a single reduced-form object. Traders are assumed to have constant absolute risk aversion preferences, implying that optimal trade sizes are proportional to the discrepancy between a trader's belief and the current market price on the log-odds scale. When combined with the LMSR pricing

rule, this behaviour induces a local adjustment equation in which changes in log-odds prices are proportional to belief–price discrepancies. The proportionality constant is a composite parameter

$$\kappa = \frac{\tau}{b},$$

where b reflects the mechanical depth of the AMM and τ captures aggregate trader risk tolerance. Only this ratio is identifiable from transaction data. The parameter κ therefore represents risk-adjusted liquidity, measuring the responsiveness of prices to incoming order flow under automated market making.

The empirical contribution of the paper is to estimate κ from observed price dynamics across markets and over time. Trading activity is partitioned into local time windows, and changes in log-odds prices are related to lagged price movements through a reduced-form dynamic representation in which κ governs the degree of price inertia. Estimation relies solely on observed prices and does not require explicit modelling or identification of latent belief processes.

Applying this framework to a large collection of prediction markets yields distributions of window-level estimates of κ that vary systematically across market categories and over time. Some markets exhibit higher values of κ, corresponding to more inertial price behaviour, while others display lower values, indicating sharper price responses to modest order flow. Within markets, estimates of κ fluctuate at high frequency and exhibit rapid mean reversion, consistent with episodic tightening and loosening of effective liquidity rather than persistent regime shifts.

By construction, κ provides a mechanism-consistent analogue to classical liquidity measures. It captures how market design and trader risk interact to determine observed price responsiveness in automated prediction markets. This interpretation allows liquidity to be analysed in a way that is comparable across markets and over time, while remaining grounded in the institutional features that distinguish AMM-based trading from traditional microstructure settings.

1.4 INTEGRATION AND THEMES

Although the three papers address different empirical settings, they are unified by a common methodological objective. Each paper seeks to recover latent informational structure from data that are incomplete, irregular, or systematically filtered by the mechanism through which they are generated. Observable data are therefore not

treated as direct measurements of beliefs or information, but as outcomes of a data generating process that must be modelled explicitly in order to support meaningful inference.

This problem arises in distinct forms across the thesis. In forecasting tournaments, the analyst observes an irregular sequence of probability forecasts whose distribution may shift over time in response to new information or changes in participation. In prediction market order books, only a strategically selected subset of trader beliefs is visible, producing a sparse and biased snapshot of the underlying belief population. In AMM-based markets, observed price changes reflect both information arrival and the mechanical sensitivity imposed by the pricing rule. In each case, naive analysis of the raw data risks conflating genuine belief dynamics with artefacts introduced by sampling, visibility, or market design.

A second unifying theme is the treatment of probabilistic signals as inherently dynamic objects. Forecast sequences evolve through periods of inactivity punctuated by bursts of updating. Order books thicken and thin as participation varies and as traders respond to information. AMM prices move continuously, but the responsiveness of those prices to trades varies over time with effective liquidity conditions. Static summaries are therefore insufficient. Each paper introduces a method that explicitly models temporal evolution, allowing the analyst to distinguish persistent structure from short-lived fluctuations.

These shared concerns are reflected in three recurring methodological principles.

1. **Inference on latent structure from partial data.** In all three papers, the object of interest is not directly observed. In *Kairosis*, the latent object is the location of a change point in the distribution of submitted forecasts. In the order book analysis, it is the full cross-sectional distribution of trader beliefs that gives rise to a selectively visible set of standing orders. In the AMM liquidity model, it is the composite parameter κ, which governs how beliefs translate into price adjustments but cannot be observed directly. In each case, inference proceeds by specifying a generative model that links the latent structure to the observed data.

2. **Mechanism-aware modelling.** The thesis consistently treats observed data as the output of a mechanism that constrains what can be seen. Forecast submissions arrive asynchronously and in bursts, which *Kairosis* accommodates through forecaster-time and change-point modelling. Order book data are

shaped by strategic visibility decisions, which are modelled explicitly through rejection filtering. AMM prices are determined mechanically by a cost function, requiring liquidity to be interpreted through a structural lens rather than via classical microstructure measures. In each setting, modelling the mechanism is essential for separating belief dynamics from artefacts of the data generating process.

3. **Dynamic probabilistic inference.** All three papers emphasise time variation in probabilistic signals. *Kairosis* produces time-varying forecast weights driven by posterior evidence of structural change. The order book framework generates trajectories of (α_t, β_t) that track the evolution of belief dispersion over a market's lifetime. The AMM liquidity analysis produces window-level estimates of κ_t that reveal short-run fluctuations and mean reversion in effective liquidity. These dynamic representations allow probabilistic information to be interpreted as an evolving process rather than as a static quantity.

Taken together, the three papers provide a coherent statistical perspective on how probabilistic beliefs are expressed, filtered, and transformed in modern forecasting and prediction market environments. While the empirical settings differ, the unifying contribution lies in the use of mechanism-aware, generative models to extract latent informational structure from imperfect data. This perspective clarifies how observed probabilities should be interpreted, how they respond to information flow, and how institutional design shapes their behaviour.

1.5 THESIS STRUCTURE

The remainder of the thesis is organised as follows. Chapter 2 introduces the mathematical preliminaries and modelling conventions used throughout the thesis. Chapter 3 presents the *Kairosis* framework for dynamic forecast aggregation and evaluates its performance using forecasting tournament data. Chapter 4 develops the belief inference framework for sparse prediction market order books and applies it to Polymarket snapshots. Chapter 5 introduces the structural liquidity model for AMM-based prediction markets and provides empirical estimates across a large cross-section of markets. Chapter 6 concludes with a synthesis of the results and discusses implications and directions for future research.

— 2 —

Preliminaries

This chapter introduces the mathematical objects, modelling conventions, and notation used throughout the thesis. The thesis consists of three related but methodologically distinct papers: a framework for dynamic forecast aggregation, a model for belief inference from sparse order books, and an empirical analysis of risk-adjusted liquidity in automated market makers. Although each paper addresses a different empirical setting, they rely on a common set of probabilistic representations, behavioural assumptions, and price dynamics.

The purpose of this chapter is therefore not to present new results, but to define the shared mathematical and statistical tools that recur across later chapters. In particular, the chapter introduces the probabilistic notation, belief representations, and reduced-form market structures used throughout the thesis, and briefly situates these objects within the forecasting, prediction market, and market microstructure literatures. The chapter also introduces the main datasets used in the empirical analysis and establishes common modelling conventions used throughout the dissertation.

Section 2.1 introduces probability notation, log-odds representations, scoring rules, and belief distributions. Section 2.2 outlines the components underlying the kairosis framework for dynamic forecast aggregation. Section 2.3 describes the structure of prediction-market order books and the belief-filtering mechanism induced by selective visibility. Section 2.4 develops the pricing and behavioural ingredients required for understanding risk-adjusted liquidity in automated market makers.

2.1 SHARED PROBABILITY TOOLS

Throughout the thesis, all random variables are assumed to be defined on a common probability space $(\Omega, \mathcal{F}, P)$. Binary outcomes are represented using Bernoulli random variables taking values in $\{0, 1\}$, while forecast sequences and market observables are treated as stochastic observations indexed either by calendar time or by forecast arrival order.

2.1.1 Probability and Log Odds

Let $Y : \Omega \to \{0, 1\}$ denote a Bernoulli random variable representing the realisation of a binary event, with

$$\pi = \Pr(Y = 1).$$

Throughout the thesis, both probability and log-odds representations are used. The logit transform is defined as

$$q \equiv \log\Big(\frac{\pi}{1-\pi}\Big), \qquad \pi = \frac{e^q}{1+e^q}.$$

The log-odds scale is unbounded and additive under many economic mechanisms. Small changes in probability near 0 or 1 correspond to large changes in log-odds, reflecting heightened informational sensitivity near extreme outcomes. For this reason, most structural models in later chapters are expressed on the logit scale.

2.1.2 Forecast Sequences

A forecast submitted at time t_r is denoted by (p_r, t_r), where $p_r \in [0, 1]$. The collection of forecasts associated with a single event is

$$\mathcal{D} = \{(p_r, t_r)\}_{r=1}^{R}, \qquad t_1 < \cdots < t_R.$$

The sequence $\mathcal{D}$ is treated as a stochastic collection of forecasts generated from an evolving latent belief distribution indexed by either time or forecast arrival order. Throughout the thesis, conditioning on $\mathcal{D}$ should therefore be interpreted as conditioning on the observed forecast sequence and the sigma-field generated by these observations.

For modelling convenience, the index r is often treated as a pseudo-time coordinate, referred to as *forecaster time*. This representation emphasises the ordering and density of information arrivals rather than absolute clock time.

2.1.3 Proper Scoring Rules

A scoring rule assigns a numerical score to a probabilistic forecast based on the realised outcome. A scoring rule is said to be *proper* if the expected score is optimised by reporting the true underlying belief distribution, and *strictly proper* if truthful reporting is uniquely optimal [39].

Forecast accuracy is evaluated using proper scoring rules. Two strictly proper rules used throughout the thesis are

$$\text{Brier}(p, Y) = -(p - Y)^2, \qquad \text{Log}(p, Y) = Y \log(p) + (1 - Y) \log(1 - p).$$

Strict propriety ensures that truthful reporting of beliefs is optimal in expectation.

2.1.4 Beta Distributions

The Beta distribution provides a flexible model for cross-sectional belief heterogeneity. For $\alpha, \beta > 0$,

$$p \sim \text{Beta}(\alpha, \beta), \qquad f(p \mid \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}.$$

The mean and variance are

$$\mathbb{E}[p] = \frac{\alpha}{\alpha + \beta}, \qquad \text{Var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

Small values of α and β produce polarised belief distributions, while large, symmetric values correspond to strong consensus.

2.2 PRELIMINARIES FOR PAPER 1: DYNAMIC FORECAST AGGREGATION

This section introduces the probabilistic and statistical framework underlying the kairosis aggregation methodology developed in Hassoun, MacKay, and Powell [46]. The framework models sequences of probabilistic forecasts as observations generated from evolving latent belief distributions, with structural changes potentially occurring over forecast arrival time. More broadly, the methodology relates to the literature on forecast aggregation, proper scoring rules, and dynamic belief updating [28, 36, 39].

2.2.1 Forecast Sequences and Representation

Consider a sequence of probabilistic forecasts for a binary event,

$$\mathcal{D} = \{(p_r, t_r)\}_{r=1}^{R},$$

where $p_r \in [0, 1]$ denotes the forecast submitted at time t_r, and $t_1 < t_2 < \cdots < t_R$. Forecasts are indexed by their arrival order r, which we refer to as *forecaster time*. This representation emphasises the ordering and density of information arrivals rather than absolute clock time.

Within the kairosis framework, the sequence $\mathcal{D}$ is treated as a stochastic sequence of forecasts generated from a time-varying latent belief distribution. Structural changes in this latent distribution are assumed to occur at unknown points in forecaster time. Throughout, conditioning on $\mathcal{D}$ should be interpreted as conditioning on the sigma-field generated by the observed forecast sequence.

The methodology considered here is retrospective rather than sequential: the goal is to identify structural changes in belief distributions using the full forecast history associated with a resolved event, rather than performing real-time online detection as in sequential change-point procedures [54, 13].

2.2.2 Discretisation via Binning

To compare belief distributions across time, the unit interval is partitioned into K disjoint bins $\mathcal{B}_1, \ldots, \mathcal{B}_K$. Each forecast p_r is assigned to a bin according to

$$p_r \in \mathcal{B}_k.$$

Let n_k denote the number of forecasts falling in bin $\mathcal{B}_k$ over a specified subset of the sequence. Binning provides a nonparametric summary of the cross-sectional belief distribution and allows changes in shape or location to be detected without imposing a parametric distributional assumption on the forecast sequence. Similar discretisation approaches are common in Bayesian change-point detection and nonparametric density comparison procedures [12, 2].

2.2.3 Dirichlet–Categorical Likelihood

Let $(n_1, \ldots, n_K)$ denote the observed bin counts over a given segment of the forecast sequence. Following the notation introduced in Section 2.1, a Dirichlet prior with

parameters $(\alpha_1, \ldots, \alpha_K)$ is placed on the latent categorical probabilities associated with the partition $\{\mathcal{B}_1, \ldots, \mathcal{B}_K\}$.

The Dirichlet–categorical model is standard in Bayesian modelling of discrete probability vectors due to its conjugacy properties [35, 16].

Under this model, the marginal likelihood of the observed counts is

$$P(n_1, \ldots, n_K \mid \alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\Gamma\left(\sum_{k=1}^{K} (n_k + \alpha_k)\right)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}.$$

The latent probabilities correspond to the unobserved probabilities that a forecast generated from the underlying belief distribution falls into each bin $\mathcal{B}_k$. The resulting likelihood therefore depends only on the empirical distribution of forecasts within the segment and serves as the basic comparison object for detecting structural changes in beliefs.

2.2.4 Candidate Change Points and Prior

Potential change points are restricted to forecast arrival indices $r = 1, \ldots, R$. Let r_{CP} denote the random variable representing the location of the change point in forecaster time. A candidate change point at index r divides the sequence into a "before" segment $\{1, \ldots, r\}$ and an "after" segment $\{r + 1, \ldots, R\}$.

A geometric prior is placed on the change-point location,

$$P(r_{\mathrm{CP}} = r) = p(1 - p)^{R-r},$$

which assigns greater weight to more recent indices and regularises against splits far in the past. The geometric prior is used primarily for parsimony and to encode a weak preference for more recent structural changes. Bayesian change-point procedures commonly impose simple priors over candidate break locations in order to regularise inference [12, 2].

2.2.5 Posterior Mass and Weighting Function

For each candidate split r, the posterior mass is proportional to the product of the marginal likelihoods of the two segments and the prior,

$$P(r_{\mathrm{CP}} = r \mid \mathcal{D}) \propto L_{\mathrm{before}}(r)\, L_{\mathrm{after}}(r)\, p(1 - p)^{R-r},$$

where $L_{\text{before}}(r)$ and $L_{\text{after}}(r)$ are the Dirichlet–categorical likelihoods computed from the corresponding bin counts.

The cumulative posterior distribution,

$$W(s) = \Pr(r_{\text{CP}} \leq s),$$

defines a weighting function over forecaster time. This function is used to downweight earlier forecasts when aggregating predictions, without explicitly selecting a single change point.

The weighting interpretation is related to broader approaches in dynamic forecast aggregation that assign greater influence to recent or regime-consistent forecasts [74].

2.2.6 Role of Forecaster Time

All inference is conducted in forecaster time rather than calendar time. Periods with dense forecast activity correspond to rapid progression in forecaster time, while long gaps between forecasts correspond to compressed regions.

This construction ensures that aggregation weights respond to informational density rather than clock-time spacing. Implicitly, the framework assumes that forecast arrival intensity contains information about the evolution of beliefs, while remaining agnostic about the latent evolution of beliefs between forecast submissions. Similar distinctions between event time and calendar time are common in financial econometrics and market microstructure modelling [31, 44].

2.3 PRELIMINARIES FOR PAPER 2: BELIEF INFERENCE FROM ORDER BOOKS

This section introduces the probabilistic and structural assumptions underlying the order-book belief inference framework developed in Chapter 4. The framework treats observed order books as selectively filtered representations of an underlying latent belief distribution. More broadly, the approach is motivated by the literature on prediction markets as mechanisms for information aggregation [87, 6] and by the market microstructure literature on the informational content of order flow and limit order books [61, 44].

2.3.1 Structure of a Limit Order Book

A limit order book consists of a bid side and an ask side. Let

$$b_1 \geq b_2 \geq \dots \quad \text{and} \quad a_1 \leq a_2 \leq \dots$$

denote the visible bid and ask prices, with associated quantities $q_i^{(b)}$ and $q_i^{(a)}$. The best bid and ask define the midpoint price

$$\bar{p} = \frac{b_1 + a_1}{2}.$$

Limit order books are a standard object in market microstructure theory [61, 44]. In prediction markets, however, order books are often substantially sparser than those observed in traditional financial markets. Only a limited number of price levels may be visible at a given time, and substantial regions around the midpoint may contain few or no visible orders. Consequently, observed order prices provide only partial information about the latent cross-sectional distribution of trader beliefs.

2.3.2 Selective Visibility and Filtering

Observed orders are not assumed to represent an unbiased sample from the latent belief distribution. Traders with beliefs close to the prevailing midpoint face relatively low expected gains from posting visible limit orders and may instead choose not to reveal their beliefs through displayed liquidity. This idea is related to the broader literature on strategic order placement and liquidity provision in electronic markets [33, 71].

To model this selective visibility, a simple threshold mechanism is introduced:

$$p \text{ is observed} \iff |p - \bar{p}| \geq \delta,$$

where $\delta > 0$ denotes a visibility threshold. Beliefs sufficiently far from the midpoint generate visible orders, while beliefs near the midpoint remain latent.

This filtering mechanism induces a deterministic map from the latent belief distribution to an observed truncated distribution of order prices. The observed order book should therefore be interpreted as a selectively revealed subset of beliefs rather than a direct representation of the full latent distribution.

The threshold formulation is intended as a reduced-form approximation to the economic intuition that moderate beliefs generate lower incentives for visible liquidity provision. Alternative filtering mechanisms are possible, but the simple threshold rule is adopted here for tractability and interpretability.

2.3.3 Latent Belief Model

Latent trader beliefs are modelled as draws from a Beta distribution,

$$p \sim \text{Beta}(\alpha, \beta),$$

where α and β parameterise both the central tendency and dispersion of beliefs across traders.

The Beta family is commonly used in Bayesian modelling for probabilities and belief heterogeneity due to its flexibility on the unit interval [35, 16]. It is flexible enough to represent consensus, polarization, and skewness while remaining parsimonious.

Throughout, beliefs are defined on the probability scale, with the midpoint price $\bar{p}$ interpreted as an observed market summary statistic reflecting prevailing market consensus under the modelling assumptions adopted here. As in much of the prediction market literature, this interpretation is strongest under conditions approximating risk neutrality and adequate market liquidity [87, 55].

2.3.4 Filtering and the Forward Map

Given a candidate parameter pair (α, β), the observable distribution of order prices is generated by the following forward process:

1. Draw latent beliefs $p_1, \ldots, p_M \sim \text{Beta}(\alpha, \beta)$.
2. Apply the visibility filter $|p_i - \bar{p}| \geq \delta$.
3. Retain only beliefs that pass the filter.

This procedure produces a truncated and typically asymmetric observed distribution concentrated away from the midpoint. The observed order prices are treated as draws from this filtered distribution.

Conditional on $(\alpha, \beta, \bar{p}, \delta)$ and the filtering mechanism, the forward simulation map is deterministic and does not require explicit modelling of trader arrival times or order sizes.

2.3.5 Inverse Problem and Identification

The inferential task is to recover (α, β) from a finite sample of observed order prices, given $\bar{p}$ and δ. Because the filtering operation removes central mass from the latent distribution, the resulting likelihood is generally not available in closed form.

Identification relies on two constraints:

1. The filtered distribution must approximately match the empirical distribution of observed order prices.
2. The latent mean must remain broadly consistent with the midpoint price,

$$\mathbb{E}[p] = \frac{\alpha}{\alpha + \beta} \approx \bar{p}.$$

The midpoint constraint anchors the latent distribution and helps prevent parameter shifts that would otherwise generate observationally similar filtered samples after truncation.

2.3.6 Rejection-Sampling Estimation

Inference proceeds via forward simulation and rejection sampling. For each candidate (α, β):

1. Simulate a large latent sample from $\text{Beta}(\alpha, \beta)$.
2. Apply the visibility filter.
3. Compare the simulated filtered sample to the observed order prices using summary statistics or distance metrics.

Simulation-based inference procedures of this type are closely related to rejection sampling and approximate Bayesian computation methods [73, 14].

Parameter values generating filtered samples inconsistent with the observed order book are rejected. Accepted draws form an approximation to the posterior distribution or admissible parameter region associated with (α, β) under the assumed filtering mechanism.

This simulation-based procedure avoids direct modelling of the latent belief process while retaining the selective-visibility structure induced by sparse order books.

2.3.7 Interpretation

The resulting estimates should be interpreted as describing the latent cross-sectional distribution of trader beliefs that is jointly consistent with the observed order book, the midpoint price, and the assumed filtering mechanism. The framework attempts

to distinguish latent belief heterogeneity from visibility effects and therefore treats order-book sparsity as an informational feature rather than purely as sampling noise.

More broadly, the framework is intended as a reduced-form approach for extracting additional information from sparse prediction markets, rather than as a complete structural model of trader behaviour or market equilibrium.

2.4 PRELIMINARIES FOR PAPER 3: STRUCTURAL LIQUIDITY IN AMM MARKETS

This section introduces the pricing, behavioural, and reduced-form time-series framework underlying the analysis of structural liquidity in automated market makers developed in Chapter 5. The framework combines logarithmic market scoring rule (LMSR) pricing [42, 43] with risk-averse trader behaviour and reduced-form stochastic price dynamics. More broadly, the approach is related to the literature on automated market making, market scoring rules, and information aggregation in prediction markets [6, 27].

2.4.1 LMSR Pricing

The logarithmic market scoring rule (LMSR) cost function introduced by Hanson [42] is

$$C(q_N, q_Y) = b \log\left(e^{q_N/b} + e^{q_Y/b}\right),$$

where q_Y and q_N denote outstanding shares on the *Yes* and *No* outcomes respectively.

The instantaneous market price is

$$p = \frac{e^{q_Y/b}}{e^{q_N/b} + e^{q_Y/b}}, \qquad x \equiv \log\left(\frac{p}{1-p}\right) = \frac{q_Y - q_N}{b}.$$

The quantity x therefore represents the market price on the logit scale introduced in Section 2.1. The parameter $b > 0$ governs the curvature of the LMSR cost function and therefore the mechanical sensitivity of prices to order flow. Larger values of b correspond to deeper markets with smoother price responses, while smaller values imply sharper price movements for a given trade size.

The LMSR mechanism is attractive because it guarantees bounded loss for the market maker while maintaining continuous prices and liquidity [43].

2.4.2 CARA Utility and Local Logit Adjustment

Traders are assumed to have constant absolute risk aversion (CARA) utility,

$$U(w) = -\exp(-w/\tau),$$

where $\tau > 0$ denotes trader risk tolerance. CARA utility is standard in microeconomic models of risky choice and market behaviour [57].

Let π denote a trader's subjective belief about the event probability, with corresponding logit

$$e = \log\Big(\frac{\pi}{1-\pi}\Big).$$

Under a small-stakes approximation, maximisation of expected utility against the LMSR pricing rule implies that optimal trade size is approximately proportional to the discrepancy between trader beliefs and prevailing market prices on the logit scale,

$$\Delta Q^\star \propto \tau\Big(e - x\Big).$$

Combining this behavioural response with the LMSR mapping between order flow and logit prices yields a local adjustment rule of the form

$$x_{t+1} \;\approx\; x_t + \kappa\Big(e_t - x_t\Big), \qquad \kappa \equiv \frac{\tau}{b}. \tag{2.1}$$

The composite parameter κ captures *risk-adjusted liquidity*: it governs how aggressively market logit prices respond to discrepancies between trader beliefs and prevailing prices. From transaction data alone, the liquidity parameter b and the trader risk-tolerance parameter τ are not separately identified; only their ratio κ is identifiable within the reduced-form framework considered here.

The interpretation of κ as a reduced-form object is important. The model is not intended as a complete structural equilibrium description of trader behaviour, but rather as a tractable approximation linking observed price responsiveness to latent behavioural and liquidity primitives.

2.4.3 Induced Reduced-Form Price Dynamics

Equation (2.1) defines a linear stochastic difference equation once trader beliefs are allowed to evolve over time. Writing $x_t \equiv x_t^{\text{market}}$ and letting $\{e_t\}$ denote an unobserved belief process, the market logit follows

$$x_{t+1} = (1-\kappa)x_t + \kappa e_t. \tag{2.2}$$

This representation highlights that observed price dynamics confound belief dynamics with market liquidity. When $\kappa < 1$, prices adjust sluggishly and act as a smoothed version of beliefs; when $\kappa > 1$, prices may over-adjust and exhibit oscillatory behaviour. The boundary case $\kappa = 1$ corresponds to one-step tracking of beliefs.

If trader beliefs follow a stationary autoregressive or ARMA process [41, 20], the induced price process inherits a higher-order ARMA structure whose coefficients depend on κ. Consequently, persistence and serial dependence in observed logit prices may arise either from slow-moving beliefs or from high effective liquidity.

This lack of identification motivates treating κ as a reduced-form descriptor of price responsiveness rather than as a structural estimate of belief dynamics. The resulting framework is therefore intended primarily as a tool for interpreting and comparing observed market dynamics.

2.4.4 Window-Level Representation and Estimation Motivation

For empirical implementation, trading is partitioned into contiguous time windows w. Let $x^{\text{market}}_{w,\text{start}}$ and $x^{\text{market}}_{w,\text{end}}$ denote the logit price at the beginning and end of window w, and define

$$\Delta x^{\text{market}}_{w} = x^{\text{market}}_{w,\text{end}} - x^{\text{market}}_{w,\text{start}}.$$

Window-level representations exploit the reduced-form dynamics implied by (2.2) while remaining agnostic about the latent evolution of trader beliefs. In this formulation, κ is interpreted as a local, time-varying measure of logit price responsiveness: smaller values correspond to greater inertia and smoothing, while larger values indicate sharper, potentially overshooting responses to order flow.

The use of reduced-form window-level dynamics follows the broader econometric tradition of modelling observable price processes while treating latent beliefs or information flows as unobserved state variables [41].

These reduced-form implications underpin the price-only estimation strategies developed in Chapter 5, which recover κ from observed logit price dynamics without requiring direct observation or explicit modelling of trader beliefs.

2.5 DATASETS

The empirical analyses developed throughout the thesis rely on three distinct datasets corresponding to the three substantive application areas studied in the subsequent chapters. In all three cases, substantial data engineering, preprocessing, and organisational work were required prior to statistical analysis.

The dynamic forecast aggregation analysis in Chapter 3 uses probabilistic forecasting data obtained from the Metaculus forecasting platform. The raw data consist of highly granular time-stamped probability forecasts submitted by individual forecasters for binary geopolitical, economic, technological, and social events. Each observation contains a forecast probability, submission timestamp, user identifier, question identifier, and associated event metadata. The resulting datasets form irregularly spaced sequences of probabilistic forecasts whose density and timing vary substantially across events.

Considerable preprocessing was required to transform the raw forecasting data into a form suitable for dynamic aggregation analysis. Forecast histories were collected and organised at the event level, duplicate and malformed entries were removed, timestamps were standardised, and forecast sequences were aligned with event-resolution information. The resulting data structure preserves the full temporal ordering of forecasts and allows the analysis of belief evolution over forecaster time rather than calendar time. The final dataset contains large numbers of sequential forecast updates observed at fine temporal granularity, making it suitable for studying structural changes in collective belief distributions.

The belief inference analysis in Chapter 4 uses high-frequency prediction-market order book data collected from Polymarket. Unlike many conventional financial datasets, historical order book states are not publicly archived in a directly usable format. Consequently, bespoke live data-ingestion infrastructure was developed specifically for this thesis in order to collect and maintain the required data.

This infrastructure involved maintaining continuously running server-side data collection processes during active market periods. Exchange endpoints were queried repeatedly in real time in order to capture evolving order book states, including visible bid and ask prices, associated quantities, spreads, midpoint prices, and timestamps. The resulting data streams were stored and organised as sequences of order book snapshots observed through time.

Substantial ETL-style preprocessing was subsequently required before statistical

analysis could be performed. Raw order book data were cleaned, synchronised across collection intervals, transformed into structured snapshot objects, and organised into market-specific time series. Additional processing was required to identify inactive periods, handle missing or irregular observations, and standardise market representations across contracts. The resulting dataset provides a detailed record of sparse prediction-market order book dynamics and forms the basis for the latent belief inference framework developed in Chapter 4. Additional exploratory summaries and preprocessing details are provided in the Appendix.

The structural liquidity analysis in Chapter 5 uses transaction-level and market-state data obtained from Polymarket and related Polygon blockchain data sources through GraphQL-based APIs. These data include transaction histories, market prices, liquidity states, trading timestamps, and associated market metadata for LMSR-style prediction markets operating on blockchain-based infrastructure.

Collecting and organising these data required integrating information originating from multiple heterogeneous sources. Market-state information, trade histories, price series, and blockchain transaction records were queried separately and then merged into unified market-level datasets. Additional preprocessing was required to reconcile inconsistent identifiers, standardise timestamps, remove duplicate events, and construct contiguous market histories suitable for time-series analysis.

The resulting dataset provides detailed observations of prediction-market trading activity and automated market maker dynamics at high temporal granularity. In particular, the data allow the analysis of reduced-form price responsiveness, persistence, and liquidity behaviour across evolving market states and trading environments.

Together, these datasets provide detailed observations of probabilistic forecasting behaviour, prediction-market microstructure, and automated market maker dynamics across a range of forecasting and trading environments. The construction and preprocessing of these datasets formed a substantial component of the overall research contribution developed throughout the thesis.

2.6 SUMMARY

This chapter has defined the probability representations, belief models, market microstructure assumptions, and reduced-form price dynamics used throughout the thesis. These preliminaries provide the technical foundation for the dynamic

aggregation, belief inference, and risk-adjusted liquidity analyses developed in the chapters that follow.

— 3 —

Kairosis: A method for dynamical probability forecast aggregation informed by Bayesian change point detection

3.1 INTRODUCTION

Geopolitical forecasting tournaments have become increasingly popular over the last decade, notable providers include the Good Judgment Project and Metaculus. A typical question from Metaculus is that of Figure 3.1, "Will Donald Trump be president of the USA in 2019?". From when the question opened (May 17, 2017), forecasters submitted probability forecasts (on a scale of 0 to 1), until the question was resolved on Feb 1, 2019, although here we show only the first seven months' forecasts. A forecaster may make as many or as few forecasts as they wish, at any time they wish; thus we observe only the overall evolving distribution of forecasts, rather than multiple time-series of individual forecasters' evolving beliefs.

After resolution, the forecasts are scored. If forecasts are considered "static", taking no account of *when* the forecast is submitted, a simple proper probability score, such as the Brier (quadratic, [21]) or Log (logarithmic, [40]) score, can be used. Proper scores are optimized by, and therefore incentivize forecasters to submit their best estimates of, the true probability – although propriety fails if rewards are not proportional to the score, for example if the prize goes to the overall winner (see for example [72], [22]). But prescience is clearly valuable [76], and Metaculus, for example, weights the score by how long it was submitted before resolution. Indeed it is clear just from a visual inspection that the distribution of forecasts is not

stationary; it changes both smoothly and sharply at certain points, as does the density of forecasts submitted. The reason is obvious: news occurs much in this way, and new information is continually informing the forecasting process.

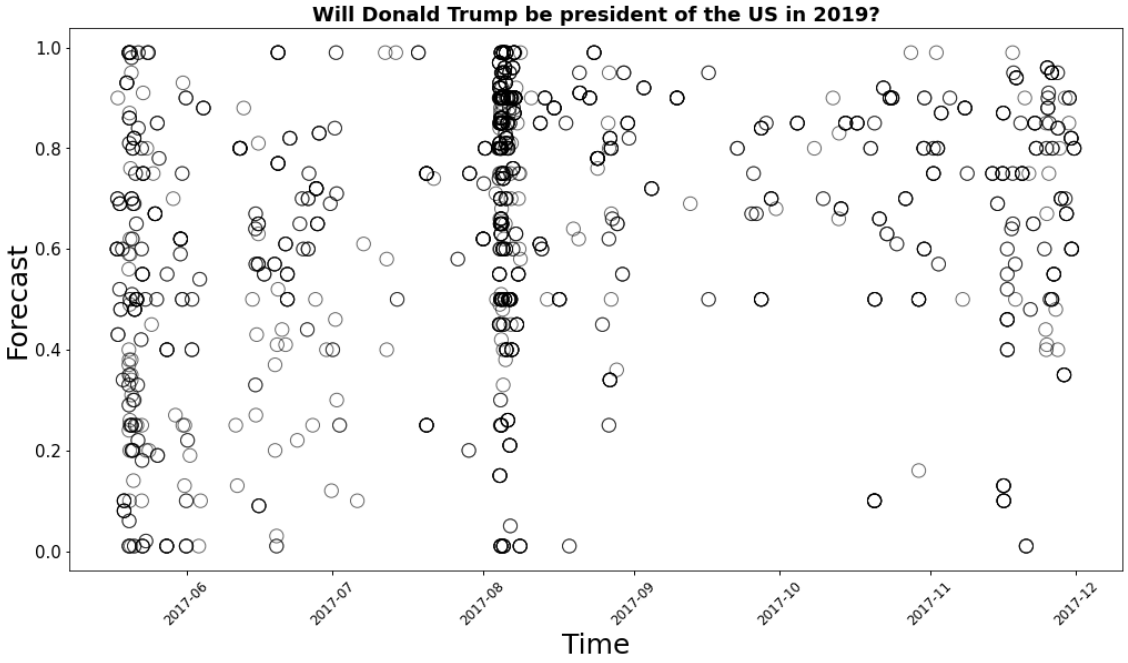


Figure 3.1: **Metaculus forecasters predict Trump presidency continuing to 2019**. Probabilistic forecasts, unattached to specific forecasters, are recorded over 7 months in 2017. Forecast values of zero and one correspond to certainty that Trump will and will not be President in 2019, respectively. Such data are the starting point of the kairosis calculations.

Forecast aggregation, by which we can access the wisdom of crowds [81], is well developed for static probability forecasts. With no information beyond the raw distribution of forecasts one can use simple measures of central tendency such as the mean or median, more subtle measures such as the extremized mean [7], or more exotic statistics of the distribution [66]. If it is possible to measure information heterogeneity or forecaster quality, much more can be done, typically by constructing various forms of weighted pool [68, 29, 74, 24]. Even more is possible when one asks the forecasters about others' likely views [62, 67]. For a recent review see [86].

In contrast, the study of dynamic forecasting problems such as those of Figure 3.1 is in its infancy. Suppose that, at any given time within the question window, we wish to construct the best possible current forecast from all forecasts already submitted. We know nothing about the knowledgeability of the individual forecasters,

or the evidence informing their forecasts. Clearly, static techniques, which aggregate all extant forecasts regardless of submission time, are inappropriate and would be expected to be sub-optimal. But what should be done instead? The state of the art is summarized by [49], who begin with the central assumption that forecasts should tend to improve over time; two suggestions are to discount the past by weighting exponentially, or to select the most recent 20% of forecasts. Yet these neglect the information which a visual inspection of the plot immediately shows is present, concerning past trends and events: the distribution of forecasts is clearly evolving, with moments of change and intense forecasting. At its simplest, we could instead try to identify the most recent "change point" (at which the underlying statistical distribution of forecasts changed), and use only forecasts made since then. This effectively assumes that at the change point significant new information became available but thereafter no more emerged.

However, as [49] note, "it is possible to envision a hybrid method ... of the selection and weighting approaches", and it is our purpose in this chapter to provide this. Our technique is informed by, but not equivalent to, Bayesian change point detection, and can be viewed as something like an interpolation between exponential discounting and the "most recent change point" method suggested above. It achieves this by using exponential discounting as its prior (starting) position, and then constructing for each possible change point a likelihood that this splits the forecasts into two sets, before and after the supposed change point, with distinct distributions. This gives us a posterior likelihood for every possible change point, whose probability density is then integrated to give a cumulative mass function, which is used to weight past forecasts. The practical effect of our method is to create a weighting with multiple downward (as we move into the past) steps, each corresponding to significant changes in the distribution of forecasts. At one extreme, if there is no obvious time at which the distribution of forecasts changes, the weighting stays close to the original exponential discounting of past forecasts. At the other extreme a single, obvious change point between very different distributions of forecasts is effectively a horizon, behind which old forecasts add nothing.

It is clear from Figure 3.1 that forecasts are not made at a uniform rate in chronological time; rather there are intense bursts of activity, typically due to the publication of relevant news, interspersed by quieter periods. If R forecasts are made at chronological times t_r, $r = 1, \ldots, R$, then the forecast order $1, \ldots, R$ effectively creates a "forecaster clock" that dilates and magnifies the moments at which forecasts

are being made at a high rate, and it is in this "forecaster time" that we search for change points. However, intense forecasting activity need not imply that the information landscape and distribution of forecasters' beliefs are changing. To conceptualize the moments of change, we note that the ancient Greeks distinguished $\chi\rho\acute{o}\nu o\varsigma$ (*chronos*), chronological or calendar time, from $\kappa\alpha\iota\rho\acute{o}\varsigma$ (*kairos*), the "time" of critical moments of lived experience [78], and our weighting function can then be thought of as the transformation from chronos to kairos (an example of which is presented in Figure 3.3b). "Kairos" has a nice interpretation as "a moment of time when a prophecy was pronounced" [83], which fits comfortably with our conjecture that change points in forecast distributions are being driven by the arrival of new information. "Kairosis" then makes for a correspondingly concise name for our method.

As noted above, dynamical probability forecast aggregation is a nascent topic whose recent developments are nicely summarised in [49]. Many of the most prominent approaches involve tracking individual forecasters over time. For example, [48] use structured regression models with smooth exponential or logarithmic time dependence to identify skilled forecasters. Other work posits properties that a good time-series probability forecast should satisfy and uses these to diagnose improvability and inefficiency [69]. Related studies track individuals and analyse the rationality of their Bayesian updates [8]. Hierarchical models have also been used to analyse infrequently updated forecasts that are intractable using standard time-series techniques [75]. In a slightly different context, [85, 84] advocate structured regression models and Bayesian change-point detection for the historical analysis of time series.

In contrast, in the Metaculus tournament data we consider in our current work, individual forecasters are not identifiable and may make as many or as few forecasts as they wish. Our problem is essentially to extract the best possible current estimate of the crowd view from the dynamics of the evolving distribution of forecasts. Although it is not currently the focus of our attention, it would be perfectly possible to develop our method to incorporate additional weights based on forecasters' skill. Similarly, kairosis could be modified to work with other aggregation techniques from the static wisdom-of-crowds literature – using, for example, different measures of central tendency of the kairosis-weighted probability distribution, such as an extremized mean [10] or its skew-adjusted variant [66].

The chapter is structured as follows. In Section 3.2 we describe our method in detail, illustrating in Figures 3.1–3.3 its application to the question of the Trump

presidency in 2019. In Section 3.3, after establishing a set of metrics and comparators, we report empirical results across 650 forecasting problems posed by Metaculus. The main part of the chapter finishes with a discussion of possible developments and applications. In 3.A we provide an additional illustrative example of computing a forecast-weighting function with our kairosis methodology in order to cement the ideas presented in the body of the chapter. In 3.B we present the preliminary results of the extension of kairosis to point forecasts of continuous variables. In 3.C we report the results of a sensitivity analysis that accompanies the findings of Section 3.3.

3.2 METHODS

In this section we expand on our proposed method for aggregating subjective probability forecasts and explain the calculations involved. The result is a set of aggregation weights derived from the cumulative mass function (CMF) of a posterior distribution over times at which significant changes in forecaster behaviour are thought to occur.

3.2.1 Deriving a distribution over change point locations

Formally, we use Bayes' theorem to obtain a posterior mass function for the time of the most recent change point. Let t_{CP} denote the random variable corresponding to the time of the most recent change point. In this framework, there exists a kairos: a most recent critical event or structural change occurring at some realised time $t_{\text{CP}} = t_r$. Our objective is therefore to construct the posterior probability mass function over candidate change-point times t_r. The corresponding CMF provides us with the posterior probability that the most recent change point occurred earlier than the given t. Equivalently, this is the posterior probability that a forecast immediately following t was made after the most recent change point and so ought to contribute to our post-change point aggregated forecast, while those made before t should not.

The framework considered here is retrospective rather than sequential. That is, the methodology uses the full realised sequence of forecasts associated with a resolved event in order to infer the posterior distribution of the most recent change point. This differs from classical sequential change-point detection procedures [54, 13], where the objective is online detection as observations arrive through time.

To make the case more formally we can introduce a binary weighting function

$$w(s \mid t_{\text{CP}} = t) = \begin{cases} 0 & s < t, \\ 1 & s \geq t \end{cases} \tag{3.1}$$

that tells us how to weight a forecast made at time s conditional on the most recent change point's having occurred at time t. Then, taking an expectation over possible values of t_{CP}, we are led to the expression

$$\mathbb{E}\left[w(s \mid t_{\text{CP}})\right] := \sum_{r=1}^{R} w(s \mid t_{\text{CP}} = t_r)\, P(t_{\text{CP}} = t_r) = \sum_{t_r \leq s} P(t_{\text{CP}} = t_r), \tag{3.2}$$

which is exactly the CMF for t_{CP} evaluated at s.

Now, returning to our central example, suppose it is October 10, 2017. Our data are R forecasts made at times t_r, $r = 1, \ldots, R$, up to this date and we are tasked with submitting an optimal aggregated probability forecast to answer the question of the 2019 US President. It is reasonable to assume that most of the data is relevant, but to what extent is not immediately apparent. One might, for example, propose that more recent data is likely to be more informative than that from the distant past. On the assumption that critical events invalidating preceding forecasts occur independently at a constant rate over time, we are led naturally to the idea of exponentially discounting old forecasts.

We therefore begin with a prior assumption that between consecutive forecasts there is a constant probability $p \in [0, 1]$ that a change point-inducing event occurs. Under the assumption that change points occur independently between successive forecast arrivals, this motivates a geometric prior distribution over candidate change-point times,

$$P(t_{\text{CP}} = t_r) = p(1-p)^{R-r},$$

since for t_r to represent the most recent change point there must have been $R - r$ subsequent inter-forecast intervals without an additional change point. The geometric prior therefore provides a parsimonious way of encoding the idea that more recent change points are a priori more likely than distant ones. We then update this prior distribution using the observed forecast sequence and Bayes' theorem,

$$P(t_{\text{CP}} = t_r | \text{Forecasts}) = \frac{P(\text{Forecasts} | t_{\text{CP}} = t_r) P(t_{\text{CP}} = t_r)}{P(\text{Forecasts})}. \tag{3.3}$$

The next key ingredient for our methodology is thus to specify the distribution of the forecasts given a candidate change point. As introduced in Chapter 2, we use

the compound Dirichlet-categorical distribution to describe the number of forecasts falling in different sub-intervals ("bins") of $[0, 1]$. Following ideas related to Bayesian hypothesis testing discussed by Holmes, Caron, Griffin, and Stephens [50], we use the probabilities assigned to the observed bin counts to construct $P(\text{Forecasts} \mid t_{\text{CP}} = t_r)$ appearing in (3.3).

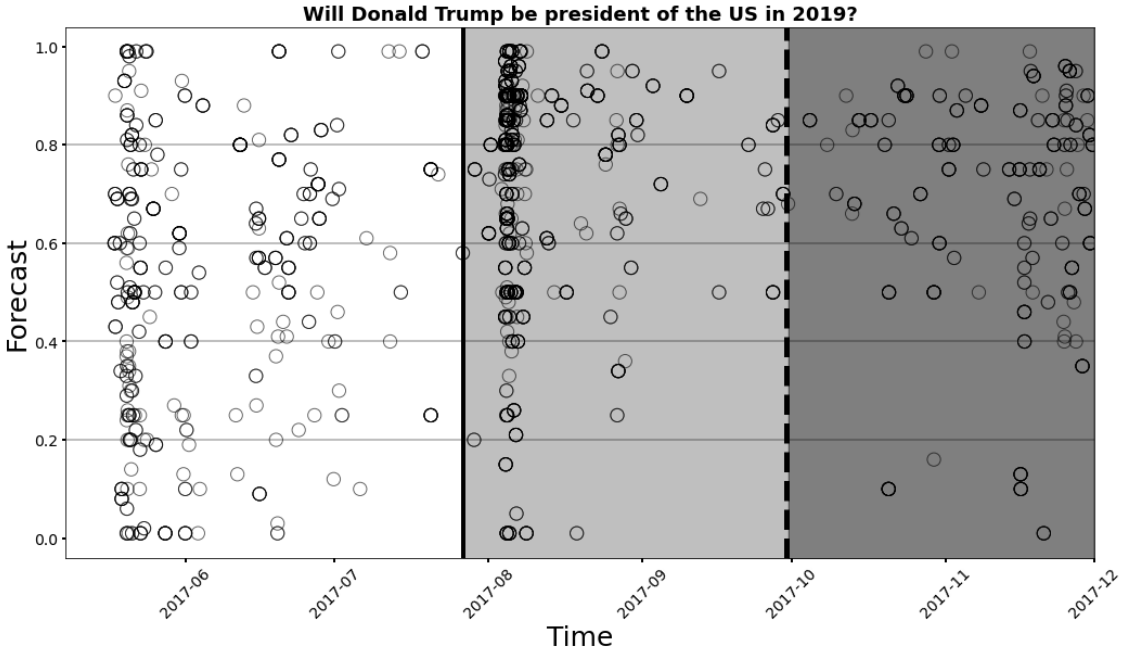


Figure 3.2: **Forecasts are partitioned to measure their relative frequencies during different intervals**. A forecast aggregator needs to make her own forecast at Oct 10, 2017 (dashed vertical line). She is unaware of forecasts in the darkest shaded region. She considers the relative probabilities of change points having occurred immediately before each of the individual forecasts were made. The solid vertical line illustrates one candidate. Her calculations are based on relative frequencies of forecasts in a set of bins (separated by horizontal lines).

The Dirichlet-categorical distribution is a compound distribution from which we can produce samples in two stages: first, drawing latent probabilities from an underlying Dirichlet distribution, then using these probabilities to assign labels to objects. The conjugacy between the Dirichlet and categorical distributions make it a natural choice for Bayesian modelling of categorical data and, importantly, leads us to a closed-form expression for its probability mass function.

The Dirichlet-categorical distribution assigns probability mass

$$P(n_1, \ldots, n_K) = \frac{\Gamma(\sum \alpha_k)}{\Gamma(\sum n_k + \alpha_k)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}, \tag{3.4}$$

to the event in which bin counts n_k are observed for bin labels $k = 1, \ldots, K$. The gamma function, an analytic extension of the factorial, is $\Gamma(x) = \int_0^\infty s^{x-1}e^{-s}\,ds$. The $\alpha_k \geq 0$ are parameters for the Dirichlet distribution that is effectively averaged over. They are usefully interpreted as pseudo-counts, reflecting approximate *a priori* beliefs for the true values of the bin probabilities – thus, as we shall require later, small values of α_k correspond to uninformed priors, while large α_k impose strong, informative prior beliefs. For readers familiar with information entropy, it is useful to note that if the bin proportions n_k/N converge to fixed limiting frequencies as $N = \sum_{k=1}^K n_k \to \infty$, then Stirling's approximation implies

$$\log P(n_1, \ldots, n_K) = N \sum_{k=1}^K \frac{n_k}{N} \log\left(\frac{n_k}{N}\right) + o(N).$$

Thus, asymptotically, the log-probability is proportional to the negative entropy of the empirical bin distribution.

Now, to propose a change point at time $t_{\text{CP}} = t_r$ (that is, immediately after the rth forecast) is to suppose that forecasts before and after t_r are more appropriately described by different latent belief distributions. Operationally, this is implemented by modelling the two segments using separate Dirichlet-categorical distributions. This should not be interpreted as implying that forecasters completely ignore all previously available information after a change point. Rather, the modelling assumption is intended as a reduced-form approximation capturing the possibility that a sufficiently important event may substantially alter the distribution of subsequent forecasts. Thus, for a single change point at t_r, expression (3.3) becomes

$$\begin{aligned} P(t_{\text{CP}} = t_r \mid \text{Forecasts}) &\propto P(\text{Forecasts} \mid t_{\text{CP}} = t_r)\, P(t_{\text{CP}} = t_r) \\ &= \frac{\Gamma(\sum \alpha_k)}{\Gamma(\sum n_k + \alpha_k)} \prod_{k=1}^K \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \\ &\quad \times \frac{\Gamma(\sum \alpha'_k)}{\Gamma(\sum n'_k + \alpha'_k)} \prod_{k=1}^K \frac{\Gamma(n'_k + \alpha'_k)}{\Gamma(\alpha'_k)} \\ &\quad \times p(1-p)^{R-r} \end{aligned} \tag{3.5}$$

where the unprimed and primed letters denote counts n_k and pseudo-counts α_k up to and after t_r.

We then evaluate (3.5) for every candidate change point $r = 1, \ldots, R$, taking us from calendar time 17 May 2017 to 10 Oct 2017 in our example. With each evaluation, we are asking the question "What is the probability that this set of

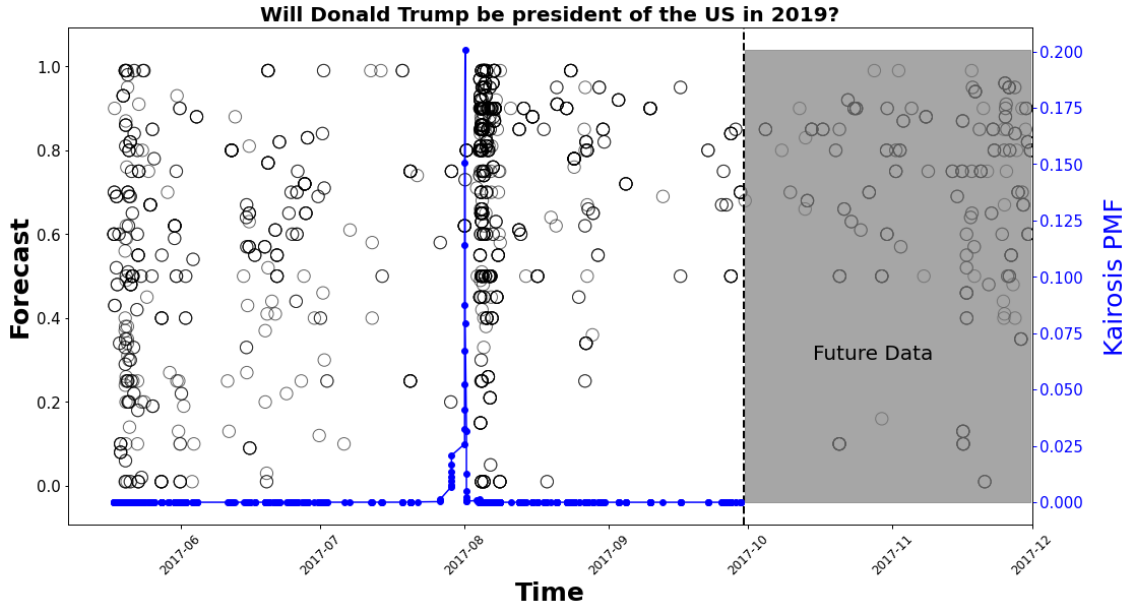
forecasts is actually drawn from two different distributions, one up to and another after our candidate t_r?" A visualisation of a single step in the process is shown in Figure 3.2 where the two periods are highlighted with no shading and light gray shading, respectively. Having computed (3.5) for each time point, we can normalize it to derive our posterior mass function for the location of the change point. The corresponding CMF then provides weights for our aggregated forecast.

Because the kairosis weighting procedure combines several probabilistic components and sequential calculations, Appendix 3.A provides a fully worked illustrative example on simulated data with a simplified set of three possible change points in order to make the mechanics of the method more transparent.

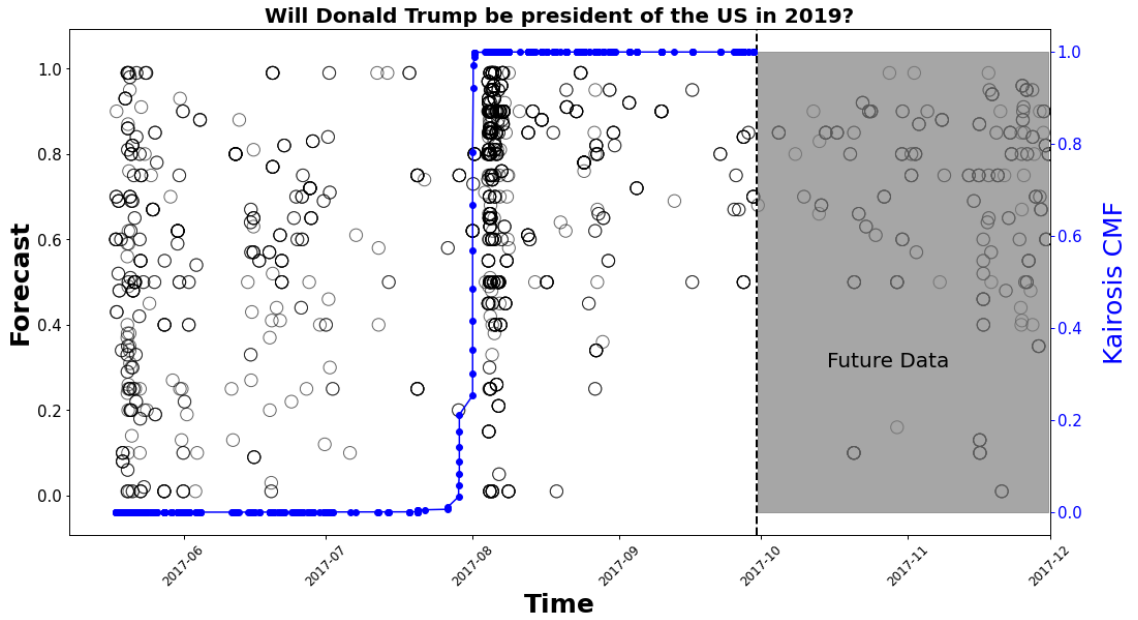
The mass and cumulative mass functions for our running example are illustrated in Figure 3.3. The dominant feature here is a flurry of forecasting activity in early August 2017. The "clock" of forecaster time is running fast here, increasing the concentration in chronological time of potential change points. Our posterior probability for a change point accounts for this activity and, using the Dirichlet-categorical term in (3.5), considers whether the distribution of forecasts actually changes here. Indeed, it appears to shift upwards, and the probability of a change point is deemed to be high. The resulting CMF features a large, sudden rise with the effect that our aggregation weights also rise. With the vertical scale from 0 to 1 as shown, all forecasts after early August receive weights between 0.8 and 1.0, while those before are given weights less than 0.25. This relative downweighting of earlier forecasts by a factor of 3 to 4 is then kairosis's belief of what is appropriate given its degree of confidence in the time of the most recent change point.

It is perhaps worth noting that the timing of this probable change point coincides with the announcement of the handing-over, from the Trump campaign team to the US Senate Judiciary Committee, of documents relating to suspected Russian collusion in the 2016 presidential election. Although the link between the announcement and the probable change point obviously cannot be ascertained here, it does provide a plausible explanation for the data.

In addition to their primary use as weighting functions, we find it useful to interpret CMFs such as those illustrated in Figure 3.3b as functions nonlinearly transforming chronological time to a scaled version of our notional kairos time. This concept guides our intuition when reconciling contextual information, observed forecaster data and aggregation weights.



(a)



(b)

Figure 3.3: **Kairosis weights are computed from the cumulative mass function (CMF) of the posterior distribution for the most recent change point**. Values are calculated for (a, upper) the mass distribution function and (b, lower) the cumulative mass function for the posterior distribution over potential times for the most recent change point. They are interpolated and superposed on Subfigures 3.3a and 3.3b, respectively, with their vertical axis labels on the right of the plots. For the CMF this happens to be the same $[0, 1]$ interval as for the forecasts, but the PMF is on a dissociated scale from 0 to just above its maximum value. The CMF then provides the kairosis weights for aggregating the forecasts.

3.2.2 Parameter Selection

The calculations described in the preceding section rely on the specification of a small number of parameters and modelling choices:

1. the binning of $[0, 1]$ from which the counts n_k are derived,
2. the change point occurrence rate p, and
3. the $\{\alpha_k, \alpha'_k\}$ that parameterize the prior distributions for the bin probabilities before and after a putative change point.

When considering the specification of forecast bins, we emphasize that the binning is only for the purpose of assessing the likelihoods of a change point. Once the change point CMF has been constructed, it is used to weight the original, precise forecasts. In our numerical experiments we have chosen to partition $[0, 1]$ into five equally-sized intervals, representing a fairly coarse-grained discretization of the forecasts.

Remembering that our forecaster clock is ticking with the arrival of each forecast, we can think of the prior change point occurrence rate p in terms of its reciprocal $1/p$, which describes the expected number of forecasts per change point. Seen this way round, the parameter can be understood as describing the responsiveness of the population of forecasters, which in the Metaculus data remains fairly constant from question to question even though the question topics may vary considerably. In the numerical comparisons of Section 3.3 we have specified $1/p = 10$, meaning that we expect a change point to motivate approximately 10 forecasters to contribute updated forecasts. Sensitivity analyses presented in 3.C show that this choice generally leads to near-optimal Brier scores for the kairosis-aggregated forecasts and for the aggregated forecast with geometrically decaying weights – beyond the avoidance of very low values, there is no need for fine tuning of $1/p$ to achieve good results. We note also that the prior distribution for the change point location tends towards a uniform distribution as $1/p$ becomes very large so that all values of $1/p \gg R$ lead to approximately the same posterior inferences.

Recall that the α_k and α'_k are pseudo-counts quantifying our prior beliefs for the proportion of forecasts falling in each bin, to either side of a putative change point. They appear in the likelihood function for the change point (equation (3.5)) via the quantities $\alpha_k + n_k$ and $\alpha'_k + n'_k$. First, consider the situation *after* the most recent change point. The forecasts here are understood to come from a single distribution and we expect this distribution to have low entropy, meaning that forecasts are

concentrated in a small number of bins. When this is observed to be the case for a candidate change point, then our posterior probability for its being the most recent change point is increased. Small values of α'_k, equal for all bins, correspond to a prior that allocates high probability to such low-entropy forecast distributions, and we choose $\alpha'_k = 1$. Next, consider the situation *before* the most recent change point. This can be more complicated, since the forecasts here could have resulted from multiple inter-change point distributions. We cannot expect them to come from a single distribution characterized by high levels of agreement among forecasters. We therefore encode the anticipated diversity of historical forecasts, which grows as that history increases in duration, by setting $\alpha_k = \lambda t_{\text{CP}}$ where $\lambda > 0$ is a positive scaling factor. An effect of this choice is to limit, but not ignore, the influence of distant history on our estimate for the most recent change point. In the calculations of Section 3.3 we have chosen to set $\lambda = 0.2$. This choice is motivated by the idea that for every forecast in the distant past (i.e. before the most recent change point) we add another pseudo-forecast that is effectively spread out over the five bins. Our sensitivity analyses indicate a high level of robustness to choices of $\lambda > 0.1$.

The overall results of the sensitivity analysis are conveyed by Figure 3.7. To summarize: other than the avoidance of very high values of p (corresponding to a prior belief in very frequent change points) and very low values of λ (almost no stabilization of pre-change-point forecasts) there is no need for fine tuning of p or λ in order for kairosis to achieve its results.

3.3 RESULTS

To test our kairosis methodology we study its performance for 650 Metaculus forecasting questions. These questions vary both in the length of time the question is open and in the number of forecasts received, with a mean of 145 days and 893 forecasts. The questions span diverse topics, including international conflict (*e.g.* "Will Russia expand by means of armed conflict before 2020?"), energy (*e.g.* "Will radical new low energy nuclear reaction technologies prove effective before 2019?"), and business and finance (*e.g.* "Will there be a financial crisis in China in 2017?").

The mean and median forecasts across the 650 questions were 0.436 and 0.444, respectively. The standard deviation for the forecasts (computed for each question and then averaged over questions) was 0.181, quantifying the level of agreement

typically seen between forecasts. The mean Brier and Log scores (defined below), again calculated over all questions and forecasts, were -0.179 and -0.560, respectively. The mean Brier and Log scores (averaging over questions) for the median forecasts (computed within each question) were -0.138 and -0.438, respectively, whose higher values reflect the "crowd wisdom" phenomenon that follows from the use of concave scoring functions.

3.3.1 Scoring aggregated forecasts

To evaluate forecasts and aggregates of forecasts for binary event outcomes we consider positively oriented Brier and Log scores,

$$\begin{aligned} S_{\text{Brier}}(X, p) &= -(p - X)^2 \\ S_{\text{Log}}(X, p) &= X \log(p) + (1 - X) \log(1 - p) \end{aligned} \tag{3.6}$$

where $p \in [0, 1]$ denotes a forecast for the outcome variable X which takes value zero if the forecast question resolves as "no" and one if it resolves as "yes". We contextualize these raw scores, with subscripts removed so as to refer to either Brier or Log, using a skill score,

$$S_{\text{Skill}}(X, p, p_0) = \frac{S(X, p) - S(X, p_0)}{S(X, X) - S(X, p_0)}, \tag{3.7}$$

where p_0 is a benchmark or reference forecast and the $S(X, X)$ in the denominator is the optimal score, given by a perfect "oracle" forecaster, which is zero for the Log and Brier scores. The skill score serves to shift and scale the raw scores so that a skill score of zero represents no improvement over the benchmark and a skill score of one represents unimprovable forecasting ability. A comprehensive and authoritative discussion of probability scores can be found in [39] where it is noted that, in general, skill scores are not *strictly proper* unless $S(X, p_0)$ is independent of outcome, which holds in the binary case only if $p_0 = 0.5$. Thus they should not generally be used as rewards in forecasting competitions. Nevertheless, in the context of retrospective analyses they remain a useful tool for making comparisons between forecasters and forecasting methodologies.

Our final stage of score processing involves the aggregation of skill scores over the time interval during which forecasts can be made. Specifically, for each question and each aggregation method we compute skill scores for the aggregation of forecasts

made prior to certain times. We have chosen to use three of these, equidistant in calendar time, that we associate with early-, middle- and late-stage forecasts. Weighted and unweighted mean averages of the skill scores at these three times are reported in Table 3.1 (described in more detail below), where the weights are linearly decreasing in calendar time so that early, prescient forecasts are rewarded more highly.

3.3.2 Performance Evaluation

To assess the effectiveness of kairosis, we compare its performance against three competitor methods. Each of the these methods (1)-(4) can be considered as providing a weighting for aggregating individual forecasts. A universal forecast of 0.5 provides a more primitive comparison.

1. a uniform weighting (*i.e.* leading to unweighted aggregate forecasts);
2. a kairosis-weighting (with five bins, $\alpha_k = \lambda t_{\mathrm{CP}}, \alpha'_k = 1$, $\lambda = 0.2$ concentration parameters, and $p = 1/10$ in the geometric decay prior);
3. a binary weighting that effectively discards the oldest 80% of forecasts (as proposed by [49]);
4. a weighting that decays exponentially in forecaster time (the decay rate being specified to match the prior distribution for the change point location used in the kairosis-weighting);
5. a universal forecast of 0.5 for all questions.

As noted earlier, kairosis can be seen as being positioned between methods (3) and (4). A likelihood function that is constant over time, which is approached if the $\{\alpha_k, \alpha'_k\}$ were to be made very large, makes for a posterior distribution for the change point location that is the same as the exponential prior. The exponential prior then leads to exponentially decaying forecast aggregation weights. Alternatively, by sending the parameter p to zero the exponential prior becomes flat and by sending $\{\alpha_k, \alpha'_k\}$ to zero we heighten our sensitivity to the bin counts so that large changes (which may or may not fall around the 80% mark) result in a step-like CMF and a step-like weighting function.

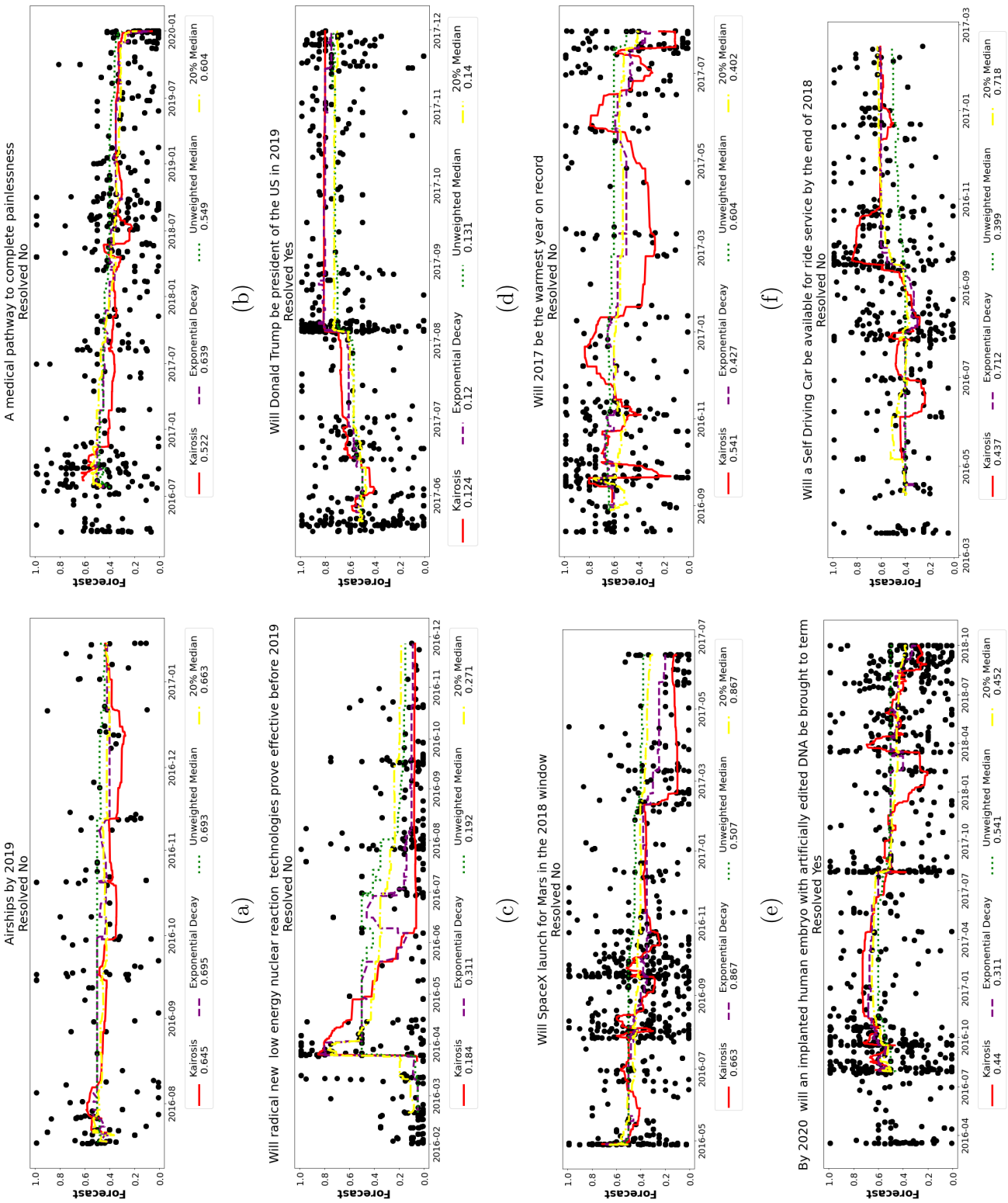
In each case (except (5), where they are equal) we compute both the weighted mean and the weighted median (that is, for forecasts $i = 1, \ldots, R$ ordered by

increasing weights w_i, the forecast f_j with j the smallest integer such that $\sum_{i=1}^{j} w_i > \frac{1}{2}$).

In Table 3.1 we present the four different aggregated skill scores (using Log and Brier raw scores, and uniform and time-decreasing weights) for the eight different forecast aggregation methods averaged across the 650 questions, along with the universal forecast of 0.5. The skill scores use the unweighted median forecast aggregation as the benchmark, so that the entries in the top row of the table are all necessarily zero. (Note that the universal 0.5 performs vastly worse than all the aggregation methods.) The best (highest) score in each column is placed in a box, and on each of the four measures this is the kairosis-weighted median. Notice that the unweighted mean performs worse than benchmark on all scores. The kairosis median is the only method to perform better than benchmark on all four scores.

			Aggregate skill scores			
		Brier scores	From raw Brier scores		From raw Log scores	
Forecast weighting	Forecast aggregate	Unweighted Brier scores	Unweighted over time	Weighted over time	Unweighted over time	Weighted over time
Uniform	Median	-0.143 (0.006)	0.000 (0)	0.000 (0)	0.000 (0)	0.000 (0)
	Mean	-0.147 (0.005)	-0.657 (0.136)	-0.652 (0.135)	-0.199 (0.022)	-0.198 (0.022)
Kairosis	Median	-0.137 (0.006)	0.060 (0.009)	0.054 (0.009)	0.046 (0.006)	0.042 (0.006)
	Mean	-0.140 (0.005)	-0.540 (0.134)	-0.545 (0.133)	-0.139 (0.022)	-0.143 (0.022)
Most recent 20%	Median	-0.143 (0.006)	-0.135 (0.067)	-0.159 (0.069)	-0.011 (0.015)	-0.023 (0.015)
	Mean	-0.144 (0.005)	-0.694 (0.347)	-0.726 (0.357)	-0.146 (0.029)	-0.157 (0.030)
Exponential decay	Median	-0.151 (0.006)	-0.211 (0.034)	-0.252 (0.038)	-0.040 (0.015)	-0.061 (0.017)
	Mean	-0.149 (0.006)	-0.447 (0.075)	-0.476 (0.077)	-0.124 (0.018)	-0.138 (0.019)
All Forecasts 0.5	All Forecasts 0.5	-0.250 (0.000)	-18.658 (5.681)	-18.525 (5.677)	-2.170 (0.198)	-2.154 (0.197)

Table 3.1: **Performance comparison for forecast aggregation methods.** Scores for aggregated forecasts are averaged over 650 forecast questions and three forecast times (early-, middle- and late-stage). Rows index methods for forecast aggregation and columns index variants of the skill score. Table entries are skill scores benchmarked against the unweighted median, so that positive values indicate better-than-benchmark performance and negative values worse-than-benchmark performance. For each skill score variant (i.e. each column) the best forecast is boxed. Standard deviations are presented in parentheses.


Airships by 2019
Resolved No
Kairosis 0.645
Exponential Decay 0.695
Unweighted Median 0.693
20% Median 0.663
(a)
A medical pathway to complete painlessness
Resolved No
Kairosis 0.522
Exponential Decay 0.639
Unweighted Median 0.549
20% Median 0.604
(b)
Will radical new low energy nuclear reaction technologies prove effective before 2019
Resolved No
Kairosis 0.184
Exponential Decay 0.311
Unweighted Median 0.192
20% Median 0.271
(c)
Will Donald Trump be president of the US in 2019
Resolved Yes
Kairosis 0.124
Exponential Decay 0.12
Unweighted Median 0.131
20% Median 0.14
(d)
Will SpaceX launch for Mars in the 2018 window
Resolved No
Kairosis 0.663
Exponential Decay 0.867
Unweighted Median 0.507
20% Median 0.867
(e)
Will 2017 be the warmest year on record
Resolved No
Kairosis 0.541
Exponential Decay 0.427
Unweighted Median 0.604
20% Median 0.402
(f)
By 2020 will an implanted human embryo with artificially edited DNA be brought to term
Resolved Yes
Kairosis 0.44
Exponential Decay 0.311
Unweighted Median 0.541
20% Median 0.452
(g)
Will a Self Driving Car be available for ride service by the end of 2018
Resolved No
Kairosis 0.437
Exponential Decay 0.712
Unweighted Median 0.399
20% Median 0.718
(h)
Forecast

3.3.3 Remarks on kairosis and crowd inaccuracy

The existence of shared biases within a crowd of forecasters imposes a natural limit on the effectiveness of any aggregate forecast. In the context of the Metaculus questions, the shared bias can be attributed to the forecasters' all inferring event probabilities from subsets of a common, and ultimately inconclusive, set of relevant data. In this sense the forecasts available to us ought to be considered partial observations of the common data rather than of the event itself. Kairosis allows us to adapt to shifting information landscapes but clearly cannot estimate event outcomes to arbitrary accuracy.

The two key questions now are whether the crowd of Metaculus forecasters possesses a significant amount of information relevant to a particular question, and whether a significant proportion of that information can be exploited via kairosis but not with simpler aggregates. The answers vary appreciably between questions. In Figure 3.4 we take a closer look at the operational dynamics of kairosis on the crowd forecasts as compared with the other methods for eight questions. Sub-figures (a)-(d) illustrate instances where kairosis adapts effectively to directional changes in crowd opinion which prove to be correct. We observe in each case the capacity of kairosis to react quickly but stably to significant changes, and also to respond to steady trends. Conversely, sub-figures (e)-(h) demonstrate scenarios where kairosis accurately follows the crowd's movements, even though the crowd itself was incorrect. Such cases are often characterized by the exaggerated movements of an uncertain crowd (which are tracked by kairosis), perhaps due to overreaction or news events not adding real information, even as a question's closing date approaches.

3.4 DISCUSSION

Our results indicate the broad feasibility of using change-point methods for dynamical probability forecast aggregation. They do so by identifying points in time at which the distribution of forecasts changes, which we attribute primarily to new information emerging and informing the views of the forecasters. The concept of kairos helps us attribute meaning to the change point CMF that provides weights for our aggregated forecasts, linking a classical concept to a computational method by way of a Bayesian model.

Looking forward, we believe it would be particularly useful to combine our work with more sophisticated change point detection methods, specifically online methods such as those proposed in [2]. A sequential, online approach in which the change point CMF and/or the forecast aggregation weights are updated rather than recomputed may be necessary to scale up and speed up our calculations to larger systems. We suspect that such an approach is also key to generalizing our method to situations in which multiple change points are identified, a problem which otherwise threatens a combinatorial explosion in the number of likelihood evaluations. Many smaller variations to the kairosis methodology are also possible, which we invite the reader to experiment with using the Python code provided in the Supplementary Materials. For instance, we used kairos-informed weights to construct a weighted median, but other measures of central tendency could be used instead. We used only raw forecast data, but measures of forecaster skill could easily be used, say, to weight the counts appearing in (3.4) so that our change point calculations are more sensitive to the most skilled.

Kairosis can also be adapted to forecasts on non-binary domains. In 3.B we test the concept by applying kairosis to point forecasts in Metaculus questions on continuous domains: in this form, kairosis works better than other methods, but by a smaller margin. Kairosis could also be used to analyse distributional forecasts of continuous variables where the probabilities are elicited via a set of bins.

The change point model underlying kairosis can also be used as a means to analyze the development of opinions among a population without directly linking this to the outcome of a forecasting question. This would be particularly useful if we were interested, for example, in quantifying the effect of certain events on those opinions. Another possibility would be to infer the degree of collective wisdom of the crowd from its dynamic behaviour, potentially via its under- or over-responsiveness to current events, which could then be used to discount the crowd altogether in favour, say, of some baseline event probability. It would be particularly interesting to reconcile this idea with recent work on herd dynamics among forecasters [52].

In summary, kairosis is an effective new method for dynamic probability forecast aggregation. It is built from a coherent, tractable underlying Bayesian model and makes no assumptions on the distribution of forecasts. We have demonstrated its potential to quickly account for sudden changes in the beliefs of a population of forecasters and anticipate that its good performance will translate to other domains, such as psephology and marketing, involving sentiment and behaviour tracking.

3.A A WORKED EXAMPLE OF COMPUTING THE KAIROSIS WEIGHTING FUNCTION

In this example we step through our kairosis calculations for an artificial example, whose simulated data has been designed to exemplify the types of phenomena we are interested in. For the sake of exposition the example involves only three candidate points in forecaster time for the most recent change point t_{CP}. For each candidate value we compute (up to proportionality) a posterior probability according to

$$\underbrace{P(t_{\mathrm{CP}} = t_r \mid \text{Forecasts})}_{\text{posterior}} \propto \underbrace{P(\text{Forecasts} \mid t_{\mathrm{CP}} = t_r)}_{\text{likelihood}} \underbrace{P(t_{\mathrm{CP}} = t_r)}_{\text{prior}} \quad t_r \in \{t_1, t_2, t_3\}.$$

We specify a prior probability that decays geometrically as explained in Section 3.2.1. The likelihood term factorizes into two parts on the basis that a change point is assumed to force a change in the distribution of the forecasts up to and after it so that

$$\begin{aligned} P(\text{Forecasts} \mid t_{\mathrm{CP}} = t_r) = & P(\text{Forecasts made up to } t \mid t_{\mathrm{CP}} = t_r) \\ & \times P(\text{Forecasts made after } t \mid t_{\mathrm{CP}} = t_r). \end{aligned}$$

The factors here are based on the bin counts up to and after the proposed change point t_r. They are computed using the probability mass function of the Dirichlet-categorical distribution whose functional form is provided in equation (3.4). This mass function is maximized when all the counts fall within a single bin and minimized when they are evenly distributed across them all. Intermediate values are attained for intermediate degrees of forecast dispersion.

In Figure 3.5 we plot histograms of the pre- and post-change point bin counts for the three candidate time points in our example. The log-likelihoods induced by these counts, which combine additively when computing the log-posterior, are largest when the candidate change point splits the forecasts into subsets both of which are concentrated on a small number of bins. Importantly, these do not need to be the same bins for both the subsets. The log-likelihood is smaller when a set of forecasts is more evenly distributed over the bins.

The values that contribute to the posterior cumulative mass function, which provides us with a weighting function for aggregation, can be found in Table 3.2. We can see, for example, that the earliest and latest candidate change points are assigned very small posterior probabilities. This is owing to the empirical distribution

of forecasts on one side of these time points containing a substantial number of both high and low forecasts, leading to relatively large negative log likelihoods (of -50.667 and -51.985). The greatest posterior probability is calculated for the second candidate time point, which succeeds in partitioning the forecasts into two subsets that are both reasonably concentrated, a feature quantified by the correspondingly moderate contributions (of -26.321 and -34.430) to the log likelihood.

The three candidate change points lead us to a piecewise linear weighting function with three step discontinuities, as shown in Figure 3.6. When a candidate time point is assigned a large posterior probability of being a change point the values of the weighting function to either side are pushed apart, making the weights lower to the left and higher to the right.

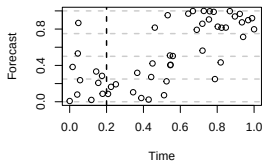
	t= 0.2	t= 0.5	t= 0.8
Log-prior	-5.290	-4.644	-3.547
Loglike. for earlier forecasts	-14.792	-26.321	-51.985
Loglike. for later forecasts	-50.667	-34.430	-13.125
Unnormalized log-posterior	-70.749	-65.395	-68.657
Posterior mass funct.	0.005	0.959	0.037
Posterior cumulative mass funct.	0.005	0.963	1.000

Table 3.2: Intermediate calculations accompanying 3.A.

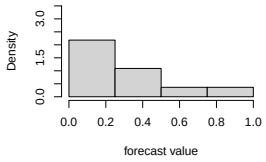
3.B NON-PROBABILISTIC QUESTIONS

As an initial test of the applicability of kairosis beyond probability forecasting for true/false questions, we also applied our method to Metaculus questions which required point forecasting of results on continuous domains. This preliminary analysis involved minimal changes to kairosis, but the results are promising.

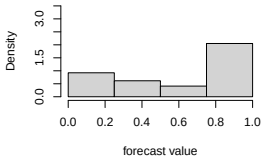
The crucial alteration of the method is in the binning of forecasts. Probability forecasts are all on the same, fixed interval $[0, 1]$, whereas with general point forecasts the domain varies from question to question, necessitating bins which vary not only between questions but also with present (forecaster) time R. In our experimental calculations, whose results are presented below, we adopted a simple adaptive binning strategy which involved partitioning the range of possible forecast values according to the quantiles of the forecasts observed up until the aggregated forecast is needed. This partition defines the bin structure that is used to compare forecast distributions to



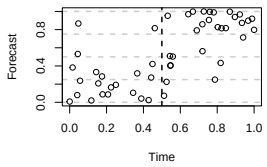
(a) Simulated forecasts partitioned at $t = t_1$.



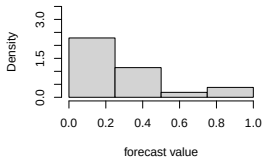
(b) Empirical distribution of pre-change point forecasts for $t = t_1$.



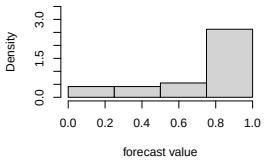
(c) Empirical distribution of post-change point forecasts for $t = t_1$.



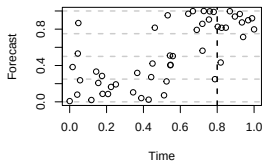
(d) Simulated forecasts partitioned at $t = t_2$.



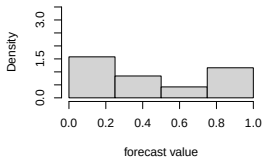
(e) Empirical distribution of pre-change point forecasts for $t = t_2$.



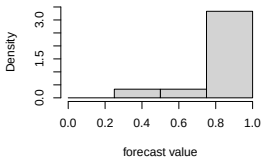
(f) Empirical distribution of post-change point forecasts for $t = t_2$.



(g) Simulated forecasts partitioned at $t = t_3$.



(h) Empirical distribution of pre-change point forecasts for $t = t_3$.



(i) Empirical distribution of post-change point forecasts for $t = t_3$.

Figure 3.5: **Illustrative kairosis calculation for a simulated forecasting example**. The panels show simulated forecast trajectories and the corresponding empirical distributions of forecasts before and after each candidate change point described in Section 3.A. These plots illustrate how forecast concentration on either side of a proposed change point drives the likelihood terms used in the kairosis weighting procedure.

either side of each candidate change point which, following the calculations described in the main chapter, leads to our kairosis weighting function used for aggregation.

Across more than 200 questions with an average open time of 85 days and 2321 forecasts per question, kairosis consistently outperformed both the benchmark and other competing methods. However, it is worth noting that the standard deviations in this scenario are considerably larger than those observed in the probability forecasting

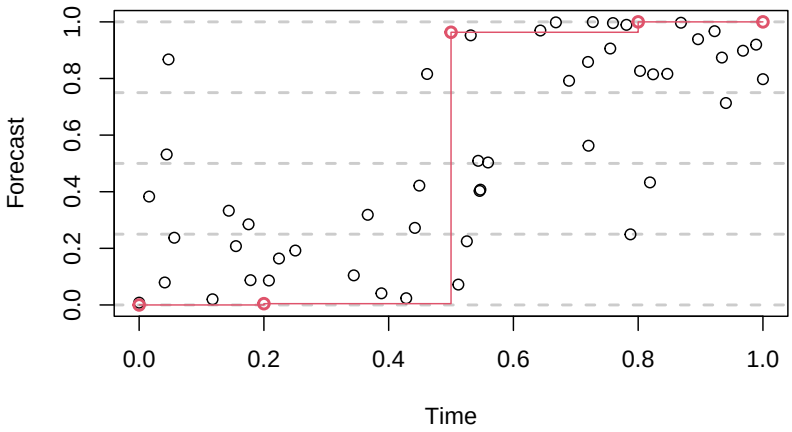


Figure 3.6: **Kairosis weighting function derived from the simulated forecasting example**. The piecewise linear weighting function is constructed from posterior change-point probabilities computed for the artificial data described in Section 3.A. Larger discontinuities correspond to candidate change points with higher posterior mass.

questions. Specifically, we observe only a two-standard-deviation difference from the benchmark, in contrast to the more substantial margins seen in probability forecasting. Table 3.1 presents a comparison of the unweighted and weighted skill scores, along with their standard deviations, for each of the proposed methods.

3.C SENSITIVITY ANALYSIS

We now assess the sensitivity of our kairosis calculations to changes in the parameters p, which governs the geometric prior on change point locations, and λ, which governs the rate at which pseudo-counts $\alpha_k = \lambda t_{\text{CP}}$ are added to the Dirichlet prior for bin probabilities before the supposed most recent change point. Specifically, we calculate Brier scores for aggregated forecasts using weights derived from our kairosis calculations and average these over the Metaculus questions. The calculations are repeated for a range of values of $\{p, \lambda\}$ and the results plotted in Figure 3.7.

We observe high levels of robustness for all $\lambda > 0.1$ and $p < 0.2$. Larger values of p encode strong prior beliefs that change points occur very often so that the prior probabilities for locations of the last change point decay extremely quickly. The

		Aggregate skill scores	
		From raw Brier (neg. squared error) scores	
Forecast weighting	Forecast aggregate	Unweighted over time	Weighted over time
Uniform	Median	0.000(0)	0.000(0)
	Mean	-1.459 (0.481)	-1.363 (0.458)
Kairosis	Median	0.042 (0.027)	0.047 (0.031)
	Mean	-2.655 (0.926)	-2.467 (0.884)
Most recent 20%	Median	-0.709 (0.377)	-0.728 (0.387)
	Mean	-3.848 (1.329)	-3.366 (1.079)
Exponential decay	Median	-0.098 (0.129)	-0.068 (0.091)
	Mean	-2.695 (0.933)	-1.363 (0.849)

Table 3.3: Performance comparison for forecast aggregation methods averaged over 200 non-probability forecast questions. Table entries are skill scores benchmarked against the unweighted median, so positive values indicate better-than-benchmark performance, and negative values indicate worse-than-benchmark performance. For each skill score variant, the best forecast is boxed. Standard deviations are presented in parentheses.

result is an aggregator that places the majority of its weight on only the most recent few forecasts. In this sense the dangers of mis-specifying p are the same as those for mis-specifying the decay parameter for an aggregator using a simple exponential weighting function. We also observe some (less severe) loss in performance for very small values of λ. Indeed, when λ approaches zero we lose some robustness to large movements in the distribution of forecasts before the proposed time of the last change point. We see that the kairosis method is most successful when, via a non-zero specification of λ, it treats the recent past and the distant past differently.

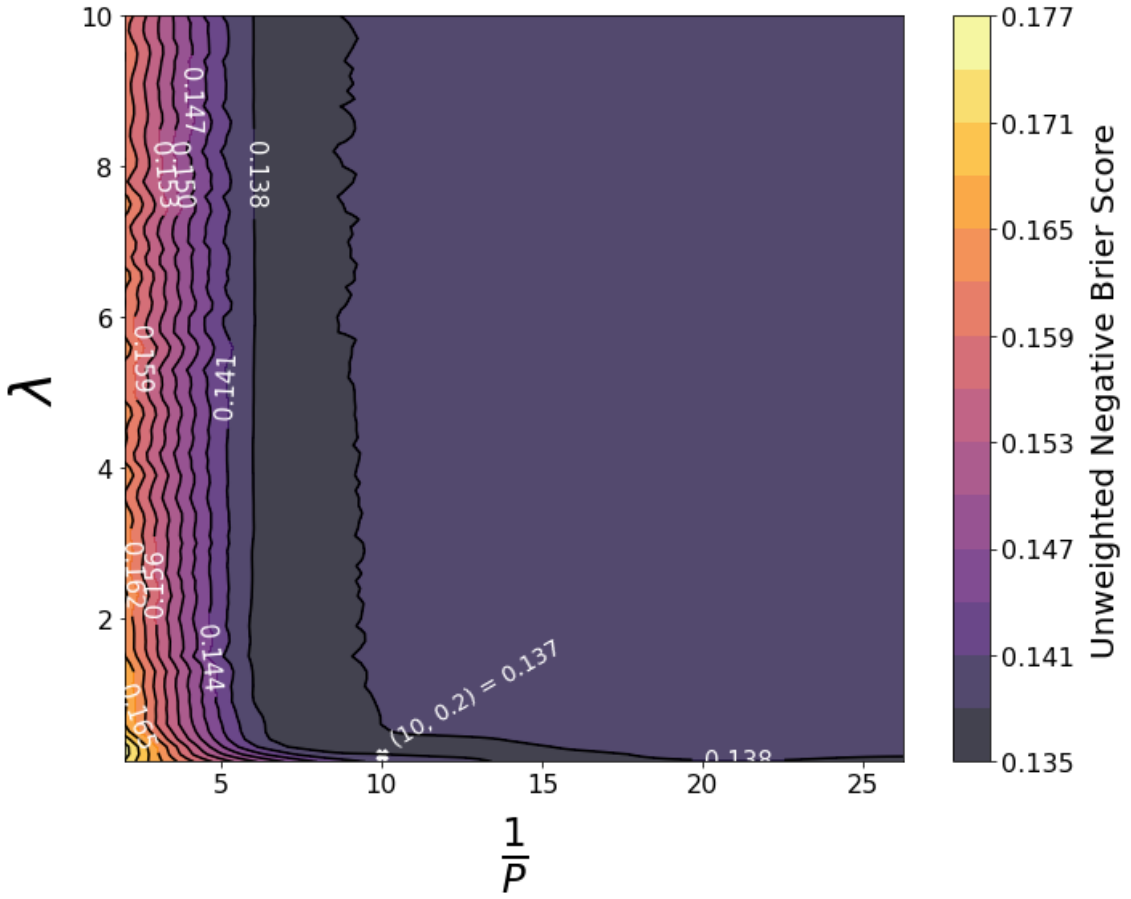


Figure 3.7: **Sensitivity of the kairosis model to the parameters p and λ.** The contour plot shows how aggregated forecast performance varies across values of the geometric prior parameter p and the Dirichlet pseudo-count parameter λ. Regions of stable performance indicate robustness of the kairosis weighting scheme to moderate parameter misspecification.

— 4 —

Belief Inference from Sparse Order Books in Prediction Markets

4.1 INTRODUCTION

Prediction markets aim to aggregate diverse beliefs into a market price that reflects the crowd's perceived probability of an event. In an ideal liquid market, the price of a binary option is often interpreted as the collective probability that the event will occur [87]. In practice, however, many prediction markets are thin, with low liquidity and sparse order books, particularly for niche or rapidly created contracts where participation is limited [55, 23]. In such settings, extracting meaningful information about market sentiment and confidence is challenging: a small number of trades or standing orders can move prices substantially, and the price alone may not fully reflect the range or dispersion of traders' beliefs [55].

Thin markets pose several related difficulties. First, the common assumption that the market price equals the average belief of participants may fail when trading activity is limited or when wealth, risk tolerance, or strategic behavior is unevenly distributed across traders [55]. In such cases, prices may be disproportionately influenced by a small number of aggressive traders, revealing little about the broader population's beliefs. Second, market microstructure frictions imply that only a subset of beliefs is ever observed. Traders whose beliefs lie close to the prevailing price often have little incentive to post orders, either because expected payoffs are small or because execution is unlikely. As a result, moderate beliefs remain largely unobserved, and the visible order book is biased toward more extreme views. Finally, the market price alone does not distinguish between qualitatively different belief

structures: two markets may trade at the same price while exhibiting very different levels of consensus or polarization among participants.

This chapter addresses the problem of inferring the latent distribution of trader beliefs from sparse order book data. We develop a Bayesian inference framework in which traders' subjective probabilities are modeled as draws from a Beta(α, β) distribution. Due to strategic trading behavior and market frictions, however, the observed order book reflects only a truncated and selectively visible subset of these beliefs: standing orders are typically placed only when private beliefs deviate sufficiently from the current midpoint price. The order book therefore represents a filtered and extremized sample of the underlying belief population.

To recover the latent belief distribution, we explicitly model this selection mechanism. Candidate (α, β) parameters are used to simulate draws from a Beta distribution, after which a rejection filter is applied to replicate the market's selective visibility of beliefs. The parameters are adjusted so that the filtered simulated sample matches salient features of the observed order book, while remaining consistent with the observed market price as an anchor for the latent mean. This inversion yields an estimate of the full belief distribution using only snapshot order book data, without requiring trader-level observations.

We apply this framework to real-world prediction market data from Polymarket, focusing on snapshots of the contract corresponding to Game 2 of the 2025 NBA Playoff series between the New York Knicks and the Boston Celtics. By estimating (α_t, β_t) across successive snapshots, we recover time-varying belief dynamics and examine how confidence, polarization, and directional skew evolve in response to in-game events. The results illustrate how identical market prices can arise from qualitatively different belief structures and demonstrate how modeling latent beliefs provides interpretability beyond price movements alone.

4.2 BACKGROUND AND LITERATURE REVIEW

4.2.1 Belief Elicitation in Prediction Markets

Prediction markets have long been studied as effective mechanisms for eliciting and aggregating beliefs about uncertain events [87]. In a prediction market for a binary event, traders buy and sell contracts that pay \$1 if the event occurs (and \$0 otherwise). Under certain conditions, such as approximately risk-neutral traders,

sufficient liquidity, and relatively homogeneous information, the market price of this contract can be interpreted as the crowd's collective probability that the event will happen. Traditional economic theory suggests that if traders are risk-neutral and have homogeneous information, the market price will equal the mean belief or expectation of the event's probability across traders [87]. This makes prediction markets useful as forecasting tools and as means to extract the "wisdom of the crowd." Indeed, empirical studies have shown that prediction market prices often outperform expert opinions and polls in forecasting real-world events (e.g., election outcomes, sports results) [87].

However, the extent to which market prices reveal the underlying distribution of beliefs is a nuanced question. Manski [55] presented a formal analysis showing that even in a rational expectations equilibrium, the market price may *not* correspond to the average belief unless specific conditions hold. In fact, if traders have heterogeneous beliefs and there are no arbitrage opportunities, a wide range of belief distributions can be consistent with a single market price. Manski demonstrated that without additional assumptions (such as risk neutrality or symmetric information), the price of a binary option reveals only that a fraction of traders believes the probability is at least as high as the price and another fraction believes it is at most as high, but it does not directly yield the mean or median belief. For example, a price of 0.60 could mean every trader believes the probability is exactly 60%, or it could reflect a population split between traders who are certain the event will occur and traders who are certain it will not. In the latter case, the observed price can arise through a simple market-clearing mechanism: if traders with belief 1 place buy orders and traders with belief 0 place sell orders of comparable size, the equilibrium price reflects the relative mass or aggressiveness of these opposing positions rather than the average belief. Subsequent work by Wolfers and Zitzewitz and others provides sufficient conditions under which prices do coincide with mean beliefs (for instance, under CARA utility and independent signals) [87], but in general additional structure is required to infer belief distributions from market data.

One common approach to enriching the information from prediction markets is to design market mechanisms that elicit not just a single price, but a distribution over outcomes. For example, some prediction markets allow trading in multiple outcome shares to recover a full probability distribution (e.g. contracts for each percentile of an outcome). In our context of binary markets, however, we are limited to a single price. Our approach is to utilize the *order book*—the collection of outstanding limit

orders—to glean more information. The idea of using order book data for belief elicitation relates to the concept of a *belief curve* or demand curve: how does the quantity of shares traders are willing to buy or sell vary with price? If one can recover that demand curve, it in principle reflects the cumulative distribution of beliefs among traders [37].While we do not directly invert the entire demand curve in this chapter, we adopt the spirit of that approach by assuming a parametric form (a Beta distribution) for beliefs and fitting its parameters using observed standing orders. Our work relates to prior studies in economics and decision theory that infer subjective probability distributions from betting behavior [25]. We extend this line of inquiry by modeling how market structure selectively reveals more extreme beliefs, with moderate beliefs often unrepresented in sparse order books.

4.2.2 Market Microstructure and Order Book Dynamics

Market microstructure theory studies how the trading process and market design affect price formation and information revelation [61]. A key aspect of microstructure in order-driven markets is the presence of a *limit order book*, where traders post bids (offers to buy) and asks (offers to sell) at specified prices. The difference between the highest bid and lowest ask is the *bid-ask spread*, and the midpoint of this spread is often taken as the market's current price or consensus value. Order books provide a richer picture than just the last traded price: they display the depth (volume) available at various prices and thus reflect the willingness of traders to trade at prices away from the current equilibrium.

In a thick, liquid market (e.g., a major stock or a high-volume futures contract), the order book is typically dense and the bid-ask spread is very tight (often just a minimal tick). In such cases, the midpoint price is a reliable indicator of the market consensus, and the volume on either side may indicate the supply and demand elasticity. In a thin market, by contrast, the order book is sparse: there may be only a few orders on each side, and the spread can be wide. Some prices may have no orders at all until far from the midpoint. This sparseness can occur in prediction markets, especially for niche or novel events, where the number of participants is small. The limited number of orders means that the market is prone to volatility and that the visible orders represent only a subset of potential trading interest.

An important feature of market microstructure is that traders do not place orders continuously at every price that matches their belief. Instead, practical

considerations such as transaction costs, queue priority, and strategic behavior lead to sparse and selective order placement. As a result, beliefs near the current price may be underrepresented in the visible book, especially in low-liquidity environments. This generates a systematic bias toward more extreme orders, which must be accounted for when inferring latent beliefs from order flow. Empirical studies have observed that limit orders tend to cluster at discrete price levels and often remain sparse near the midpoint unless new information arrives [17].

In the context of our model, these microstructure considerations justify the introduction of a filtering mechanism for visible orders. We assume that there is a symmetric band around the midpoint price within which traders typically do not place limit orders, because the market price is "close enough" to their belief. This band reflects trading frictions, hesitation, or simply a lack of incentive to act when one's belief is near consensus. Only traders whose beliefs are sufficiently far from the market price will post orders, seeing a potential profit opportunity by quoting prices that diverge from current consensus. As a result, the visible order book reflects a filtered subset of the underlying belief distribution, biased toward more extreme values.

Our approach connects with microstructure literature by formally modeling this selective visibility and using it to infer latent belief distributions. While most classical models [53, 38] focus on price formation under asymmetric information, our interest lies in recovering the cross-sectional distribution of private valuations (beliefs) from the structure of the standing orders. Related ideas appear in work estimating investor preferences or utility from order flow, but our focus is on probabilistic beliefs in a prediction market setting.

4.2.3 Bayesian Inference and Beta Distributions in Markets

Modeling belief distributions naturally motivates the use of Bayesian statistics, although alternative frequentist approaches could also be used. In Bayesian inference, a Beta distribution is the conjugate prior for a Bernoulli/Binomial likelihood and naturally represents uncertainty over probabilities. If we think of each trader as having a subjective probability (belief) about a binary event, it is sensible to model the population of such beliefs with a Beta distribution [35]. The Beta(α, β) family can represent a wide variety of shapes: for example, $\alpha = \beta = 1$ is a uniform distribution (complete uncertainty or maximum entropy over $[0, 1]$); $\alpha > 1, \beta > 1$ yields a single-

peaked distribution indicating consensus around an interior probability; $\alpha < 1, \beta < 1$ produces a U-shaped distribution indicating polarization towards 0 and 1; and if one of α or β is much larger than the other, the distribution is skewed heavily towards one end (indicating one outcome is widely seen as much more likely).

Beta distributions have been used in prior research as models for subjective probabilities and as posteriors updated by observing event outcomes or poll data [35]. In prediction markets, one can think of α and β (minus 1) as representing the effective number of independent signals or pieces of evidence that traders collectively have in favor of or against the event. For instance, a Beta(α, β) can be seen as the posterior belief distribution after $\alpha - 1$ imagined "successes" and $\beta - 1$ "failures." Some researchers have used Beta distributions to model the evolution of odds in sports betting or betting exchange markets, where odds (probabilities) update as bets come in, treating the bets as information updates. In our setting, however, we are not updating beliefs over time via Bayes' rule (that will come when we discuss temporal dynamics); rather, we are at a fixed time and we posit that the cross-sectional distribution of beliefs across traders is Beta-distributed.

The use of a parametric Bayesian model (like Beta) combined with observed market data falls under the umbrella of *inverse inference*: we have a forward model that given (α, β) would produce some distribution of orders/prices, and we invert this relationship to find (α, β) that best explains the data. This is akin to statistical inference with incomplete data, since we do not observe all beliefs, only those that result in orders. Thus, our inference procedure can be viewed as a parametric inference problem with truncated data, implemented here using a Bayesian-inspired Beta framework, which we discuss next.

4.2.4 Polymarket and Prior Work

Polymarket is a decentralized prediction market platform that gained popularity in the early 2020s. Unlike older prediction markets such as Intrade or the Iowa Electronic Markets (which were centralized) or purely automated market maker platforms like Hanson's LMSR-based markets [42], Polymarket operates on the blockchain (Polygon network) and uses a hybrid model that effectively functions as an order book for binary outcome shares. Every market on Polymarket has two outcomes (traditionally labeled "Yes" and "No"). Shares are priced between \$0 and \$1 (specifically, in USDC stablecoin), and a "Yes" share will pay \$1 if the event occurs

(and 0 if not), while a "No" share is essentially the complementary asset (paying 1 if the event does not occur). Polymarket requires that every matched trade involves the purchase of one Yes and one No share for $1 total, ensuring the market is always fully collateralized [23]. This means if a trader buys a Yes share for $0.60 from someone, that someone is effectively buying the corresponding No share for $0.40, and $1 is exchanged for the pair of shares. The price of Yes ($0.60) thus implies the price of No is $0.40, consistent with a probability interpretation (the two must sum to 1 in a fair binary market).

Polymarket presents an order book interface where users can place limit orders to buy or sell shares at specific prices. The best (highest) bid and best (lowest) ask determine the current market spread, and the midpoint of these is often displayed as the market's implied probability [23]. If the bid-ask spread becomes too wide (e.g., greater than $0.10), Polymarket instead highlights the last traded price as the market price, to avoid misleading users with a very stale midpoint. For active markets, however, the spread is usually small and midpoint is a reasonable consensus estimate.

Despite Polymarket's prominence in the blockchain space, academic literature specifically focusing on Polymarket is still nascent. There have been some descriptive analyses and commentary on Polymarket (including regulatory perspectives, given that Polymarket had to geo-block US users after a CFTC fine). Buckley [23] provides an overview of blockchain-based prediction markets and mentions Polymarket as a case study of a decentralized information market. Another line of research has examined user behavior on Polymarket in relation to political events, for example analyzing whether traders' political biases affect their betting (e.g., how Democrats vs. Republicans trade on election markets). These studies leverage the transparent data from Polymarket's blockchain (every trade and order can be recorded) but have not, to our knowledge, tried to infer latent belief distributions from the order book. Our work appears to be among the first to formalize a model for the cross-sectional distribution of beliefs using order book snapshots in Polymarket or similar markets.

More broadly, there is related work on inferring distributions of opinions or risk preferences in financial markets. For example, some researchers have used options markets to derive the risk-neutral distribution of future asset prices. Others have used survey data in combination with market prices to estimate the dispersion of beliefs among investors. Our approach is distinctive in that it uses high-frequency, granular market data (the limit order book) to extract a belief distribution in a

context where no direct survey or multiple-outcome data is available. It contributes to the literature by demonstrating a method to extract additional information from thin markets beyond the midpoint price alone. In particular, the framework attempts to recover features such as belief dispersion, directional skew, and the degree of consensus among participants. Such information may be useful for market organizers or observers seeking to monitor liquidity conditions, detect periods of polarization or uncertainty, or better understand how information is incorporated into prices.

4.2.5 Order Book Visibility and Examples

The Polymarket order book is public and updates in real time. For our analysis, we extract snapshots of the order book at selected time points. Each snapshot contains a static list of outstanding limit orders. We treat these observed orders as the result of a filtering process, where only a subset of trader beliefs are revealed. These filtered orders form the basis for our inference of the underlying belief distribution.

Figure 4.1 illustrates a stylized sparse prediction market order book. In thin markets, bids and asks may cluster away from the midpoint, leaving a relatively empty trading region near the current consensus price. This motivates the idea that visible orders represent only a filtered subset of underlying trader beliefs.

Consider an example market: *"Will Candidate X win the 2024 election?"* Suppose at a certain snapshot on a date, the best bid for Yes is \$0.47 and the best ask is \$0.53, yielding a midpoint price $p = 0.50$. The order book might look like (Yes side perspective):

Bids	**Price**	**Volume**
Bid	\$0.47	1000
Bid	\$0.45	500
Bid	\$0.40	800
Bid	\$0.30	1000
(etc.)		

Asks	**Price**	**Volume**
Ask	\$0.53	1000
Ask	\$0.55	700
Ask	\$0.60	500
Ask	\$0.70	300
...	...	...

Table 4.1: Example order book snapshot (Yes shares) for a binary election market.

4.3 ESTIMATION PROCEDURE

Our objective is to recover the latent distribution of trader beliefs that gives rise to the observed order book in a binary prediction market. We assume that this latent

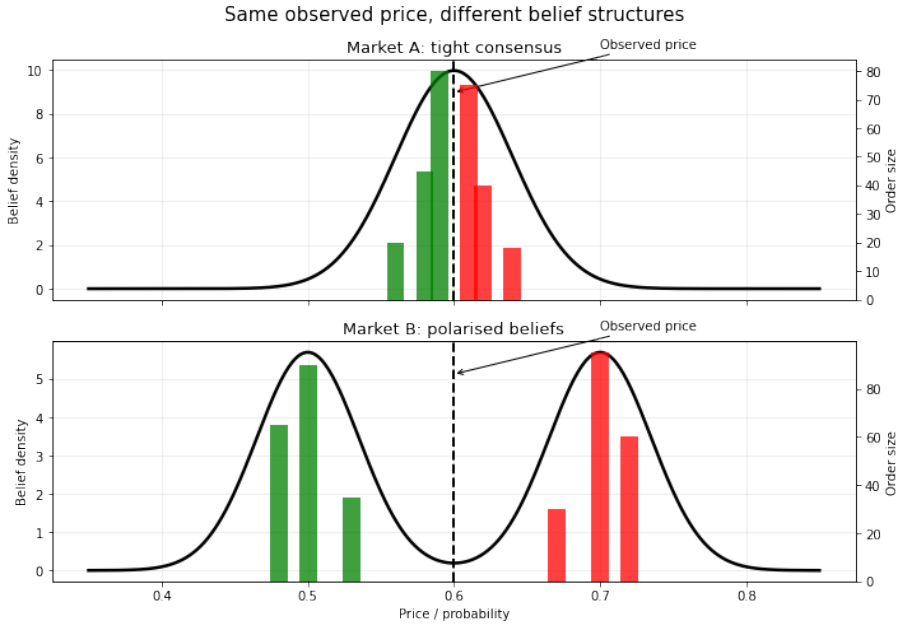


Figure 4.1: Example of a sparse binary prediction market order book with an empty region near the midpoint price. Visible orders tend to cluster away from the consensus region, motivating the filtering framework developed in this chapter.

distribution takes the form of a Beta distribution, denoted $\pi_{\text{latent}}(p)$ over the unit interval:

$$\pi_{\text{latent}}(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1},$$

where $B(\alpha, \beta)$ is the Beta function ensuring proper normalization. However, not all latent beliefs are expressed in the market due to strategic behavior, frictions, or execution priority. To account for this, we introduce a rejection mechanism that filters the latent distribution, yielding a filtered or observed distribution $\pi_{\text{obs}}(p)$ that aligns with the volume-weighted prices visible in the limit order book.

To model this filtering, we define a rejection function $f(p)$, which attenuates the probability mass near active trading regions. The rejection function takes the form:

$$f(p) = f_{\max} - p^{\alpha_r - 1}(1-p)^{\beta_r - 1},$$

where (α_r, β_r) parameterize the rejection intensity and skew, and $f_{\max}$ is the maximum of the unnormalized Beta kernel:

$$f_{\max} = \max_{q \in [0,1]} q^{\alpha_r - 1}(1-q)^{\beta_r - 1}.$$

The term $f_{\max}$ ensures that the rejection function remains non-negative over the unit interval. Economically, it represents the maximum filtering intensity around

the active trading region, where beliefs are least likely to remain visible as standing orders due to immediate execution or limited trading incentives.

This form ensures that $f(p)$ is strictly positive and highest away from the trading region, thereby modeling the suppression of beliefs that would otherwise lead to immediate execution. The filtered belief distribution is defined as the product of the latent belief and the rejection function:

$$\pi_{\text{obs}}(p) = \frac{1}{Z}\,\pi_{\text{latent}}(p) \cdot f(p),$$

where Z is the normalizing constant given by:

$$Z = \int_0^1 \pi_{\text{latent}}(p) \cdot f(p)\,dp = f_{\max} - \frac{B(\alpha + \alpha_r - 1, \beta + \beta_r - 1)}{B(\alpha, \beta)}.$$

This closed-form expression avoids the need for numerical integration and enables efficient likelihood evaluation.

Figure 4.2 illustrates the rejection-based filtering mechanism. Beliefs near the midpoint are less likely to appear as visible standing orders, while more extreme beliefs survive the filtering process and become observable in the order book.

4.3.1 Likelihood Construction

Given a set of N observed price–volume pairs $\{(p_i, V_i)\}_{i=1}^{N}$ extracted from snapshots of the order book, we construct a log-likelihood function based on the filtered belief model. The likelihood weights each observed price by its corresponding traded volume:

$$\log L(\alpha, \beta, \alpha_r, \beta_r) = -\log Z \cdot \left(\sum_{i=1}^{N} V_i\right) + \sum_{i=1}^{N} V_i \left[\log \pi_{\text{latent}}(p_i) + \log f(p_i)\right].$$

This formulation emphasizes price points with higher volume, ensuring that more active trading levels exert greater influence on parameter estimation.

4.3.2 Parameter Initialization and Optimization

All four parameters $(\alpha, \beta, \alpha_r, \beta_r)$ are estimated jointly via maximum likelihood. To improve convergence, initial values for the latent parameters (α, β) are derived from the empirical price-volume histogram using the method of moments. Specifically, letting $x_i = p_i$ and $w_i = V_i / \sum_j V_j$ denote volume-normalized prices, we compute:

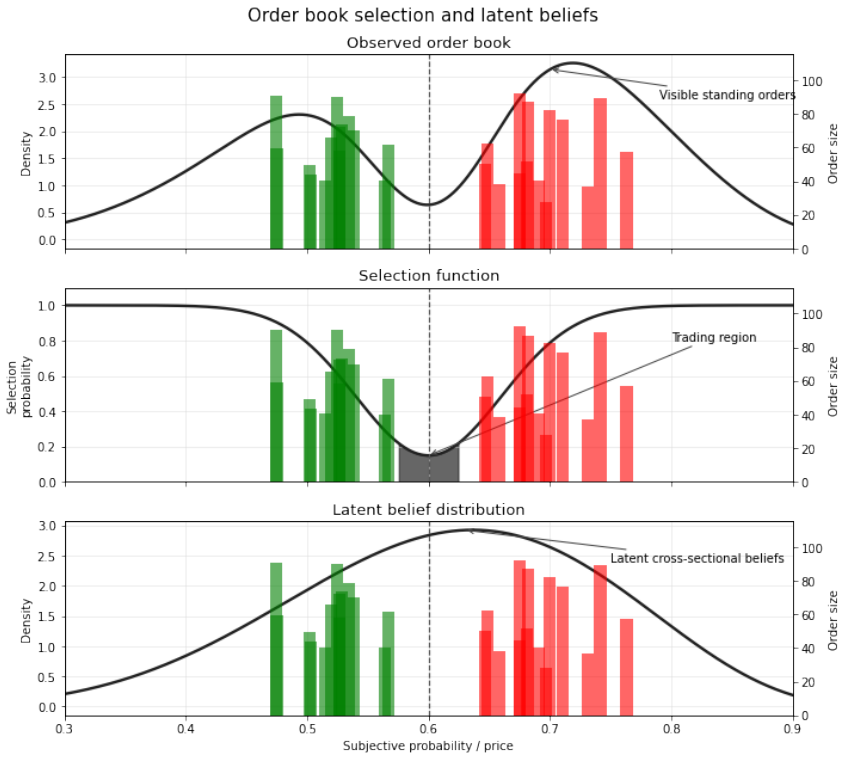


Figure 4.2: Illustration of the filtering mechanism used in the model. Beliefs near the midpoint are less likely to appear as visible orders, while more extreme beliefs survive the filtering process and become observable in the order book.

$$\mu = \sum_{i=1}^{N} w_i x_i, \quad \text{Var} = \sum_{i=1}^{N} w_i (x_i - \mu)^2,$$

where N is the number of observed price levels in the snapshot, $x_i = p_i$ denotes the observed price at level i, and $w_i = V_i / \sum_{j=1}^{N} V_j$ is the corresponding normalized volume weight.

Using the standard method-of-moments parameterization for the Beta distribution, we define the concentration parameter

$$\text{mass} = \frac{\mu(1-\mu)}{\text{Var}} - 1,$$

and initialize the latent Beta parameters via

$$\alpha = \mu \cdot \text{mass}, \qquad \beta = (1-\mu) \cdot \text{mass}.$$

Initial values for the rejection parameters (α_r, β_r) may be seeded to match (α, β) or inferred from the apparent mode of trading activity. All parameters are

bounded below by a small constant (e.g., 1.01) to ensure valid Beta densities and avoid singularities in the log-likelihood. Optimization is performed using bounded numerical routines such as `L-BFGS-B`. Gradient expressions are optionally computed using the identity:

$$\frac{\partial}{\partial x} \log B(x, y) = \psi(x) - \psi(x + y),$$

where $\psi(\cdot)$ is the digamma function.

4.3.3 Structural Mode Constraint

Rather than treating the alignment of the rejection kernel's mode with the trading midpoint as a regularization term, we impose it as a hard constraint. Specifically, we require that the mode of the rejection distribution coincides with the observed trading midpoint p_{trade}. The mode of a Beta distribution with shape parameters $\alpha_r > 1$ and $\beta_r > 1$ is given by:

$$\text{mode}_r = \frac{\alpha_r - 1}{\alpha_r + \beta_r - 2}.$$

Solving this equation for β_r in terms of α_r and p_{trade}, we obtain:

$$\beta_r = \frac{\alpha_r(1 - p_{\text{trade}}) + 2p_{\text{trade}} - 1}{p_{\text{trade}}}.$$

We treat p_{trade} as observed, typically equal to the midpoint price between the best bid and ask in a given order book snapshot. As a result, the rejection kernel becomes a one-parameter Beta distribution indexed by α_r, with β_r determined via this identity. This reparameterization reduces the dimensionality of the optimization problem and eliminates instability due to misalignment of the rejection filter.

4.3.4 Revised Parameter Estimation

With the rejection kernel structurally aligned to the observed trading midpoint, the estimation problem reduces to three free parameters: $(\alpha, \beta, \alpha_r)$. The fourth shape parameter, β_r, is determined analytically by enforcing that the mode of the rejection distribution equals the trading midpoint p_{trade}. Specifically, we impose the identity:

$$\frac{\alpha_r - 1}{\alpha_r + \beta_r - 2} = p_{\text{trade}} \quad \Rightarrow \quad \beta_r = \frac{\alpha_r(1 - p_{\text{trade}}) + 2p_{\text{trade}} - 1}{p_{\text{trade}}}.$$

Given a candidate triplet $(\alpha, \beta, \alpha_r)$, we compute the latent belief density $\pi_{\text{latent}}(p)$ as a Beta distribution, and construct the rejection function as:

$$f(p) = f_{\max} - p^{\alpha_r - 1}(1-p)^{\beta_r - 1},$$

where $f_{\max}$ is the peak value of the unnormalized rejection kernel. The filtered distribution is then formed as the product $\pi_{\text{latent}}(p) \cdot f(p)$, and normalized by a closed-form constant:

$$Z = f_{\max} - \frac{B(\alpha + \alpha_r - 1, \beta + \beta_r - 1)}{B(\alpha, \beta)}, \quad \text{so that} \quad \pi_{\text{obs}}(p) = \frac{1}{Z}\pi_{\text{latent}}(p) f(p).$$

We fit the parameters by maximizing the volume-weighted log-likelihood:

$$\log L(\alpha, \beta, \alpha_r) = -\log Z \cdot \left(\sum_{i=1}^{N} V_i\right) + \sum_{i=1}^{N} V_i \left[\log \pi_{\text{latent}}(p_i) + \log f(p_i)\right],$$

where (p_i, V_i) are the observed price-volume pairs from the limit order book.

To improve numerical stability and avoid pathological fits, we incorporate two soft regularization terms into the objective function. The first discourages excessively concentrated rejection kernels. In practice, very large values of (α_r, β_r) can produce near-degenerate filtering regions that overfit small fluctuations in sparse order books. We therefore penalize the overall concentration of the rejection distribution:

$$\text{Sharpness Penalty} = \lambda_1(\alpha_r + \beta_r).$$

The second regularization term is intended to discourage the rejection distribution from collapsing toward the latent belief distribution itself. Without this constraint, the filtering mechanism may absorb structure intended to be captured by the latent belief model, reducing interpretability. We therefore penalize excessive similarity between the two parameter vectors using the reciprocal form

$$\text{Divergence Penalty} = \lambda_2 \left[(\alpha_r - \alpha)^2 + (\beta_r - \beta)^2\right]^{-1}.$$

The total objective function to be minimized is therefore

$$\mathcal{L}_{\text{total}}(\alpha, \beta, \alpha_r) = -\log L(\alpha, \beta, \alpha_r) + \lambda_1(\alpha_r + \beta_r) + \lambda_2 \left[(\alpha_r - \alpha)^2 + (\beta_r - \beta)^2\right]^{-1}.$$

In practice, we use small values of λ_1 and λ_2 (e.g., 10^{-3} and 5×10^{-4}, respectively) so that the likelihood remains the dominant component of the objective while still preventing unstable parameter estimates. Optimization is performed using a constrained numerical routine such as `L-BFGS-B`, with all parameters bounded below by a small positive constant (e.g., 1.01) to ensure valid Beta shapes and numerical stability. More generally, the regularization parameters could be calibrated using out-of-sample likelihood performance, cross-validation across snapshots, or other stability-based model selection procedures.

This reparameterized and regularized estimation framework yields a filtered belief distribution that is both consistent with observed market structure and robust to overfitting, enabling interpretable inference of latent beliefs from noisy limit order book data.

4.3.5 Summary

This estimation framework integrates latent belief modeling, rejection filtering, and closed-form normalization into a coherent likelihood-based inference procedure. By jointly estimating both the underlying distribution of trader beliefs and the market-imposed filtering mechanism, we obtain a principled and tractable representation of belief heterogeneity in prediction markets. The approach ensures stability, interpretability, and efficiency, making it suitable for repeated estimation across high-frequency snapshots.

4.4 TEMPORAL DYNAMICS OF BELIEF PROXIES

To move beyond static snapshots, we study how belief estimates evolve over time. For each order book snapshot, we compute three quantities: the mean of the latent belief distribution (μ_{latent}), the mean of the filtered distribution based on visible orders (μ_{filtered}), and the mean after applying the rejection sharpening step ($\mu_{\text{rejection}}$). These are estimated at regular intervals, such as hourly, giving us three time series to analyze across T snapshots.

The latent mean, $\mu_{\text{latent}}(t)$, corresponds to the market midpoint and acts as a rough consensus estimate. The filtered mean, $\mu_{\text{filtered}}(t)$, reflects the volume-weighted average of the standing orders, showing where most of the visible belief mass lies.

The rejection-sharpened mean, $\mu_{\text{rejection}}(t)$, pushes this further by highlighting how strongly directional the visible orders are once we adjust for filtering.

To study how these quantities relate to one another over time, we fit a vector autoregressive (VAR) model:

$$\mathbf{y}_t = A_1\mathbf{y}_{t-1} + A_2\mathbf{y}_{t-2} + \cdots + A_k\mathbf{y}_{t-k} + \varepsilon_t,$$

where

$$\mathbf{y}_t = \begin{bmatrix} \mu_{\text{latent}}(t) \\ \mu_{\text{filtered}}(t) \\ \mu_{\text{rejection}}(t) \end{bmatrix},$$

and each A_i is a 3×3 coefficient matrix. We choose the lag length k using AIC or BIC. This model captures how each component responds to past values of the others.

We use Granger causality tests to explore directional relationships among the three belief proxies, where one series is said to Granger-cause another if its lagged values improve predictive accuracy in a vector autoregressive model. The results support a forward flow of influence: directional pressure in the tails (captured by $\mu_{\text{rejection}}$) helps predict both the filtered distribution and the eventual movement of the market midpoint.

Specifically, we find that $\mu_{\text{rejection}}$ Granger-causes μ_{filtered} at the 10% level, and also Granger-causes μ_{latent} at the 5% level. In turn, μ_{filtered} significantly predicts μ_{latent}, consistent with the idea that accumulated order flow precedes changes in the consensus price. We also find marginal evidence of weak feedback from μ_{filtered} to $\mu_{\text{rejection}}$, but no evidence that the midpoint price Granger-causes either of the other two quantities. A full summary of these pairwise tests is provided in Table 4.2.

Table 4.2: Pairwise Granger causality tests among belief proxies. The null hypothesis in each case is that lagged values of the "Cause" variable do not improve prediction of the "Effect" variable within the VAR framework.

Cause	**Effect**	**F-statistic**	**p-value**
μ_{latent}	μ_{filtered}	0.4375	0.9112
$\mu_{\text{rejection}}$	μ_{filtered}	1.9510	0.0550
μ_{filtered}	μ_{latent}	2.3183	0.0217
$\mu_{\text{rejection}}$	μ_{latent}	2.2938	0.0232
μ_{filtered}	$\mu_{\text{rejection}}$	1.6058	0.1261
μ_{latent}	$\mu_{\text{rejection}}$	0.4361	0.9120

Taken together, these results support a directional structure where belief pressure accumulates in the tails, influences the visible distribution of orders, and ultimately drives price movements. This is consistent with strategic behavior by informed traders who place extreme orders to express directional views. These orders shift the balance of visible volume, which in turn creates pressure on the midpoint to adjust.

As an example, suppose that at time $t = 0$, a wave of bids appears at prices such as \$0.30 and \$0.35. This lowers $\mu_{\text{rejection}}$, indicating strong bearish sentiment among committed traders. Over time, these orders dominate the standing book, pulling down μ_{filtered}. The midpoint then follows, adjusting downward as the imbalance persists and shapes expectations.

Our VAR and Granger analysis suggests that price formation in prediction markets is influenced by directional belief pressure embedded in the tails of the order book. Changes in the filtered and rejection-sharpened means often precede shifts in the market midpoint, providing early signals of evolving sentiment. This underscores the importance of incorporating full order book data when studying belief dynamics in real time.

4.5 CASE STUDY: KNICKS VS. CELTICS (MAY 7, 2025)

The empirical analysis focuses on Game 2 of the 2025 Eastern Conference Semifinals between the New York Knicks and the Boston Celtics, held on May 7, 2025, at TD Garden. This series received substantial media attention prior to Game 2 following the Knicks' dramatic overtime comeback victory in Game 1, where New York erased a 20-point deficit and stole home-court advantage from Boston [65]. Contemporary media coverage emphasized both the Celtics' status as strong favourites and the increased uncertainty introduced by the upset.

Despite the upset in Game 1, pre-game markets strongly favored Boston. In European-style fixed-odds markets, a quoted price of o corresponds to an implied probability $p = 1/o$. The Celtics were priced at $o = 1.18$, implying a win probability of approximately $p \approx 0.85$, while the Knicks were priced at $o = 5.00$, implying $p = 0.20$. These prices reflected both Boston's dominant regular-season performance and widespread confidence in their ability to recover when playing at home. Expectations about the margin of victory were similarly lopsided: the point-spread market opened

at Boston -10.5, indicating that the Celtics were expected to win by roughly ten points.

Polymarket order book data reflected this pre-game sentiment. Early snapshots showed significant ask-side pressure and strong skew in the distribution of standing limit orders, with much of the volume concentrated at high price levels that corresponded to Boston-favored outcomes. The market microstructure exhibited substantial ask-side imbalance and positive skew in the distribution of visible order volume, reinforcing the prevailing public narrative.

Data for this analysis were collected in real time using Polymarket's GraphQL API. A custom pipeline was developed for this project to capture raw order book snapshots at minute-level resolution during live play. This enabled a temporally precise reconstruction of belief dynamics throughout the game, offering greater granularity than standard post hoc data summaries.

4.5.1 Latent vs. Observable Belief Structures

A core objective of this chapter is to distinguish between the observed signals in the order book and the latent belief distribution that drives them. In prediction markets with limit order books, the beliefs that influence market behavior are only partially visible. First, only standing limit orders are observable at a given time, omitting the flow of transient or executed orders. Second, among the visible orders, only those placed far enough from the midpoint are likely to remain unfilled, introducing a form of selection bias toward more extreme beliefs.

To study this structure, we work with three belief proxies per snapshot. The first, μ_{latent}, corresponds to the midpoint price and serves as a coarse indicator of market consensus. The second, μ_{filtered}, is computed from the volume-weighted mean of visible orders and reflects the concentration of belief among passive traders. The third, $\mu_{\text{rejection}}$, sharpens this estimate by adjusting for asymmetries and sparsity in the book, yielding a posterior that approximates the latent belief distribution under a rejection-filtered model.

Together, these quantities allow us to decompose the belief formation process and distinguish between what is visible in the market and what is likely driving order placement beneath the surface.

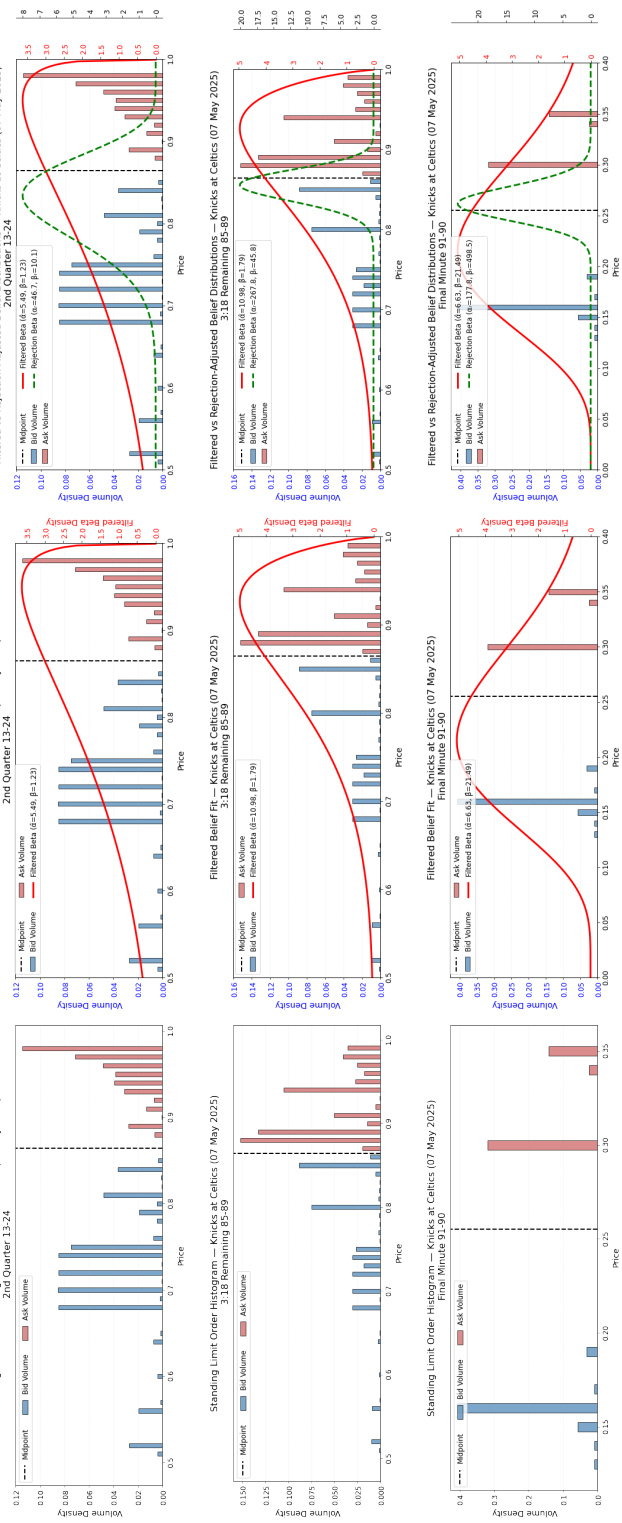


Figure 4.3: **Full evolution of inferred belief distributions during the game.** Rows correspond to early game, closing time, and the final minute. Columns show successive stages of the inference procedure: the observed order book histogram, the belief distribution implied by visible orders, and the inferred latent belief distribution after correcting for selection effects.

This decomposition allows us to distinguish between what is directly visible in the order book, what is structurally filtered by the market mechanism, and what must be inferred as the latent belief distribution underlying observed behavior. We apply this framework to 100 snapshots collected between 8:00 PM and 10:30 PM ET during Game 2 of the Eastern Conference Semifinals, yielding a time series of fitted belief distributions that evolve alongside the game state and market sentiment. Figure 4.3 presents a 3×3 panel of representative snapshots that illustrate this progression.

In this layout, each row corresponds to a different phase of the game—early play, the closing stretch, and the final minute—while each column represents a progressively deeper layer of inference: beginning with the raw histogram of visible order volume, moving to the fitted Beta distribution (with mean μ_{filtered}), and finally to the rejection-adjusted belief mean ($\mu_{\text{rejection}}$).

The left column displays the full shape of the order book, capturing asymmetries, sparsity, and concentration of volume in the tails. The middle column smooths this structure into a fitted Beta distribution, offering a cleaner summary of belief mass. The right column adjusts for spread and imbalance, sharpening the estimate to better approximate the latent belief distribution implied by trader behavior under market frictions.

This visual sequence is qualitatively consistent with the layered structure of the inference pipeline. No formal goodness-of-fit statistic is claimed here; the purpose of the figure is primarily illustrative, showing how the inferred distributions evolve across successive stages of the filtering framework. As the game progresses and information asymmetry narrows, belief distributions become more concentrated and aligned. In early phases, belief mass is dispersed and skewed, reflecting uncertainty and divergence in expectations. By the final minute, beliefs converge sharply around the true outcome, with the rejection-adjusted proxy leading this tightening.

Overall, this triadic structure provides a dynamic lens on belief formation. Each layer of inference reveals a distinct aspect of market behavior, from liquidity provision to probabilistic expectations to directional conviction, allowing us to disentangle the surface-level structure of the book from the underlying belief dynamics driving it.

4.5.2 Static Snapshot Analysis

We highlight three representative snapshots from the game to illustrate how inferred beliefs evolve in response to game state, market structure, and emerging momentum

shifts. Each snapshot corresponds to a distinct narrative moment, with visibly different belief dynamics.

Early Second Quarter (20:03 ET) At this point, the Celtics held a commanding 24–13 lead after a dominant first quarter. The order book was heavily skewed toward high prices, with substantial volume resting on the ask side. This structure reflects strong confidence in a Boston win, both from public expectation and early game flow. The filtered Beta distribution was fitted with parameters $(\hat{\alpha}, \hat{\beta}) = (5.49,\ 1.23)$, corresponding to a filtered mean near $\mu_{\text{filtered}} \approx 0.82$. After rejection sharpening, which incorporated the narrow spread and ask-side imbalance, the adjusted belief distribution tightened considerably, with fitted parameters $(\alpha_r, \beta_r) = (46.7,\ 10.1)$ and $\mu_{\text{rejection}} \approx 0.82$. The resulting posterior was sharply right-skewed, reflecting concentrated belief in a near-certain Boston outcome. No signs of contrarian positioning were yet visible.

Mid-Fourth Quarter (21:14 ET) By the mid-fourth quarter, the Knicks had narrowed the gap to a few possessions, and market behavior began to shift. The order book showed a more balanced structure, with visible depth on both sides and less skew toward Boston. The filtered belief distribution widened, with fitted parameters $(\hat{\alpha}, \hat{\beta}) = (10.98,\ 1.79)$ and $\mu_{\text{filtered}} \approx 0.86$, still favoring Boston but with less certainty. After applying the rejection filter, the adjusted parameters $(\alpha_r, \beta_r) = (267.8,\ 45.8)$ yielded $\mu_{\text{rejection}} \approx 0.85$, slightly lower and less sharp than earlier. This moment reflects a turning point in sentiment: while prices continued to favor Boston, the distribution shows early adjustment among traders who began anticipating a potential shift in momentum before it became visible on the scoreboard.

Final Minute (21:23 ET) With under a minute remaining, the Knicks had completed a dramatic comeback and taken a 91–90 lead. The order book remained active, with residual ask-side volume suggesting continued, albeit diminished, belief in a Boston recovery. However, the belief structure had changed fundamentally. The filtered Beta distribution was now $(\hat{\alpha}, \hat{\beta}) = (6.63,\ 21.49)$, giving $\mu_{\text{filtered}} \approx 0.24$, and the rejection-adjusted fit was $(\alpha_r, \beta_r) = (177.8,\ 498.5)$, yielding $\mu_{\text{rejection}} \approx 0.26$. Belief mass had decisively shifted toward a Knicks victory. The narrowing and convergence of these estimates illustrate the market's ability to rapidly incorporate both score and clock-based information. Unlike earlier moments, there was little

skew or dispersion: belief was now concentrated around the likely outcome, with limited remaining asymmetry or disagreement. This snapshot captures how inference from book structure can track sentiment as it crystalizes near resolution.

4.5.3 Dynamic Evolution of Beliefs

We now examine how belief estimates evolve throughout the game. Figure 4.4 shows the smoothed trajectories of μ_{latent}, μ_{filtered}, and $\mu_{\text{rejection}}$ across 100 snapshots from Game 2, plotted alongside the score difference. These dynamics reveal how market belief adapts to in-game developments.

Several patterns stand out:

- **Early divergence (Q1):** In the opening phase, μ_{filtered} and $\mu_{\text{rejection}}$ are closely aligned and remain elevated, reflecting confident belief in a Boston win. In contrast, μ_{latent} remains noticeably lower, lagging behind both the other proxies and the score difference trend.
- **Coupling under stability (Q2):** As the game stabilizes during the second quarter, all three measures begin to converge. μ_{latent} catches up to the higher level of the other estimates, and belief proxies move together in line with Boston maintaining a lead.
- **Breakdown and divergence (Q3):** During the third quarter, μ_{latent} drops sharply, decoupling from the smoother and relatively flat paths of μ_{filtered} and $\mu_{\text{rejection}}$. This suggests greater sensitivity of the latent measure to shifts in book structure, and better alignment with changes in score differential.
- **Sharp alignment at resolution (Q4):** In the final moments, all three belief measures decline rapidly as the score tightens, eventually converging near 0.5. Notably, μ_{latent} closely tracks the falling score difference throughout the fourth quarter, while the filtered and rejection-adjusted estimates respond more gradually. This suggests that the latent proxy, despite its volatility, may capture real-time sentiment shifts more responsively than the smoothed belief estimates.

These dynamics illustrate belief–price dislocations, where informed sentiment shifts ahead of the midpoint. This anticipatory behaviour is especially clear in the fourth quarter, where $\mu_{\text{rejection}}$ declines before the Knicks take the lead suggesting that some traders act on faster-updating or privately interpreted signals (e.g., possession

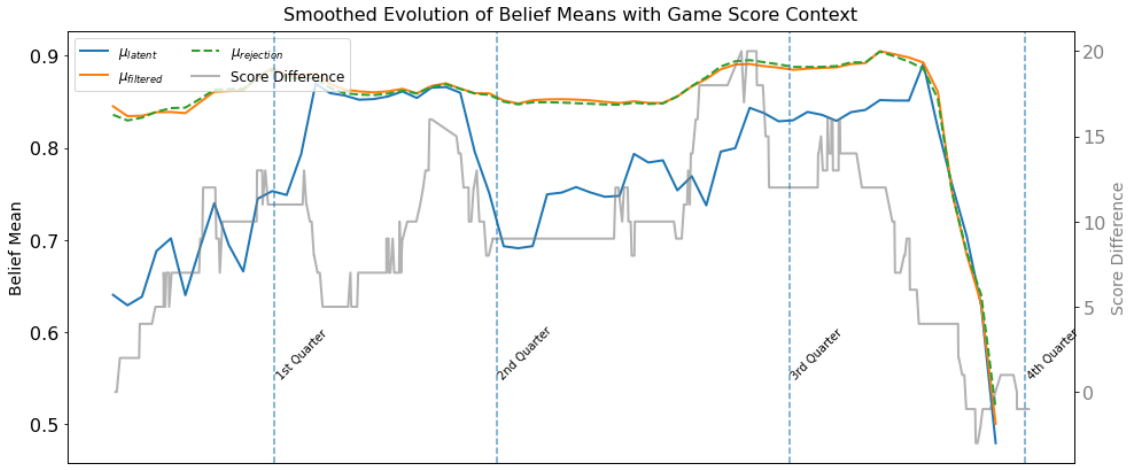


Figure 4.4: **Smoothed evolution of belief means inferred from Polymarket order book snapshots**. Belief estimates shown are the latent mean (μ_{latent}), filtered mean (μ_{filtered}), and rejection-sharpened mean ($\mu_{\text{rejection}}$), overlaid with the score difference for Game 2.

quality or momentum shifts). The eventual convergence of all belief measures near resolution reflects a collapse of information asymmetry as the outcome becomes clear. Together, this layered decomposition provides a more nuanced view of belief formation under uncertainty than any single price-based measure.

4.5.4 Spread and Imbalance Effects on Inferred Beliefs

We analyze how two microstructure features—the bid-ask spread and order book imbalance—affect the shape of the inferred belief distribution, as captured by the parameters of the rejection-adjusted Beta fit.

We focus on two structural properties:

- **Concentration:** $\alpha_r + \beta_r$, reflecting how tightly beliefs are clustered.
- **Skew:** $\alpha_r - \beta_r$, reflecting directional bias toward 0 or 1.

Spread and Belief Concentration Figure 4.5a shows a negative relationship between bid-ask spread and belief concentration. Wider spreads are associated with flatter belief distributions. This pattern is consistent with reduced liquidity: when trading is costly, traders are less willing to reveal precise beliefs, resulting in more diffuse positioning.

Imbalance and Belief Skew Figure 4.5b shows that order book imbalance is associated with systematic shifts in belief skew. When bid volume dominates, inferred beliefs skew toward 1, and when ask volume dominates, they skew toward 0. This suggests that imbalance encodes directional sentiment not captured by midpoint prices.

Imbalance is defined as:

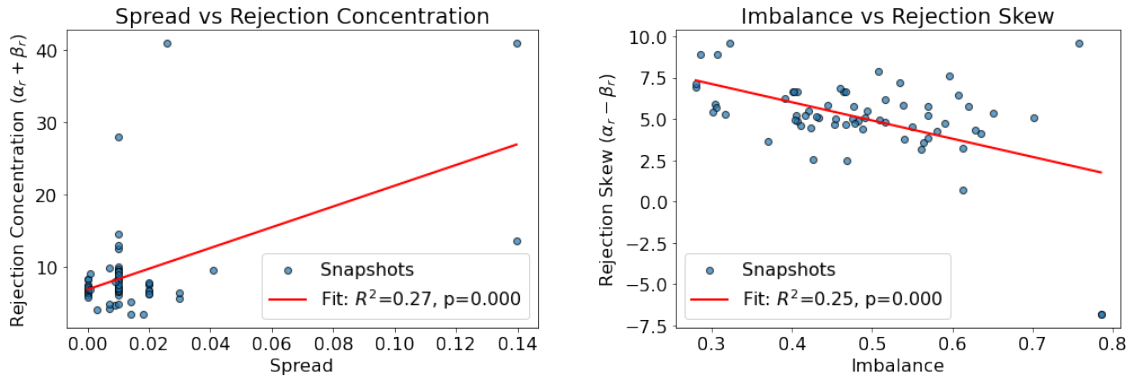
$$\text{Imbalance} = \frac{\text{Bid Volume} - \text{Ask Volume}}{\text{Bid Volume} + \text{Ask Volume}},$$

ranging from –1 (ask dominance) to +1 (bid dominance), with 0 indicating balance.

Implications These results support a key assumption of the rejection-based inference framework: that observed orders are strategically filtered expressions of belief. Specifically:

- **Wider spreads** are linked to lower concentration, reflecting reduced willingness to express precise beliefs under higher frictions.
- **Directional imbalance** shifts the shape of the belief distribution, revealing asymmetric positioning.

Statistical regressions confirm that both relationships are significant, though not deterministic. Together, these findings validate that belief structures are shaped not only by information, but by the market conditions through which that information is expressed.



(a) Spread vs belief concentration ($\alpha_r + \beta_r$). (b) Imbalance vs belief skew ($\alpha_r - \beta_r$).

Figure 4.5: **Microstructure effects on the shape of inferred belief distributions**. Wider bid–ask spreads reduce concentration, while order imbalance induces skewness in the inferred belief distribution.

4.6 DISCUSSION AND CONCLUSION

This chapter presented a structural method for inferring belief distributions from visible order book data in binary prediction markets. Using a Beta distribution framework and incorporating information about spread and order book asymmetry, we proposed a three-step estimation procedure to move from observed trading data to latent market sentiment. The approach was applied to real-time Polymarket data during a live NBA playoff game to illustrate how beliefs evolve over time in response to game events and market structure.

The method proceeds by first fitting a Beta distribution to the volume-weighted histogram of visible standing orders, providing a filtered approximation of belief away from the midpoint. A rejection mechanism, incorporating the bid-ask spread and volume imbalance, is then used to model strategic behaviour in order placement. Finally, inverse simulation is used to recover the latent belief distribution consistent with this observed structure under the assumed filtering constraints.

The case study results show that belief concentration tends to increase as resolution approaches, consistent with greater consensus as more information becomes available. Additionally, divergence among belief proxies—particularly during key game transitions—suggests that latent sentiment may shift ahead of price-based signals. This is most apparent in the fourth quarter of the case study, where the rejection-sharpened belief begins declining before the market midpoint adjusts, implying some degree of anticipatory behaviour among participants.

These dynamics also respond to features of the order book microstructure. Wider spreads are associated with lower belief concentration, and volume imbalance corresponds to directional skew. These patterns support the view that standing orders reflect constrained expressions of belief, shaped by risk, cost, and strategic intent.

While the model remains a simplification and is reliant on structural assumptions, it provides a transparent and tractable way to extract belief information from market data without requiring access to individual-level trading histories. The approach could be adapted to other binary prediction markets, and may be of use in applications such as sentiment monitoring, real-time risk aggregation, or behavioural analysis of trading activity.

Overall, the method contributes to a more granular understanding of belief formation in markets where prices alone provide an incomplete signal. Further extensions may consider more flexible distributions, multi-event inference, or real-

time implementation in monitoring systems.

4.6.1 Limitations and Future Directions

This study has several limitations that suggest directions for further work:

- The Beta distribution, while analytically convenient, may be insufficient to capture more complex belief structures, such as multimodality or heavy tails. Nonparametric approaches or mixture models may offer greater flexibility.
- The rejection mechanism is currently specified using fixed parametric forms based on spread and imbalance. Future work could seek to estimate this filtering process from data or allow its parameters to vary over time.
- Granger causality analysis identifies temporal precedence but does not establish structural causation. Further investigation would require more rigorous causal inference methods, potentially involving exogenous variation or structural models.
- The framework abstracts from the role of market makers, automated agents, and strategic liquidity provision. These actors may influence the order book independently of belief, and their effects warrant more detailed treatment.

Extensions could include hierarchical or panel models for cross-market inference, nonparametric belief recovery techniques, and generalisation to settings with multiple outcomes or continuous state spaces.

— 5 —

Market liquidity and risk aversion among traders inform the connection between beliefs and prices

5.1 INTRODUCTION

Prediction markets provide a natural laboratory for studying how information, risk, and liquidity interact in price formation. Participants trade outcome-contingent contracts whose prices can be interpreted as market-implied probabilities for future events. Unlike forecasting tournaments, in which predictions are scored after resolution, these markets generate a continuous stream of probabilistic forecasts through trading itself: every order to buy or sell represents a marginal adjustment in collective belief [87, 60]. Platforms such as Kalshi, Polymarket, and PredictIt host hundreds of such markets at any time. A typical example of intraday prediction-market activity is shown in Figure 5.1, drawn from the market "Will Bitcoin close above its open price on 22 May 2025?". Over a single trading day, more than 5,000 individual transactions were executed on the "Yes" side alone, representing nearly one million USDC in matched volume and over thirty million shares traded. From the market's opening the day prior until resolution at midnight, traders buy and sell "Yes" contracts priced between \$0 and \$1, with each price corresponding to the implied probability of Bitcoin closing above its opening level. A trader who buys at \$0.40 receives \$1.00 if the event occurs and nothing otherwise. For such a trade to be favourable on average, the trader must believe that the event is more likely than its price implies, meaning that the event probability exceeds 40%. As orders arrive, prices shift by

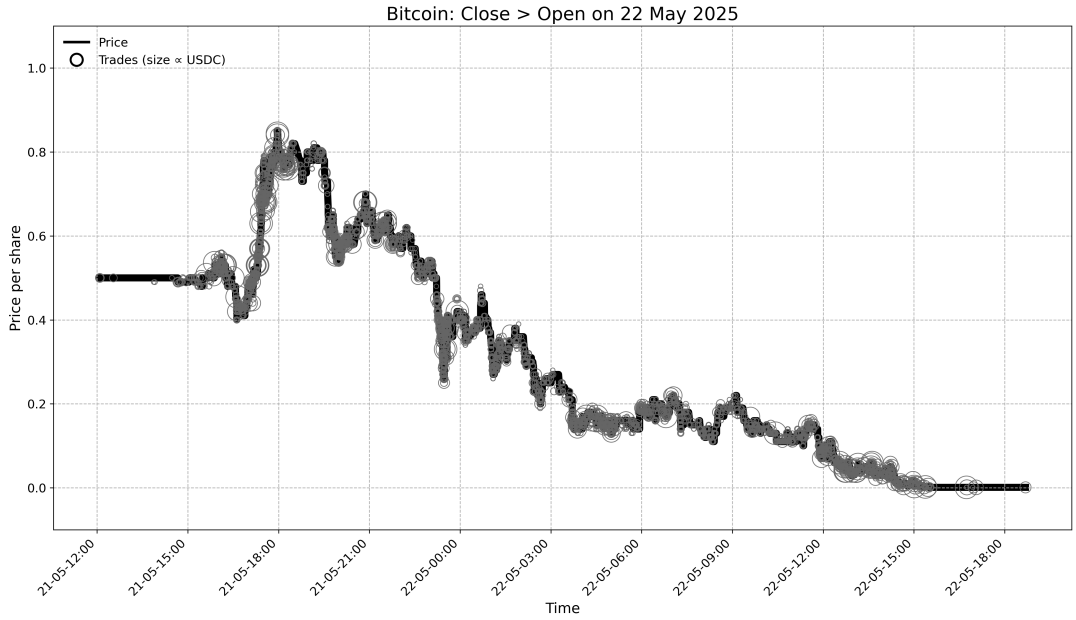


Figure 5.1: **Intraday trading dynamics for a Bitcoin prediction market**. Each hollow circle represents an individual trade, with marker size proportional to USDC volume, while the line traces the evolving market price throughout the trading day.

amounts that depend on how much liquidity the market provides. When depth is high, large trades move prices little; when liquidity thins, even modest trades can generate sharp jumps. Liquidity therefore measures the conditional responsiveness of price to trading activity and serves as a central object of inference in market microstructure analysis.

In traditional financial venues such as the NYSE, NASDAQ, or LSE, liquidity arises endogenously from a dense population of buyers and sellers operating under a continuous double auction (CDA). In a CDA, traders post limit and market orders to a central book, and prices evolve as opposing orders meet [77, 63]. The mechanism depends on large numbers of active participants and professional liquidity providers, firms such as Citadel Securities or Jane Street, that continuously quote two-sided prices and absorb temporary order-flow imbalances [32, 11]. With billions of shares traded daily on U.S. exchanges [59], this density sustains narrow spreads and efficient price discovery through bilateral competition. In this environment, market impact reflects how information diffuses through the limit-order book as traders reveal private signals or rebalance portfolios.

Prediction markets, by contrast, operate at smaller scale. Most individual markets contain only a few dozen to a few thousand traders and cannot rely on

continual matching of opposing orders. A pure CDA would frequently stall because no counterparty is available at the current price. To guarantee immediacy of trade, the markets themselves therefore adopt an automated market maker (AMM), a mechanism that stands ready to buy or sell at any time, quoting prices according to a deterministic cost function. The most widely used implementation is the logarithmic market scoring rule (LMSR) [43, 1], which sets prices as the gradient of a convex potential function. This design ensures that execution is always possible and that the quoted price changes smoothly with cumulative net order flow. The cost of permanent liquidity, however, is structural: prices in AMM markets are no longer outcomes of strategic interaction but rather are deterministic responses to the mechanism's internal state. Price variation thus reflects how the market maker's balance sheet evolves, not necessarily how information diffuses among traders.

Empirically, this distinction is visible in modern platforms such as Polymarket. From a sample of more than five million trades, we estimate that roughly seventy percent of total volume is transacted directly with the platform's automated market maker, while the remaining thirty percent consists of trades between participants. This hybrid structure means that although peer-to-peer matching occurs, most price formation is governed by the cost-function mechanism rather than bilateral bargaining. In practical terms, Polymarket behaves as a cost-function AMM that occasionally intermediates user-to-user orders; it should therefore be analysed as an automated rather than a double-auction market.

This difference has important empirical implications. Classical microstructure statistics, Kyle's price-impact coefficient λ [53], Hasbrouck's VAR decomposition of information content [45], and Amihud's illiquidity ratio [3], were developed for continuous double-auction markets in which prices emerge from competitive order matching. In an AMM, liquidity and impact are embedded directly in a cost function, so standard measures lose their interpretation. "Impact" no longer signals informational advantage or asymmetric response; it simply reflects how the mechanism updates its internal inventory. Depth, spread, and volatility are determined mechanically rather than strategically. As a result, the familiar language of market microstructure, order-flow autocorrelation, informational efficiency, permanent versus transitory impact, does not translate cleanly to automated markets. Analysing their dynamics therefore requires a framework grounded in the mechanism itself.

An AMM essentially prints contracts that oblige it to pay the bearer one unit of currency if a certain event occurs. A convex function $H(q)$ is used to determine

a notional cost of printing q contracts. For reasons explored in depth by [43], particularly the guaranteed availability of liquidity, bounded worst-case loss for the market maker, and smooth price adjustment in response to order flow, it is useful to consider cost and price functions of the form

$$H(q) = b \log\left(1 + e^{q/b}\right), \qquad H'(q) = \frac{1}{1 + e^{-q/b}}. \tag{5.1}$$

The logarithmic form of $H(q)$ is closely related to the logarithmic scoring rule used in probabilistic forecasting and prediction markets. In particular, the resulting pricing rule inherits several desirable properties discussed by [43], including smooth price adjustment, bounded loss for the market maker, and incentive compatibility for probabilistic forecasting. Markets using the pricing function in (5.1) are therefore commonly referred to as logarithmic market scoring rule (LMSR) markets. For further discussion of alternative scoring rules, readers are referred to [43] and [39]. Note that buying (selling a negative amount of) a 'yes' contract (paying out if an event does occur) at price p is equivalent to selling a 'no' contract (paying out when the event does not occur) at price $1 - p$ since in both cases the AMM's net income is

$$-p + X = (1 - p) - (1 - X), \tag{5.2}$$

where X is a binary indicator function taking value one when the event does occur. Because of this the AMM need not consider the prices of 'yes' and 'no' contracts separately. Note also that the parameter b in (5.1) controls how responsive prices are to order flow: smaller b implying greater sensitivity and larger b producing more stable prices.

A risk-tolerant trader interacting with an AMM will buy contracts until the market price exceeds the price she values them at. A risk-averse trader, however, is liable to stop buying before that happens. The degree to which the trader affects the price is thus determined by her belief about the contract's value, but also by her tolerance for risk and the AMM's liquidity parameter. Precise formulations of this relationship in the context of LMSRs are the subject of [26], who view the market as an opinion-pooling device in which trades aggregate information through liquidity/risk-weighted updates. Related lines of work have extended this view to more complex market environments, including continuous double-auction hybrids [64], generalised cost-function and exponential-family market makers [34], and multi-agent simulations of heterogeneous Bayesian traders interacting with such mechanisms

[56]. Collectively, these studies highlight how design parameters shape information aggregation, liquidity and trader compensation in thin or algorithmic markets.

In the work below we build on the framework of [26], fully accounting for the fact that traders' beliefs are mediated by a combination of liquidity and risk-aversion. By assuming certain forms for the market maker's pricing function and the traders' utility functions, we can formulate the mediation in a precise way and reason precisely about the effects it would have on observable price dynamics. Our contribution is to use these ideas in the reverse direction, estimating the degree of mediation directly from observed price dynamics. We then, empirically, relate the degree of mediation to the accuracy of the price as a forecast of the underlying event, finding that a moderate level of mediation is a good thing.

The remainder of the paper proceeds as follows. Section 5.2 reviews classical measures of market microstructure and explains why they are well suited to continuous double-auction settings but less appropriate for automated market makers. Section 5.3 introduces the theoretical framework linking the LMSR liquidity parameter b to traders' CARA risk tolerance τ, deriving the local micro-adjustment equation that connects log-odds movements to belief–price discrepancies. Section 5.4 describes how this relationship can be estimated empirically from order-book and trade-level data, defining the composite parameter $\kappa = \tau/b$ that governs logit price dynamics. Section 5.5 presents empirical results across a sample of prediction markets, documenting the distribution and short-run behaviour of κ and its variation across market categories. Section 5.6 concludes with implications for market design and interpretation of heterogeneity in automated trading environments, and outlines directions for further work.

5.2 CLASSICAL AND MECHANISM-BASED MICROSTRUCTURE

This section reviews established measures of market microstructure and explains why they are suited to continuous double-auction (CDA) settings but not to automated market makers (AMMs).

5.2.1 Classical Measures

Liquidity has long been quantified through reduced-form measures that describe how order flow translates into price movement. Among the earliest and most influential is Kyle's λ [53], which formalizes the link between trade size and price impact. In Kyle's model a trader's order for Δq_t contracts moves the price by Δp_t according to

$$\Delta p_t = \lambda \Delta q_t, \tag{5.3}$$

where the changes

$$\Delta p_t = p_{t+1} - p_t, \qquad \Delta q_t = q_{t+1} - q_t \tag{5.4}$$

are measured over a suitable period of time. Kyle's λ quantifies rate of change of price with respect to trade volume, the reciprocal being understood as the demand at each price or market-depth. A smaller λ indicates that prices absorb trades easily (high liquidity and market-depth), whereas a larger λ implies that even small orders move prices sharply (low liquidity and market-depth). Kyle's impact coefficient, which remains the canonical benchmark for translating information flow into price adjustment, is often estimated by regressing price changes on trading volumes. Originally introduced in the context of insider trading and informational asymmetry, Kyle's λ is widely interpreted as a measure of price sensitivity to order flow. In our setting it serves primarily as a benchmark liquidity statistic against which mechanism-based AMM measures can be contrasted.

Amihud's illiquidity ratio [3] quantifies illiquidity in terms of daily returns R_t, which helps reduce price-scale effects, and daily trade volumes V_t,

$$A = \frac{1}{T} \sum_{t=1}^{T} \frac{|R_t|}{V_t}. \tag{5.5}$$

where

$$R_t = \frac{p_{t+1} - p_t}{p_t}, \qquad V_t = \sum_{h=0}^{H} |q_{t,h+1} - q_{t,h}|, \tag{5.6}$$

and the sum over h in the definition of V_t adds together all buy and sell trades over the course of a day. High values signal that modest trading volume produces large price movements, indicating illiquidity. Its simplicity and compatibility with widely available returns and volume data have made it popular, though it exhibits high variance when either price or volume are very small.

Hasbrouck [45] takes a modelling approach to liquidity estimation, using a vector-autoregression (VAR) model to capture the dependence of price/volume pairs on lagged observations. He uses the substantial theory for VAR models to quantify the expected effects, both instantaneous and delayed, on prices from random shocks and recent trades. Our own work also makes use of autoregressive models to formalize the connection between price and trade volume. Unlike the VAR framework of [45], which models several jointly evolving quantities simultaneously, we focus on a reduced-form univariate specification tied to a single mechanism-based quantity, allowing for simpler interpretation and estimation in the AMM setting. Additionally, we restrict attention to a single scalar quantity on which the connection hinges, allowing us to more easily measure and plot this quantity.

A further contribution comes from Roll's spread estimator [70], which infers transaction costs from the negative serial correlation of price changes. In efficient markets with bid–ask bounce, Roll argues that consecutive returns tend to alternate in sign, leading him to the expression

$$S \approx 2\sqrt{-\mathrm{Cov}(\Delta p_t, \Delta p_{t-1})}, \tag{5.7}$$

where S is an implicit, or effective, bid–ask spread whose size is understood to capture market illiquidity. This covariance-based estimator is attractive because it uses only price data, although it does assumes that other forms of serial dependence forms are negligible.

Taken together, these measures capture complementary aspects of liquidity, price impact, trading cost, and depth, yet all implicitly rely on the institutional setting of the continuous double auction. They are typically descriptive rather than structural: measuring correlations between flow and price, with only superficial linkage to the processes by which market makers adjust prices and traders specify volumes. This limitation motivates our shift to a mechanism-based framework, in which liquidity and trader behaviour are linked directly through model parameters that can be compared across markets.

5.2.2 Liquidity Estimation in the context of Prediction Markets

The use of traditional liquidity measures in the context of prediction market data, rather than the conventional financial markets for which they were designed, raises several questions and challenges. Prices on the prediction market are bounded

between zero and one, trades often occur with an automated market maker rather than another participant, and observed high-frequency price variation is liable to be affected by a trading platform's internal mechanics.

Conceptually, our objectives when analyzing the prediction market data are also slightly different from those of the originators of the classical measures. In the context of the prediction markets we are interested in the connections between prices and beliefs, and beliefs and event outcomes. We aim to use liquidity to quantify the strength of the first connection with an eye to eventually predicting outcomes.

The presence of an automated market maker (AMM) is the feature we focus on particularly in the work below. The AMM fundamentally changes how liquidity is generated in prediction markets and how its prices react to trading activity. Consider a typical trade in a liquid blue-chip stock on the London Stock Exchange, such as British Petroleum (BP.L). During normal trading hours, even a large order can usually be matched quickly at prices close to the midpoint, as ample counterparties stand ready to provide liquidity. In contrast, in thin or over-the-counter markets, traders may need to sell at a steep discount to find buyers. These outcomes reflect the familiar economics of supply and demand operating through a continuous double auction. The AMM is an algorithm that trades on behalf of the market platform, matching a limited quantity of trades where there is currently no available counterparty. The size of this limited quantity, which has implications for the platform's exposure to risk, is encoded as a parameter in the algorithm and plays a key role in determining price dynamics[43]. Because the AMM updates prices according to a deterministic cost function, the relationship between inventory changes and quoted prices can be described analytically. This provides a more structured environment for modelling price dynamics than settings in which prices emerge purely from decentralized bilateral matching. One of our objectives with the current paper is to test this hypothesis, leveraging our knowledge of AMM algorithms to target quantities that best capture the properties and implications of price movements.

First among these implications is whether and how changing prices genuinely reflect the beliefs of market participants. By explicitly recognizing the role of the AMM in our calculations we are better placed to isolate the connection between prices and beliefs, and to use knowledge of this connection to quantify the connection between prices and the events to which the market has attached those prices.

5.3 THEORETICAL FRAMEWORK: LMSR DYNAMICS AND CARA PREFERENCES

Having established the limitations of CDA-based liquidity measures, we now turn to the pricing mechanism that underpins most modern prediction markets: the logarithmic market scoring rule (LMSR) introduced by [42, 43]. The LMSR defines prices algorithmically through a convex cost function, guaranteeing continuous liquidity even when few counterparties are active. Rather than balancing supply and demand through bilateral matching, the market maker itself quotes buy and sell prices according to its internal inventory state. This design makes the market's "depth" an explicit model parameter rather than an emergent property, allowing liquidity to be expressed analytically through the curvature of the cost function. We outline these dynamics below.

5.3.1 LMSR Price Dynamics

As discussed in Section 5.1, an AMM is assumed to set the price p for a contract according to the derivative of a convex cost function

$$H(q) = b \log\left(1 + e^{q/b}\right), \qquad p = H'(q) = \frac{1}{1 + e^{-q/b}}. \tag{5.8}$$

where q is the volume of contracts the AMM has sold so far and b is a illiquidity parameter. The precise choice of the cost function H is motivated by its connections to the logarithmic scoring rule for forecasts. Considering the logit-transformed price

$$logit(p) = \log\left(\frac{p}{1-p}\right),$$

which we denote by x for convenience, we arrive at a very simple relationship

$$x = logit(p) = q/b, \tag{5.9}$$

from which it is clear that larger values of b reduce the connection between trade volumes and (logit) price.

5.3.2 CARA Preferences and Trader Behaviour

The simple logit–price relationship in (5.9) naturally raises the question of what determines the order flow that pushes q one way or another. How and why do traders

decide the volumes of their trades? Among possible behavioural specifications, we follow [26] because CARA preferences lead to analytically tractable demand expressions and naturally interact with the LMSR pricing rule and assume that traders behave according to constant absolute risk aversion (CARA) preferences. According to this assumption a trader's utility for money and coefficient of absolute risk aversion are described by the functions

$$U(w) = -\exp(-w/\tau), \qquad -U''(w)/U'(w) = \tau^{-1}. \tag{5.10}$$

Readers are referred to [57] for a standard treatment of CARA utility and measures of absolute risk aversion.

The constant coefficient of absolute risk aversion implies that for these traders there is a disconnect between their total wealth and their appetite for risk. Such behaviour is relevant in the context of short-horizon traders who make isolated decisions rather than managing a long-term portfolio. In practice, a Polymarket participant often decides to buy or sell a small number of shares in a single market when they notice a price that seems misaligned with their belief, without planning a sequence of future trades or considering the effect on other positions.

This interpretation is broadly consistent with the observed trading data. Across 5.2 million executed trades, 74.7% of wallet–market interactions consist of a single trade and 90.8% consist of at most two trades, corresponding to either a one-shot action or a simple open–close of a position. Only 5.1% of interactions involve three or four trades, and fewer than 4% involve five trades or more. Each trade is therefore typically a one-shot response to current information rather than part of a dynamic optimisation problem. The CARA assumption provides a reasonable reduced-form description of this behaviour: it models traders who size their positions according to both their beliefs relative to the market price and their tolerance for risk, without needing to account for wealth effects or intertemporal strategy. While other utility forms, such as logarithmic utility, could in principle be used, CARA provides a simple and empirically realistic description of how most traders behave on platforms like Polymarket.

5.3.3 CARA Demand under a Small-Stakes Approximation

Consider now a trader with CARA utility whose wealth changes by a small quantity dw. A second-order Taylor expansion about her current wealth leads us to the

expression

$$\begin{aligned} u(w_0+dw) =& u(w_0)+u'(w_0)\ dw+\frac{1}{2}\ u''(w_0)\ dw^2+O(dw^3) \\ =& u(w_0)+u'(w_0)\left[dw+\frac{u''(w_0)}{2u'(w_0)}\ dw^2\right]+O(dw^3) \\ =& u(w_0)+u'(w_0)\left[dw-\frac{1}{2\tau}\ dw^2\right]+O(dw^3). \end{aligned} \tag{5.11}$$

In the limit in which the third order terms are negligibly small, the trader maximizes her expected utility by maximizing the expectation of the term in the square brackets in (5.11). In the present context, the trader's profit or loss is a result of buying a small quantity dq of a contract at price p and cashing it in for price X when the underlying event is resolved so that $dw = dq(X-p)$. If $\tau > 0$, her expected utility is maximized at trade volume

$$\text{argmax}_{dq}\ E\left[dq(X-p)-\frac{1}{2\tau}dq^2(X-p)^2\right]=\tau\frac{E(X-p)}{E\left[(X-p)^2\right]}, \tag{5.12}$$

which, in accordance with the linear relationship (5.9), drives the AMM's logit price up by

$$dx = dq/b = \frac{\tau}{b}\frac{E(X-p)}{E\left[(X-p)^2\right]} = \frac{\tau}{b}\frac{\pi-p}{\pi(1-\pi)+(\pi-p)^2}, \tag{5.13}$$

where π denotes the trader's expectation for X. If we now consider limits in which $(\pi-p)^2$ is small, then we can write the change in logit price and the trader's logit expectation as

$$dx = \frac{\tau}{b}\frac{\pi-p}{\pi(1-\pi)}+O((\pi-p)^2) \tag{5.14}$$

and, via a first-order Taylor expansion,

$$e := \log\left(\frac{\pi}{1-\pi}\right) = x+\frac{\pi-p}{\pi(1-\pi)}+O((\pi-p)^2), \tag{5.15}$$

from which we derive the expression

$$dx = \frac{\tau}{b}(e-x)+O((e-x)^2) \tag{5.16}$$

Equation (5.16) shows that price responsiveness arises from the interaction of two distinct components: market design, through the liquidity parameter b, and trader behaviour, through risk tolerance τ. Neither parameter alone determines observed

price movements; only their ratio matters for logit-price dynamics. We therefore define the composite risk–liquidity parameter

$$\kappa = \frac{\tau}{b}. \tag{5.17}$$

Large values of κ correspond to rapid logit adjustment; either because traders are aggressive or liquidity is thin, while small values imply slower, more inertial price dynamics. Interpreted this way, κ captures how trader preferences and AMM design jointly shape market logit-price dynamics, which is the object estimated in the subsequent empirical analysis.

5.4 ESTIMATION OF THE RISK–LIQUIDITY COEFFICIENT FROM LOGIT PRICE DYNAMICS

Having established notation for a contract's logit-price and for a trader's logit-expectation for the binary variable X underlying the contract

$$x = logit(p), \qquad e = logit(\pi), \qquad \pi = E(X), \tag{5.18}$$

we derived the local adjustment rule

$$dx \approx \kappa\,(e - x)\,, \tag{5.19}$$

where $\kappa = \tau/b$ is the composite risk-liquidity parameter and the approximate nature of the relationship hinged on trades being small. Equation (5.19) admits a natural interpretation as a discrete-time dynamical system. Writing x_t for the AMM's logit-price immediately prior to a trade, the post-trade logit satisfies

$$x_{t+1} \approx x_t + \kappa\,(e_t - x_t) = (1 - \kappa)x_t + \kappa e_t. \tag{5.20}$$

According to this model, a κ coefficient of one means that the logit-prices are the same as the logit-expectations but delayed by one time interval. Values of κ less than one turn the logit-prices into a smoothed version of the logit-expectations and values greater than one lead to over-shooting, oscillatory behaviour of the kind anticipated by Roll in [70].

To operationalize (5.20), we sample logit-prices at T uniform intervals over the period during which the contract is traded. In doing so, the quantity e_t becomes

a interval-level proxy for the representative logit-expectation of traders who move prices during interval t.

The logit-price can now be seen as a discrete-time autoregressive process of order one, whose innovation terms are provided by the traders' logit-expectations. While the core mechanism generating persistence is autoregressive, we allow for an additional moving-average component in the expectation process in order to accommodate transient microstructure effects, discretization noise, and short-run execution irregularities present in high-frequency trading data. The resulting ARMA specification therefore extends, rather than replaces, the earlier AR interpretation. It can be shown (in Appendix 5.D) that if the process e_t is a stationary autoregressive moving-average (ARMA) process of order (p,q), i.e.

$$e_t = \varepsilon_t + \sum_{i=1}^{p} \phi_i e_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \qquad \varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2), \tag{5.21}$$

then x_t is an ARMA$(p+1, q)$ process

$$x_t = \kappa \varepsilon'_t + \sum_{i=1}^{p+1} \phi'_i x_{t-i} + \sum_{j=1}^{q} \theta'_j \varepsilon'_{t-j}, \qquad \varepsilon'_t \overset{\text{iid}}{\sim} (0, \sigma^2), \tag{5.22}$$

with coefficients $\theta'_j = \kappa\theta_j$ and

$$\phi'_1 = \phi_1 + 1 - \kappa, \qquad \phi'_{p+1} = -(1-\kappa)\phi_p \tag{5.23}$$

$$\phi'_i = \phi_i - (1-\kappa)\phi_{i-1}, \qquad i = 2, \ldots, p. \tag{5.24}$$

Equations (5.23)-(5.24) show how κ is confounded with the ARMA coefficients of the e_t process. It is not possible, for example, to distinguish between a series of logit-prices that varies slowly because it is driven by a logit-expectation process that changes slowly, and series of logit-prices that varies slowly because the liquidity is low. We can, however, estimate κ if certain assumptions are made about e_t.

If e_t is assumed to be an ARMA(1,q) process then x_t is an ARMA(2,q) process with autoregression coefficients

$$\phi'_1 = \phi_1 + 1 - \kappa, \qquad \phi'_2 = -(1-\kappa)\phi_1, \tag{5.25}$$

implying that

$$0 = u^2 - \phi'_1 u - \phi'_2, \qquad u = 1 - \kappa \tag{5.26}$$

so that κ can be written as a function of the autoregression coefficients

$$\kappa = 1 - \frac{1}{2}\left(\phi_1' - \sqrt{\phi_1'^2 + 4\phi_2'}\right). \tag{5.27}$$

We note also that if e_t is a random walk with $\phi_1 = 1$, a special case of the assumption above, then $\kappa = 2 - \phi_1'$. Equation (5.27) naturally suggests a plug-in estimator for κ resulting from substituting standard least-squares estimators for the autoregression coefficients ϕ_1' and ϕ_2'. It is this estimator that we compute below in Section 5.5.

The ARMA coefficients are estimated using standard conditional least-squares methods under the assumption that the innovation process is covariance-stationary with finite second moments. Residual diagnostics from the fitted models generally exhibited limited remaining serial dependence, although mild heteroskedasticity is unsurprising given the high-frequency and event-driven nature of prediction market data. The models are therefore interpreted primarily as reduced-form approximations intended to capture short-run dependence structure rather than as complete structural descriptions of the trading process.

The assumed order of the e_t ARMA process here implies that, after accounting for the MA component that describes transient (possibly non-white) noise terms, its future values are independent of past values given the present value. There are no additional autoregressive terms corresponding to longer-memory dependence in the expectation process. This sort of behaviour is broadly consistent with traders reacting primarily to current information rather than exhibiting persistent higher-order dependence in expectations. Observable momentum-type effects in series of logit-prices, which do depend on past adjustments, are understood to be the result of the liquidity parameter κ.

5.4.1 Risk-liquidity corrections for logit-price

Equation (5.20) suggests that trader expectations can be recovered from logit-prices according to

$$e_t = x_t + \kappa^{-1}(x_{t+1} - x_t). \tag{5.28}$$

This equation tell us that the traders' expectations are not instantaneously captured by the logit-prices because of the combined liquidity-risk effect. We can, however, compensate for this by looking at where and how quickly the logit-prices are moving. Having estimated κ from an observed series of logit-prices, we can plug it into (5.28) to construct a proxy for the unobserved trader expectations.

5.5 EMPIRICAL RESULTS

The parameter κ, derived and discussed in Sections 3 and 4, arises from the interaction of AMM price setting and trader risk aversion, and governs the responsiveness of logit prices to trading activity. We estimate κ using equation (5.27), plugging in estimates of autoregression coefficients computed from series of logit prices. The resulting estimates provide a reduced-form description of logit price dynamics under automated market making.

In Subsection 5.5.1 we simulate data from the models discussed in previous sections, illustrating how variation in κ is expected to manifest in observable logit price dynamics. In Subsection 5.5.2 we analyze price data from Polymarket through the lens of this framework, estimating κ across markets and examining its empirical behaviour.

5.5.1 Simulated Data

To illustrate the effects of differing values of κ and assess our ability to estimate it from prices, we conduct auxiliary simulation checks in which latent trader beliefs evolve according to ARMA processes and logit prices follow from equation (5.20). In these simulations, κ is not point-identified from prices alone, but it systematically influences observable features of logit price dynamics.

Holding belief dynamics fixed, smaller values of κ generate more persistent logit price series, exhibiting lower variance in first differences. When belief persistence is moderate, price-based estimators co-move with the true value of κ; as belief persistence approaches unit-root behaviour, this relationship weakens, reflecting the fact that price dynamics become highly persistent for all values of κ and differences in smoothing are difficult to distinguish in finite samples.

These diagnostics motivate treating empirical estimates of κ as reduced-form measures of price inertia rather than as direct estimates of latent belief dynamics.

This interpretation has two important implications for the empirical analysis. First, κ should be understood as a local, time-varying descriptor of price responsiveness rather than a fixed market characteristic. Second, variation in κ does not admit a unique decomposition into mechanism-driven and belief-driven components. Differences in estimated κ may reflect changes in participation, shifts in trader conviction, or transient order-book imbalances, even when the underlying market

design is unchanged.

For these reasons, we do not interpret κ as defining a small number of latent, persistent liquidity regimes. Instead, our empirical analysis treats κ as a continuous, reduced-form descriptor of price responsiveness whose distribution and behaviour vary across markets and over time. This framing preserves the mechanism-based interpretation of κ while avoiding structural assumptions about belief dynamics that are not identified from transaction data.

Accordingly, the remainder of Section 5 examines how variation in κ relates to economically meaningful outcomes, focusing in particular on the quality of market prices as probability forecasts.

5.5.2 Polymarket data

We begin by summarising trading activity across major Polymarket categories to provide context for subsequent analyses of belief dispersion.

Table 5.1: Summary of trading activity by category. Reported are total trades, mean and median shares traded, mean and median trade volume (in USDC), number of unique makers, and trade counts by side.

Category	Trades	Mean shares	Median shares	Mean ($)	Median ($)	# Traders	BUY	SELL
NBA	1,551,275	1,023.506	50.000	266.179	20.000	75,677	1,279,534	271,741
Trump	368,560	902.167	38.655	142.433	16.839	36,139	239,446	129,114
Bitcoin	280,695	1,861.814	85.730	221.498	28.997	16,790	201,338	79,357
Premier League	180,104	998.643	59.357	297.157	25.000	18,014	148,727	31,377
Temperature	129,372	466.808	31.345	43.590	13.000	8,778	98,976	30,396
Elon	72,524	952.774	58.140	82.166	20.000	3,774	51,642	20,882
Super Bowl	4,017	692.578	41.026	73.936	20.000	689	3,041	976

The dataset spans seven major Polymarket categories, which differ substantially in liquidity, participant composition, and information structure:

- **NBA**: Markets linked to the 2024–25 season, including matches and season-long propositions. This is the most liquid category, with over 1.55 million trades and approximately 75,000 unique participants, reflecting a continuous flow of information.
- **Trump**: Political markets related to electoral outcomes, policy developments, and legal events. These markets account for roughly 368,000 trades and exhibit concentrated liquidity driven by news cycles and partisan information flows.

- **Bitcoin**: Contracts tied to price thresholds, regulatory decisions, and adoption milestones. Around 281,000 trades are observed, with relatively large average trade sizes indicative of more capitalised participants.
- **Premier League**: Match outcomes and season propositions for the 2024–25 English Premier League. These markets generate approximately 180,000 trades across 18,000 traders, with information arriving episodically around weekly fixtures.
- **Temperature**: Weather- and climate-based threshold markets. These contribute about 129,000 trades, with smaller trade sizes and belief updating driven primarily by gradual forecast revisions.
- **Elon**: Markets referencing Elon Musk and associated firms. Trading volume is modest (around 72,500 trades) and typically concentrated around discrete announcements or earnings events.
- **Super Bowl LIX**: Short-horizon sports markets associated with a single major event. Despite low overall volume (just over 4,000 trades), these markets display sharp liquidity spikes around key announcements and game developments.

5.5.3 Forecast Accuracy and the Risk–Liquidity Parameter

This subsection examines how the risk–liquidity parameter κ relates to the quality of market prices as probability forecasts. While κ is estimated purely from price dynamics, its interpretation motivates examining whether it is systematically related to how effectively prices aggregate information. We assess this link directly by evaluating forecast accuracy at resolution.

5.5.3.1 Evaluation framework

For each market, forecast accuracy is evaluated using the Brier score computed at resolution. At the level of a window w, the forecast is taken to be the end-of-window market price $p^{\text{market}}_{w,\text{end}}$, interpreted as a probability forecast for the event outcome. Let $Y \in \{0, 1\}$ denote the realised outcome. The window-level Brier score is

$$\text{Brier}_w = \left(p^{\text{market}}_{w,\text{end}} - Y\right)^2.$$

Windows are grouped into bins based on the estimated value of $\widehat{\kappa}_w$, and average Brier scores are computed within each bin. This approach allows us to examine how

forecast accuracy varies systematically with the degree of logit price inertia captured by κ, without imposing additional structure on belief dynamics.

5.5.3.2 Binned results

Table 5.2 reports average Brier scores across quintiles of $\widehat{\kappa}$. Forecast accuracy is highest for windows in which κ is close to one. Brier scores increase for both smaller and larger values of κ, indicating poorer forecast performance in regimes where prices either adjust too strongly or too weakly to order flow.

Table 5.2: Forecast accuracy by κ quintile (end-of-window prices). Lower Brier scores indicate more accurate probability forecasts.

κ bin	Mean κ	Mean Brier	N windows	s.e.
$(0, 0.78]$	0.43	0.1860	6,403	0.0025
$(0.78, 1.00]$	0.97	**0.1813**	7,428	0.0023
$(1.00, 1.34]$	1.12	0.1888	5,379	0.0027
$(1.34, 2.00]$	1.64	0.1971	6,402	0.0025
$(2.00, 6.84]$	2.83	0.1987	6,403	0.0025

Figure 5.2 visualises this relationship for the quintile specification by plotting mean Brier scores against mean κ, with error bars reflecting standard errors.

To examine this pattern at finer resolution, Table 5.3 reports the same analysis using deciles of $\widehat{\kappa}$. The minimum average Brier score again occurs in bins centred very close to $\kappa = 1$, and forecast accuracy deteriorates smoothly as κ moves away from this region in either direction.

Table 5.3: Forecast accuracy by κ decile (end-of-window prices).

κ bin	Mean κ	Mean Brier	N windows	s.e.
$(0.00, 0.48]$	0.23	0.1861	3,202	0.0036
$(0.48, 0.78]$	0.63	0.1858	3,201	0.0035
$(0.78, 1.00)$	0.92	**0.1802**	3,202	0.0035
$[1.00, 1.00]$	1.00	0.1822	4,226	0.0031
$(1.00, 1.05]$	1.01	0.1880	2,177	0.0041
$(1.05, 1.33]$	1.20	0.1893	3,202	0.0035
$(1.33, 1.58]$	1.46	0.1957	3,200	0.0035
$(1.58, 2.00]$	1.81	0.1986	3,202	0.0035
$(2.00, 2.46]$	2.09	0.1969	3,201	0.0035
$(2.46, 6.84]$	3.57	0.2005	3,202	0.0035

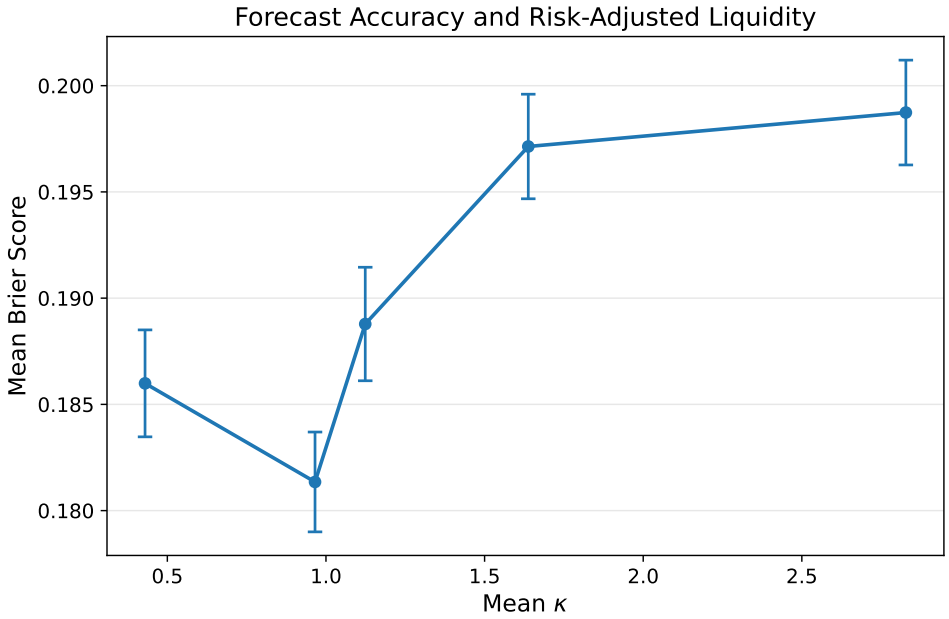


Figure 5.2: **Mean Brier score as a function of mean liquidity parameter κ using quintile bins**. Error bars denote ± 1 standard error.

Figure 5.3 confirms that this pattern is robust when the distribution of κ is partitioned more finely.

5.5.4 Interpretation of Forecast Accuracy Results

The risk–liquidity parameter κ governs the inertia of logit price adjustment implied by the interaction of the LMSR pricing rule and trader risk aversion. Empirically, forecast accuracy varies systematically with κ. Windows in which κ is close to one achieve the lowest Brier scores, while both smaller and larger values are associated with reduced accuracy.

When κ is small, prices adjust slowly and exhibit substantial inertia, incorporating information with delay. When κ is large, prices respond strongly to individual trades and may over-adjust, increasing sensitivity to transient noise. The region around $\kappa \approx 1$ represents a balance between responsiveness and inertia, and it is in this regime that market prices perform best as probability forecasts.

These results establish a direct empirical relationship between logit price dynamics under automated market making and ex post forecast quality. They show that the dynamic behaviour captured by κ is economically meaningful and that intermediate

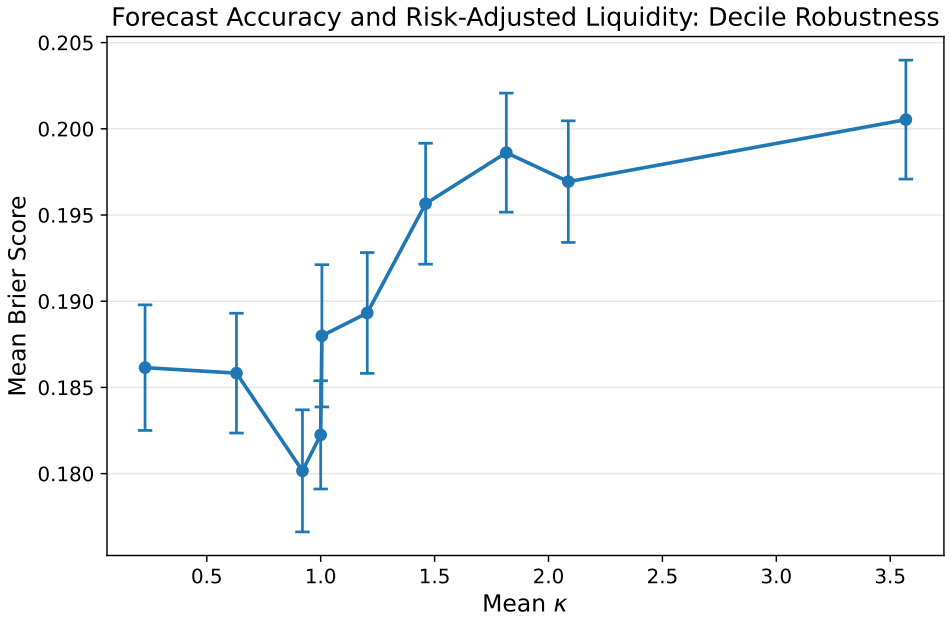


Figure 5.3: **Mean Brier score as a function of mean liquidity parameter κ using decile bins**. Error bars denote ± 1 standard error.

levels of price responsiveness are associated with superior information aggregation in prediction markets.

5.6 DISCUSSION AND IMPLICATIONS

This paper examined how the risk–liquidity parameter κ varies across prediction markets operated by automated market makers. By estimating κ at high temporal resolution and analysing its short-horizon behaviour, the results show that price responsiveness differs systematically across markets and over time. The framework links a mechanism-based pricing rule to observable price dynamics, providing a reduced-form way to characterise how markets absorb order flow under automated liquidity provision. This section interprets the empirical role of κ, discusses heterogeneity across market domains, outlines methodological and practical implications, and identifies directions for future work.

5.6.1 Interpretation of $\boldsymbol{\kappa}$ and heterogeneity

The parameter κ summarises the joint influence of market design and trader risk tolerance on logit price adjustment under an automated market maker. In the underlying model, $\kappa = \tau/b$ represents the ratio of aggregate trader risk tolerance to the AMM liquidity parameter. Smaller values of κ correspond to greater inertia in logit price adjustment, while larger values imply more rapid responses to order imbalance. Empirically, estimates of κ vary over time within markets and differ substantially across market categories, indicating that price responsiveness is not constant even when the mechanism itself is fixed.

Rather than drifting arbitrarily, κ tends to fluctuate around locally stable levels over short horizons, reflecting how automated pricing rules interact with changing participation, attention, and information flow. This behaviour should be interpreted as a reduced-form outcome of trading conditions rather than as direct evidence about the persistence of latent beliefs or trader composition. In particular, variation in κ may arise from episodic bursts of activity, imbalanced order flow, or transient changes in market depth, even when the underlying AMM parameters remain unchanged.

Cross-market heterogeneity in κ highlights systematic differences in trading environments. Cryptocurrency-related markets exhibit relatively moderate levels of κ with limited dispersion, consistent with deeper liquidity and frequent participation. Political and event-driven markets display greater variation, reflecting concentrated trading around discrete information arrivals. Sports and cultural markets occupy intermediate positions, with generally stable price responsiveness punctuated by scheduled events. Taken together, these patterns suggest that κ captures meaningful differences in how markets process information under automated pricing, while remaining agnostic about the specific sources of belief formation.

5.6.2 Methodological and practical implications

Methodologically, the analysis demonstrates that informative measures of price responsiveness can be recovered directly from price dynamics without imposing parametric assumptions on latent belief processes. The window-based estimation of κ provides a mechanism-consistent analogue to classical liquidity measures, such as Kyle's λ or Amihud's illiquidity ratio, while avoiding reliance on transaction-level regressions or ex post price impact estimates. Because κ is derived from the AMM

pricing rule itself, it offers a natural way to connect structural models of automated liquidity provision with empirical observation.

From a practical perspective, κ serves as a diagnostic indicator of market conditions. Tracking its evolution can help identify periods in which prices become unusually sensitive or unusually inert to order flow. While such episodes are often short-lived, sustained changes in κ may signal shifts in participation, attention, or information arrival. Market designers could use this information to assess whether liquidity parameters remain appropriate as trading conditions evolve, or to guide the design of adaptive mechanisms that respond to changes in market activity.

The framework also has implications for simulation and forecasting. Treating κ as a time-varying reduced-form descriptor of price responsiveness allows it to be incorporated into stress tests or counterfactual analyses of automated markets. More broadly, the results highlight that even in fully algorithmic settings, liquidity conditions are shaped by the interaction of market design and trader behaviour, rather than being fixed primitives.

5.6.3 Caveats and Directions for Future Work

The results in Section 5.2 provide evidence of a systematic relationship between the risk–liquidity parameter κ and the quality of market prices as probability forecasts. This relationship is supported by both the theoretical interpretation of κ as a reduced-form dynamic parameter and the empirical patterns observed in the data. Nevertheless, several caveats delimit the scope of the findings and point to natural directions for future work.

First, κ is estimated using a specific time-domain approach based on local price dynamics. Alternative estimators may yield complementary perspectives on price inertia, including methods based on frequency-domain representations or wavelet decompositions. Comparing the stability and interpretability of κ across such estimators would help clarify the extent to which the observed relationship depends on the chosen estimation procedure.

Second, while the analysis aggregates across a large number of windows, it does not explicitly assess the sensitivity of κ estimates to isolated shocks or bursts of trading activity. If κ were highly sensitive to a small number of extreme events, this could affect its interpretation as a persistent characteristic of market behaviour. Formal sensitivity analyses could help distinguish persistent dynamics from transient

effects.

Third, the framework treats latent beliefs as locally stationary within windows. In practice, belief dynamics may exhibit non-stationarity, particularly as markets approach resolution and trading intensity increases. Allowing for time-varying belief processes or explicitly modelling changes in belief variance over the market lifecycle may further refine the interpretation of κ.

Finally, the analysis focuses on binary prediction markets. Extending the framework to other domains, such as markets whose payoffs depend on prices crossing thresholds or ranges, may require modified interpretations of both κ and forecast accuracy. Assessing the applicability of the present results across such settings remains an open question.

These considerations do not undermine the empirical link documented in Section 5.2, but they clarify the conditions under which it should be interpreted and highlight opportunities for future work to test the robustness and generality of the relationship between market liquidity dynamics and forecast quality.

5.7 SUMMARY

This paper developed and applied a mechanism-consistent framework for measuring risk-adjusted liquidity in prediction markets operated by automated market makers. By deriving a mechanism-consistent adjustment rule involving $\kappa = \tau/b$ and estimating this parameter directly from price dynamics, the analysis linked AMM design parameters to observed variation in price responsiveness across markets and over time. Empirical results show that κ varies systematically across domains and within markets, reflecting changes in participation and information flow rather than fixed liquidity regimes. Together, these findings position κ as a useful reduced-form descriptor of how automated markets absorb order flow, providing both a descriptive measure of market quality and a foundation for future work on adaptive liquidity design in algorithmic trading environments.

5.A ILLUSTRATIVE COMPARISON WITH CLASSICAL LIQUIDITY PROXIES

This appendix reports a limited descriptive comparison between the risk–liquidity parameter κ and selected classical liquidity proxies, specifically Kyle's λ and Amihud's illiquidity ratio. The purpose of this exercise is not to establish equivalence, superiority, or substitution across measures, but to document how a mechanism-based measure behaves relative to reduced-form proxies developed for continuous double-auction markets.

Kyle's λ and Amihud's illiquidity are constructed from realised price and volume data and are commonly interpreted as indicators of market depth or price impact in environments where prices emerge from competitive order matching. In contrast, κ captures logit price inertia induced by an automated market maker interacting with risk-averse traders. As a result, these quantities do not share a common structural interpretation.

Empirically, markets summarised by κ exhibit substantial overlap when projected into the space defined by classical proxies, and no stable clustering structure emerges across measures. This outcome is expected: κ reflects mechanism-driven smoothing and trader behaviour, while classical measures reflect realised impact conditional on observed trades. Differences across measures should therefore be interpreted as reflecting distinct economic objects rather than measurement error or mis-specification.

For these reasons, comparisons with classical liquidity proxies are not used in the main analysis. The primary empirical results focus on within-κ structure—its level, dispersion, and dynamics—which are directly interpretable within the automated market-making framework.

5.B ILLUSTRATIVE MEAN-REVERTING BEHAVIOUR IN $\widehat{\kappa}$

To illustrate the short-term behaviour of the estimated risk-adjusted liquidity parameter $\widehat{\kappa}$, a small set of representative markets were examined for evidence of mean reversion. For each market, two simple regressions were estimated: an autoregressive model,

$$\widehat{\kappa}_t = \phi_1 \widehat{\kappa}_{t-1} + \varepsilon_t,$$

and a mean-reversion specification,

$$\Delta\widehat{\kappa}_t = \beta_{\Delta\kappa\leftarrow\kappa_{t-1}}\widehat{\kappa}_{t-1} + \eta_t,$$

where a negative $\beta_{\Delta\kappa\leftarrow\kappa_{t-1}}$ indicates a tendency for $\widehat{\kappa}$ to move back toward its equilibrium level following short-lived changes in market conditions. To summarise both the strength and reliability of this adjustment, the coefficient of determination R^2 from each regression was combined with the slope parameter into a simple index:

$$\text{score} = (-\beta_{\Delta\kappa\leftarrow\kappa_{t-1}})R^2.$$

Table 5.4: Selected markets showing mean-reverting behaviour in the estimated risk-adjusted liquidity parameter $\widehat{\kappa}$. Each row reports autoregressive and mean-reversion coefficients, their associated R^2 values, and the composite score score $= (-\beta_{\Delta\kappa\leftarrow\kappa_{t-1}})R^2$.

Market	ϕ_1	$\beta_{\Delta\kappa\leftarrow\kappa_{t-1}}$	R^2	Score
Trump blanket EU tariff (by Jun 30)	−0.04	−1.04	0.52	0.54
Trump tariffs on Canada (before May)	0.05	−0.95	0.47	0.45
Qatar gifts Trump AF1 (May)	0.11	−0.89	0.44	0.39
NBA: MIA–OKC (Feb 2025)	0.10	−0.90	0.45	0.41
Strategy buys Bitcoin (Mar 2024)	0.12	−0.88	0.43	0.38
Elon tweets 625–649 (Jan–Feb)	0.00	−1.00	0.50	0.50

Across these markets, the estimates show a clear pattern of short-run mean reversion. The coefficients on lagged κ in the $\Delta\widehat{\kappa}_t$ regressions fall in a narrow band around -0.9 to -1.0, indicating that deviations in $\widehat{\kappa}$ are typically followed by adjustments of similar magnitude in the opposite direction. The R^2 values, which range from roughly 0.43 to 0.52, suggest that a meaningful share of the one-step change in $\widehat{\kappa}$ is explained by its previous level, though a substantial amount of variation remains idiosyncratic.

The autoregressive coefficients ϕ_1 are small and centred near zero, consistent with limited persistence in the level of $\widehat{\kappa}$. This indicates that movements in effective liquidity conditions are short-lived: shifts tend not to propagate beyond one or two windows, and there is little evidence of sustained drift over longer horizons.

Taken together, these results imply that $\widehat{\kappa}$ behaves as a high-frequency, mean-reverting process. Markets experiencing brief tightening or loosening of liquidity tend to return toward local average conditions relatively quickly. While the strength of this reversion varies modestly across domains, the overall pattern is similar: adjustments in $\widehat{\kappa}$ are rapid, and shocks do not persist over extended periods.

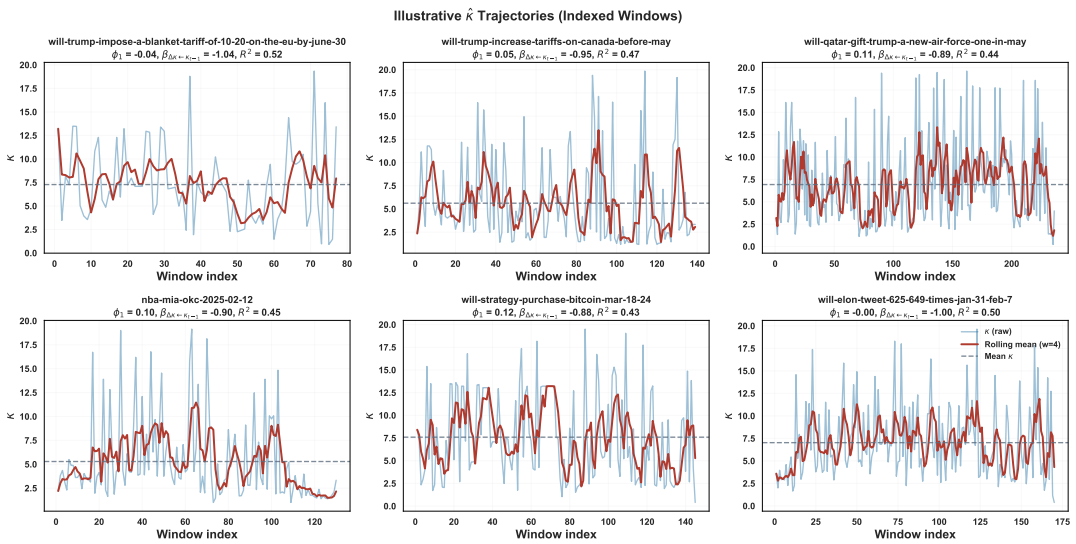


Figure 5.4: **Time series of estimated risk-adjusted liquidity parameter κ across illustrative markets**. The blue line shows raw estimates, the red line an eight-window rolling mean, and the dashed line the average level of κ for each market. Estimates exhibit short-term fluctuations around a stable mean, consistent with temporary liquidity adjustments.

5.C SCOPE OF CROSS-MARKET COMPARISONS

The empirical analysis above focuses exclusively on the internal structure of the risk–liquidity parameter κ: its level, dispersion, and short-run dynamics within and across markets. This emphasis is deliberate. The parameter κ is derived from the pricing rule of an automated market maker interacting with risk-averse traders, and its interpretation is therefore inherently mechanism-specific.

Classical liquidity proxies such as Kyle's λ or Amihud's illiquidity ratio were developed for continuous double-auction environments in which prices emerge from competitive order matching. In automated markets, where prices are set deterministically by a cost function, these measures do not share a common structural interpretation with κ. Direct comparisons between κ and classical proxies therefore do not admit a natural economic null and risk conflating distinct sources of price persistence and impact.

For this reason, the main analysis does not rely on clustering or cross-metric comparisons between κ and classical liquidity measures. Instead, we treat κ as a self-contained, reduced-form descriptor of logit price inertia under automated market making and characterise how its distribution and dynamics vary across information

environments.

For completeness, an illustrative comparison with selected classical liquidity proxies is reported in Appendix 5.A. These results are descriptive only and are not used in subsequent interpretation.

5.D A NOTE ON ARMA PROCESSES

We proposed a particular discrete-time dynamic model according to which traders' logit-expectations are understood to drive logit-prices, mediated by a liquidity parameter κ. The unobserved logit-expectations can be eliminated from equation (5.20) by substituting them for lagged logit-prices as follows,

$$\begin{aligned} x_t &= (1-\kappa)\, x_{t-1} + \kappa e_{t-1} \\ &= (1-\kappa)\, x_{t-1} + \kappa \left(\sum_{i=1}^{p} \phi_i e_{t-1-i} + \sum_{j=0}^{q} \theta_j \epsilon_{t-1-j} \right) \\ &= (1-\kappa)\, x_{t-1} + \left(\sum_{i=1}^{p} \phi_i (x_{t-i} - (1-\kappa) x_{t-1-i}) + \kappa \sum_{j=0}^{q} \theta_j \epsilon_{t-1-j} \right). \end{aligned} \tag{5.29}$$

Equation (5.29) reveals x_t to be another ARMA process, whose terms can be rearranged into the form

$$x_t = \sum_{i=1}^{p+1} \phi'_i x_{t-i} + \sum_{j=0}^{q} \theta'_j \epsilon_{t-j} \tag{5.30}$$

with $\theta'_j = \kappa \theta_j$ and

$$\phi'_1 = \phi_1 + 1 - \kappa, \qquad \phi'_{p+1} = -(1-\kappa)\phi_p \tag{5.31}$$

$$\phi'_i = \phi_i - (1-\kappa)\phi_{i-1}, \qquad i = 2, \ldots, p. \tag{5.32}$$

— 6 —

Conclusion and Future Work

This thesis has examined how probabilistic beliefs are aggregated and expressed in forecasting tournaments and prediction markets, with particular attention to settings in which the observed data are incomplete, irregular, or shaped by the institutional mechanisms that generate them. In such environments, probabilities are not direct observations of beliefs, but outcomes of processes that filter, transform, or constrain how information enters the data. As a result, standard interpretations of forecasts or prices as direct summaries of collective belief can be misleading without explicit modelling of the underlying data-generating structure.

Across three empirical contexts—dynamic forecast aggregation, belief inference from sparse order books, and liquidity in automated market makers—this thesis develops statistical frameworks that link observed probabilistic data to the mechanisms through which they are produced. The work does not propose a single model of belief formation. Instead, it adopts a modular approach that addresses distinct inference problems arising in different institutional settings. In each case, the objective is to construct model-based representations of latent informational structure that are consistent with both the observed data and the rules governing forecast submission, order placement, or price formation.

A common methodological principle throughout the thesis is that inference must be conditioned on the data-generating mechanism. Forecasts submitted asynchronously by human forecasters, standing orders observed in thin limit order books, and prices produced by automated market makers each reflect systematic selection and transformation effects. These effects imply that observed probabilities cannot be interpreted directly as beliefs without additional structure. The models developed in this thesis therefore incorporate these mechanisms explicitly, allowing the

analysis to distinguish between information-driven variation and artefacts induced by institutional design, and to study the evolution of probabilistic signals within clearly stated modelling assumptions.

6.1 SUMMARY OF CONTRIBUTIONS

The first contribution of the thesis is the development of *Kairosis*, a Bayesian method for aggregating probability forecasts submitted over time. The method models the evolution of the forecast distribution using a change-point structure and derives aggregation weights from the posterior distribution over candidate structural breaks, rather than imposing an externally chosen discounting rule. Applied to forecasting tournament data, *Kairosis* produces aggregate forecasts that are competitive with standard static and recency-weighted benchmarks under proper scoring rules, and in many cases achieves lower average scores. These results provide evidence that allowing aggregation weights to vary with estimated changes in the forecast distribution can improve performance in settings with irregular submission patterns and episodic information arrival.

The second contribution concerns the interpretation of prices and order books in thin prediction markets. The thesis develops a simulation-based inversion procedure that treats the visible order book as a selectively filtered sample from a latent cross-sectional distribution of trader beliefs. The approach jointly models the belief distribution and the visibility mechanism that governs which beliefs generate standing orders. When applied to Polymarket data, the resulting estimates show that markets with similar prices can be associated with substantially different inferred belief distributions. This finding illustrates the limitations of midpoint prices as sufficient statistics for market beliefs in sparse order-book environments.

The third contribution addresses the measurement of liquidity in automated market maker prediction markets. In these settings, price responsiveness is determined mechanically by the market-making rule rather than emerging endogenously from order flow competition. The thesis develops a parsimonious, mechanism-consistent model that relates observed price dynamics to belief proxies through a single composite parameter capturing risk-adjusted price responsiveness. Estimation of this parameter across markets and over time documents substantial heterogeneity in

price inertia and yields a reduced-form summary measure that facilitates comparison across markets operating under similar automated rules.

6.2 COMMON MODELLING PRINCIPLES

A common feature across the thesis is the treatment of observed probabilities as outcomes of explicit data-generating mechanisms rather than as direct measurements of belief. Forecast submissions, limit order books, and prices generated by automated market makers each reflect partial and selectively observed information, shaped by institutional design, incentives, and participation constraints. The methods developed in the thesis therefore model these mechanisms directly, with inference carried out conditional on the assumed structure linking latent beliefs to observable data.

A second shared feature is the explicit treatment of temporal structure. In each empirical setting, probabilistic signals evolve over time in ways that are not adequately summarised by static statistics. The proposed frameworks produce time-varying quantities, including aggregation weights, belief-distribution parameters, and measures of price responsiveness, which allow changes in informational conditions to be examined directly. This temporal perspective is particularly relevant in environments characterised by clustered activity and irregular information arrival, where static summaries can obscure meaningful variation.

6.3 LIMITATIONS

The results presented in this thesis are conditional on modelling assumptions that are chosen for tractability and interpretability rather than completeness. In all three empirical settings, beliefs, visibility, and behavioural responses are represented using stylised structures that abstract from many features of real-world decision making. Cross-sectional heterogeneity in trader or forecaster behaviour is captured only through aggregate distributions, and strategic interaction is treated in reduced form rather than modelled explicitly. These choices reflect a deliberate trade-off between realism and transparency.

From a statistical perspective, identification relies on assumptions linking latent informational objects to observable outcomes. In the order-book setting, belief

inference depends on assumptions about which beliefs appear as standing orders and how this visibility evolves over time. In the automated market maker setting, liquidity is characterised using a behavioural model that aggregates key features of market design and trader behaviour into a single reduced-form parameter. While these assumptions are economically motivated, alternative specifications could yield different numerical values. The results are therefore interpreted comparatively, focusing on how inferred quantities vary across settings and over time rather than as absolute measures.

Empirically, the scope of the analysis is constrained by data availability and platform-specific design. The forecasting and market data examined in the thesis are drawn from a limited number of platforms with particular institutional features. Consequently, the extent to which the findings generalise to other forecasting environments, market mechanisms, or liquidity regimes remains an open question. The estimated quantities should therefore be interpreted as model-dependent summaries rather than as universal measures of belief or market quality.

6.4 FUTURE WORK

The frameworks developed in this thesis admit several natural extensions. In the context of dynamic forecast aggregation, future work could examine how change-point-based weighting interacts with forecaster-specific information when individual histories, expertise measures, or incentive structures are observable. This would allow temporal variation and cross-sectional heterogeneity to be combined within a unified aggregation framework, albeit at the cost of increased model complexity.

In the order-book setting, belief inference could be extended by incorporating richer models of order placement, execution, and cancellation. Allowing the visibility mechanism to depend on market conditions or strategic considerations may help distinguish belief-driven sparsity from behaviour arising from execution risk or adverse selection. Such extensions would require stronger behavioural assumptions and more detailed data.

For automated market maker environments, further work could investigate how estimated price responsiveness relates to realised market outcomes, such as volatility, trading volume, or forecast accuracy, under alternative design parameters. The role of trader composition and participation dynamics in shaping liquidity estimates

also remains an open area for investigation. More broadly, the modelling approach adopted in this thesis could be applied to other environments in which probabilities are generated mechanically or strategically, provided that the underlying data-generating mechanisms can be specified with sufficient clarity.

6.5 FINAL REMARKS

This thesis has taken the view that probabilistic outputs in forecasting systems and prediction markets should be interpreted as products of explicit mechanisms rather than as direct expressions of belief. By modelling these mechanisms directly, the work places inference about belief aggregation, order-book structure, and liquidity on a clearer statistical footing. Although the models considered are necessarily simplified, they provide structured and internally consistent ways to interpret observed probabilities and their evolution over time in environments where direct observation of beliefs is not possible. The contribution of the thesis is therefore methodological in nature, offering tools for disciplined analysis rather than definitive claims about belief formation or market efficiency.

— A —

Microstructure of Decentralised Prediction Markets

Abstract

This appendix examines the trading microstructure of a large-scale decentralised prediction market, focusing on how participant heterogeneity relates to price volatility. Using over five million trades from Polymarket, we apply behavioural clustering to classify wallet-level traders and analyse their role in shaping market outcomes. The results indicate that decentralised prediction markets exhibit structural parallels to traditional exchanges, including skewed participation and functionally distinct trader roles. Volatility clustering is associated with shifts in participant composition, particularly increases in retail activity. These findings provide additional context for understanding behaviour and instability in decentralised markets

A.1 INTRODUCTION

Prediction markets are exchange platforms where participants trade contracts whose payouts depend on uncertain future events. In theory, market prices aggregate diverse information and can serve as probabilistic forecasts of real-world outcomes [87]. A rich literature documents the forecasting accuracy of prediction markets in domains ranging from elections and sports to economic indicators [15, 87]. However, achieving efficient information aggregation in practice depends critically on market design and participant behavior. Liquidity constraints and trader heterogeneity can hinder the idealized performance of these markets [82]. This issue is especially pronounced

in *decentralized* prediction markets,platforms like Augur and Polymarket,which are built on blockchain technology and operate without traditional intermediaries.

A key innovation enabling decentralized prediction markets is the use of *automated market makers* (AMMs) to provide continuous liquidity. In contrast to order book exchanges (which require sufficient active buyers and sellers at all times), an AMM is a smart contract that is always willing to quote a price for a trade, based on a predefined pricing function [42, 5]. The AMM model addresses the classic "thin market" problem in prediction markets [42]: traders need not coordinate their presence or negotiate prices with each other, as they trade directly against the liquidity pool maintained by the AMM. Polymarket, the focus of this appendix, employs such a design—specifically using liquidity pools funded by users who earn fees in return for providing collateral (USDC) to facilitate trading. This mechanism, similar to the constant-product market maker popularized in decentralized finance, ensures that for each prediction market there is always a counterparty available for trade, albeit at the cost of price impact that grows with trade size [5].

While automated liquidity provision can improve access and immediacy, it also introduces new microstructural dynamics. The price updates algorithmically with each trade, meaning that large trades can move prices significantly, and liquidity providers (including the AMM itself) may bear inventory risk and impermanent loss as markets evolve. Moreover, decentralized prediction markets inherit the anonymity and global reach of cryptocurrency markets: participants range from casual bettors to sophisticated arbitrageurs, all interacting on-chain via pseudonymous wallets. This heterogeneity in trader types and motives can lead to complex price dynamics, including the possibility of volatility clustering, periods of unusually high price variability associated with bursts of trading activity or information flow.

In this appendix, we examine how the composition of traders in a decentralized prediction market influences market microstructure and price behavior. Using a comprehensive dataset of over five million trades from Polymarket (a leading decentralized prediction market platform), we analyze trading activity at the wallet-address level to identify distinct classes of market participants. We then investigate how these participant types contribute to phenomena such as volatility clustering and price efficiency. Our approach draws on insights from financial market microstructure theory, which emphasizes the roles of different trader archetypes, market makers, informed traders, noise traders, etc., in shaping liquidity and volatility [53, 18, 38]. By mapping blockchain addresses to behavioral profiles, we bridge the gap between

traditional microstructure models and the new context of blockchain-based markets.

The remainder of the appendix is organized as follows. In the next section, we situate our work in the relevant literature on prediction market design, automated market making, and trader heterogeneity in markets. We then present an exploratory data analysis of Polymarket activity to characterize the overall trading landscape and concentration. Following that, we employ clustering techniques to classify traders into behavioral types and interpret these clusters as analogues of familiar market roles (e.g. liquidity providers vs speculators). We analyze the temporal patterns of trading volume by cluster and assess their impact on price volatility, including a cross-sectional regression of market-level volatility on trader composition and a detailed case study of a single market with extreme activity. Finally, we discuss the implications of our findings for prediction market efficiency and design, and how the presence of automated market makers and diverse participants can drive market outcomes.

A.2 LITERATURE REVIEW

Prediction markets have been studied extensively as mechanisms for aggregating information and forecasting events. Early seminal work demonstrated that market-generated probabilities can outperform polls and expert opinions in predicting election results and economic indicators [15, 87]. The theoretical appeal of prediction markets lies in the Hayekian idea that prices encapsulate dispersed information held by many individuals. Wolfers and Zitzewitz [87] provide a survey of this literature, highlighting cases in which prediction market prices closely track real-world event probabilities. In practice, however, real-money prediction markets often face low liquidity and participation constraints. For example, Rhode and Strumpf [79] note that market inefficiencies and arbitrage opportunities can persist in thinly traded contracts, and that external shocks or constraints (such as regulatory limits on participation) can impair market accuracy.

One stream of research has explored the design of market makers to address liquidity issues in prediction markets. Traditional exchanges such as TradeSports or Betfair rely on a continuous double auction, with human or algorithmic market makers posting bids and asks. When trading interest is sparse, these order book markets can suffer from wide bid–ask spreads or prolonged periods of inactivity, a phenomenon

often referred to as the "thin market problem." Hanson [42] introduced the concept of a *market scoring rule*, an automated market maker that prices trades so as to maintain consistency with an underlying cost function. His Logarithmic Market Scoring Rule (LMSR) became a cornerstone for many subsequent designs, guaranteeing bounded losses for the market maker while providing continuous liquidity.

Decentralized platforms have adopted similar approaches. For example, Augur's initial implementation and other blockchain-based prediction markets allow users to trade directly against a smart contract rather than with other traders. In parallel, developments in decentralized finance (DeFi) have produced constant-function market makers, such as the constant-product rule used by Uniswap [5], which has also been adapted to binary outcome markets by maintaining a fixed product of outcome token holdings. These automated mechanisms simplify the trading process and attract liquidity providers, who supply capital to the market in exchange for fees and thereby serve as passive market makers.

The presence of automated market makers in prediction markets influences the microstructure in important ways. Unlike a human market maker who might adjust quotes based on judgment or inventory, an AMM follows a fixed algorithm for price adjustment. This can lead to predictable price impact: large trades move prices according to the formula, which savvy traders can anticipate. As a result, informed traders with superior knowledge or faster reaction to news may strategically trade against the AMM (and the liquidity providers behind it) for profit, potentially akin to picking off stale quotes in a traditional market [82]. Indeed, [82], studying TradeSports data, found that liquidity, while generally facilitating trading, did not always improve informational efficiency; at times it even increased mispricing, as naive limit order traders were picked off by the more informed. This finding echoes the broader point [18] that markets comprise both informed traders and "noise traders" (who trade on speculation, error, or entertainment), and that an influx of uninformed liquidity can increase volatility and delay price discovery.

Another relevant body of work is the general financial market microstructure literature on trader composition and market quality. Classic models [38, 53] formalize how market makers set prices in the presence of informed and uninformed trading. The key insight is that prices adjust based on the inferred mix of trader types behind each order flow: a higher probability of informed trading leads to larger quotes and more volatility in price updates (to protect market makers from adverse selection). In modern electronic markets, it has been observed that a small number of professional

or algorithmic traders account for a large share of volume and liquidity provision. Empirical studies show that the introduction of algorithmic trading can tighten spreads and improve liquidity [47], and that a single high-frequency market maker can dominate liquidity supply on a new venue [58]. These participants often trade in high frequency and volume, but in ways that mitigate price volatility (e.g., by rapidly arbitraging away mispricings or supplying depth). On the other hand, retail traders, typically trading smaller sizes and less frequently, collectively contribute a large number of transactions that may be driven by noise or behavioral biases. [9] and [19] document that retail order flow in equity and option markets tends to be uninformed and can even predictably lose money to more sophisticated players, implying that their trades might temporarily push prices away from fundamental values.

In the context of prediction markets, we can analogize these roles: some wallet addresses behave like market makers or arbitrageurs (frequently trading, often providing liquidity or correcting mispricings), while others behave like casual speculators or bettors (infrequently trading small amounts, perhaps chasing trends or reacting to news in slower, less informed ways). There is also an intermediate category of informed speculators, traders who may not provide liquidity but who trade in larger size or with better timing than the average retail participant, potentially because they possess superior information or analytic acumen about the event in question. Understanding the mix of these participant types is crucial because their interaction determines market outcomes: how quickly prices respond to new information, how much prices fluctuate day-to-day, and how much volume the market can sustain. For instance, a market dominated by uninformed enthusiasm might show excessive volatility and a higher incidence of mispricing (as suggested by noise trader theory [18]), whereas a market with robust liquidity provision by market makers might exhibit smoother price trajectories and tighter spreads [47].

Despite the growing interest in decentralized prediction markets, relatively few studies have empirically examined their microstructure. Most existing analyses focus on predictive accuracy or the presence of biases like the favorite-longshot bias (overpricing of low-probability outcomes) [80]. Our work contributes to the literature by providing one of the first detailed explorations of trader composition in a blockchain-based prediction market and linking it to volatility clustering, a phenomenon well documented in financial markets (e.g., ARCH/GARCH effects) but less studied in prediction markets. By leveraging a complete transaction history

from Polymarket, we can classify traders in a data-driven way and quantify how each group's activity correlates with price behavior. This approach allows us to connect strands from prediction market research, automated market making design, and traditional market microstructure in order to better understand how these novel markets function and where inefficiencies or instabilities may arise.

A.3 POLYMARKET TRADING ACTIVITY

We begin by characterizing the structure of trading activity on Polymarket at the wallet level, aggregating across all 5.17 million observed trades spanning 1,811 unique prediction markets. These markets cover a diverse range of topics, including sports (e.g. Super Bowl, NBA, Premier League), climate and science (e.g. temperature records, COVID-19 trends), politics (elections, Donald Trump-related contracts), pop culture, and cryptocurrencies, reflecting a blend of political, financial, and entertainment-based forecasting questions. By analyzing the entire ecosystem of trades, we can discern broad participation patterns and identify the most active entities in this decentralized market.

In this exploratory analysis, we decompose the dataset across several behavioral and structural dimensions:

- **Volume Concentration:** We identify high-volume participants and measure concentration of volume and trades. A particular focus is given to the role of Polymarket's automated market maker contracts (AMMs), which are expected to be among the top accounts by volume given their liquidity provision role.
- **Participant Typology:** We differentiate between different classes of participants such as infrastructure (the AMM itself), heavy users ("whales" or high-frequency traders), and one-off or incidental participants. Criteria include total trade count, total volume, number of distinct markets engaged in, and the fraction of their trades acting as a maker (providing liquidity) versus taker (consuming liquidity).
- **Market Breadth:** We examine how widely each participant spreads their activity across markets. Some traders may specialize in particular categories (only sports, or only crypto-related markets), while others might participate broadly across many unrelated events.

- **Distributional Summary:** We report key summary statistics (mean, median, dispersion) of trades per wallet and volume per wallet to highlight the skewness and heterogeneity in participation. Often, such distributions are heavy-tailed, indicating that a small fraction of users account for a large fraction of activity.

Table A.1 displays the top five wallets by total trading volume (in millions of USDC) and their corresponding activity metrics. For context, we include each wallet's share of overall volume and trades, the percentage of all markets in which they traded, the percentage of their trades where they acted as the liquidity provider (maker), and an indicator of whether the wallet is an AMM contract. The dominance of a few key participants is immediately clear. The single largest participant, with address `0x4bfb...982e`, is an AMM and alone accounts for over 22% of total volume and more than 17% of all trades in the dataset. The second-largest (`0xc5d5...f80a`) is another AMM contract, contributing an additional 3.2% of volume and 3.5% of trades. Combined, these two automated agents provided roughly a quarter of all trading volume and one-fifth of all trade executions. Neither ever acts as a maker (0% maker trades), consistent with the fact that they are always passively taking liquidity and never offering new prices liquidity to other market participants. Furthermore, each spans essentially all markets (by design, the AMM is present in every market), underscoring their role as the backbone of liquidity across the platform.

By contrast, the next three largest traders (by volume) are human or non-AMM wallets. Their volumes (ranging from \$24 million to \$36 million) are an order of magnitude smaller than the top AMM, and their share of trades is below 0.5% each. Notably, these top human traders exhibit a high percentage of maker trades (50–82%), suggesting that they often provide liquidity via limit orders, perhaps to earn trading fees or profit from spread capture. They also participate in a significant number of markets (between 6.8% and 26.3% of all markets), indicating diversified interest or possibly acting as arbitrageurs who seek value across many contracts. In summary, the upper tail of the participant distribution consists of a mix of the platform's built-in liquidity providers and a few very active human traders who behave in liquidity-supplying or high-engagement ways.

For completeness, we note that the bottom deciles of participants are the opposite extreme: the majority of wallets (tens of thousands of addresses) only executed a handful of trades (often just one or two) and in only one market. These are likely casual users testing the platform or taking a one-off position on a specific event.

Individually they contribute negligible volume, but collectively they account for a large portion of the participant count. This long tail of sporadic users has little influence on aggregate market outcomes, though their presence underscores the open-access nature of a decentralized platform where barriers to entry are low.

Wallet	Volume	Trades	% Volume	% Trades	% Markets	% Maker	AMM?
0x4bfb...982e	304.8	899,127	22.12%	17.38%	59.53%	0.00%	Yes
0xc5d5...f80a	44.2	181,459	3.21%	3.51%	40.47%	0.00%	Yes
0x3078...3e11	36.1	22,587	2.62%	0.44%	16.01%	50.58%	No
0xee00...cea1	24.5	11,209	1.78%	0.22%	26.28%	82.29%	No
0xae36...8fb5	24.2	12,528	1.76%	0.24%	6.79%	54.80%	No

Table A.1: Top 5 wallets by total trading volume (in millions of USDC). Percent columns represent each wallet's share of the overall dataset's activity. The two largest participants are Polymarket's automated market maker contracts, which together account for about 25.3% of volume and 20.9% of trades.

To put these figures in perspective, Table A.2 provides summary statistics for the Polymarket dataset as a whole. The platform had over 151,000 unique wallet addresses that engaged in trading during the sample period, executing a total of more than 5.17 million trades and generating roughly $1.378 billion in notional volume. The average (mean) number of trades per wallet is 34.2, but this is heavily skewed by the ultra-active traders; the median number of trades per wallet is only 4. Similarly, the mean volume per wallet is about $9,115, whereas the median volume is a mere $50. This stark disparity between mean and median underlines the skewed distribution of activity, most wallets are lightly used, while a small fraction are extremely active. The standard deviation values are enormous relative to the means (2,404 trades and $816k volume), further highlighting the presence of high-end outliers. There were 1,811 distinct markets in this period, meaning on average each wallet participated in only a tiny fraction of available markets (the median of 1 market per wallet, not shown in table, confirms that most users only bet on a single question).

This exploratory analysis highlights a duality common in many markets: a small core of very active agents (including the platform's own liquidity engine and a handful of heavy traders) versus a long tail of casual participants. The dominance of AMM accounts in volume and their ubiquity across markets suggests that any analysis of price formation must account for the AMM's presence as a counterparty in the majority of trades. Meanwhile, the heterogeneity among human traders, in terms of their activity levels and strategies, motivates a more detailed segmentation. In the

Metric	Value	Description
Total unique wallets	151,191	Individual wallet addresses that traded
Total trades	5,173,094	Count of all buy or sell transactions
Total volume	$1.378 billion	Aggregate notional trading volume (USDC)
Total markets	1,811	Distinct prediction questions (contracts)
Mean trades per wallet	34.2	Median: 4; Std. Dev.: 2,404.8
Mean volume per wallet	$9,114.8	Median: $50; Std. Dev.: $816,665

Table A.2: Summary statistics for overall Polymarket activity. The distributions of trades and volume per participant are highly skewed, as evidenced by the gap between mean and median values, indicating that most participants are small and a few are extremely active.

next section, we formalize this by clustering traders into behavioral categories, which will allow us to examine how different types of participants contribute to market dynamics such as volatility and liquidity provision.

A.4 PARTICIPANT TYPOLOGY AND BEHAVIORAL CLUSTERING

While the aggregate statistics above reveal concentration and heterogeneity, they do not directly tell us about distinct *patterns* of trader behavior. In this section, we identify recurring behavioral profiles at the wallet level by applying unsupervised learning (clustering) to the trading metrics of non-AMM participants. The goal is to see if traders naturally group into categories that resemble the roles known in microstructure theory, such as market makers, informed speculators, noise traders, etc., and to quantify the characteristics of each group.

A.4.1 Cluster Identification Methodology

For each non-AMM wallet (i.e., excluding the two known AMM addresses), we construct a feature vector capturing key dimensions of activity:

1. Total trading volume (USDC) across all markets.
2. Total number of trades executed.
3. Number of distinct markets participated in.

4. Percentage of trades where the wallet was the maker (i.e., provided liquidity via a standing order).
5. Average trade size (volume per trade).
6. Category diversity index (number of different topical categories of markets traded, as a proxy for how broad or specialized the trader's interests are).

Because the distributions of volume, trades, and trade size are highly skewed, we apply logarithmic transformations to those features to reduce outlier influence and then standardize all features to have mean 0 and unit variance. We then use k-means clustering on this standardized data. After experimenting with different values of k, we settled on $k = 4$ clusters as a balance between granularity and interpretability; this choice was supported by an "elbow" in the within-cluster variance plot and by the distinct profiles that emerged.

Table A.3 reports the summary statistics of the resulting four clusters. The clusters, labeled 0 through 3, vary dramatically in size (number of wallets) and in their average behavioral metrics: - Cluster 0 is by far the largest group with 75,756 wallets (about half of all non-AMM traders). However, it has the lowest average volume (only $42.77) and very few trades on average (3.14), covering just 1.3 markets on average. This group's maker/taker split is roughly 51/49, indicating no strong tendency either way. The small size and low activity suggest this cluster consists of *casual or retail traders* who make a few trades in one or two markets, likely with no strategic role. - Cluster 1 has 18,593 wallets. These have higher average volume ($381) but notably very low average trade count (2.63) and markets (1.37). What distinguishes them is the extremely high maker percentage: 97% of their trades are as a liquidity provider. Their average trade size is also relatively large ($148, which is high given they only do 2 trades on average). This indicates *passive liquidity providers*, traders who seldom trade but when they do, it's by posting orders (perhaps maintaining tight spreads or making markets in a selective way). - Cluster 2 includes 44,628 wallets with moderate activity: average volume around $2,062, about 12.4 trades on average across 4 markets. They have a balanced maker ratio (49% maker). Average trade size ($317) is larger than Cluster 1 and roughly similar order of magnitude as Cluster 3 (described next). This cluster seems to represent *tactical traders* or *informed speculators*, they participate more than casuals, spread across a few markets, and use both market and limit orders interchangeably. - Cluster 3 is the elite group of 12,212 wallets. These have staggering average volume ($75.9k) and

trades (266) across an average of 33.5 markets. They engage with a broad swath of the platform and trade very frequently. Their maker share (55%) suggests they both take and provide liquidity roughly in balance. They also have a high average trade size ($286, close to that of Cluster 2). These are likely *institutional-level or systematic traders*, perhaps professional arbitrageurs, syndicates, or bots that operate at scale across many markets.

Cluster	Wallets	Avg Volume	Avg Trades	Avg Markets	% Maker	% Taker	Avg Trade Size	Log Size	Cat. Count
0	75,756	$42.77	3.14	1.31	51%	49%	$16.00	2.05	2.99
1	18,593	$380.96	2.63	1.37	97%	3%	$148.01	3.23	2.99
2	44,628	$2,061.98	12.40	3.96	49%	51%	$317.50	3.40	3.01
3	12,212	$75,885.62	266.35	33.53	55%	45%	$286.31	3.99	3.02

Table A.3: Behavioral clustering of non-AMM wallets using k-means ($k = 4$). Volume is in USDC. Log Size is the $\log_{10}$ of average trade size (included as a feature in clustering). Category count is an index of market diversity (normalized in preprocessing). The clusters highlight a large group of low-activity retail traders (Cluster 0), a group of almost exclusively passive makers (Cluster 1), a mid-tier of active speculators (Cluster 2), and a small group of extremely active, broad-reaching traders (Cluster 3).

A.4.2 Cluster Interpretation: Roles in the Market

The four clusters can be mapped to intuitive market participant archetypes as follows:

- **Cluster 0 – Retail Bettors (Casual Participants):** This cluster represents the modal user in count but not in influence. These traders have minimal engagement, typically a handful of trades in a single market, often as a one-off bet. Their roughly equal mix of maker and taker trades suggests they are not deliberately providing liquidity; rather, they may place whichever order type is convenient (e.g., taking the best available price, or occasionally posting an order if they want a better price). They mirror the retail flow in traditional markets: numerous individuals whose trading is likely driven by personal beliefs or entertainment, not by consistent strategy. In aggregate, they contribute little volume (less than 1% of total), akin to retail traders on stock markets who individually have negligible impact, though collectively their presence provides liquidity demand for the market.

- **Cluster 1 – Passive Liquidity Providers (Selective Makers):** Participants in this cluster behave like market makers or liquidity providers on a small scale. With nearly all their trades being maker orders, they prefer to earn the spread or fee by offering prices to others. The fact that they trade only a couple of times on average but in size suggests they deploy capital strategically. They might be arbitrageurs

or users running automated strategies that post resting orders on outcomes they deem mispriced. They could also include participants who provided liquidity to the AMM pools (if such actions were recorded as trades) or who act similarly to liquidity providers in order books. Economically, this group stabilizes markets by adding depth: their presence can reduce short-term volatility because they buy when others are selling and vice versa, as long as their orders are in place.

- **Cluster 2 – Active Speculators (Opportunistic Traders):** This group is somewhat in the middle: they trade more frequently and in larger size than casuals, and engage in multiple markets, yet they are not as ubiquitous or continuous as Cluster 3. Their balanced maker/taker ratio indicates they both demand and supply liquidity, likely depending on the situation. We interpret them as informed or tactical traders, perhaps individuals or small funds with domain expertise in certain events. They might take positions based on information (acting as takers when news breaks to exploit a mispricing), but also leave orders in the book to capitalize if the price swings (acting as makers opportunistically). They are an important group for price discovery, as their trades are more likely to carry information than those of Cluster 0, but they are not providing as much liquidity as a dedicated market maker would.

- **Cluster 3 – High-Volume Institutions (Whales/Systematic Traders):** This cluster, though small in number, dominates the trading volume. These participants operate at a scale reminiscent of institutional traders or high-frequency trading firms in traditional markets. They engage with dozens of markets (suggesting no narrow focus but rather a systematic coverage) and execute hundreds of trades, potentially daily. The fact that their maker share is around 55% implies they both take and provide liquidity actively; this could correspond to algorithmic strategies that continuously update quotes (market making) while also removing liquidity when profitable opportunities arise (taking). Their large volume suggests they are behind most of the big moves and arbitrage flows on the platform. They likely have the greatest impact on price efficiency, ensuring that prices across related markets stay consistent and that obvious arbitrages are removed. They may also coordinate across Polymarket and other venues (or the wider crypto market) to trade on relevant information.

To visualize how these clusters differ, Figure A.1 plots a two-dimensional projection of wallets by two key behavioral features: average trade size (on the horizontal axis, proxying the capital scale of the trader) and percentage of trades as maker (vertical axis, indicating their liquidity provision tendency). Each point is a wallet

colored by cluster membership, and the cluster centroids are marked and annotated with the role labels discussed above. We see clear separation: Retail bettors (red points, Cluster 0) cluster near the bottom-left, with small trade sizes and mixed liquidity roles; Passive Liquidity providers (orange, Cluster 1) stand out at the bottom-right, with high maker percentages but moderate trade sizes; Active Speculators (green, Cluster 2) occupy the mid-region with intermediate sizes and a spread of maker ratios; and the High-Volume Institutions (blue, Cluster 3) are towards the top, indicating relatively high maker activity combined with large trade sizes (and many points spread widely across due to sheer number of trades they do).

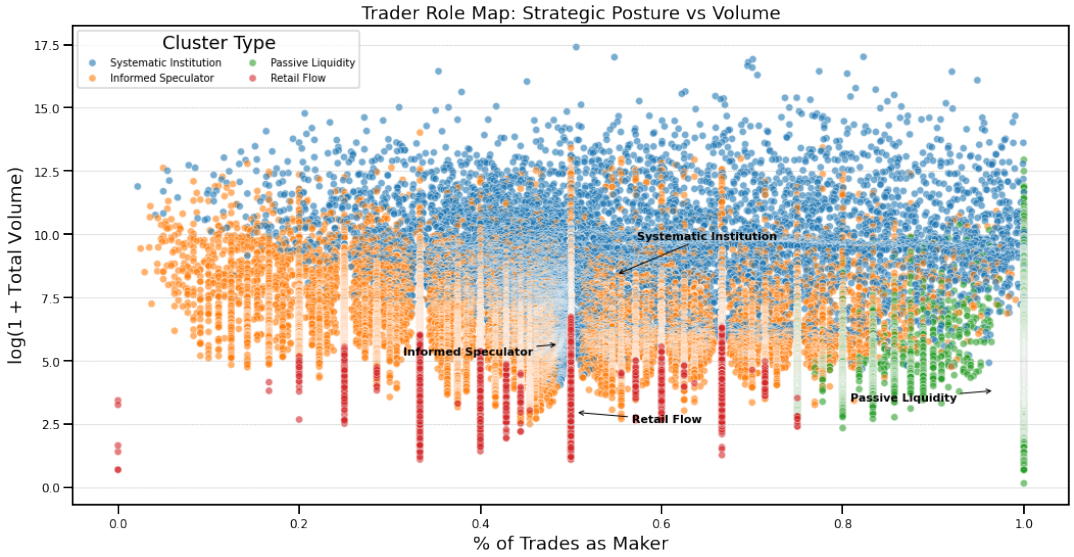


Figure A.1: **Trader wallets embedded in feature space and clustered by behavioral type**. Each point represents a wallet, positioned by average trade size and maker share of trades, and colored by cluster assignment. Cluster centroids are annotated with interpreted trader roles, illustrating distinct behavioral regimes.

The delineation of clusters provides a framework for analyzing market microstructure in Polymarket. We have effectively decomposed the anonymous on-chain traders into functional categories akin to roles one would find on a traditional exchange. In the following subsections, we leverage this classification to study how each trader type contributes to market outcomes, particularly focusing on trading volume distribution, time-series patterns of activity, and the relationship between trader composition and price volatility.

A.4.3 Trading Volume Distribution and Temporal Dynamics

One striking aspect of modern electronic markets is the disproportionate share of volume contributed by a small number of participants, often professional liquidity providers or institutional traders. Our clustering confirms a similar phenomenon on Polymarket. Figure A.2 summarizes the share of total traded volume (left side) and share of total trade count (right side) attributable to each cluster, across the entire dataset. The contrast is stark: the high-volume institutions (Cluster 3, in blue) account for over 90% of the *volume* traded, yet they constitute less than 9% of the *trades*. Inversely, the retail cluster (Cluster 0, in red) makes up more than half of all trades executed (about 54%), but contributes under 1% of total volume. The remaining clusters sit in between: Active Speculators (Cluster 2, green) handle roughly 7-8% of volume and about 26% of trades, whereas Passive Liquidity providers (Cluster 1, orange) contribute about 2% of volume and a negligible fraction of trades (their count is low because they trade infrequently, but when they do, it's often relatively sizable).

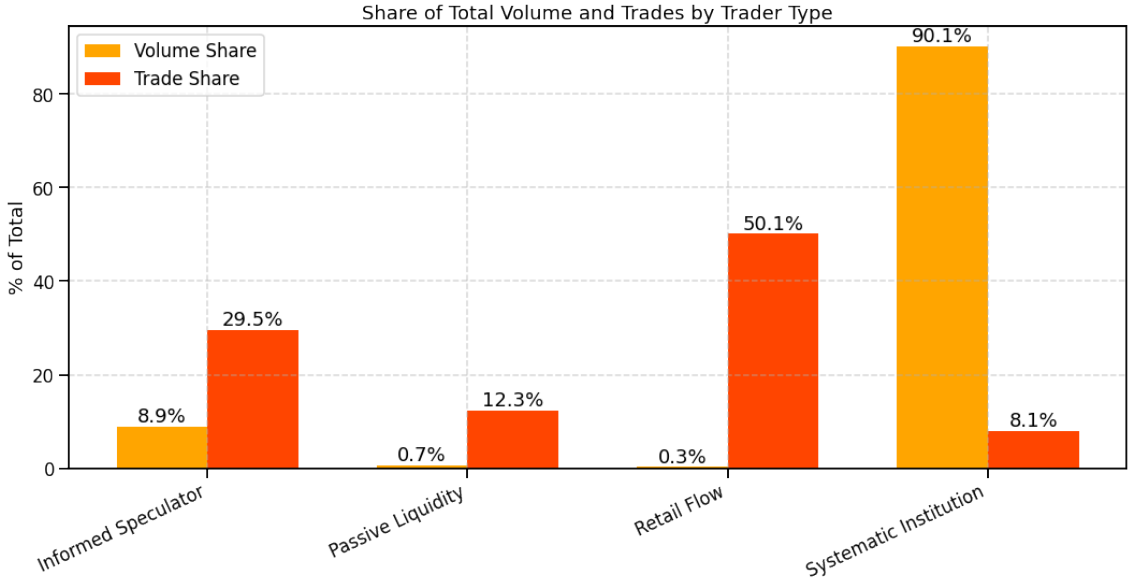


Figure A.2: **Share of total executed volume and number of trades by trader cluster**. Systematic institutional traders dominate executed volume while accounting for a small fraction of trades, whereas retail traders execute many trades but contribute little volume, reflecting a barbell market structure.

This barbell distribution of volume vs. trades parallels observations in equity markets and futures: institutional algorithmic traders or large funds account for the

bulk of dollar volume, while retail orders make up a large portion of transactions (often smaller trades like odd-lots in equities). The implications are significant. Most price formation (which involves significant volume changing hands and moving prices) is driven by the heavy hitters in Cluster 3, whereas the high trade count of Cluster 0 suggests that a lot of message traffic and minor price fluctuations could be driven by retail trades that are low volume and possibly uninformed. The presence of many small trades can create short-term noise in the order flow, which market makers need to absorb or fade, while the large trades by institutions likely reflect more informed shifts or liquidity-driven repositioning.

Next, we examine how the activity of different clusters varies over time and in response to market events. Figure A.3 shows the evolution of daily trading volume shares for each cluster, and distinguishes between whether those trades were passive (maker) or aggressive (taker). Several patterns emerge from this time-series view. First, the blue area (Cluster 3) dominates the maker side on almost every day, indicating that the large systematic traders are the primary suppliers of liquidity day-to-day. This aligns with the idea that professional or algorithmic traders consistently provide depth, stepping in as counterparties to others' trades. On the taker side (the lower panel of the figure), Cluster 3 still holds a large share for much of the time, but we see noticeable periods where the green (Cluster 2) and red (Cluster 0) shares spike. These likely correspond to specific events or news triggers where informed speculators and retail bettors flood in to take positions, consuming liquidity provided by the market makers.

For instance, certain spikes in the red area in the taker volume share might correspond to popular events (like a highly publicized political development or a viral social media moment) that prompted many casual users to trade on Polymarket that day. The informed speculator cluster (green) might spike around times of scheduled news (e.g., an election night or a major sports finals), where those traders act quickly on new information to take positions before prices fully adjust.

In sum, the temporal analysis supports a microstructural interpretation with the following features:

- **Cluster 3 (heavyweights):** These traders are persistently active throughout the sample and participate predominantly on the passive side of the market. Their continuous presence provides baseline liquidity and helps stabilise trading conditions over time.

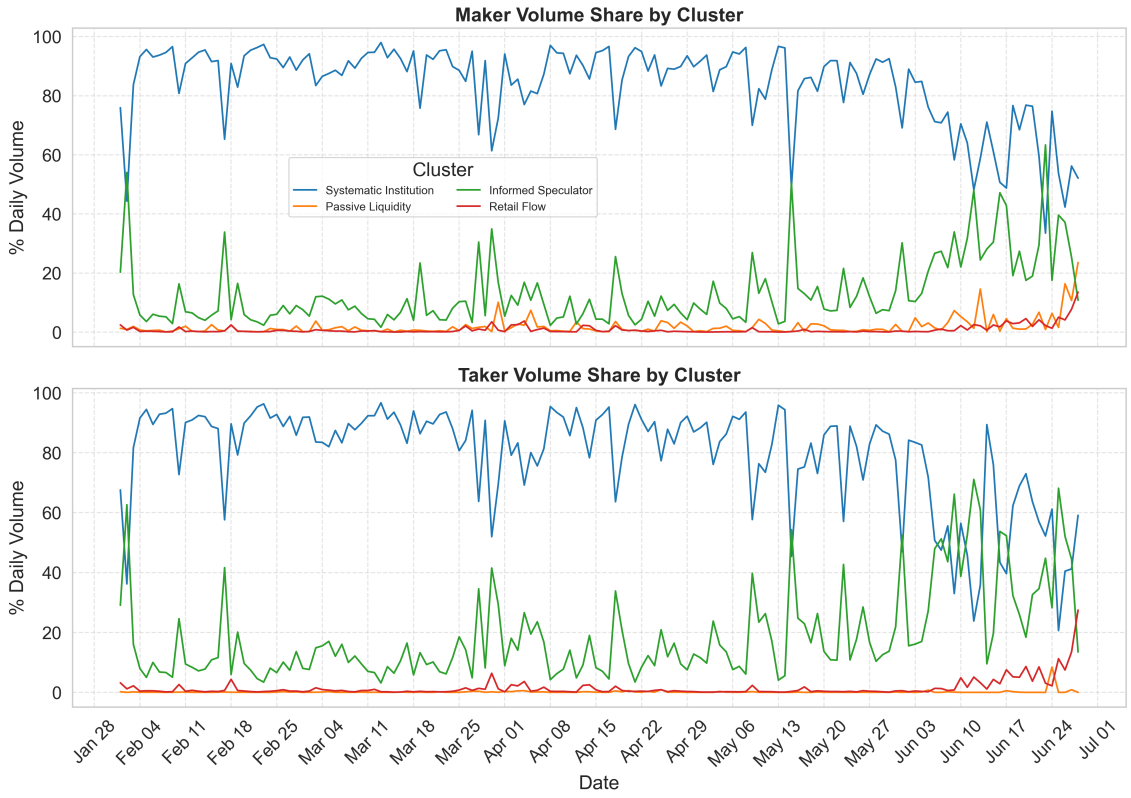


Figure A.3: **Daily trading volume by trader cluster and trade role**. The top panel shows maker (liquidity-supplying) volume and the bottom panel taker (liquidity-demanding) volume. Systematic institutions provide most liquidity, while surges in informed and retail trading coincide with periods of heightened activity.

- **Cluster 2 (informed mid-tier traders):** Activity from this group increases sharply at key moments, consistent with selective participation when new information arrives. Their tendency to take liquidity suggests an important role in short-run price discovery.
- **Cluster 0 (retail traders):** Retail traders are present at a steady background level but exhibit episodic surges in activity during periods of heightened attention or uncertainty. These bursts are associated with increased order flow and short-term volatility.
- **Cluster 1 (passive low-frequency traders):** Due to their low trading intensity, this group is not prominent in daily aggregate volume measures. Their influence is more likely reflected in micro-level order book characteristics such as depth and bid–ask spreads.

Taken together, the time-varying composition of trader activity suggests a link

between market participation and short-run volatility and efficiency. We investigate this relationship directly in the next section by relating trader composition to volatility measures.

A.4.4 Trader Composition and Price Volatility

A core question in market microstructure is how the mix of traders (with different motives and information) influences price variability. If a market is dominated by informed trading, one might expect prices to move quickly to reflect new information, potentially causing volatility in the short run but leading to accurate prices in the long run. Conversely, if a market has a high proportion of noise or liquidity-driven trading, prices might exhibit excessive fluctuations not tied to fundamentals. Given our clustering of trader types, we can quantitatively assess these effects in the Polymarket data.

We conduct a cross-sectional regression analysis at the level of individual markets. For each prediction market (contract) i, we compute its realized price volatility over the course of its trading period. Specifically, we take the standard deviation of the daily price (closing price or last trade price) within that market from inception to resolution. This serves as the dependent variable. The key independent variables are the fraction of that market's total trading volume contributed by each cluster (excluding AMMs, since AMM volume is by design half of every trade and present in all markets). In addition, we control for the total volume in the market (since more active markets might have different dynamics), the total number of trades (a proxy for level of activity and fragmentation of trades), and the number of unique wallets that traded in that market (as a proxy for breadth of participation or interest).

Table A.4 presents the Ordinary Least Squares regression results. Each coefficient can be interpreted as the estimated change in volatility (in standard deviation units of price, where price is between 0 and 1 for a probability) associated with a 1-unit change in the corresponding independent variable, holding others constant. Since cluster volume shares are fractions that sum to 1 (for four clusters), one of them is effectively redundant. We include all four for completeness but note that the coefficients are to be interpreted relative to the omitted effect captured in the intercept.

The regression results reveal several noteworthy relationships: - The coefficient on **Retail Flow share** is positive and highly significant (0.576, $p < 0.001$). This

Table A.4: OLS regression of per-market price volatility on trader composition (volume share by cluster) and controls. Dependent variable: standard deviation of daily market price.

Independent Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	***p*-Value**
Intercept	0.361	0.020	18.00	0.000
Systematic Institution (Clust 3)	−0.032	0.021	−1.53	0.127
Informed Speculator (Clust 2)	−0.021	0.026	−0.79	0.428
Passive Liquidity (Clust 1)	−0.163	0.067	−2.44	0.015
Retail Flow (Clust 0)	**0.576**	0.073	**7.84**	0.000
Total Volume (USDC)	-3.20×10^{-9}	3.17×10^{-9}	−1.01	0.312
Total Trades	6.11×10^{-6}	1.22×10^{-6}	**4.99**	0.000
Unique Wallets	-8.33×10^{-5}	7.84×10^{-6}	**-10.63**	0.000
R^2				0.287

suggests that markets in which a greater proportion of volume comes from Cluster 0 (retail traders) tend to have higher price volatility. In economic terms, a 10 percentage point increase in the retail volume share is associated with roughly a 0.058 increase in the standard deviation of price (which is quite substantial given that prices range 0–1). This finding aligns well with the hypothesis that noise traders contribute to excess volatility [18]. Retail traders may be chasing trends or reacting to events in a non-coordinated way, causing prices to swing more widely. - The share of **Passive Liquidity (Cluster 1)** has a negative coefficient (-0.163) that is statistically significant ($p = 0.015$). This implies that markets with more volume from passive makers experience lower volatility, all else equal. Intuitively, when a larger chunk of trading is due to standing orders by these liquidity providers, the order book is likely thicker and better able to absorb imbalances, dampening price fluctuations. This echoes the idea that liquidity provision stabilizes prices and that liquidity providers require compensation (the spread) for reducing volatility [47]. - The coefficients for **Systematic Institutions** and **Informed Speculators** are negative but not statistically different from zero at conventional levels. This could mean that, conditional on the other factors, the presence of these traders doesn't have a clear linear relationship with volatility across markets. One interpretation is that these traders may gravitate to certain kinds of markets (perhaps higher profile ones that inherently have more news and volatility, offsetting their stabilizing actions). Alternatively, their effect may be already captured by other variables: for example, markets with many systematic traders might also simply be the markets

with more volume or participants (hence the unique wallets variable or total trades could be soaking up some explanatory power). - Among the control variables, **Total Trades** has a positive and significant coefficient, indicating that markets with more transaction traffic tend to have higher volatility. This could reflect that more active markets see more intraday fluctuation (or it could be mechanical: if price jumps a lot, more trades may happen as traders react). - Interestingly, **Unique Wallets** (number of distinct traders in the market) has a significant negative coefficient. This suggests that markets that attract a broader set of participants (higher diversity of traders) have lower volatility. One possible explanation is that when more traders are involved, there is a balance of opinions and sufficient liquidity on both sides to prevent erratic moves, no single trader can dominate the order flow. It might also indicate that niche markets with only a few traders can have wild price swings (due to low competition and potential collusion or manipulation), whereas widely followed markets (e.g., major political events) have a crowd that keeps prices in check to some extent.

These findings complement the general understanding of prediction markets with a microstructure perspective. Prior studies of prediction markets often assume that more trading (volume) or more traders leads to more efficient (hence less volatile) prices, but our results nuance that view: it's not just the quantity of trading that matters, but who is trading. A market buzzing with uninformed or reactive trading can be volatile even if it's "liquid" in a raw volume sense, whereas a market supported by diligent liquidity providers can remain calm despite high trading interest.

A.4.5 Implications for Market Design and Efficiency

Our analysis indicates that heterogeneity across trader types on Polymarket has tangible effects on market quality. In particular, the strong association between retail participation and volatility connects to a broader theoretical literature on noise traders. Noise trader models [18, 30] argue that trading by relatively uninformed participants can generate excess volatility and risk, potentially deterring arbitrage by more informed traders due to the unpredictability of noise-driven order flow. In a prediction market setting, if surges in activity originate disproportionately from retail traders—such as during periods of social media attention or bandwagon effects—prices may be pushed away from underlying probabilities. Relative to better-capitalised or more informed participants, these traders contribute limited

stabilising liquidity, so correcting the resulting mispricing may take time. During this adjustment period, the market exhibits elevated volatility and reduced accuracy.

On the other hand, the presence of passive liquidity providers and systematic traders can enhance efficiency. These actors resemble the "stabilizing speculation" concept proposed by [38], where some traders conjecture about future order flow and step in to provide liquidity, thereby smoothing out prices. In Polymarket, Cluster 1 and Cluster 3 traders fulfill this role in different ways: Cluster 1 by literally posting static liquidity (limit orders or pool liquidity) that buffers price movements, and Cluster 3 by being quick to arbitrage and supply continuous liquidity as part of their strategies. The negative association of their activity with volatility suggests they help enforce the law of one price and keep prices closer to fair values, consistent with microstructure theories where competitive market makers reduce price variance in the face of noisy order flow [38].

It is also important to consider the role of the AMM itself in this ecosystem. Although we excluded the AMM volume shares from the regression (since it's omnipresent), the AMM is essentially a passive liquidity provider across all markets. The fact that volatility varies widely across markets despite every market having an AMM implies that the AMM alone doesn't guarantee uniform outcomes, human trader composition still matters. The AMM sets a baseline of liquidity, but if a market is mostly traded by retail users (bidding prices up and down) the AMM's pricing rule will follow along, adjusting to each trade and thus not preventing volatility. In contrast, if a market has additional liquidity from Cluster 1 or quick corrections from Cluster 3, those can counteract or slow down the price swings that the AMM would otherwise allow. This interplay between algorithmic and human liquidity is a unique aspect of hybrid markets like Polymarket, and our findings highlight that simply having an automated market maker is not a panacea for volatility, active human liquidity is still crucial.

A.4.6 Case Study: A Volatile Geopolitical Market

To make the analysis more concrete, we consider into the most traded market in our dataset: *"Will Ukraine agree to give Trump rare earth metals before April?"*. This politically-charged market attracted over $11 million in volume, far exceeding the typical market on Polymarket. It was essentially a bet on a geopolitical scenario involving the former U.S. president and Ukraine, touching on international relations

and commodities (rare earth metals) , a complex and speculative proposition that likely drew attention due to its sensational framing.

Figure A.4 shows the price trajectory of the "Ukraine/Trump rare earth" market over time (from its inception to its resolution deadline). The price is expressed as the probability (0 to 1 in USDC terms) of the event occurring. We observe that the market price was far from static: there were periods of steady drift (both upward and downward trends can be seen), punctuated by sudden jumps or crashes. For instance, at one point the price spiked sharply, perhaps due to a piece of news or rumor suggesting the event might happen, only to fall back down within days as countervailing information or skepticism set in. Such patterns indicate a mix of information-driven trading and possibly overreaction and correction, which is fertile ground for examining the role of different trader types.

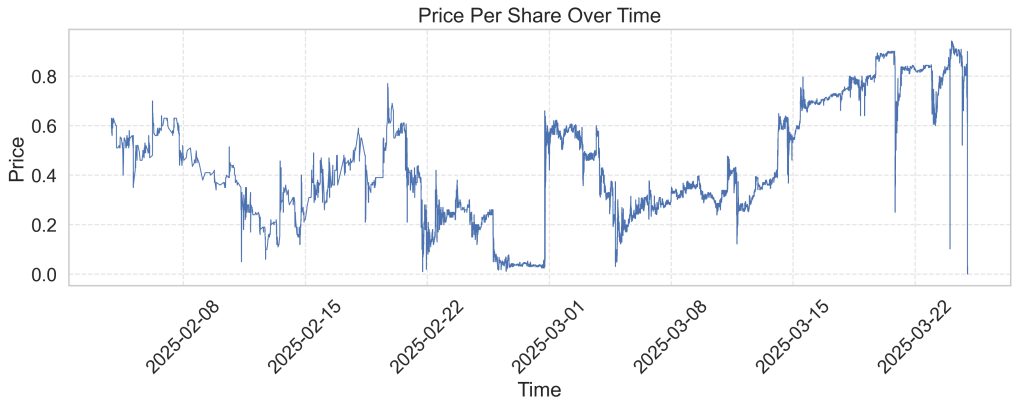


Figure A.4: **Price trajectory for the Ukraine–Trump rare earth metals prediction market**. The implied probability exhibits trending behavior punctuated by sharp movements, consistent with episodic information arrivals and speculative trading.

Next, Figure A.5 shows daily trading volume in this market, disaggregated by trader cluster (excluding the AMM, which mechanically intermediates all trades). Several patterns stand out:

- **High-volume periods:** The largest spikes in total trading volume, occurring in late February through early March, coincide with substantial contributions from the *Systematic Institution* cluster (blue) and the *Informed Speculator* cluster (green). These episodes are consistent with periods of increased information flow or heightened market attention, during which sophisticated traders actively reposition or absorb incoming order flow.

- **Passive liquidity during volatility:** During these same high-volume periods, the *Passive Liquidity* cluster (orange) contributes little observable volume. This suggests that passive market participants play a limited role during volatile episodes, either because resting orders are withdrawn as conditions become uncertain or because their activity is small relative to that of larger traders.
- **Retail participation:** The *Retail Flow* cluster (red) is present throughout the sample but contributes relatively little daily volume. On certain spike days, retail activity increases, likely reflecting a large number of small trades rather than a few large positions. The fact that retail traders never dominate volume indicates that large price movements are driven primarily by higher-capital participants rather than by retail flow alone.

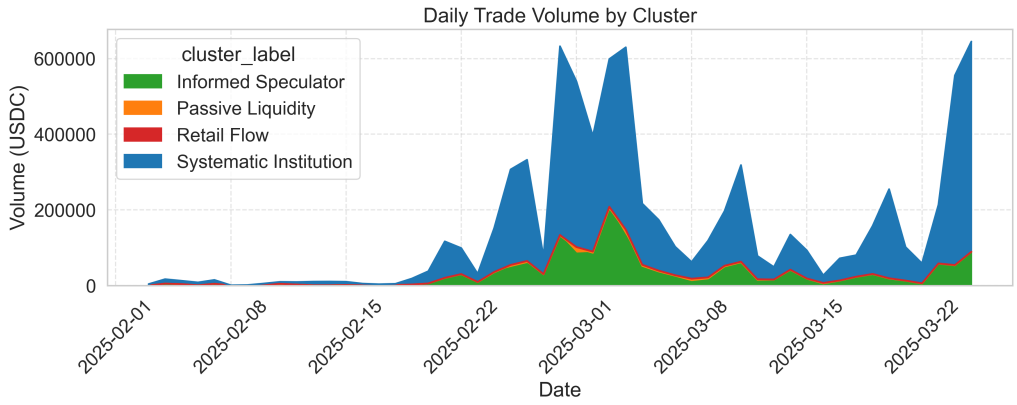


Figure A.5: **Daily trading volume in the Ukraine–Trump rare earth market by trader cluster**. High-volume and informed traders dominate activity on the most volatile days, while passive liquidity provision is limited. Retail participation remains persistent but comparatively small.

From this case study, we infer that major price movements in this market were led by the actions of bigger, likely better-informed traders (Clusters 2 and 3). The scarcity of passive liquidity on volatile days meant that when a wave of orders hit (whether informed or just excited noise), there were few standing orders to absorb them, so prices moved dramatically until the aggressive flow subsided or reversed. The retail cluster's increased activity during the turmoil might have further fueled the swings , for example, if the price started rising, some retail traders could pile on hoping to ride momentum, only to contribute to an overshoot that then crashes.

This narrative matches classic descriptions of how markets behave around news: the initial movers (often informed or professional) take positions, which moves

the price; momentum and noise traders may join, pushing it further; then as more information comes or as pros realize it overshot, they push it back, etc. The prediction market context is no different, except that here an AMM is always a counterparty, adjusting price continuously, and participants are pseudonymous wallets rather than known institutions.

To further dissect the dynamics, we also separate the notion of trade frequency versus trade volume in this market. Figure A.6 shows the daily number of trades (transactions) by each cluster in the same market. Now a different picture emerges for the retail cluster: while their volume share was small, their trade count is relatively large on certain days, reflecting many tiny trades. On March 8, for example, the red bar spikes in trade count terms, even though in volume it was small, this was a day where thousands of micro-transactions (perhaps users betting a few dollars each) took place, predominantly by retail traders. Such frenetic small-scale activity can create short-term volatility (a kind of churning) without contributing much to moving the fundamental price if larger traders are on the other side ensuring it doesn't deviate too far. But if those larger traders momentarily withdraw, even small trades can swing the price in the AMM environment.

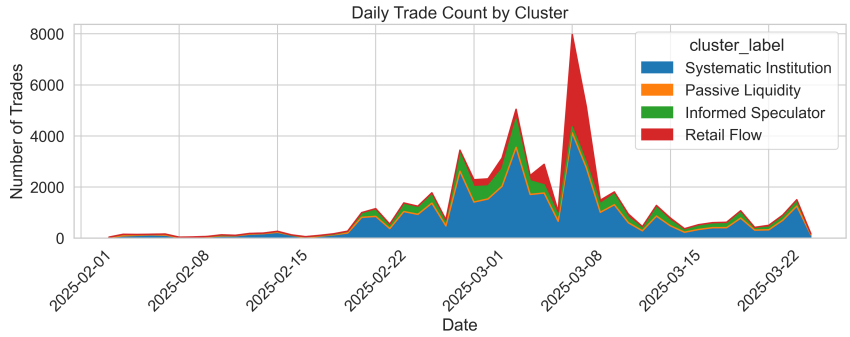


Figure A.6: **Daily number of trades in the Ukraine–Trump rare earth market by trader cluster**. Retail traders account for a large number of trades on certain days despite small volume, whereas institutional and informed traders execute fewer but larger trades.

A.4.7 Volatility Clustering and Trader Behavior

The term "volatility clustering" refers to the empirical observation that high-volatility events in financial time series tend to be temporally clustered, after a big move, prices often continue to be volatile for a while (a hallmark of ARCH/GARCH processes;

see [4]). One explanation for this is the presence of herd behavior or the slow dissemination of information, where an initial shock is followed by secondary waves of trading as more people react. In our context, we can ask: when we see volatility clustering, is it associated with particular types of traders being active?

We approach this by computing a high-frequency volatility measure for the Ukraine Nuclear Deal market and correlating it with the prevailing type of active traders. Specifically, we calculate a rolling volatility σ_t based on 100-trade windows (regardless of clock time) to capture local price variance. Figure A.7 plots this rolling volatility (smoothed with an exponential moving average for clarity) alongside the composition of taker-side volume by cluster in the same rolling window (also smoothed). This visualization allows us to see, for example, if volatility spikes coincide with an increased share of retail trading in that window, or vice versa.

In the figure, one can notice that certain volatility spikes align with shifts in the dominant taker type. A prominent example is annotated around March 8, 2025: a sharp rise in volatility (black line) accompanies a surge in the red area, meaning retail traders formed a larger fraction of the aggressive order flow at that time. Historical records confirm that around March 7-8, 2025, there was a notable geopolitical incident (a missile strike in the Russia-Ukraine conflict) that could have triggered speculative activity in related prediction markets. It appears that many retail traders might have rushed into the market upon hearing of those events, causing a flurry of trades and price fluctuations. This is consistent with the idea of event-driven noise trading: when something newsworthy happens, uninformed traders often overreact or trade on emotion, leading to heightened volatility.

Conversely, periods where the blue cluster (institutions) dominates tend to show more subdued volatility. When the professionals are the main actors, price changes may be more orderly as they possibly enforce arbitrage bounds and provide liquidity.

To further quantify this, we consider a simple classification: for each 100-trade window, identify which cluster contributed the most taker volume, and then examine the average volatility in windows dominated by each cluster. Figure A.8 summarizes this by plotting the mean volatility (and perhaps a confidence interval) for windows where retail was dominant versus where institutions were dominant. The difference is pronounced: windows dominated by retail takers have significantly higher volatility on average than those dominated by systematic takers. In fact, retail-dominated windows had roughly double the volatility of institution-dominated windows in this market. Windows dominated by informed speculators fell in between, leaning higher,

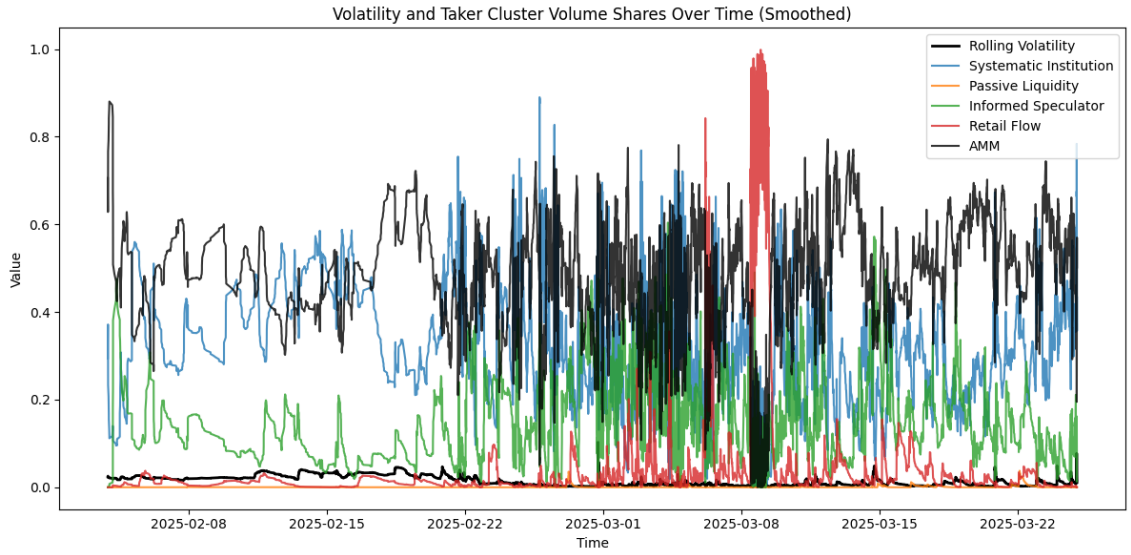


Figure A.7: **Rolling trade-level volatility and composition of taker-side trading**. Volatility is computed over a 100-trade window with EWMA smoothing. Spikes in volatility coincide with increased retail participation on the taker side, while calmer periods are dominated by systematic and informed traders.

whereas those with passive liquidity dominance were too few (since passive rarely "dominates" taker flow) to draw strong conclusions but tended to be low-volatility when they did occur.

Finally, Table A.5 provides a numerical comparison of one of the most extreme retail-driven days (March 8, 2025) with the rest of the days in this market. On March 8, retail traders executed 3,615 trades (versus an average of about 101 trades on other days), and their average trade size was only \$1.06 (versus \$9.64 on other days). In other words, an army of small bets flooded in. The total volume they traded was \$3,823 (compared to a daily average of \$454 outside that day). Interestingly, the price change on March 8 was not a large net move (the average price traded was 0.343 that day vs 0.441 on others, and daily return volatility actually measured lower, 1.77% vs 12.44% on average). The lower return volatility on that specific day, despite high intraday churn, might indicate that although many trades happened, they oscillated around a price level that had already adjusted (perhaps the news had spiked the price prior, and on the 8th itself, lots of trading happened around the new equilibrium). Nonetheless, the presence of thousands of micro-trades is a hallmark of volatility clustering: even if the day's net change was small, the intraday path was likely noisy.

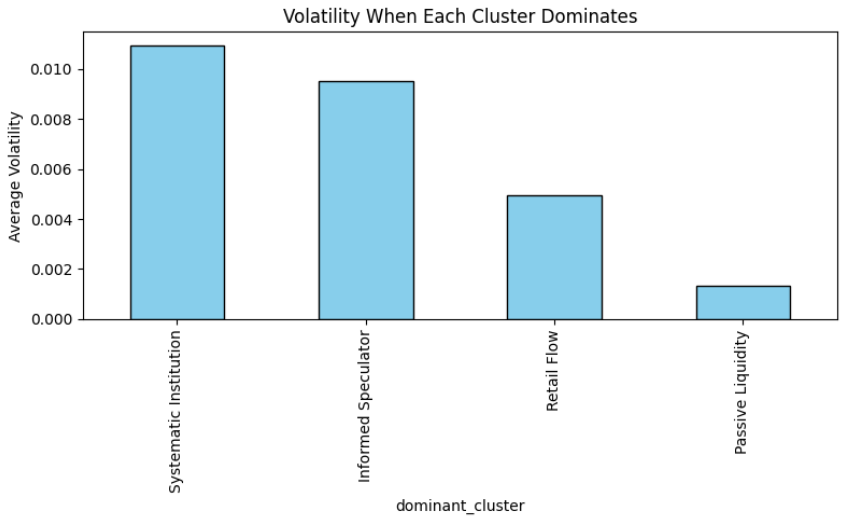


Figure A.8: **Average intra-window volatility grouped by dominant taker-side trader cluster**. Windows dominated by retail flow exhibit the highest volatility, while those dominated by systematic institutional traders show the lowest, linking trader composition to market stability.

Table A.5: Retail Trader Activity on March 8, 2025 vs. Other Days in the Ukraine/Trump Market

Metric	**March 8, 2025 (retail)**	**Other days (retail, avg)**
Trades (count)	3,615	100.6
Avg Trade Size	$1.058	$9.637
Total Volume	$3,823.7	$454.4
Avg Price	0.343	0.441
Daily Price Std	0.018	0.046
Return Volatility (%)	1.77%	12.44%

The case study underscores a few key points: First, that retail traders can dramatically increase market activity in response to salient events, but their trades tend to be small and numerous, contributing more to *volume of transactions* than to *volume of capital*. Second, that high-frequency metrics can show volatility clustering associated with these bursts of activity. And third, that the market's ability to absorb such bursts without large price dislocations depends on the presence of liquidity (whether from the AMM or other traders). In this example, it appears the price had already moved prior to the biggest retail influx, suggesting the informed traders moved first and set a new price, after which retail piled in mostly on one side of the trade (likely buying, given the news, which the AMM or other sellers accommodated

at the new higher price).

A.5 CONCLUSION

This study has examined the trading activity within a large-scale, decentralised prediction market, using data from Polymarket. The results suggest that, while the underlying infrastructure differs from traditional financial exchanges, many features of market microstructure are preserved. Participation remains highly skewed, trader roles appear to be functionally distinct, and shifts in trader composition can be associated with changes in volatility and pricing behaviour.

Automated market makers (AMMs), which underpin trading in these markets, introduce a structural difference in execution—every trade is matched algorithmically rather than through bilateral interaction. Nevertheless, the data indicate that conventional economic mechanisms, such as liquidity provision and information-based trading, continue to play an important role.

Although further work is required to generalise these findings across different markets and time periods, the analysis suggests that monitoring trader composition may provide useful information about market stability. In particular, episodes of elevated retail participation were found to coincide with shifts in volatility patterns, raising the possibility that certain forms of trading activity are predictive of short-term market dynamics.

These observations may be of interest to designers and operators of decentralised markets. While this work does not make normative claims, it points to the potential value of on-chain analytics in identifying emerging regimes of volatility or participation. Future research might explore whether simple interventions—such as adjusting fee parameters or liquidity incentives—can contribute to more stable price discovery without undermining openness or participation.

References

[1] J. Abernethy, Y. Chen, and J. W. Vaughan. "Cost Function Market Makers for Measurable Spaces". In: *Proceedings of the 2013 Conference on Economics and Computation (EC '13)*. ACM, 2013.

[2] R. P. Adams and D. J. MacKay. "Bayesian online changepoint detection". In: *arXiv preprint arXiv:0710.3742* (2007).

[3] Y. Amihud. "Illiquidity and Stock Returns: Cross-Section and Time-Series Effects". In: *Journal of Financial Markets* 5.1 (2002), pp. 31–56.

[4] T. G. Andersen, T. Bollerslev, F. X. Diebold, and P. Labys. "Modeling and Forecasting Realized Volatility". In: *Econometrica* 71.2 (2003), pp. 579–625.

[5] G. Angeris and T. Chitra. "An analysis of Uniswap markets". In: *arXiv preprint arXiv:2006.06344* (2020).

[6] K. J. Arrow, R. Forsythe, M. Gorham, R. Hahn, R. Hanson, J. Ledyard, S. Levmore, J. List, S. Ross, A. Roth, et al. "The promise of prediction markets". In: *Science* 320.5878 (2008), pp. 877–878.

[7] P. Atanasov, P. Rescober, E. Stone, S. A. Swift, E. Servan-Schreiber, P. Tetlock, L. Ungar, and B. Mellers. "Distilling the wisdom of crowds: Prediction markets vs. prediction polls". In: *Management Science* 63.3 (2017), pp. 691–706.

[8] N. Augenblick and M. Rabin. "Belief movement, uncertainty reduction, and rational updating". In: *The Quarterly Journal of Economics* 136.2 (2021), pp. 933–985.

[9] B. M. Barber, S. Lin, and T. Odean. "Resolving a paradox: Retail trades positively predict returns but are not profitable". In: *Journal of Financial and Quantitative Analysis* 59.6 (2024), pp. 2547–2581.

[10] J. Baron, B. A. Mellers, P. E. Tetlock, E. Stone, and L. H. Ungar. “Two reasons to make aggregated probability forecasts more extreme”. In: *Decision Analysis* 11.2 (2014), pp. 133–145.

[11] Barron’s. *Citadel’s Ken Griffin on Markets, the Fed, and Building His Firm for the Next Century. Barron’s News Report*. Reports Citadel Securities executes approximately 25% of U.S. equity trades and $652 bn in daily notional volume. Sept. 2025.

[12] D. Barry and J. A. Hartigan. “A Bayesian Analysis for Change Point Problems”. In: *Journal of the American Statistical Association* 88.421 (1993), pp. 309–319.

[13] M. Basseville and I. V. Nikiforov. *Detection of Abrupt Changes: Theory and Application*. Prentice Hall, 1993.

[14] M. A. Beaumont. “Approximate Bayesian Computation in Evolution and Ecology”. In: *Annual Review of Ecology, Evolution, and Systematics* 41 (2010), pp. 379–406.

[15] J. Berg, F. Nelson, and T. Rietz. “Prediction markets”. In: *The New Palgrave Dictionary of Economics*. Palgrave Macmillan, 2008.

[16] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. 1st ed. Wiley, 2009.

[17] B. Biais, P. Hillion, and C. Spatt. “An empirical analysis of the limit order book and the order flow in the Paris Bourse”. In: *The Journal of Finance* 50.5 (1995), pp. 1655–1689.

[18] F. Black. “Noise”. In: *The Journal of Finance* 41.3 (1986), pp. 528–543.

[19] E. Boehmer, C. M. Jones, X. Zhang, and X. Zhang. “Tracking Retail Investor Activity”. In: *Journal of Finance* 76.5 (2021), pp. 2249–2305.

[20] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time Series Analysis: Forecasting and Control*. 5th ed. Wiley, 2015.

[21] G. W. Brier. “Verification of forecasts expressed in terms of probability”. In: *Monthly Weather Review* 78.1 (1950), pp. 1–3.

[22] J. Broecker and L. A. Smith. “Scoring Probabilistic Forecasts: The Importance of Being Proper”. In: *Weather and Forecasting* 22 (2007), pp. 382–388.

[23] W. Buckley and D. Bell. “Decentralized Prediction Markets: A Case Study of Polymarket”. In: *Journal of Prediction Markets* 16.2 (2022), pp. 47–70.

[24] D. V. Budescu and E. Chen. “Identifying expertise to extract the wisdom of crowds”. In: *Management Science* 61.2 (2015), pp. 267–280.

[25] A. Caplin and J. Leahy. “Aggregation and imperfect competition: On the existence of equilibrium prices”. In: *Econometrica* 59.1 (1991), pp. 25–59.

[26] M. Chakraborty and S. Das. “Market Scoring Rules Act as Opinion Pools for Risk-Averse Agents”. In: *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*. 2015, pp. 3222–3230.

[27] Y. Chen and D. M. Pennock. “Designing Markets for Prediction”. In: *AI Magazine* 31.4 (2010), pp. 42–52.

[28] R. T. Clemen. “Combining Forecasts: A Review and Annotated Bibliography”. In: *International Journal of Forecasting* 5.4 (1989), pp. 559–583.

[29] M. Clements and D. I. Harvey. “Combining probability forecasts”. In: *International Journal of Forecasting* 27.2 (2011), pp. 208–223.

[30] J. B. De Long, A. Shleifer, L. H. Summers, and R. J. Waldmann. “Noise trader risk in financial markets”. In: *Journal of political Economy* 98.4 (1990), pp. 703–738.

[31] R. F. Engle. “The Econometrics of Ultra-High-Frequency Data”. In: *Econometrica* 68.1 (2000), pp. 1–22.

[32] Financial Times. *Jane Street is big. Like, really, really big. Financial Times.* Reports Jane Street held roughly 10.4% of North American equity market share in 2023. June 2024.

[33] T. Foucault. “Order Flow Composition and Trading Costs in a Dynamic Limit Order Market”. In: *Journal of Financial Markets* 2.2 (1999), pp. 99–134.

[34] M. Gangur, J. Walsh, and K. Larson. “Automated Market Makers for Prediction Markets: A Survey”. In: *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2020)*. 2020, pp. 1863–1865.

[35] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. 3rd. Boca Raton: CRC Press, 2013. ISBN: 9781439840955.

[36] C. Genest and J. V. Zidek. “Combining Probability Distributions: A Critique and an Annotated Bibliography”. In: *Statistical Science* 1.1 (1986), pp. 114–135.

[37] S. Gjerstad. “Risk aversion, beliefs, and prediction market prices”. In: *Constitutional Political Economy* 16.4 (2005), pp. 349–364.

[38] L. R. Glosten and P. R. Milgrom. “Bid, ask and transaction prices in a specialist market with heterogeneously informed traders”. In: *Journal of Financial Economics* 14.1 (1985), pp. 71–100.

[39] T. Gneiting and A. E. Raftery. “Strictly proper scoring rules, prediction, and estimation”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378.

[40] I. J. Good. “Rational decisions”. In: *Journal of the Royal Statistical Society, Series B* 14.1 (1952), pp. 107–114.

[41] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[42] R. Hanson. “Combinatorial information market design”. In: *Information Systems Frontiers*. Vol. 5. Springer, 2003, pp. 107–119.

[43] R. Hanson. “Logarithmic market scoring rules for modular combinatorial information aggregation”. In: *Journal of Prediction Markets* 1.1 (2007), pp. 3–15.

[44] J. Hasbrouck. *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press, 2007.

[45] J. Hasbrouck. “Measuring the Information Content of Stock Trades”. In: *The Journal of Finance* 46.1 (1991), pp. 179–207.

[46] Z. Hassoun, N. MacKay, and B. Powell. “Kairosis: A method for dynamical probability forecast aggregation informed by Bayesian change-point detection”. In: *International journal of forecasting* (2025).

[47] T. Hendershott, C. M. Jones, and A. J. Menkveld. “Does algorithmic trading improve liquidity?” In: *The Journal of Finance* 66.1 (2011), pp. 1–33.

[48] M. Himmelstein, P. Atanasov, and D. V. Budescu. “Forecasting forecaster accuracy: Contributions of past performance and individual differences.” In: *Judgment and Decision Making* 16.2 (2021), pp. 323–362.

[49] M. Himmelstein, P. Atanasov, D. V. Budescu, and Y. Han. “The wisdom of timely crowds”. In: *Judgment in Predictive Analytics*. Springer, 2023, pp. 215–242.

[50] C. C. Holmes, F. Caron, J. E. Griffin, and D. A. Stephens. “Two-sample Bayesian nonparametric hypothesis testing”. In: *Bayesian Analysis* 10.2 (2015), pp. 297–320.

[51] D. Johnstone. “Economic Darwinism: Who has the Best Probabilities?” In: *Theory and Decision* 62.1 (2007), pp. 47–96.

[52] J. Keppo and V. A. Satopää. “Bayesian herd detection for dynamic data”. In: *International Journal of Forecasting* 40.1 (2024), pp. 285–301.

[53] A. S. Kyle. “Continuous auctions and insider trading”. In: *Econometrica* 53.6 (1985), pp. 1315–1335.

[54] R. S. Liptser and A. N. Shiryaev. *Statistics of random processes: General theory*. Vol. 394. Springer, 1977.

[55] C. F. Manski. “Interpreting the predictions of prediction markets”. In: *Economics Letters* 91.3 (2006), pp. 425–429.

[56] R. Martin, M. Chakraborty, and S. Kutty. “Timing Is Money: The Impact of Arrival Order in Beta–Bernoulli Prediction Markets”. In: *Proceedings of the 2nd ACM International Conference on AI in Finance (ICAIF 2021)*. 2021, pp. 1–9.

[57] A. Mas-Colell, M. D. Whinston, and J. R. Green. *Microeconomic Theory*. Oxford University Press, 1995.

[58] A. J. Menkveld. “High frequency trading and the new market makers”. In: *Journal of Financial Markets* 16.4 (2013), pp. 712–740.

[59] Nasdaq. *U.S. Equities Market Share and Average Daily Share Volume*. Page shows average daily share volume (billions) by tape. 2025.

[60] H. Ng, L. Peng, Y. Tao, and D. Zhou. “Price Discovery and Trading in Modern Prediction Markets”. In: *SSRN Working Paper* (Jan. 2026). Available at SSRN: https://ssrn.com/abstract=5331995.

[61] M. O’Hara. *Market Microstructure Theory*. Cambridge, MA: Blackwell Publishers, 1995.

[62] A. B. Palley and J. B. Soll. “Extracting the wisdom of crowds when information is shared”. In: *Management Science* 65.5 (2019), pp. 2291–2309.

[63] C. A. Parlour and D. J. Seppi. “Limit Order Markets: A Survey”. In: *Handbook of Financial Intermediation and Banking*. Ed. by A. Boot and A. Thakor. Elsevier, 2008, pp. 63–95.

[64] J. Peabody, S. Das, and M. T. Young. “Price Evolution in a Continuous Double Auction Prediction Market with a Scoring-Rule Based Market Maker”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*. 2015, pp. 924–930.

[65] Posting and Toasting. *Playoff Game Preview: Knicks at Celtics, Game Two, May 7, 2025*. 2025. URL: https://www.postingandtoasting.com/2025/5/7/24425544/playoff-game-preview-knicks-at-celtics-game-two-may-7-2025 (visited on 05/15/2026).

[66] B. Powell, V. A. Satopää, N. MacKay, and P. E. Tetlock. “Skew-adjusted extremized mean: A simple method for identifying and learning from contrarian minorities in groups of forecasters”. In: *Decision* 11.1 (2022), pp. 173–193.

[67] D. Prelec. “A Bayesian truth serum for subjective data”. In: *Science* 306.5695 (2004), pp. 462–466.

[68] R. Ranjan and T. Gneiting. “Combining probability forecasts”. In: *Journal of the Royal Statistical Society* B72.1 (2010), pp. 71–91.

[69] E. Regnier. “Probability forecasts made at multiple lead times”. In: *Management Science* 64.5 (2018), pp. 2407–2426.

[70] R. Roll. “A Simple Implicit Measure of the Effective Bid-Ask Spread in an Efficient Market”. In: *Journal of Finance* 39.4 (1984).

[71] I. Rosu. “A Dynamic Model of the Limit Order Book”. In: *Review of Financial Studies* 22.11 (2009), pp. 4601–4641.

[72] M. S. Roulston. “Performance targets and the Brier score”. In: *Meteorological Applications* 14 (2007), pp. 185–194.

[73] D. B. Rubin. “Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician”. In: *The Annals of Statistics* 12.4 (1984), pp. 1151–1172.

[74] V. A. Satopää, J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. "Combining multiple probability predictions using a simple logit model". In: *International Journal of Forecasting* 30.2 (2014), pp. 344–356.

[75] V. A. Satopää, S. T. Jensen, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. "Probability aggregation in time-series: Dynamic hierarchical modeling of sparse expert beliefs". In: *Annals of Applied Statistics* 8.2 (2014), pp. 1256–1280.

[76] B. A. Schuler, J. P. Murmann, M. Beisemann, and V. Satopää. "Individual foresight: Concept, operationalization, and correlates". In: *International Journal of Forecasting* 41.4 (2025), pp. 1521–1538.

[77] E. Smith, J. D. Farmer, L. Gillemot, and S. Krishnamurthy. "Statistical Theory of the Continuous Double Auction". In: *Quantitative Finance* 3.6 (2003), pp. 481–514.

[78] J. E. Smith. "Time, Times, and the 'Right Time': 'Chronos' and 'Kairos'". In: *The Monist* 53.1 (1969), pp. 1–13.

[79] E. Snowberg, J. Wolfers, and E. Zitzewitz. *Prediction Markets for Economic Forecasting*. Tech. rep. NBER Working Paper No. 18222, 2012.

[80] E. Snowberg, J. Wolfers, and E. Zitzewitz. "Prediction markets, information aggregation, and new directions". In: *NBER Working Paper No. 14614* (2010).

[81] J. Surowiecki. *The Wisdom of Crowds*. Anchor, 2005.

[82] P. C. Tetlock. "Liquidity bias in judgmental forecasting: Evidence from prediction markets". In: *Journal of Economic Behavior & Organization* 67.1 (2008), pp. 253–268.

[83] P. Tzamalikos. *Origen: Philosophy of History & Eschatology*. Brill, 2007.

[84] G. Wawro and I. Katznelson. *Time counts: quantitative analysis for historical social science*. Princeton University Press, 2022.

[85] G. J. Wawro and I. Katznelson. "Designing historical social scientific inquiry: How parameter heterogeneity can bridge the methodological divide between quantitative and qualitative approaches". In: *American Journal of Political Science* 58.2 (2014), pp. 526–546.

[86] R. L. Winkler, Y. Grushka-Cockayne, K. C. Lichtendahl Jr, and V. R. R. Jose. "Probability forecasts and their combination: A research perspective." In: *Decision Analysis* 16.4 (2019), pp. 239–260.

[87] J. Wolfers and E. Zitzewitz. “Prediction markets”. In: *Journal of Economic Perspectives* 18.2 (2004), pp. 107–126.