THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Essays in Market Microstructure

Benjamin Chen

Thesis submitted to the Department of Finance of the London School of Economics and Political Science for the degree of
Doctor of Philosophy

July 2025

To my late grandfather, Chen Xizhong 陈希仲*,*
who, through his art and life, taught me the quiet power of observation – to see the extraordinary in the ordinary, and to truly live and breathe the world as it unfolds. This dissertation is dedicated to his loving memory.

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of approximately 46,742 words.

Statement of Conjoint Work. I confirm that Chapter 3 is jointly co-authored with Emre Ozdenoren, Jiahua Xu and Kathy Yuan. I contributed 25% of this work.

Acknowledgements

My PhD years, though challenging, have been among the most rewarding of my life. This journey would not have been possible without the support and encouragement of many remarkable individuals whom I was fortunate to meet along the way. I am deeply grateful to my advisor, Kathy Yuan, for her patient guidance, invaluable advice, and constant encouragement.

Special appreciation also goes to my co-authors, Emre Ozdenoren and Jiahua Xu. Working alongside them has been an exceptionally rewarding experience.

In addition, I benefited enormously from insightful discussions with faculty members at the LSE, including Ulf Axelson, Cynthia Balloch, Mike Burkart, Constantin Charles, Thummim Cho, Daniel Ferreira, Christian Julliard, Peter Kondor, Dong Lou, Igor Makarov, Ian Martin, Martin Oehmke, Walker Ray, Dimitri Vayanos, Hongda Zhong, and Jean-Pierre Zigrand.

My time at LSE would not have been the same without the many great friends I made. My cohort has been particularly important: Oliver Ashtari, Marcus Gomes de Castro, Yang Guo, Xinchen Ma, Robert Rogers, and Elizaveta Smorodenkova. Your friendship, insightful discussions, and shared experiences have been an integral part of my academic journey.

I would also like to acknowledge the financial support from the London School of Economics, which made this research possible.

Above all, my family has been my greatest source of strength. Without them, I would never have gotten to where I am today.

Abstract

In the first chapter, we investigate the hidden costs associated with Guaranteed Volume-Weighted Average Price (G-VWAP) contracts. Using a continuous-time mean–variance model incorporating both permanent and temporary price impacts, we demonstrate that brokers offering guaranteed execution at seemingly attractive terms exploit their market power through strategic timing of trades. Higher permanent price impact encourages brokers to front-load trades, thereby increasing execution prices and embedding hidden costs within the VWAP benchmark. In contrast, increased temporary impact flattens the broker's trading path, discouraging rapid trades.

In the second chapter, we study the implications of inverted exchanges on liquidity provision, particularly in the presence of high-frequency traders. Inverted exchanges mitigate inefficiencies arising from tick-size constraints by enabling a finer grid. Inverted venues solve the mismatch between an HFT's price priority and a liquidity demander's time priority. The model yields testable predictions on HFT activity and relative exchange trading volumes, which we confirm using high-frequency data.

In the third chapter (co-authored with Emre Ozdenoren, Jiahua Xu and Kathy Yuan), we examine dominant currencies in Decentralized Finance. Using data collected from Uniswap, we analyze the swapping routes between currency pairs. In line with the dominant currency paradigm, we find that safety is a leading dominance attribute during bust periods, while liquidity is more important during booms. We also find that an active money market, market size, and a currency's correlation with transaction costs are important determinants for dominance, suggesting essential design choices for future Central Bank Digital Currencies.

Contents

List of Figures

List of Tables

Chapter 1

Guaranteed VWAP and the Costs of Free Execution

We develop a continuous-time model of broker execution under a Guaranteed Volume-Weighted Average Price (VWAP) contract, highlighting the hidden costs that arise from the broker's market power. Incorporating both permanent and temporary price impacts into a mean–variance framework, we show that higher permanent impact endogenously induces front-loaded trading, as the broker accelerates purchases early on to mitigate long-term upward price pressure. By contrast, increased temporary impact flattens the broker's trading path, discouraging rapid trades and thus spreading execution more evenly over time. These comparative statics underscore how VWAP contracts can mask significant execution costs, ultimately paid by end clients. Our findings shed light on the broker's economic incentive to offer seemingly attractive VWAP guarantees at low or zero fees: by judiciously timing trades, brokers can profit from price impact while still matching the VWAP benchmark.

1.1 Introduction

Over the past two decades, Volume-Weighted Average Price (VWAP) has become an integral benchmark in institutional equity trading, now accounting for approximately 30–40% of institutional order flows across major financial markets. VWAP represents the average price at which a stock trades over a specified time interval, weighted by trading volume, offering investors a transparent metric for evaluating execution quality. This benchmark's popularity has surged due to advancements in electronic trading platforms, increased transparency, and dedicated matching services provided by exchanges, making VWAP central to contemporary financial market structures.

In response to this trend, brokers increasingly offer Guaranteed VWAP (G-VWAP) execution contracts, promising clients execution at or better than the VWAP benchmark, often at minimal or zero explicit fees. This practice presents a fundamental economic puzzle: How can brokers sustainably offer VWAP guarantees at such attractive terms? Understanding the underlying economic incentives and hidden costs in these arrangements is essential, as traditional benchmark evaluation methods often overlook subtle execution costs, thereby obscuring brokers' implicit profit margins.

This paper addresses these critical issues by developing a continuous-time optimal execution model explicitly incorporating both permanent and temporary price impacts within a mean–variance framework. Permanent price impact captures how accumulating inventory permanently affects the equilibrium price level, while temporary impact represents transient deviations in price due to the immediacy and aggressiveness of trading. Thus, our model aligns closely with observations in market microstructure, extensively documented by seminal works such as Almgren and Chriss (2001) and Bertsimas and Lo (1998).

Specifically, we consider a scenario in which a risk-averse broker commits to acquiring a fixed quantity of shares, N, over a predetermined time horizon, $[0, T]$, under a Guaranteed VWAP contract. We model market prices as evolving stochastically, influenced by the broker's inventory and instantaneous trading speed, which jointly determine both permanent and temporary price impacts. Our goal is to characterize explicitly how brokers strategically adjust their execution strategies to optimize their expected profitability, balancing execution costs against risk exposure due to price fluctuations.

Our analysis yields explicit solutions for the broker's optimal trading strategy, revealing key insights into how execution dynamics respond to different forms of price impact. An increase in permanent impact induces brokers to "front-load" their trades, aggressively accumulating shares early to minimize exposure to rising prices throughout the execution interval. Conversely, increased temporary price impact encourages smoother trading schedules, with brokers spreading execution more evenly over time to avoid significant transient price deviations. When there is no temporary market impact, the model simplifies considerably, and the broker opportunistically pushes up the price at times of high volume. The only penalty comes from the inventory risk (deviations from the naive schedule from which the broker optimally deviates point-wise). To satisfy the boundary conditions, they can simply execute a block trade and unwind the additional inventory. Effectively, the temporary price impact prevents outsized VWAP manipulation. These comparative statics contrast sharply with a "naïve" VWAP approach, where brokers simply trade proportionally to market volume, ignoring strategic adjustments to market conditions and inadvertently inflating execution costs.

We further quantify the hidden execution costs of G-VWAP under constant volume.

G-VWAP contracts pay brokers based on the realized market VWAP, incentivizing them to strategically shape the execution path. Even when explicit fees are absent, brokers earn implicit profits by manipulating the timing of trades. By front-loading trades, brokers deliberately elevate early market prices, thus increasing the final VWAP against which they deliver shares, securing profits from hidden costs embedded in execution prices. Risk aversion is the only factor preventing the broker from driving the VWAP benchmark arbitrarily high.

In practice, portfolio managers and traders frequently aim to minimize the average implementation shortfall (IS)—the difference between average execution price and the arrival price—particularly when urgency in order execution is minimal. Under these conditions, traders commonly utilize VWAP algorithms, as these algorithms distribute order execution evenly throughout the trading day without significant regard for execution risk. Traditional IS algorithms, conversely, do not specifically aim to minimize average IS. Rather, their primary objective is to balance execution risk against potential market impact, typically executing orders more rapidly to mitigate risk. This often results in higher average IS. Consequently, traders with low urgency requirements often employ VWAP algorithms in ways that deviate from their original purpose, which is to closely follow the VWAP benchmark. Institutional investors often benchmark their execution against VWAP believing it represents fair market pricing: our findings challenge this assumption by highlighting substantial hidden costs. Investors relying on guaranteed VWAP benchmarks without scrutiny unknowingly subsidize broker profits through elevated market prices induced by strategic execution. Investors should thus carefully assess broker incentives or prefer alternative benchmarks less susceptible to manipulation.

The remainder of the paper is organized as follows. In Section 3.2, we review the related literature. In Section 1.3, we present the model setup. In Section 1.4, we solve the model and analyze the determinants of front-loading. Section 1.5 extends the model to non-linear market impact. Section 1.6 extends the model with a stochastic volume profile. In Section 1.7, we quantify our results with a numerical analysis. Section 3.8 concludes. All proofs are provided in the appendix.

1.2 Related Literature

Our analysis contributes to the growing literature on optimal trade execution and the economic incentives underlying widely-used trading benchmarks, particularly VWAP contracts. Closely related to our work is the paper by Baldauf et al. (2024), which examines optimal contracting in block trading contexts. They analyze agency conflicts arising when clients outsource large executions to dealers, emphasizing how price impact and hidden actions create subtle contractual dynamics. Our model extends this line of reasoning by specifically analyzing brokers' incentives under Guaranteed VWAP (G-VWAP) arrangements, a popular yet understudied form of contract.

Early seminal contributions to the literature on execution costs include Bertsimas and Lo (1998) and Almgren and Chriss (2001), who introduced foundational continuous-time frameworks for analyzing optimal trading strategies considering both permanent and temporary price impacts. Further extensions by Almgren and Chriss (1997) focused on optimal liquidation strategies, providing insights into strategic trade-offs between execution speed and price impact. Our approach builds upon this fundamental structure but shifts the focus explicitly to VWAP-based contractual arrangements and the hidden execution costs inherent in these benchmarks.

Previous literature has extensively studied optimal VWAP execution strategies, such as Konishi (2002), McCulloch and Kazakov (2012), and Frei and Westray (2015). These studies primarily emphasized the trader's perspective, focusing on the minimization of trading costs relative to a known benchmark without explicitly modeling broker incentives under guaranteed benchmarks. In contrast, our model incorporates a broker-centric perspective, explicitly characterizing how brokers strategically adjust their execution paths under G-VWAP contracts to exploit price impacts for implicit profit.

Price manipulation and strategic trading behavior have been analyzed extensively in various market contexts. Bernhardt and Taub (2008) study front-running dynamics, analyzing informed traders' incentives to trade ahead of large liquidity demands, a concept related to—but distinct from—the strategic front-loading behavior we document under VWAP contracts. Röell (1990) addresses dual-capacity trading and its implications for market quality, providing a foundational understanding of how intermediaries' dual roles can affect execution outcomes. Similarly, Alfonsi and Acevedo (2014) and Klöck et al. (2011) explore optimal execution and manipulation within limit order book and dark pool contexts. While these papers examine manipulative behavior broadly, our analysis provides specific economic rationale and quantitative insights into brokers' strategic deviations from naïve VWAP benchmarks.

Recent research on benchmark design, notably Duffie and Dworczak (2021), emphasizes the robustness of financial benchmarks against manipulative behavior. While they provide a general framework for designing manipulation-resistant benchmarks, we contribute to this dialogue by illustrating precisely how guaranteed benchmarks like VWAP can be manipulated through strategic execution timing, reinforcing the necessity of thoughtful benchmark design.

Finally, our paper complements studies on principal trading arrangements by Baldauf et al. (2021) and optimal benchmark choices by Frei and Mitra (2020), as well as the incorporation of order-flow dynamics into optimal execution strategies discussed in Cartea and Jaimungal (2015). Collectively, this literature highlights the complex interactions between execution strategies, benchmark choices, and market microstructure.

Our work significantly advances understanding of guaranteed benchmark execution, specifically revealing hidden cost mechanisms underlying seemingly attractive broker offerings. By highlighting brokers' incentives to deviate strategically from naïve execution, we provide practical insights for institutional investors, regulators, and trading platform designers seeking to enhance execution transparency and market efficiency.

1.3 Model Setup

1.3.1 Trading Environment

We model a continuous-time trading environment over a fixed execution interval $[0, T]$. A broker is contracted to purchase a total quantity of shares, denoted by N, on behalf of a client. The broker begins the execution window without inventory and must exactly achieve the target inventory level N by the terminal time T. Formally, we define the broker's inventory at time $t \in [0, T]$ as $x(t)$, which satisfies the boundary conditions

$$x(0) = 0, \quad x(T) = N, \tag{1.1}$$

The broker's *control* is the *trading speed*, $u(t) := \dot{x}(t)$, for $0 \leq t \leq T$. Thus, $x(t)$ evolves via

$$x(t) \ = \ x(0) \ + \ \int_0^t u(s)\, ds.$$

The broker executes trades continuously in a limit order market, with transaction prices subject to both permanent and temporary price impacts. Throughout our analysis, we assume that the broker is risk-averse, seeking to minimize expected execution costs while also controlling the variance of those costs.

1.3.2 Market Model

We adopt a standard linear price-impact model in continuous time, which closely follows the canonical market microstructure framework of Almgren and Chriss (2001). Specifically, the market price of the security at time t, denoted by $S(t)$, evolves according to:

$$S(t) = S(0) + \lambda\, x(t) + \eta\, \dot{x}(t) + \sigma\, B_t, \tag{1.2}$$

where

- $S(0)$ is the initial market price at time $t = 0$.
- $\lambda > 0$ captures the *permanent impact* per unit of inventory,
- $\eta > 0$ captures the *temporary impact* per unit of trading speed,
- $\sigma > 0$ is the price volatility coefficient,
- B_t is a standard Brownian motion with zero drift.

The parameter $\lambda > 0$ captures the permanent price impact of accumulated inventory. Intuitively, as the broker accumulates shares, their inventory permanently shifts the equilibrium market price upward. The parameter $\eta > 0$ measures temporary price impact, which captures transient price deviations directly proportional to the instantaneous trading speed. High trading speed temporarily pushes prices away from fundamental levels, whereas slower trading mitigates such transient deviations.

Finally, $\sigma > 0$ is the volatility parameter, and B_t is a standard Brownian motion with zero drift, representing random fluctuations in the market price unrelated to the broker's activity.

1.3.3 Guaranteed VWAP

We introduce a deterministic[1] market trading volume profile $V(t) \geq 0$ for $t \in [0, T]$. Let $V_c(t) = \int_0^t V(u)du$ be the cumulative volume function. Thus, the total market volume over the execution interval is given by

$$V_c(T) = \int_0^T V(t)\,dt, \tag{1.3}$$

The Volume-Weighted Average Price (VWAP) over the interval $[0, T]$ is defined naturally as the total value traded divided by the total volume:

$$\mathrm{VWAP}_{[0,T]} = \frac{\displaystyle\int_0^T S(t)\,V(t)\,dt}{\displaystyle\int_0^T V(t)\,dt} = \frac{1}{V_c(T)} \int_0^T S(t)\,V(t)\,dt, \tag{1.4}$$

where $S(t)$ is the observed market price at time t.

[1] For analytical tractability, we have assumed that the volume is deterministic. In reality, practitioners often use volume forecasts to execute VWAP orders. Here, we can interpret the volume function as the expected volume. This simplification neglects the volume variance risk, which we fully acknowledge. In Section 1.6, we show that the stochastic volume model has similar front-loading properties.

1.3.4 Mean-Variance Objective

Under the Guaranteed VWAP (G-VWAP) contract, the broker commits to delivering the acquired N shares at the *realized VWAP*, receiving total revenue:

$$N \times \mathrm{VWAP}_{[0,T]} \;=\; \frac{N}{V_c(T)} \int_0^T S(t)\, V(t)\, dt,$$

The broker's cost arises from purchasing these shares on the market, paying the instantaneous market price $S(t)$ whenever trading at rate $\dot{x}(t)$. Thus, the broker's profit-and-loss (PnL) from executing the VWAP order is stochastic and depends crucially on their chosen trading strategy. The broker's objective is to maximize a standard mean–variance utility criterion [2] that captures both the expected cost and the riskiness (variance) of the execution strategy:

$$\max_{x(\cdot)} \Big\{ \mathbb{E}[\Pi[x]] \;-\; \tfrac{\gamma}{2}\, \mathrm{Var}[\Pi[x]] \Big\}, \tag{1.5}$$

where $\Pi[x]$ denotes the broker's PnL under the trading strategy $x(\cdot)$, and $\gamma > 0$ is the broker's risk-aversion coefficient. In the following, we contrast this optimization against a "naïve" VWAP execution strategy, under which the broker simply trades in direct proportion to the market's volume profile:

1.3.5 Optimization Problem

The broker seeks to balance expected execution costs against risk (variance) in a standard mean-variance optimization problem.

Expected Cost. Under the assumed linear price impact model (1.2), the broker's expected cost of acquiring the shares is given by:

$$\int_0^T \mathbb{E}[S(t)]\dot{x}(t)dt = \int_0^T \left[S(0) + \lambda x(t) + \eta \dot{x}(t)\right] \dot{x}(t)dt, \tag{1.6}$$

where $x(t)$ denotes the broker's inventory at time t.

[2] Mean-variance utility is standard in optimal-execution models because it generates tractable linear-quadratic control problems (Almgren and Chriss, 1997). When combined with linear or power-law price impact, the optimal schedule satisfies the Huberman and Stanzl (2004) no-dynamic-arbitrage condition: the broker's round-trip trades yield non-positive expected PnL.

Profit Variance. Applying Itô's isometry, the variance of the broker's profit-and-loss can be expressed as:

$$\mathrm{Var}[\Pi[x]] = \sigma^2 \int_0^T \left(\frac{N}{V_c(T)} \int_0^t V(u), du - x(t) \right)^2 dt. \tag{1.7}$$

Consequently, the broker's optimization problem can be stated as maximizing the following mean-variance functional:

$$\mathcal{J}[x] = \int_0^T \left(\frac{N}{V_c(T)} V(t) - \dot{x}(t) \right) \bar{S}(t) dt - \frac{\gamma}{2}\sigma^2 \int_0^T \left(\frac{N}{V_c(T)} \int_0^t V(u) du - x(t) \right)^2 dt, \tag{1.8}$$

where $\bar{S}(t)$ represents the deterministic component of the price, specifically $S(0) + \lambda x(t) + \eta \dot{x}(t)$.

1.4 Optimal Trading Schedule

Before deriving the broker's optimal execution path, we first examine a simpler benchmark known as the *naive VWAP approach*. Under this strategy, the broker executes trades in exact proportion to the market's volume profile, thereby accumulating shares at the same rate as the overall market. Although such a policy may appear intuitive, it ignores potential benefits of adjusting execution speed in response to market impact parameters.

1.4.1 Naive VWAP Approach

To implement the naive VWAP policy, the broker matches the volume curve by maintaining the trading speed

$$\dot{x}(t) \;=\; \frac{N}{V_c(T)} V(t), \quad 0 \;\le\; t \;\le\; T,$$

where N is the total number of shares the broker must purchase over the interval $[0, T]$. Integrating this rate from 0 to t yields

$$x(t) \;=\; \int_0^t \dot{x}(u)\, du \;=\; \frac{N}{V_c(T)} \int_0^t V(u)\, du,$$

thereby ensuring that, at any time t, the broker's inventory tracks the cumulative market volume proportionally.

Under this naive schedule, the expected cost of purchasing the N shares reflects both fundamental and impact-driven components. Formally, the expected VWAP paid by the broker is

$$\mathbb{E}[\text{VWAP}] \;=\; \int_0^T \mathbb{E}[S(t)]\,\dot{x}(t)\,dt,$$

where $\mathbb{E}[S(t)] = S(0) + \lambda\,x(t) + \eta\,\dot{x}(t)$ accounts for permanent impact (λ) arising from the broker's inventory level and temporary impact (η) due to instantaneous trading speed. Substituting $\dot{x}(t) = \frac{N}{V_c(T)}\,V(t)$ and rearranging, one can identify three distinct cost components:

Base Cost Term:

$$\int_0^T S(0)\,\frac{N}{V_c(T)}\,V(t)\,dt \;=\; S(0)\,N.$$

This term is simply the initial price multiplied by the total number of shares purchased.

Permanent Impact Term:

$$\lambda\,\frac{N}{V_c(T)}\;\cdot\;\frac{N}{V_c(T)}\int_0^T \Big[\int_0^t V(u)\,du\Big] V(t)\,dt.$$

Here, the broker's growing inventory shifts the equilibrium price upward, magnifying subsequent trading costs.

Temporary Impact Term:

$$\eta\,\frac{N}{V_c(T)}\;\cdot\;\frac{N}{V_c(T)}\int_0^T V(t)^2\,dt.$$

This cost reflects short-lived price deviations induced by trading at each instant.

Proposition 1 (Naive VWAP Cost)**.** Under a naive VWAP strategy, the expected VWAP per share is given by

$$\mathbb{E}\big[\text{VWAP}_{\text{naive}}\big] \;=\; S(0) \;+\; \frac{N}{V_c(T)^2}\left(\lambda\int_0^T \big[\int_0^t V(u)\,du\big]\,V(t)\,dt \;+\; \eta\int_0^T V(t)^2\,dt\right).$$

The key insight is that, although naive VWAP execution may appear to eliminate tracking error relative to the market's volume curve, it does not control for price impacts or mitigate the resulting risk exposure. Consequently, brokers who rely exclusively on naive VWAP strategies either charge a risk premium or face exposure to potentially volatile execution costs.

1.4.2 General Solution

We now characterize the broker's optimal trading path, which balances immediate price impact against prolonged upward price pressure. This balance leads to systematic deviations from the naive VWAP schedule, especially in the presence of nontrivial permanent and temporary impact parameters. The theorem below states the solution in closed form.

Theorem 1 (Optimal Trading Schedule)**.** Define

$$\alpha \;:=\; \sqrt{\tfrac{\gamma\,\sigma^2}{2\,\eta}},$$

and consider the mean–variance problem in (1.8) with boundary conditions $x(0) = 0$ and $x(T) = N$. The unique optimal trading schedule $x^*(t)$ is given by the trading speed

$$\dot{x}^*(t) \;=\; \dot{x}_p(t) \;+\; \Big[N \;-\; x_p(T)\Big]\,\frac{\alpha\,\cosh\!\big(\alpha\,t\big)}{\sinh\!\big(\alpha\,T\big)},$$

and the corresponding inventory

$$x^*(t) \;=\; x_p(t) \;+\; \Big[N \;-\; x_p(T)\Big]\,\frac{\sinh\!\big(\alpha\,t\big)}{\sinh\!\big(\alpha\,T\big)}. \tag{1.9}$$

Here, $x_p(t)$ is a particular solution obtained through the Green's function method:

$$x_p(t) \;=\; \int_0^T G(t,s)\,\Big[\tfrac{1}{2\,\eta}\,f(s)\Big]\,ds,$$

where

- $G(t,s)$ is the Green's function for the second-order operator $\frac{d^2}{dt^2} - \alpha^2$, given by

$$G(t,s) \;=\; \frac{1}{\alpha\,\sinh\!\big(\alpha\,T\big)} \times \begin{cases} \sinh\!\big(\alpha\,t\big)\,\sinh\!\big(\alpha\,(T-s)\big), & \text{if } t \le s, \\ \sinh\!\big(\alpha\,s\big)\,\sinh\!\big(\alpha\,(T-t)\big), & \text{if } t > s, \end{cases}$$

- the forcing term $f(t)$ is

$$f(t) \;=\; -\,\eta\,\frac{N}{V_c(T)}\,V'(t) \;+\; \lambda\,\frac{N}{V_c(T)}\,V(t) \;-\; \gamma\,\sigma^2\,\frac{N}{V_c(T)}\,V_c(t).$$

This representation explicitly shows how the broker's inventory and trading speed adjust over time to trade off immediate (temporary) impact versus long-term (permanent) price pressure, while also accounting for execution risk.

Notice that this solution generalizes Almgren and Chriss (2001) framework, as our model retrieves their optimal trading for a Market On Close (MOC) benchmark, the price at which the trading session closes. In effect, it is equivalent to a VWAP benchmark where the volume is a Dirac mass at time T. That is, we set

$$V(t) = 0 \quad \text{for } t < T, \qquad V(T) = \delta_T, \tag{1.10}$$

so that the cumulative volume

$$V_c(t) = \int_0^t V(u)\,du$$

satisfies $V_c(t) = 0$ for $t < T$ and jumps at T. Consequently, the Volume-Weighted Average Price (VWAP) becomes

$$\text{VWAP} = \frac{\int_0^T S(t)V(t)\,dt}{\int_0^T V(t)\,dt} = S(T). \tag{1.11}$$

Corollary 1 (Almgren-Chriss)**.** Under the MOC benchmark, we obtain the optimal inventory trajectory:

$$x^*(t) = N\left[1 - \frac{\sinh\big(\alpha(T-t)\big)}{\sinh\big(\alpha T\big)}\right].$$

Instead of a liquidation scenario, this results corresponds to the optimal inventory when the brokers needs to acquire shares instead. More generally, for any volume profile, once $x^*(t)$ is determined, both the expected benchmark price and the execution risk can be directly computed.

Proposition 2 (Optimal VWAP Schedule)**.** Given the optimal trading schedule $x^*(t)$, the expected VWAP per share satisfies

$$\mathbb{E}[\text{VWAP}^*] \;=\; S(0) \;+\; \frac{1}{V_c(T)}\Big[\eta\, N\, V(T) \;+\; \int_0^T x^*(t)\,\big(\lambda\, V(t) \;-\; \eta\, V'(t)\big)\,dt\Big],$$

and the variance of the broker's profit is

$$\text{Var}\Big[\Pi\big[x^*\big]\Big] \;=\; \sigma^2 \int_0^T \Big(\frac{N}{V_c(T)}\int_0^t V(u)\,du \;-\; x^*(t)\Big)^2 dt.$$

1.4.3 Discussion

In the following, we characterize the behavior of the optimal inventory.

Effect of Permanent Price Impact λ The parameter $\lambda > 0$ governs *permanent* price impact, whereby an accumulated inventory $x(t)$ permanently shifts the price dynamics upward by $\lambda\, x(t)$. Although the boundary condition $x^*(T) = N$ remains fixed regardless of λ, the path $x^*(t)$ for $t < T$ generally shifts in response to changes in λ. Differentiating with respect to λ yields

$$\frac{\partial x^*(t)}{\partial \lambda} = \frac{\partial x_p(t)}{\partial \lambda} - \frac{\partial x_p(T)}{\partial \lambda}\, \frac{\sinh\big(\alpha\, t\big)}{\sinh\big(\alpha\, T\big)}.$$

Observe that $\sinh(\alpha t)/\sinh(\alpha T)$ is small when t is near 0. Consequently, early in the trading horizon, any positive increment $\partial x_p(t)/\partial\lambda$ is largely unaffected by the subtraction term, implying

$$\frac{\partial x^*(t)}{\partial \lambda} > 0 \quad \text{for small } t.$$

In other words, *an increase in* λ leads the broker to accumulate shares more aggressively at early times, thus "front-loading" the strategy. Intuitively, if permanent impact is costly, the broker prefers to purchase earlier so that the position is established before the price is too greatly affected by large inventories.

Fixing the market impact parameter, if the broker anticipates large future volume (e.g., a "volume spike") at some $s < T$, the permanent impact component $\lambda\, x(t)$ can provide an incentive to acquire shares before that high-volume window.

Note that for tractability, we have assumed that market impact and volume are independent. This is, of course, false, as temporary market impact should be lower with higher volume. However, it is not exactly clear how volume affects the permanent component. Including a volume component in λ and η would not materially change the main results of this paper, only the magnitude.

Effect of Temporary Price Impact η. The parameter $\eta > 0$ measures the *temporary* impact associated with rapid trading, reflected by the term $\eta\, \dot{x}(t)$ in the price. Unlike λ, which permanently shifts the price, η penalizes bursts of high trading speed. Recall that $\alpha = \sqrt{\frac{\gamma\sigma^2}{2\,\eta}}$, which implies α *decreases* with increasing η. When α is smaller, the hyperbolic ratio $\sinh(\alpha t)/\sinh(\alpha T)$ varies more slowly in t. Consequently, the homogeneous component of $x^*(t)$ flattens, reducing any strong front-loading or back-loading tendencies. When $\eta = 0$, the optimization problem simplifies considerably, and the optimal inventory trajectory becomes:

$$x^*(t) = \frac{N}{M} V_c(t) - \frac{\lambda N}{\gamma\sigma^2\, M}\Big[V(t) - V(0)\Big], \quad 0 \le t < T,$$

with the corresponding trading speed

$$\dot{x}^*(t) = \frac{N}{M} V(t) - \frac{\lambda N}{\gamma\sigma^2\, M} V'(t) + \frac{\lambda N}{\gamma\sigma^2\, M}\Big[V(T) - V(0)\Big]\delta(t - T),$$

where $\delta(\cdot)$ denotes the Dirac delta function. In the absence of temporary impact, the broker is no longer discouraged from trading rapidly. The only penalty comes from the inventory risk (deviations from the naive schedule). To gain from the permanent impact, the broker optimally deviates point-wise.

Remark that an impulse function appears in the solution. There is no cost to trading infinitely fast, so the optimal strategy uses a block trade to jump to the required terminal position. The size of the final block trade is given by

$$\Delta x(T) = \frac{\lambda N}{\gamma \sigma^2 M}\Big[V(T) - V(0)\Big]$$

When temporary impact is present (i.e., $\eta > 0$), the objective includes a term penalizing trading speed—roughly, a cost proportional to $\eta\, \dot{x}(t)^2$. This regularizes the problem and forces the optimal trading schedule to be smooth so that large, instantaneous trades are avoided. However, when $\eta = 0$, the speed–penalty drops out. In this case, the functional depends only on $x(t)$ (and not on $\dot{x}(t)$), so the interior optimality condition is determined point–wise. But such a point–wise optimum typically does not automatically satisfy the boundary conditions $x(0) = 0$ and $x(T) = N$.

Effect of Order Size N The total shares to be acquired, $N > 0$, enter the solution both in the boundary condition $x^*(T) = N$ and in the particular solution $x_p(t)$.

Differentiating with respect to N yields

$$\frac{\partial x^*(t)}{\partial N} \;=\; \frac{\partial x_p(t)}{\partial N} \;+\; \frac{\sinh\big(\alpha\, t\big)}{\sinh\big(\alpha\, T\big)}\Big[1 \;-\; \frac{\partial x_p(T)}{\partial N}\Big].$$

Because $x_p(t)$ is (under typical linear market models) itself linear in N, increasing N typically shifts the entire trajectory $x^*(t)$ upward uniformly in t. In effect, the shape of the trading strategy remains similar; the amplitude simply increases to accommodate a larger total acquisition.

Effect of Risk Aversion γ**.** The parameter $\gamma > 0$ enters through the mean–variance objective, controlling the penalty on variance. In the price-dynamics model, increasing γ intensifies the broker's aversion to price uncertainty.

An increase in γ also increases $\alpha = \sqrt{\gamma\sigma^2/(2\eta)}$, which, by itself, steepens the homogeneous term $\sinh(\alpha t)/\sinh(\alpha T)$ and might encourage more front-loaded trading. However, the variance penalty also contains a term that incentivizes the broker to stay closer to the cumulative market volume path $\int_0^t V(u)\, du$. Since deviating too far from the market's natural volume profile can increase risk (as measured by variance), a higher γ counterbalances the front-loading tendency. In practical terms, the broker becomes more cautious, trading in a manner that reduces volatility of PnL. Empirically, this often means a more moderate accumulation pace rather than fully exploiting short-term cost advantages.

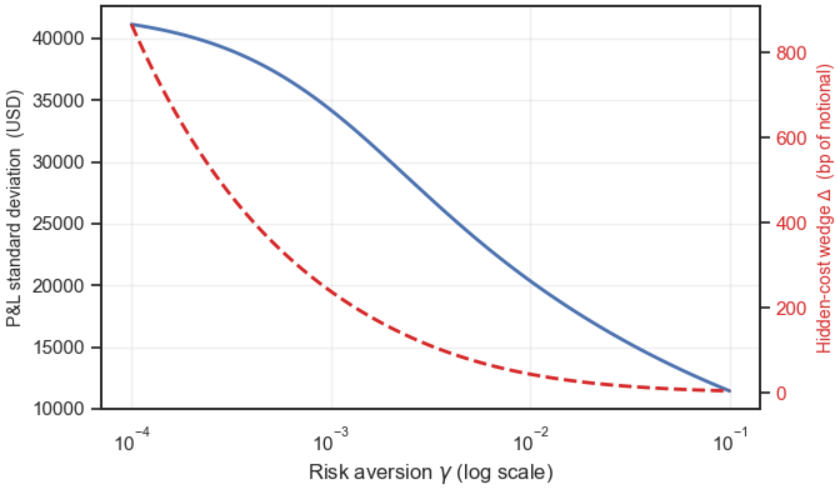


Figure 1.1: Impact of risk aversion on execution wedge, with fixed parameters $P = 20$, $\lambda = 0.4$, $\eta = 0.4$, $\sigma = 0.2$, $N = 0.50\% ADV$ and $T = 1$.

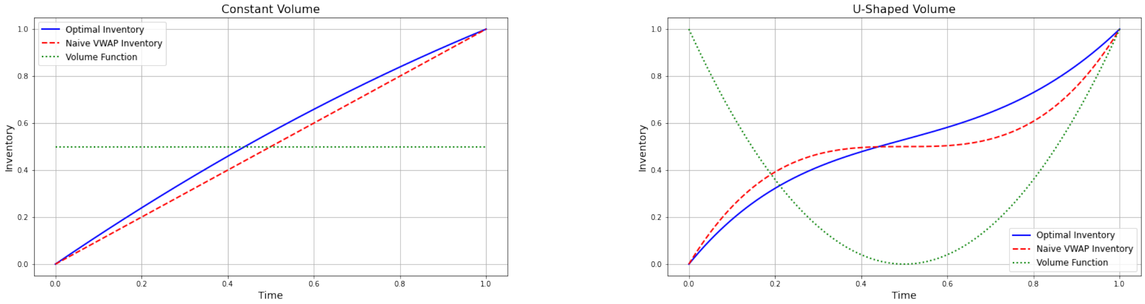
Effect of the Volume Profile $V(t)$. Finally, the volume profile $V(t)$ drives the particular solution $x_p(t)$. Periods with elevated volume have a disproportionate effect on the realized VWAP—these time windows contribute more heavily to the average price. Consequently, the broker's optimal trajectory typically places greater emphasis on acquiring shares (and thus potentially moving prices) before or around high-volume intervals. In scenarios where volume spikes are anticipated, the broker may front-load in advance to benefit from the uplift in price that accumulates when permanent impact is significant.

From an empirical standpoint, the volume profile is often U-shaped (e.g., heavier trading near market open and close). As illustrated in figure 1.2, our model suggests that if a broker can anticipate an especially active segment of the trading day, it may seek to strategically adjust its inventory in advance—if permanent impact is appreciable, doing so raises subsequent prices to the broker's advantage under a VWAP contract.

1.4.4 Hidden Execution Cost

Next, we want to study how much the broker's optimal trading distorts the final benchmark. We define the difference between the optimal and the naïve expected VWAP as

$$\Delta = E\left[\text{VWAP}^*\right] - E\left[\text{VWAP}_{\text{naive}}\right]$$



(a) Constant volume

(b) U-shaped volume

Figure 1.2: Optimal inventory with different volume patterns, with fixed parameters $\lambda = 0.4$, $\eta = 0.4$, $\sigma = 0.2$, $\gamma = 0.2$, $N = 1$ and $T = 1$.

Substituting in the above expressions, we obtain

$$\begin{aligned}\Delta = \frac{1}{V_c(T)} \Bigg\{ \lambda \left[\int_0^T x^*(t)V(t)\,dt - \int_0^T x_{\text{naive}}(t)V(t)\,dt \right] \\ + \eta \left[N\,V(T) - \frac{N}{V_c(T)} \int_0^T V(t)^2\,dt - \int_0^T x^*(t)V'(t)\,dt \right] \Bigg\}\end{aligned} \tag{1.12}$$

The key observation is that if permanent price impact is high enough, the optimal trading strategy *front-loads* execution, i.e.

$$x^*(t) > x_{\text{naive}}(t) = \frac{N}{V_c(T)} \int_0^t V(u)\,du$$

Proposition 3 (Front-loading)**.** Assume that $V(t)$ is non-decreasing and $\lambda \geq 3\eta\,\frac{V'(t)}{V(t)}$ for all $t \in [0, T]$, then the broker's inventory process under optimal trading is always higher than under the naive schedule, that is $x^*(t) > x_{naive}(t)$, $\forall t \in [0, T]$.

Although we do not view these assumptions as realistic, we can draw significant insights from this result. We motivate the monotonicity assumption through two examples. First, a common stylized fact about trading volume is its U-shaped pattern. Thus, when a broker receives an order mid-day, volume is expected to increase and culminate at market close. Secondly, irrespective of time, a large inbound order will increase the instantaneous volume, while having minimal impact on the total expected volume throughout the day. In that scenario, a weaker result than Proposition 3 is that the optimal trading leads to front-loading for a small time window τ, i.e. $x^*(t) > x_{naive}(t)$, $\forall t \in [0, \tau]$.

1.4.5 Constant Volume Case

To develop additional intuition, we consider a tractable scenario in which trading volume remains constant throughout the execution horizon. Formally, let $V(t) = V$ for all $t \in [0, T]$. In this special case, the Volume-Weighted Average Price (VWAP) coincides with a Time-Weighted Average Price (TWAP). Consequently,

$$V_c(T) = \int_0^T V(t)\,dt = VT \quad \text{and} \quad x_{\text{naive}}(t) \;=\; \frac{N}{T}\,t.$$

Under the naive trading strategy that simply scales with time (i.e., the broker buys shares uniformly over the interval), a straightforward calculation shows

$$E\big[\text{VWAP}_{\text{naive}}\big] \;=\; S(0) \;+\; \frac{N}{T}\Big(\tfrac{\lambda N}{2} \;+\; \eta\Big).$$

In contrast, under the *optimal* strategy, the expected TWAP (equivalently, the VWAP when V is constant) takes the form

$$E\big[\text{VWAP}^*\big] \;=\; S(0) + \frac{1}{V_c(T)}\Big[\eta\, N\, V + \lambda\, V \int_0^T x^*(t)\,dt\Big] \;=\; S(0) + \frac{1}{T}\Big[\eta\, N + \lambda \int_0^T x^*(t)\,dt\Big].$$

This result highlights how the *optimal* trading schedule strategically shifts more of the broker's purchases toward earlier times (often referred to as "front-loading"), raising the permanent price impact over a larger portion of the execution window. The economic intuition is that when permanent impact λ is substantial, building inventory early induces higher future prices and thus elevates the eventual VWAP. Under a Guaranteed VWAP contract, this benefits the broker by allowing the final shares to be offloaded at a higher benchmark price, effectively capturing hidden profits even if explicit fees appear minimal.

Proposition 4 (Optimal TWAP schedule)**.** Under constant volume and the optimal TWAP execution policy, the broker's inventory evolves according to

$$x(t) \;=\; \frac{N\,t}{T} \;+\; \frac{N\,\lambda}{\gamma\,\sigma^2\,T}\left[1 \;-\; \frac{1 \;-\; e^{-\alpha T}}{\sinh(\alpha T)}\,\sinh\!\big(\alpha\,t\big) \;-\; e^{-\alpha\,t}\right],$$

Proposition 5 (Hidden cost of guaranteed execution)**.** The incremental cost associated with the broker's strategic trading, defined by $\Delta = E[\text{VWAP}^*] - E[\text{VWAP}_{\text{naive}}]$, is given by

$$\Delta \;=\; \frac{N\,\lambda^2}{\gamma\,\sigma^2\,T^2}\left[T \;-\; \frac{1 \;-\; e^{-\alpha T}}{\alpha}\Big(1 \;+\; \frac{\cosh(\alpha T) \;-\; 1}{2\,\sinh(\alpha T)}\Big)\right].$$

This hidden cost Δ represents the incremental increase in the final VWAP stemming from the broker's front-loaded strategy. By purchasing more aggressively early in the day, the broker exerts permanent upward pressure on the stock price for the remainder of the execution interval, thereby boosting the ultimate average price. Although the benchmark is technically met, the client ends up facing a higher effective purchase cost over time, enabling the broker to reap additional (often opaque) profits. These findings underscore why, despite low or zero explicit fees, G-VWAP agreements can embed significant implicit execution costs that are ultimately borne by the end investor.

Corollary 2. As $\gamma \to 0$,

$$\Delta \sim \frac{N\lambda^2}{4\sqrt{2\eta\sigma^2}}\,\gamma^{-1/2}.$$

This shows that the VWAP difference diverges as $\gamma^{-1/2}$ when the risk–aversion parameter γ tends to zero, reflecting that a risk–neutral broker (i.e. $\gamma = 0$) is incentivized to front–load trades without penalty, thereby driving the benchmark price arbitrarily high.

1.5 Robustness to Concave Permanent Impact

Empirical studies (Tóth et al., 2011; Brokmann et al., 2016) show that permanent market impact is concave in trade size. To verify our qualitative results do not hinge on linearity, we generalize the linear-impact setup by letting the *permanent* component of market impact follow a power law with exponent $\rho \in (0, 1]$:

$$S(t) = S(0) + \lambda\, x(t)^{\rho} + \eta\, \dot{x}(t) + \sigma B_t, \qquad 0 \le t \le T. \tag{1.13}$$

The remaining primitives (volume curve $V(\cdot)$, risk parameter γ, etc.) are unchanged. We show that the optimal trading path continues to "front-load" for any $\rho < 1$ and provide the ODE characterizing the optimum.

Deriving the Euler–Lagrange equation. Re-write the broker's mean–variance objective, as

$$J[x] = \int_0^T \Big(\frac{N}{V_c(T)} V(t) - \dot{x}(t)\Big)\big[\lambda x(t)^{\rho} + \eta \dot{x}(t)\big]\, dt - \frac{\gamma\sigma^2}{2} \int_0^T \Big(\frac{N}{V_c(T)} V_c(t) - x(t)\Big)^2 dt, \tag{1.14}$$

subject to $x(0) = 0$ and $x(T) = N$. Setting $L(t, x, \dot{x})$ to be the integrand of (1.14) and applying the Euler–Lagrange condition $\frac{d}{dt}\partial_{\dot{x}} L - \partial_x L = 0$ yields, after straightforward algebra,

$$2\eta\, \ddot{x}(t) - \gamma\sigma^2\, x(t) = -\eta\, \frac{N}{V_c(T)} V'(t) + \lambda\rho\, x(t)^{\rho-1}\Big(\frac{N}{V_c(T)} V(t) - \dot{x}(t)\Big) - \gamma\sigma^2\, \frac{N}{V_c(T)} V_c(t). \tag{1.15}$$

For $\rho = 1$, this collapses to the linear model.

Positivity of the λ–derivative (front-loading). Let $x_\rho^\lambda(t)$ denote the solution of (1.15) for a given λ. Differentiate (1.15) with respect to λ and set $y(t) := \partial x_\rho^\lambda(t)/\partial\lambda$. We obtain the *linear* second-order ODE:

$$2\eta\,\ddot{y}(t) - \gamma\sigma^2\,y(t) = \rho\,x_\rho^\lambda(t)^{\rho-1}\Big(\frac{N}{V_c(T)}V(t) - \dot{x}_\rho^\lambda(t)\Big), \tag{1.16}$$

with homogeneous boundary conditions $y(0) = y(T) = 0$.

Because $\rho \in (0,1]$ we have $x^{\rho-1} > 0$ for $x > 0$. Moreover $V(t) \geq 0$ and, by definition of $\dot{x}$, the bracket in (1.16) satisfies $\frac{N}{V_c(T)}V(t) - \dot{x}_\rho^\lambda(t) \geq 0$ for t close to 0 (the broker cannot trade faster than the entire market instantaneously). Hence the right-hand side of (1.16) is non-negative on a neighbourhood $(0,\varepsilon)$ of the origin.

Applying the standard maximum principle for the linear ODE: $2\eta\,\ddot{y} - \gamma\sigma^2 y = f(t)$ with $f \geq 0$ shows that the minimum of y is achieved at the boundary. Given $y(0) = 0$ and $y'(0) = \frac{\rho N}{2\eta V_c(T)}V(0) > 0$, we deduce

$$y(t) = \frac{\partial x_\rho^\lambda(t)}{\partial\lambda} > 0, \qquad 0 < t < \varepsilon. \tag{1.17}$$

Thus an increase in λ *strictly raises* the optimal inventory early in the execution window, i.e.

$$x_\rho^{\lambda_2}(t) > x_\rho^{\lambda_1}(t) \quad \text{for } 0 < t < \varepsilon,\ \lambda_2 > \lambda_1. \tag{1.18}$$

Front-loading therefore persists under any concave permanent-impact exponent $\rho \in (0,1]$.

Bound on the hidden-cost wedge. Integrating (1.15) against x_ρ^λ and repeating the steps of Proposition 5 gives

$$\Delta_\rho = \frac{\lambda^2 N}{\gamma\sigma^2 T^{2-\rho}}\Big[\mathrm{C}_\rho + \mathcal{O}\big((\lambda/\eta)^{1-\rho}\big)\Big], \tag{1.19}$$

where $\mathrm{C}_\rho \in (0, \mathrm{C}_1]$ is a constant depending only on (ρ, γ, σ^2).

When $\rho = 1$, the model collapses to the linear baseline; for $\rho < 1$, permanent impact grows sub-linearly. Solving for the modified Euler-Lagrange equation, we find that front-loading remains optimal whenever $\lambda > 0$. The curvature merely moderates its intensity.

1.6 Stochastic Volume

For completeness, we now let the market volume be stochastic and revisit the broker's optimization problem. While the model becomes much harder to interpret, we show that the main economic mechanism remains. As in Frei and Westray (2015), we use a Gamma-Bridge to model the relative volume curve.

1.6.1 Modeling Volume with a Gamma Bridge

Unlike the purely deterministic volume setting of Section 1.4.5, we now let the intraday cumulative volume be random, modeled via a gamma bridge. Specifically, assume that cumulative volume $V_c(t)$ follows a gamma process with shape parameter $m > 0$ and rate $\theta = 1$, so $V_c(0) = 0$. We then define

$$\Gamma(t) \;:=\; \frac{V_c(t)}{V_c(T)},$$

so that $\Gamma(0) = 0$ and $\Gamma(T) = 1$. Intuitively, $\Gamma(t)$ represents the fraction of daily volume realized by time t, and by construction, it is always in $[0, 1]$. We assume that $\Gamma(\cdot)$ is independent of the Brownian motion $B(\cdot)$.

1.6.2 Modified Problem

Under a Guaranteed VWAP contract, the broker delivers N shares at the *realized VWAP*, which is now given by

$$\text{VWAP} \;=\; \int_0^T S(t)\,\mathrm{d}\Gamma(t).$$

Since $\Gamma(T) = 1$, this is exactly the volume-weighted average of $S(t)$ during $[0, T]$. The broker's total profit-and-loss (PnL) from the strategy $x(\cdot)$ is then

$$\Pi[x] \;=\; N\left[\text{VWAP}\right] \;-\; \int_0^T S(t)\,\dot{x}(t)\,\mathrm{d}t, \tag{1.20}$$

where the integral $\int_0^T S(t)\,\dot{x}(t)\,\mathrm{d}t$ represents the cost of purchasing N shares in the market.

Remark. In the special case where $\Gamma(t)$ were deterministic, this objective reduces to the earlier VWAP model of Section 1.4.5. Here, however, the fraction of daily volume realized up to time t is itself a random process, driven by the gamma-bridge dynamics. This randomness affects both the realized VWAP and the execution cost, so we must solve a stochastic control problem that accounts for $\Gamma(t)$ as well as $B(t)$.

As in Subsection 1.3.4, the broker maximizes a mean–variance objective:

$$\max_{x(\cdot)}\Big\{\mathbf{E}\big[\Pi[x]\big] \;-\; \frac{\gamma}{2}\,\mathrm{Var}\big(\Pi[x]\big)\Big\}, \quad \text{subject to } x(0)=0,\;\; x(T)=N. \tag{1.21}$$

The broker wants to choose a trading path $x(\cdot)$ that balances expected PnL versus its variance.

Substituting the price dynamics into (1.20):

$$\Pi[x] = N\int_0^T \Big[S(0)+\lambda\,x(t)+\eta\,u(t)+\sigma\,B(t)\Big]\,d\Gamma(t) - \int_0^T \Big[S(0)+\lambda\,x(t)+\eta\,u(t)+\sigma\,B(t)\Big]\,u(t)\,dt$$

Since the only source of randomness is $B(t)$, by Itô's isometry one can show that

$$\mathrm{Var}(\Pi[x]) = \sigma^2\int_0^T \Big(N\,y - x(t)\Big)^2 dt.$$

Thus, the objective (1.21) becomes

$$J[x] = \mathbb{E}[\Pi[x]] - \frac{\gamma\sigma^2}{2}\int_0^T \Big(N\,y - x(t)\Big)^2 dt.$$

1.6.3 HJB Equation and Value Function

For each $(t,x,y)\in[0,T]\times\mathbb{R}\times[0,1]$, define

$$V(t,x,y) \;=\; \sup_{\{\,u(s)\colon s\in[t,T]\}} \mathbf{E}\Big[\Pi\big[x_{[t,T]}\big] \;-\; \frac{\gamma}{2}\,\mathrm{Var}\big(\Pi[x_{[t,T]}]\big)\,\Big|\,x(t)=x,\;\Gamma(t)=y\Big].$$

We want to find $V(0,0,0)$ and the corresponding optimal trading speed $u^*(\cdot)$. In what follows, we provide a brief sketch of the proof for the optimal trading path, before stating the main theorem.

Infinitesimal Analysis and HJB. Using dynamic programming in continuous time, we consider a time step from t to $t+\Delta t$. The gamma bridge $\Gamma(\cdot)$ has finite variation on $[0,T]$. The Brownian motion $B(\cdot)$ affects the cost integral.

Let $\Delta\Pi[x]$ be the change in the broker's PnL over $[t,t+\Delta t]$. By taking expectations and collecting second-moment terms, we obtain a dynamic-programming identity. Then we divide by Δt and let $\Delta t\to 0$. The outcome is that the value function should satisfy the PDE:

$$\frac{\partial}{\partial t}\,V(t,x,y) \;+\; \sup_{u\in\mathbb{R}}\Big\{\mathcal{H}\big(t,x,y,u,V,\nabla V,\nabla^2 V\big)\Big\} \;+\; \mathcal{L}^{\Gamma}V(t,x,y) \;=\; 0,$$

Finally, the HJB equation is

$$V_t + \sup_{u\in\mathbb{R}} \left\{ u\, V_x - \eta\, u^2 \right\} + \frac{1-y}{T-t}\, V_y + \frac{\gamma\sigma^2}{2} \left(N\, y - x\right)^2 = 0, \tag{1.22}$$

with a terminal condition that enforces $x(T) = N$.

Quadratic Value Function Ansatz. We posit that this PDE takes the form of

$$V(t,x,y) \;=\; A(t)\, x^2 \;+\; B(t)\, x\, y \;+\; C(t)\, x \;+\; D(t)\, y^2 \;+\; E(t)\, y \;+\; F(t),$$

for some deterministic coefficient functions $\{A,B,C,D,E,F\}\colon [0,T] \to \mathbb{R}$. This ansatz is justified because the objective is a mean–variance criterion with linear/quadratic price impacts, so the PDE is quadratic in (x,u). The gamma bridge $\Gamma(\cdot)$ only appears linearly in $y = \Gamma(t)$, apart from its finite-variation generator. This leads us naturally to a polynomial form in (x,y).

From there, we obtain the unique classical solution to the HJB PDE on $[0,T)\times\mathbb{R}\times[0,1]$ that enforces $x(T) = N$. The optimal trading speed (control) is

$$u^*(s) \;=\; -\frac{1}{2\,\eta}\,\frac{\partial}{\partial x}\, V\big(s,\, X^*(s),\, \Gamma(s)\big),$$

Concretely,

$$\frac{\partial}{\partial x}\, V(s,x,y) \;=\; 2\, A(s)\, x \;+\; B(s)\, y \;+\; C(s).$$

Putting everything together, the broker's optimal trading speed is

$$u^*(s) \;=\; -\frac{1}{2\,\eta}\Big[2\, A(s)\, X^*(s) \;+\; B(s)\,\Gamma(s) \;+\; C(s)\Big].$$

Finally, after integrating $\dot{X}^*(s) = u^*(s)$ from 0 to t and re-scaling, we obtain the following theorem.

Theorem 2 (Optimal trading schedule under stochastic volume)**.** Under stochastic volume, the broker's inventory evolves according to

$$X^*(t) \;=\; \kappa \exp\Big(-\int_0^t \frac{A(r)}{\eta}\,\mathrm{d}r\Big) \;\times\; \int_0^t \Big[-\frac{B(z)\,\Gamma(z) + C(z)}{2\,\eta}\Big] \exp\Big(\int_0^z \frac{A(r)}{\eta}\,\mathrm{d}r\Big)\,\mathrm{d}z,$$

where κ is a normalizing constant to satisfy $X^*(T) = N$.

The broker continues to *front-load* trades when permanent impact increases, even though market volume is random. While the solution is much harder to interpret, as in Section 1.3, the front-loading property of the optimal inventory subsists. We now quantify the magnitude of that gain with a back-of-the-envelope calculation.

1.7 Numerical Results

To better understand our results, we translate the closed-form wedge $\Delta(\alpha)$ of Proposition 5 into dollars using the example of Ford (stock price $\approx$ \$20, ADV $\approx$ 100mm shares). A survey by Greenwich Associates puts the median low-touch electronic commission at $\approx$ 14bps in explicit commission for a guaranteed VWAP execution, and sell-side dealers frequently waive even that fee to win flow. The commission rate should increase with the size of the order. In the absence of such information and assuming that fixed fees scale up slowly with %ADV, we use the 14 bps number for our comparative analysis.

Table 1.1: Baseline parameters

Symbol	**Description**	**Units**	**Value**
λ	Permanent impact per %ADV	bp/%ADV	50
η	Temporary impact per %ADV per min	bp/%ADV/min	10
σ	Annualised volatility	%	30
γ	Broker risk aversion	—	0.010
T	Trading horizon	minutes	390
P	Share price	USD	20
ADV	Average daily volume	shares	1.0×10^8

With Δ defined as in Section 1.4.5,

$$\Delta(0.25\%) = 8.77 \text{ bp}, \qquad \Delta(0.50\%) = 17.53 \text{ bp},$$
$$\Delta(1.00\%) = 35.06 \text{ bp}, \qquad \Delta(2.00\%) = 70.13 \text{ bp}.$$

Because $\Delta \propto \lambda^2\alpha$, each additional basis-point of permanent impact or each extra tranche of shares amplifies the broker's incentive to accelerate early prints and inflate the benchmark.

Explicit commission is a flat \$2.1 ¢/share, i.e. $c = 14$ bp. The broker's expected dollar P&L on the hidden wedge is $C_{\text{imp}} = NP\Delta/10^4$:

The hidden component overtakes the flat 14 bps commission once the ticket exceeds 0.5%ADV; at 2%ADV it more than doubles the explicit fee. Hence, a broker can credibly advertise "zero or flat commission" and still expect to earn \$70K on a guaranteed-VWAP order, purely by exploiting price impact. Best-execution audits that focus on explicit fees alone will systematically understate trading cost for benchmarked flow. Regulators and asset owners should therefore evaluate all-in costs, adding benchmark slippage to booked

Table 1.2: All-in execution cost with ADV= 100mm, $P = \$20$)

	Dollar cost (USD)		Cost / notional (bp)	
Order size	Explicit	Hidden	Explicit	Hidden
0.25 % ADV ($5 m)	$5 250	$4 385	14	8.8
0.50 % ADV ($10 m)	$10 500	$17 530	14	17.5
1.00 % ADV ($20 m)	$21 000	$70 120	14	35.1
2.00 % ADV ($40 m)	$42 000	$280 520	14	70.1

commission. Such a practice would align broker incentives with end-investor welfare and reduce the cross-subsidy currently embedded in guaranteed VWAP contracts.

1.8 Conclusion

This paper develops a continuous-time model of broker execution under Guaranteed Volume-Weighted Average Price (G-VWAP) contracts, shedding light on the hidden economic incentives and costs behind these seemingly attractive benchmarks. Incorporating both permanent and temporary price impacts into a mean–variance framework, we show that brokers strategically deviate from naive VWAP approaches to exploit price dynamics. In particular, front-loading becomes optimal when permanent impact is significant, whereas more evenly distributed trading arises when temporary impact dominates.

The ability to manipulate trade timing, while still satisfying the VWAP benchmark, helps explain how brokers can offer G-VWAP at low or zero fees. By establishing inventory ahead of rising prices or flattening trades to reduce transient costs, brokers profit from implicit market distortions hidden within execution prices. These findings underscore how widely adopted VWAP guarantees can inadvertently increase total trading costs for clients, thus highlighting the importance of scrutinizing both permanent and temporary market impacts in execution arrangements.

Our framework contributes to the growing literature on optimal execution and market microstructure by emphasizing the broker's profit motive under guaranteed benchmarks. The model's quantitative insights have important implications for institutional investors and regulators alike. As recent trading platforms, such as the CBOE's BIDS VWAP-X or the LiquidNet VWAP Cross, expand the use of VWAP-based executions, our analysis suggests that regulators and platform designers should carefully evaluate how broker incentives interact with evolving market structures.

Future research could deepen these insights further by examining real broker behavior under G-VWAP using detailed transaction-level data or by extending the model to nonlinear price impacts and multiple strategic players. Such investigations would advance understanding of benchmark-driven execution strategies and inform both regulatory policies and institutional best practices.

1.9 Appendix

1.9.1 Proofs

Proof of Theorem 1

We seek to maximize the functional

$$\mathcal{J}[x] \;=\; \int_0^T \Big(\tfrac{N}{M}\,V(t) - \dot{x}(t)\Big)\,\Big(S(0) + \lambda\,x(t) + \eta\,\dot{x}(t)\Big)\,dt - \frac{\gamma}{2}\,\sigma^2 \int_0^T \Big(\tfrac{N}{M}\int_0^t V(u)\,du - x(t)\Big)^2\,dt,$$

subject to the boundary conditions $x(0) = 0$ and $x(T) = N$. The goal is to determine the function $x(t)$ that maximizes $\mathcal{J}[x]$.

Rewrite the Objective. Define

$$\bar{S}(t) \;=\; S(0) \;+\; \lambda\,x(t) \;+\; \eta\,\dot{x}(t).$$

The first term in $\mathcal{J}[x]$ then becomes

$$\int_0^T \Big(\tfrac{N}{M}\,V(t) \;-\; \dot{x}(t)\Big)\,\bar{S}(t)\,dt \;=\; \int_0^T \Big[\tfrac{N}{M}\,V(t)\,\bar{S}(t) \;-\; \dot{x}(t)\,\bar{S}(t)\Big]\,dt.$$

Meanwhile, the variance-penalty term is

$$-\;\tfrac{\gamma}{2}\,\sigma^2 \int_0^T \Big(\tfrac{N}{M}\,V_c(t) \;-\; x(t)\Big)^2\,dt.$$

Collecting these gives

$$\mathcal{J}[x] \;=\; \int_0^T L\big(t,\,x(t),\,\dot{x}(t)\big)\,dt,$$

where

$$\begin{aligned} L(t,x,\dot{x}) \;=\; & \tfrac{N}{M}\,V(t)\,S(0) \;+\; \lambda\,\tfrac{N}{M}\,V(t)\,x \;+\; \eta\,\tfrac{N}{M}\,V(t)\,\dot{x} \;-\; S(0)\,\dot{x} \;-\; \lambda\,x\,\dot{x} \;-\; \eta\,\dot{x}^2 \\ & -\;\tfrac{\gamma}{2}\,\sigma^2\Big(\tfrac{N}{M}\,V_c(t) - x\Big)^2. \end{aligned}$$

Euler–Lagrange Equation. To find the maximizer $x(t)$, we set the Euler–Lagrange equation

$$\frac{d}{dt}\Big(\tfrac{\partial L}{\partial \dot{x}}\Big) \;-\; \tfrac{\partial L}{\partial x} \;=\; 0$$

to zero. First, compute

$$\frac{\partial L}{\partial x} = \lambda \tfrac{N}{M} V(t) \;-\; \lambda \dot{x}(t) \;+\; \gamma \sigma^2 \big(\tfrac{N}{M} V_c(t) - x\big).$$

Next, the derivative with respect to $\dot{x}$ is

$$\frac{\partial L}{\partial \dot{x}} \;=\; \eta \tfrac{N}{M} V(t) \;-\; S(0) \;-\; \lambda x \;-\; 2\eta \dot{x}.$$

Taking the time derivative gives

$$\frac{d}{dt}\Big(\tfrac{\partial L}{\partial \dot{x}}\Big) \;=\; \eta \tfrac{N}{M} V'(t) \;-\; \lambda \dot{x}(t) \;-\; 2\eta \ddot{x}(t).$$

Thus, the Euler–Lagrange equation becomes

$$\big[\eta \tfrac{N}{M} V'(t) \;-\; \lambda \dot{x}(t) \;-\; 2\eta \ddot{x}(t)\big] \;-\; \big[\lambda \tfrac{N}{M} V(t) \;-\; \lambda \dot{x}(t) \;+\; \gamma \sigma^2 \big(\tfrac{N}{M} V_c(t) - x\big)\big] \;=\; 0.$$

Noting that $-\lambda \dot{x}$ and $+\lambda \dot{x}$ cancel, we rearrange to obtain

$$2\eta \ddot{x}(t) \;-\; \gamma \sigma^2 x(t) \;=\; -\eta \tfrac{N}{M} V'(t) \;+\; \lambda \tfrac{N}{M} V(t) \;-\; \gamma \sigma^2 \tfrac{N}{M} V_c(t).$$

Defining

$$f(t) \;:=\; -\eta \tfrac{N}{M} V'(t) \;+\; \lambda \tfrac{N}{M} V(t) \;-\; \gamma \sigma^2 \tfrac{N}{M} V_c(t),$$

we arrive at the key ODE:

$$2\eta \ddot{x}(t) \;-\; \gamma \sigma^2 x(t) \;=\; f(t).$$

Solving the Two-Point Boundary Value Problem. Let

$$\alpha \;=\; \sqrt{\frac{\gamma \sigma^2}{2\eta}}.$$

Then the ODE becomes

$$\ddot{x}(t) \;-\; \alpha^2 x(t) \;=\; \frac{1}{2\eta} f(t).$$

We write $x(t) = x_h(t) + x_p(t)$, where x_h solves the homogeneous equation $\ddot{x}_h - \alpha^2 x_h = 0$, and x_p is a particular solution of the inhomogeneous problem. The homogeneous solution is

$$x_h(t) \;=\; C_1 e^{\alpha t} \;+\; C_2 e^{-\alpha t},$$

often expressed in hyperbolic sine/cosine form. A convenient particular solution satisfying $x_p(0) = 0$ and $x_p(T) = 0$ is given by

$$x_p(t) \;=\; \int_0^T G(t,s)\,\Big[\tfrac{1}{2\,\eta}\,f(s)\Big]\,ds,$$

where

$$G(t,s) \;=\; \frac{1}{\alpha\,\sinh(\alpha T)}\begin{cases}\sinh(\alpha\,t)\,\sinh\big(\alpha\,(T-s)\big), & t \le s,\\ \sinh(\alpha\,s)\,\sinh\big(\alpha\,(T-t)\big), & t > s.\end{cases}$$

Boundary Conditions. The final solution must satisfy $x(0) = 0$ and $x(T) = N$. By construction, $x_p(0) = 0$ and $x_p(T) = 0$. We therefore choose

$$x_h(t) \;=\; \Big(N - x_p(T)\Big)\,\frac{\sinh(\alpha\,t)}{\sinh\big(\alpha\,T\big)},$$

so that the full solution

$$x(t) \;=\; x_p(t) \;+\; \Big(N - x_p(T)\Big)\,\frac{\sinh(\alpha\,t)}{\sinh\big(\alpha\,T\big)}$$

automatically satisfies both boundary conditions. This completes the derivation of the optimal trading schedule. □

Proof of Proposition 1

The only source of randomness is the Brownian motion B_t, so the stochastic component of the cost is

$$\sigma\frac{N}{M}\int_0^T V(t)\,B_t\,dt.$$

Thus, the variance of the execution cost is

$$\mathrm{Var}\left(\Pi_{\text{naive}}\right) = \sigma^2\frac{N^2}{M^2}\,\mathrm{Var}\left(\int_0^T V(t)\,B_t\,dt\right).$$

Since the covariance of B_t satisfies $\mathrm{Cov}(B_t, B_s) = \min(t,s)$, we obtain

$$\mathrm{Var}\left(\int_0^T V(t)\,B_t\,dt\right) = \int_0^T\int_0^T V(t)V(s)\,\min(t,s)\,dt\,ds.$$

By splitting the double integral, we obtain

$$\int_0^T \int_0^T V(t)V(s)\min(t,s)\,dt\,ds = 2\int_0^T s\,V(s)\left(\int_s^T V(t)\,dt\right)ds.$$

Since $\int_s^T V(t)\,dt = M - V_c(s)$, the expression becomes

$$\int_0^T \int_0^T V(t)V(s)\min(t,s)\,dt\,ds = 2\int_0^T s\,V(s)\Big[M - V_c(s)\Big]ds.$$

Thus, the variance is

$$\mathrm{Var}\left(\Pi_{\text{naive}}\right) = 2\,\sigma^2\,\frac{N^2}{M^2}\int_0^T s\,V(s)\Big[M - V_c(s)\Big]ds. \tag{1.23}$$

□

Proof of Proposition 2

Given a trading schedule $x(t)$ (with $x(0) = 0$ and $x(T) = N$), the expected VWAP is

$$\begin{aligned} \mathbb{E}[\mathrm{VWAP}] &= \frac{1}{M}\int_0^T \mathbb{E}[S(t)]\,V(t)\,dt & (1.24)\\ &= \frac{1}{M}\int_0^T \Big(S(0) + \lambda\,x(t) + \eta\,\dot{x}(t)\Big)V(t)\,dt. & (1.25)\\ &= S(0) + \frac{\lambda}{M}\int_0^T x(t)V(t)\,dt + \frac{\eta}{M}\int_0^T \dot{x}(t)V(t)\,dt. & (1.26) \end{aligned}$$

A useful manipulation comes from noting that

$$\frac{d}{dt}\{x(t)V(t)\} = \dot{x}(t)V(t) + x(t)V'(t),$$

so that integrating from 0 to T gives

$$\int_0^T \dot{x}(t)V(t)\,dt = x(T)V(T) - x(0)V(0) - \int_0^T x(t)V'(t)\,dt.$$

Since $x(0) = 0$ and $x(T) = N$ (and often we take $V(0) = 0$), it follows that

$$\int_0^T \dot{x}(t)V(t)\,dt = N\,V(T) - \int_0^T x(t)V'(t)\,dt.$$

Substituting back yields

$$\mathbb{E}[\text{VWAP}] = S(0) + \frac{1}{M}\left[\eta\, N\, V(T) + \int_0^T x(t)\Big(\lambda\, V(t) - \eta\, V'(t)\Big)dt\right]. \tag{2}$$

□

Proof of Corollary 1

The cost incurred is

$$\text{Cost} = \int_0^T S(t)\,\dot{x}(t)\,dt,$$

A brief calculation shows that the expected profit is

$$\begin{aligned} E[\Pi] &= N\,E[S(T)] - \int_0^T E[S(t)]\,\dot{x}(t)\,dt \\ &= N\Big(S(0) + \lambda N\Big) - \left\{S(0)N + \lambda\int_0^T x(t)\,\dot{x}(t)\,dt + \eta\int_0^T [\dot{x}(t)]^2\,dt\right\} \\ &= \lambda N^2 - \lambda\frac{N^2}{2} - \eta\int_0^T [\dot{x}(t)]^2\,dt, \end{aligned} \tag{1.27}$$

where we have used the identity

$$\int_0^T x(t)\,\dot{x}(t)\,dt = \frac{1}{2}\Big[x(T)^2 - x(0)^2\Big] = \frac{N^2}{2}.$$

Since the only source of risk is through the terminal price $S(T)$ (via the Brownian motion B_t), the variance of the profit is given by

$$\text{Var}[\Pi] = \sigma^2\int_0^T \Big(N - x(t)\Big)^2 dt.$$

Thus, the broker's mean-variance objective is to maximize

$$J[x] = E[\Pi] - \frac{\gamma}{2}\,\text{Var}[\Pi], \tag{1.28}$$

which (up to an additive constant) is equivalent to minimizing the functional

$$\mathcal{J}[x] = \eta\int_0^T [\dot{x}(t)]^2\,dt + \frac{\gamma\sigma^2}{2}\int_0^T \Big(N - x(t)\Big)^2 dt, \tag{1.29}$$

subject to the boundary conditions

$$x(0) = 0, \quad x(T) = N.$$

Define the Lagrangian

$$L(x, \dot{x}) = \eta\,[\dot{x}(t)]^2 + \frac{\gamma\sigma^2}{2}\Big(N - x(t)\Big)^2.$$

The Euler–Lagrange equation is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0.$$

A straightforward computation yields

$$\frac{\partial L}{\partial \dot{x}} = 2\eta\,\dot{x}(t), \qquad \frac{d}{dt}\,(2\eta\,\dot{x}(t)) = 2\eta\,\ddot{x}(t),$$

and

$$\frac{\partial L}{\partial x} = -\gamma\sigma^2\Big(N - x(t)\Big).$$

Thus, the Euler–Lagrange equation becomes

$$2\eta\,\ddot{x}(t) + \gamma\sigma^2\Big(N - x(t)\Big) = 0. \tag{1.30}$$

Dividing by 2η and defining

$$\alpha^2 = \frac{\gamma\sigma^2}{2\eta},$$

equation (1.30) may be rewritten as

$$\ddot{x}(t) - \alpha^2\Big(x(t) - N\Big) = 0. \tag{1.31}$$

Let

$$y(t) = x(t) - N.$$

Then $y(t)$ satisfies

$$\ddot{y}(t) - \alpha^2 y(t) = 0,$$

with boundary conditions

$$y(0) = x(0) - N = -N \quad \text{and} \quad y(T) = x(T) - N = 0.$$

The general solution of this homogeneous ordinary differential equation is

$$y(t) = C_1 e^{\alpha t} + C_2 e^{-\alpha t}.$$

Alternatively, writing the solution in hyperbolic form, we have

$$y(t) = C_3 \sinh(\alpha t) + C_4 \cosh(\alpha t).$$

The initial condition $y(0) = -N$ immediately implies

$$C_4 = -N.$$

The terminal condition $y(T) = 0$ yields

$$0 = C_3 \sinh(\alpha T) - N \cosh(\alpha T),$$

so that

$$C_3 = N \, \frac{\cosh(\alpha T)}{\sinh(\alpha T)}.$$

Thus, the solution for $y(t)$ is

$$y(t) = N \, \frac{\cosh(\alpha T)}{\sinh(\alpha T)} \sinh(\alpha t) - N \cosh(\alpha t).$$

A standard rearrangement shows that this expression can be written as

$$y(t) = -N \, \frac{\sinh\big(\alpha(T-t)\big)}{\sinh\big(\alpha T\big)}.$$

Returning to $x(t) = N + y(t)$, we obtain the optimal inventory trajectory:

$$x^*(t) = N \left[1 - \frac{\sinh\big(\alpha(T-t)\big)}{\sinh\big(\alpha T\big)} \right].$$

Differentiating, the optimal trading speed is

$$\dot{x}^*(t) = N \, \alpha \, \frac{\cosh\big(\alpha(T-t)\big)}{\sinh\big(\alpha T\big)}.$$

□

Proof of Proposition 3

Define the difference process

$$d(t) := x^*(t) - x_{\text{naive}}(t).$$

Derive the ODE for $d(t)$. The optimal inventory $x^*(t)$ satisfies the second–order ODE

$$x^{*\prime\prime}(t) - \alpha^2 x^*(t) = \frac{1}{2\eta} f(t), \tag{1.32}$$

where

$$f(t) = -\eta \, \frac{N}{M} \, V'(t) + \lambda \, \frac{N}{M} \, V(t) - \gamma\sigma^2 \, \frac{N}{M} \int_0^t V(u)\, du.$$

On the other hand, since

$$x_{\text{naive}}(t) = \frac{N}{M} \int_0^t V(u)\, du,$$

we have

$$x'_{\text{naive}}(t) = \frac{N}{M} V(t) \quad \text{and} \quad x''_{\text{naive}}(t) = \frac{N}{M} V'(t).$$

Thus, the naïve schedule satisfies

$$x''_{\text{naive}}(t) - \alpha^2 x_{\text{naive}}(t) = \frac{N}{M} V'(t) - \alpha^2 \frac{N}{M} \int_0^t V(u)\, du. \tag{1.33}$$

Subtracting (1.33) from (1.32) we obtain an ODE for $d(t)$:

$$\begin{aligned} d''(t) - \alpha^2 d(t) &= \left[x^{*\prime\prime}(t) - \alpha^2 x^*(t) \right] - \left[x''_{\text{naive}}(t) - \alpha^2 x_{\text{naive}}(t) \right] \\ &= \frac{1}{2\eta} f(t) - \left[\frac{N}{M} V'(t) - \alpha^2 \frac{N}{M} \int_0^t V(u)\, du \right]. \end{aligned}$$

Substitute the expression for $f(t)$:

$$\frac{1}{2\eta} f(t) = \frac{N}{M} \left(-\frac{1}{2} V'(t) + \frac{\lambda}{2\eta} V(t) - \frac{\gamma\sigma^2}{2\eta} \int_0^t V(u)\, du \right).$$

Thus,

$$\begin{aligned} d''(t) - \alpha^2 d(t) = \frac{N}{M} &\left[-\frac{1}{2} V'(t) + \frac{\lambda}{2\eta} V(t) - \frac{\gamma\sigma^2}{2\eta} \int_0^t V(u)\, du \right] \\ &- \frac{N}{M} V'(t) + \alpha^2 \frac{N}{M} \int_0^t V(u)\, du. \end{aligned}$$

Combine the $V'(t)$ terms:

$$-\frac{1}{2}V'(t) - V'(t) = -\frac{3}{2}V'(t).$$

Also, note that by the definition of α,

$$\alpha^2 = \frac{\gamma\sigma^2}{2\eta},$$

so that the terms involving $\int_0^t V(u)\,du$ cancel:

$$-\frac{\gamma\sigma^2}{2\eta}\int_0^t V(u)\,du + \alpha^2\int_0^t V(u)\,du = 0.$$

Therefore, we obtain

$$d''(t) - \alpha^2 d(t) = \frac{N}{M}\left[-\frac{3}{2}V'(t) + \frac{\lambda}{2\eta}V(t)\right].$$

Multiplying both sides by 2η gives

$$2\eta\left[d''(t) - \alpha^2 d(t)\right] = \frac{N}{M}\Big[-3\eta\,V'(t) + \lambda\,V(t)\Big].$$

That is,

$$d''(t) - \alpha^2 d(t) = \frac{N}{2\eta M}\Big[\lambda\,V(t) - 3\eta\,V'(t)\Big]. \tag{1.34}$$

Sufficient Condition. Assume that the market volume $V(t)$ is non-decreasing so that $V'(t) \geq 0$ and that

$$\lambda\,V(t) - 3\eta\,V'(t) \geq 0, \quad \forall\, t \in [0, T].$$

Then the forcing term on the right-hand side of (1.34) is nonnegative.

The ODE (1.34) is linear with constant coefficients. The homogeneous part

$$d_h''(t) - \alpha^2 d_h(t) = 0$$

has general solution

$$d_h(t) = C_1\cosh(\alpha t) + C_2\sinh(\alpha t).$$

Since both the optimal and the naive schedules satisfy the same boundary conditions,

$$x^*(0) = x_{\text{naive}}(0) = 0 \quad \text{and} \quad x^*(T) = x_{\text{naive}}(T) = N,$$

it follows that

$$d(0) = 0 \quad \text{and} \quad d(T) = 0.$$

Now, with a nonnegative forcing term in (1.34), the standard comparison or maximum principle for second–order ODEs implies that the minimum of $d(t)$ is attained at the boundary. Since $d(0) = d(T) = 0$, we deduce that

$$d(t) \geq 0, \quad \forall\, t \in [0, T].$$

□

Proof of Proposition 4

After rearranging constant terms, one may show that the Euler–Lagrange equation for the problem reduces to

$$2\eta\, \ddot{x}(t) + \frac{N\lambda}{T} + \gamma\sigma^2 \left(\frac{N\, t}{T} - x(t) \right) = 0,$$

Rearranging the above equation, we obtain

$$2\eta\, \ddot{x}(t) - \gamma\sigma^2\, x(t) = -\gamma\sigma^2\, \frac{N\, t}{T} - \frac{N\lambda}{T}.$$

Dividing through by 2η leads to

$$\ddot{x}(t) - \alpha^2 x(t) = -\alpha^2 \frac{N\, t}{T} - \frac{N\lambda}{2\eta\, T},$$

Solution of the ODE. The homogeneous part

$$\ddot{x}_h(t) - \alpha^2 x_h(t) = 0$$

has the general solution

$$x_h(t) = E\, e^{\alpha t} + F\, e^{-\alpha t}.$$

To solve the inhomogeneous ODE, we seek a particular solution of the form

$$x_p(t) = \frac{N}{T}\, t + C.$$

Differentiating gives $\dot{x}_p(t) = \frac{N}{T}$ and $\ddot{x}_p(t) = 0$. Substituting $x_p(t)$ into the ODE yields:

$$0 - \alpha^2 \left(\frac{N}{T}\, t + C \right) \;=\; -\alpha^2\, \frac{N\, t}{T} - \frac{N\lambda}{2\eta\, T}.$$

Matching coefficients:

- For the term in t:

$$-\alpha^2 \frac{N}{T} t = -\alpha^2 \frac{N\,t}{T} \quad \text{(automatically satisfied)}.$$

- For the constant term:

$$-\alpha^2 C = -\frac{N\lambda}{2\eta T} \quad \Longrightarrow \quad C = \frac{N\lambda}{2\eta\, T\, \alpha^2} \;=\; \frac{N\lambda}{\gamma\sigma^2\, T}..$$

Thus, the particular solution is

$$x_p(t) = \frac{N}{T}\, t + \frac{N\lambda}{\gamma\sigma^2\, T}.$$

The general solution of the ODE is then

$$x(t) = x_p(t) + x_h(t) = \frac{N}{T}\, t + \frac{N\lambda}{\gamma\sigma^2\, T} + E\, e^{\alpha t} + F\, e^{-\alpha t}.$$

Imposing the Boundary Conditions. The boundary conditions are:

$$x(0) = 0 \quad \text{and} \quad x(T) = N.$$

At $t = 0$:

$$0 = x(0) = \frac{N\lambda}{\gamma\sigma^2\, T} + E + F,$$

so that

$$E + F \;=\; -\frac{N\lambda}{\gamma\sigma^2\, T}.$$

At $t = T$:

$$N = x(T) = \frac{N}{T}\, T + \frac{N\lambda}{\gamma\sigma^2\, T} + E\, e^{\alpha T} + F\, e^{-\alpha T} = N + \frac{N\lambda}{\gamma\sigma^2\, T} + E\, e^{\alpha T} + F\, e^{-\alpha T},$$

which implies

$$E\, e^{\alpha T} + F\, e^{-\alpha T} \;=\; -\frac{N\lambda}{\gamma\sigma^2\, T}.$$

The system to solve is:

$$\begin{cases} E + F = -\dfrac{N\lambda}{\gamma\sigma^2\, T}, \\ E\, e^{\alpha T} + F\, e^{-\alpha T} = -\dfrac{N\lambda}{\gamma\sigma^2\, T}. \end{cases}$$

A short calculation yields the unique solutions:

$$\begin{aligned} E &= -\frac{N\lambda}{\gamma\sigma^2 T}\,\frac{1-e^{-\alpha T}}{2\sinh(\alpha T)}, \\ F &= -\frac{N\lambda}{\gamma\sigma^2 T}\left[1-\frac{1-e^{-\alpha T}}{2\sinh(\alpha T)}\right]. \end{aligned}$$

Substituting E and F back into the general solution, we obtain the optimal inventory trajectory:

$$x(t) = \frac{N\,t}{T} + \frac{N\lambda}{\gamma\sigma^2 T}\left[1-\frac{1-e^{-\alpha T}}{\sinh(\alpha T)}\sinh(\alpha t) - e^{-\alpha t}\right], \quad \text{with} \quad \alpha = \sqrt{\frac{\gamma\sigma^2}{2\eta}}.$$

and the optimal trading speed is given by differentiating:

$$\dot{x}(t) = \frac{N}{T} + \frac{N\lambda\alpha}{\gamma\sigma^2 T}\left[e^{-\alpha t} - \frac{1-e^{-\alpha T}}{\sinh(\alpha T)}\cosh(\alpha t)\right].$$

□

Proof of Proposition 5

Simplifying the difference, we get

$$\Delta = \frac{\lambda}{T}\left(\int_0^T x^*(t)\,dt - \frac{N\,T}{2}\right).$$

We now compute the integral of $x^*(t)$:

$$\int_0^T x^*(t)\,dt = \int_0^T \frac{N\,t}{T}\,dt + \frac{N\lambda}{\gamma\sigma^2 T}\int_0^T \left[1-\frac{1-e^{-\alpha T}}{2\sinh(\alpha T)}\sinh(\alpha t) - e^{-\alpha t}\right]dt.$$

The first term is straightforward:

$$\int_0^T \frac{N\,t}{T}\,dt = \frac{N}{T}\cdot\frac{T^2}{2} = \frac{N\,T}{2}.$$

Next, define

$$I = \int_0^T \left[1-\frac{1-e^{-\alpha T}}{2\sinh(\alpha T)}\sinh(\alpha t) - e^{-\alpha t}\right]dt.$$

We have

$$I = T - \frac{1-e^{-\alpha T}}{2\sinh(\alpha T)}\int_0^T \sinh(\alpha t)\,dt - \int_0^T e^{-\alpha t}\,dt.$$

Since

$$\int_0^T \sinh(\alpha t)\,dt = \frac{\cosh(\alpha T)-1}{\alpha} \quad \text{and} \quad \int_0^T e^{-\alpha t}\,dt = \frac{1-e^{-\alpha T}}{\alpha},$$

it follows that

$$I = T - \frac{1-e^{-\alpha T}}{2\sinh(\alpha T)} \cdot \frac{\cosh(\alpha T)-1}{\alpha} - \frac{1-e^{-\alpha T}}{\alpha}.$$

We can write this more compactly as

$$I = T - \frac{1-e^{-\alpha T}}{\alpha}\left(1 + \frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)}\right).$$

Thus, we obtain

$$\int_0^T x^*(t)\,dt = \frac{NT}{2} + \frac{N\lambda}{\gamma\sigma^2 T}\,I.$$

Substituting back into the expression for Δ,

$$\Delta = \frac{\lambda}{T}\left[\frac{NT}{2} + \frac{N\lambda}{\gamma\sigma^2 T}\,I - \frac{NT}{2}\right] = \frac{N\lambda^2}{\gamma\sigma^2 T^2}\,I.$$

That is, the closed-form expression for the VWAP difference is

$$\Delta = \frac{N\lambda^2}{\gamma\sigma^2 T^2}\left[T - \frac{1-e^{-\alpha T}}{\alpha}\left(1 + \frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)}\right)\right]$$

□

Proof of Corollary 2

Recall the expression for the VWAP difference:

$$\Delta = \frac{N\lambda^2}{\gamma\sigma^2 T^2}\left[T - \frac{1-e^{-\alpha T}}{\alpha}\left(1 + \frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)}\right)\right],$$

with

$$\alpha = \sqrt{\frac{\gamma\sigma^2}{2\eta}}.$$

As $\gamma \to 0$, we have $\alpha \to 0$. For small x (with $x = \alpha T$), recall:

$$e^{-x} \approx 1 - x + \frac{x^2}{2} - \frac{x^3}{6}, \quad \sinh(x) \approx x + \frac{x^3}{6}, \quad \cosh(x) \approx 1 + \frac{x^2}{2}.$$

Thus,

$$1 - e^{-\alpha T} \approx \alpha T - \frac{(\alpha T)^2}{2} + \frac{(\alpha T)^3}{6},$$

so that

$$\frac{1-e^{-\alpha T}}{\alpha} \approx T - \frac{\alpha T^2}{2} + \mathcal{O}(\alpha^2).$$

Also,

$$\frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)} \approx \frac{(\alpha T)^2/2}{2\alpha T} = \frac{\alpha T}{4}.$$

Therefore,

$$1 + \frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)} \approx 1 + \frac{\alpha T}{4}.$$

Multiplying, we have:

$$\frac{1-e^{-\alpha T}}{\alpha}\left(1 + \frac{\cosh(\alpha T)-1}{2\sinh(\alpha T)}\right) \approx \left(T - \frac{\alpha T^2}{2}\right)\left(1 + \frac{\alpha T}{4}\right) \approx T - \frac{\alpha T^2}{4}.$$

Thus, the bracketed term simplifies to

$$T - \left(T - \frac{\alpha T^2}{4}\right) \approx \frac{\alpha T^2}{4}.$$

Substitute back into Δ:

$$\Delta \approx \frac{N\lambda^2}{\gamma\sigma^2 T^2} \cdot \frac{\alpha T^2}{4} = \frac{N\lambda^2}{4\gamma\sigma^2}\,\alpha.$$

Using

$$\alpha = \sqrt{\frac{\gamma\sigma^2}{2\eta}},$$

we obtain

$$\Delta \approx \frac{N\lambda^2}{4\gamma\sigma^2}\sqrt{\frac{\gamma\sigma^2}{2\eta}} = \frac{N\lambda^2}{4}\sqrt{\frac{1}{2\eta\gamma\sigma^2}}.$$

Hence, as $\gamma \to 0$,

$$\Delta \sim \frac{N\lambda^2}{4\sqrt{2\eta\sigma^2}}\,\gamma^{-1/2}.$$

□

Proof of Theorem 2

The HJB Equation. Recall the value function

$$V(t,x,y) = \sup_{u(\cdot)} \mathbb{E}\left[\Pi[x] - \frac{\gamma}{2}\,\mathrm{Var}(\Pi[x]) \,\middle|\, x(t)=x,\ \Gamma(t)=y\right].$$

The state dynamics are:

- **Inventory:** $dx(t) = u(t)\,dt$.
- **Volume:** The Gamma bridge $\Gamma(t)$ has finite variation with generator (see, e.g., Émery and Yor (2004))
$$\mathcal{L}^{\Gamma}\psi(y) = \frac{1-y}{T-t}\,\psi'(y),$$
so that
$$d\Gamma(t) = \frac{1-\Gamma(t)}{T-t}\,dt.$$

In addition, the control appears in the cost via the temporary impact cost $-\eta\,u(t)^2$. Hence, by the dynamic programming principle the HJB equation is

$$V_t + \sup_{u\in\mathbb{R}}\left\{u\,V_x - \eta\,u^2\right\} + \frac{1-y}{T-t}\,V_y + \frac{\gamma\sigma^2}{2}\big(N\,y - x\big)^2 = 0, \tag{1.35}$$

with a terminal condition that enforces $x(T) = N$.
To solve the supremum in (1.35), note that the Hamiltonian

$$H(u) = u\,V_x - \eta\,u^2$$

is quadratic in u. Its first–order condition yields

$$V_x - 2\eta\,u = 0 \quad\Longrightarrow\quad u^*(t) = \frac{V_x}{2\eta}.$$

By our cost convention (minimization) we write

$$u^*(t) = -\frac{1}{2\eta}\,V_x.$$

Substituting this optimal control back into the Hamiltonian gives

$$H^* = \frac{V_x^2}{4\eta}.$$

Thus, the HJB (1.35) becomes

$$V_t - \frac{1}{4\eta}(V_x)^2 + \frac{1-y}{T-t}\,V_y + \frac{\gamma\sigma^2}{2}\big(N\,y - x\big)^2 = 0. \tag{1.36}$$

Quadratic Ansatz. We now postulate a quadratic form for the value function:

$$V(t,x,y) = A(t)x^2 + B(t)xy + C(t)x + D(t)y^2 + E(t)y + F(t), \tag{1.37}$$

where $A(t), B(t), C(t), D(t), E(t), F(t)$ are functions to be determined.

Derivatives of the Ansatz. Compute the partial derivatives:

$$\begin{aligned}
V_t &= A'(t)x^2 + B'(t)xy + C'(t)x + D'(t)y^2 + E'(t)y + F'(t),\\
V_x &= 2A(t)x + B(t)y + C(t), \quad V_{xx} = 2A(t),\\
V_y &= B(t)x + 2D(t)y + E(t), \quad V_{yy} = 2D(t).
\end{aligned}$$

Also, note that

$$\begin{aligned}
(V_x)^2 &= \Big(2A(t)x + B(t)y + C(t)\Big)^2\\
&= 4A(t)^2x^2 + 4A(t)B(t)xy + 4A(t)C(t)x + B(t)^2y^2 + 2B(t)C(t)y + C(t)^2.
\end{aligned}$$

Substituting into the HJB equation. Substitute the derivatives into (1.36):

$$\begin{aligned}
&A'(t)x^2 + B'(t)xy + C'(t)x + D'(t)y^2 + E'(t)y + F'(t)\\
&\quad - \frac{1}{4\eta}\Big[4A(t)^2x^2 + 4A(t)B(t)xy + 4A(t)C(t)x + B(t)^2y^2 + 2B(t)C(t)y + C(t)^2\Big]\\
&\quad + \frac{1-y}{T-t}\Big[B(t)x + 2D(t)y + E(t)\Big] + \frac{\gamma\sigma^2}{2}\Big(x^2 - 2Nxy + N^2y^2\Big) = 0.
\end{aligned}$$

Since this must hold for all x and y, we now match coefficients for like monomials.

Coefficient of x^2. The x^2 terms come from:

$$A'(t)x^2 - \frac{1}{4\eta}\cdot 4A(t)^2x^2 + \frac{\gamma\sigma^2}{2}\,x^2.$$

Thus,

$$A'(t) - \frac{A(t)^2}{\eta} + \frac{\gamma\sigma^2}{2} = 0.$$

This is a Riccati equation for $A(t)$.

Coefficient of xy. The xy terms arise from:

$$B'(t)xy - \frac{1}{4\eta}\cdot 4A(t)B(t)xy + \frac{B(t)}{T-t}\,xy - \gamma\sigma^2\,Nxy.$$

Thus,

$$B'(t) - \frac{A(t)B(t)}{\eta} + \frac{B(t)}{T-t} - \gamma\sigma^2\,N = 0.$$

Coefficient of x. The x terms are:

$$C'(t)x - \frac{1}{4\eta} \cdot 4A(t)C(t)x,$$

so that

$$C'(t) - \frac{A(t)C(t)}{\eta} = 0.$$

Coefficient of y^2. The y^2 terms are:

$$D'(t)y^2 - \frac{1}{4\eta}B(t)^2y^2 + \frac{2D(t)}{T-t}y^2 + \frac{\gamma\sigma^2}{2}N^2y^2.$$

Thus,

$$D'(t) - \frac{B(t)^2}{4\eta} + \frac{2D(t)}{T-t} + \frac{\gamma\sigma^2}{2}N^2 = 0.$$

Coefficient of y. The y terms come from:

$$E'(t)y - \frac{1}{4\eta} \cdot 2B(t)C(t)y + \frac{E(t)}{T-t}y,$$

so that

$$E'(t) - \frac{B(t)C(t)}{2\eta} + \frac{E(t)}{T-t} = 0.$$

Constant Term. The constant term is:

$$F'(t) - \frac{C(t)^2}{4\eta} + \frac{E(t)}{T-t} = 0.$$

In a fully explicit derivation, one finds that matching all constant-order terms from the risk-penalty expansion shows the leftover is zero.

Solution of the ODE System. We summarize the system:

$$A'(t) - \frac{A(t)^2}{\eta} + \frac{\gamma\sigma^2}{2} = 0, \tag{1.38}$$

$$B'(t) - \frac{A(t)B(t)}{\eta} + \frac{B(t)}{T-t} - \gamma\sigma^2 N = 0, \tag{1.39}$$

$$C'(t) - \frac{A(t)}{\eta}C(t) = 0, \tag{1.40}$$

$$D'(t) - \frac{B(t)^2}{4\eta} + \frac{2D(t)}{T-t} + \frac{\gamma\sigma^2}{2} N^2 = 0, \tag{1.41}$$

$$E'(t) - \frac{B(t)C(t)}{2\eta} + \frac{E(t)}{T-t} = 0, \tag{1.42}$$

$$F'(t) - \frac{C(t)^2}{4\eta} + \frac{E(t)}{T-t} = 0. \tag{1.43}$$

Solving for $A(t)$. The Riccati equation (1.38) is

$$A'(t) - \frac{A(t)^2}{\eta} + \frac{\gamma\sigma^2}{2} = 0.$$

A standard transformation shows that its solution is

$$A(t) = \eta\,\alpha\,\coth\Big(\alpha(T-t)\Big), \quad \text{with} \quad \alpha = \sqrt{\frac{\gamma\sigma^2}{2\eta}}.$$

Note that as $t \to T$, $\coth\Big(\alpha(T-t)\Big) \sim \frac{1}{\alpha(T-t)}$, ensuring that

$$\lim_{t\to T}(T-t)A(t) = \eta.$$

The other ODEs are linear and can be solved using integrating factors. For example, (1.40) implies

$$C(t) = \frac{C_0}{T-t},$$

with C_0 determined by the terminal condition.

Similarly, one may obtain (with an appropriate normalization) an explicit expression for $B(t)$ such as

$$B(t) = -2A(t) + \frac{2\eta}{T-t},$$

and the functions $D(t)$, $E(t)$, $F(t)$ can be expressed as definite integrals. (For instance, one may write)

$$D(t) = \int_t^T \left[\lambda\sigma^2 - \frac{B(s)^2}{4\eta}\right]\mu(s)\,ds,$$

$$E(t) = \frac{1}{T-t}\int_t^T \Big[B(s) + 2D(s)\Big]ds,$$

$$F(t) = \int_t^T \left\{\frac{C(s)^2}{4\eta} - \frac{E(s)}{T-s}\right\}ds + F_0,$$

where $\mu(s)$ is an integrating factor and F_0 is determined by the terminal condition.

The optimal trading inventory. Given the value function (1.37), the optimal control is

$$u^*(t) = -\frac{1}{2\eta}\,V_x(t,x,y) = -\frac{1}{2\eta}\Big(2A(t)x + B(t)y + C(t)\Big).$$

Hence, the optimal inventory $x^*(t)$ satisfies the ODE

$$\dot{x}^*(t) = -\frac{1}{2\eta}\Big(2A(t)x^*(t) + B(t)\,\Gamma(t) + C(t)\Big), \quad x^*(0) = 0, \quad x^*(T) = N. \tag{1.44}$$

This is a linear ODE. Define the integrating factor

$$M(t) = \exp\Big(\int_0^t \frac{A(s)}{\eta}\,ds\Big).$$

Multiplying (1.44) by $M(t)$ gives

$$\frac{d}{dt}\Big[M(t)x^*(t)\Big] = -\frac{M(t)}{2\eta}\Big[B(t)\,\Gamma(t) + C(t)\Big].$$

Integrate from 0 to t:

$$M(t)x^*(t) = -\frac{1}{2\eta}\int_0^t M(s)\Big[B(s)\,\Gamma(s) + C(s)\Big]ds.$$

Thus,

$$x^*(t) = \exp\Big(-\int_0^t \frac{A(s)}{\eta}\,ds\Big)\left\{-\frac{1}{2\eta}\int_0^t \exp\Big(\int_0^s \frac{A(r)}{\eta}\,dr\Big)\Big[B(s)\,\Gamma(s) + C(s)\Big]ds\right\}.$$

To enforce the terminal condition $x^*(T) = N$, introduce a scaling factor

$$\kappa = \frac{N}{x^*(T)},$$

so that the final optimal inventory is

$$x^*(t) = \kappa \exp\Big(-\int_0^t \frac{A(s)}{\eta}\,ds\Big) \left\{-\frac{1}{2\eta}\int_0^t \exp\Big(\int_0^s \frac{A(r)}{\eta}\,dr\Big)\Big[B(s)\,\Gamma(s) + C(s)\Big]ds\right\}.$$

□

1.9.2 Parameter calibration in the literature

This appendix documents the empirical and theoretical arguments that motivate the baseline choice $\gamma = 0.10$ used throughout the numerical illustrations. Table 1.3 summarizes the values explored in the high-frequency market-making literature, all of which adopt a CARA utility with coefficient γ.

Table 1.3: Published γ values in limit-order-book models

Paper	Asset	Horizon T	γ range
Avellaneda and Stoikov (2008)	US equities	1 hour	0.01 ; 0.10 ; 0.50
Stoikov and Sağlam (2009)	Index options	1 day	0.006; 0.10
Guéant et al. (2013)	EU equities	10 min	0.01

Chapter 2

Tick Size, HFT and Inverted Exchanges

Exchanges operate various fee structures to attract liquidity on their platforms. The most popular, make-take exchanges, offer rebates to liquidity makers while charging fees to liquidity takers. An intriguing development is the rise of inverted exchanges, which charge negative taker fees and positive maker fees. More than 180 billion shares trade on these exchanges every year, representing 10% of total stock trading volume. This study develops a theoretical model to analyze the implications of inverted exchanges on liquidity provision, particularly in the presence of High-Frequency Traders. We demonstrate that inverted exchanges mitigate inefficiencies arising from tick-size constraints by enabling a finer price grid. When HFTs queue up in the limit order book to provide liquidity at the nearest tick, they prevent liquidity demanders from obtaining price improvements at the mid-point. Inverted exchanges solve the mismatch between an HFT's price priority and a liquidity demander's time priority. Our model yields testable predictions on HFT activity, relative exchange trading volumes, and order book imbalance, which we confirm using high-frequency data.

2.1 Introduction

Over the past decade, financial market microstructure has evolved significantly due to advancements in trading technology and the emergence of alternative exchange fee models. One notable innovation is the rise of inverted exchanges, which differ from the conventional make-take model by charging fees to liquidity providers and offering rebates to liquidity takers. At first glance, this pricing structure appears counterintuitive, as traditional market-making theories suggest that liquidity provision should be incentivized through rebates rather than penalized. This paper examines the economic rationale behind inverted exchanges and their implications for market quality, particularly in the presence of High-Frequency Trading (HFT). HFT firms, leveraging speed and sophisticated order execution strategies, dominate liquidity provision in modern markets. In traditional make-take structures, HFTs compete for price priority by queuing at the nearest tick. While this behavior ensures efficient liquidity provision in unconstrained environments, it creates inefficiencies when tick size constraints prevent finer price adjustments. This rigidity reduces price improvement opportunities, particularly for non-HFT participants who cannot undercut HFTs. This observation raises a key research question: How do inverted exchanges alter liquidity dynamics in a market dominated by HFTs, and what are their broader implications for execution quality?

This paper addresses this question by developing a theoretical model of liquidity provision with exchange fees. Our starting point is the observation that tick size constraints can distort liquidity dynamics by limiting price improvement opportunities for non-HFT traders. Inverted fee structures offer a potential solution by enabling a more granular pricing mechanism. Intuitively, by rewarding liquidity demanders and aligning their time priority with HFTs' price priority, inverted exchanges facilitate price improvements and reduce queuing inefficiencies, effectively reshaping competition between HFTs and non-HFTs.

In our model, there are many HFTs providing liquidity to non-HFTs in each time period. Thanks to their superior technology, HFTs can exploit stale quotes faster than non-HFT participants. Thus, they act as market makers in this economy. Trading occurs because non-HFTs have an inelastic demand while risk-neutral HFTs take the other side. Under continuous pricing, perfect competition between HFTs ensures HFTs can only break even. Spreads reflect the trade-off between fundamental asset risk and profits from liquidity traders. In reality, prices are discrete and will be higher than under continuous pricing. The inability of HFTs to further compete imposes a negative externality on non-HFT participants. HFTs are able to extract intermediation rents despite perfect competition between them.

When the tick size is binding, non-HFT traders cannot price-improve in finer increments using limit orders, forcing them to either match existing quotes or cross the spread using market orders. This ensures that their orders are executed first when liquidity demand arises, capturing profits from the bid-ask spread. Thus, the tick-size constraint effectively acts as a barrier to entry for non-HFT liquidity providers. Transaction fees do not alter this outcome.

In contrast to transaction fees, exchange fees are charged based on execution, not order type. While a market order is always liquidity-taking, the fees on a limit order depends on whether it provided liquidity or not. When the limit order immediately lifts another order, it is removing liquidity and incurs a take-fee similarly to the market order. When the limit order is resting but gets matched to an order posted afterwards, it is offered a rebate. Hence, a limit order offers cheaper execution compared to market orders only when it is able to first rest in the order book.

To reflect this, we distinguish between two types of non-HFT traders: Retail Traders (RT) who exclusively submit market orders, and Mutual Funds (MF), which have the flexibility to submit either market or limit orders. Beyond the price difference, the limit order changes the nature of the fees: market orders are liquidity taking while marketable limit orders are liquidity making. Hence, a market and a limit order at the same price will be charged different fees, effectively resulting in different net execution costs. Market orders ensures execution while limit orders offer price improvements. Nonetheless, non-HFTs cannot compete on speed with limit orders and would only use these if they have guaranteed execution.

Under continuous pricing without fees, Mutual Funds are able to execute at the fundamental value, because any limit order they quote away from the fundamental value gets arbitraged away by HFTs. The bid-ask spread only reflects fundamental value jump risk: HFTs make a profit on Retail Trader flow while losing money on value jumps opposite to their quotes. This no longer stands true in discrete pricing. When price increments are discrete, HFTs quote at the nearest tick above the break even spread under continuous pricing. Similarly, Mutual Funds can only quote the ticks near the fundamental value. Thus, HFTs earn additional profits on both Mutual Funds and Retail Traders compared to the continuous pricing benchmark. When the tick size is infinitesimally small, this converges to the continuous case with no rent.

That is, when the tick size is binding, HFTs extract a liquidity rent from non-HFTs. First, as mentioned above, they benefit from the incremental profits of forced market orders from Mutual Funds which generate additional fees. Mutual Funds are no longer able to quote within the bid-ask, preventing them from posting limit orders. They are effectively behaving like Retail Traders and using market orders only, increasing the HFT's profits. Additionally, in contrast to the continuous setting, HFTs can no longer compete over prices and tighten the spread. Therefore, they extract an additional profit on all non-HFTs, equal to the difference between the equilibrium spread under continuous pricing and the constrained price.

Inverted fees level the playing field between HFTs and Mutual Funds. Instead of incurring a take-fee, liquidity takers are offered a rebate. There are three possible cases in this microstructure, depending on the asset volatility, fraction of Retail Traders and tick size. First, there is a "pooling" region when the spread is narrow relatively to the tick size, where Mutual Funds are forced to use market orders. In this region, Mutual Funds effectively behave like Retail Traders. With a low volatility and a high fraction of uninformed Retail

Traders, HFTs can break even with tight spreads, but quotes must be at least one-tick wide. Mutual Funds' limit orders cannot be priced better than existing HFT orders, and market orders are the only order type guaranteeing them execution. HFTs extract the highest profits, because they cannot compete with each other to tighten the spread. Second, there is an inverted equilibrium, Mutual Funds always post on inverted exchanges, to undercut orders on traditional venues and obtain price priority. Asset risk is low but not low enough for HFTs to out-compete the Mutual Funds. Third, conversely, there is also a maker-taker equilibrium, where Mutual Funds always post on the traditional exchange and HFTs post on the inverted exchange. Of particular note, an inverted exchange does not attract liquidity from a specific trader type.

In the baseline model, there is a single unit of liquidity in the order book at the bid and ask, as in Glosten and Milgrom (1985). In reality, HFTs can provide liquidity at deeper levels because the effective spread is higher than the break-even spread. Thus, we introduce depth, such that the marginal profits to liquidity provision become zero. HFTs first compete on prices, then order queue. When the spread is sufficiently narrow, HFTs can still extract some profits on subsequent orders resting at the top of the book. The market is cleared at a book depth for which the last marginal liquidity provider breaks even. Higher rebates increase the profits of liquidity provision,and as a consequence, book depth. However, a long queue defeats its own purpose: inverted exchanges with deep order books can no longer be used to undercut on prices.

When the tick size is severely binding, inverted exchanges should dominate overall trading volume. In reality, even for the most tick-constrained stocks, the market share of inverted venue caps peaks at 20%. Random order routing would still predict twice as much inverted volume as in practice. One way to explain this is the principal-agent problem arising from the broker's agency, choosing the route maximizing their fees irrespective of execution price. We offer an alternative theoretical explanation to reconcile the model with this observed inverted usage: fragmentation. Under market fragmentation, HFTs post liquidity on both exchanges. Because of the difference in exchange fees, the marginal profits to liquidity provision at a fixed depth are much higher on the make-take exchanges than on the inverted venue. As a consequence, equilibrium depth is higher on the traditional venue. When there is a fundamental value jump, the whole order book is swept. The make-take venue has a much deeper book and as a consequence, occupies a larger share of total trading. More concretely, this can also be interpreted as the arrival of a large institutional liquidity trader, requiring all liquidity available across all venues. Interestingly, trading volume is the same across the two types of exchanges in the absence of asset value jumps.

Finally, our model yields several testable predictions, which we validate using high-frequency trading data. First, we predict that inverted exchanges enable non-HFT participants to compete more effectively by lowering the cost of price improvement. To do so, we measure the HFT participation rate as the ratio of NBBO transactions where at least one of the counterparties is an HFT. We find that inverted usage increases with HFT activity, suggesting that inverted venues become more appealing when HFT competition is high. Second, trading on inverted exchanges should be concentrated on the side of the order book experiencing greater imbalance. Further, our model suggests that inverted usage – the inverted share of total trading volume across all exchanges – is higher when tick sizes are large relative to price levels. The empirical evidence strongly supports these predictions, as we document significant changes in trading behavior, order book dynamics, and exchange-level liquidity fragmentation.

Beyond contributing to the literature on exchange fee models and market microstructure, our findings have important policy implications for tick size design and best routing practices. The U.S. equity markets operate under a uniform minimum tick size of one cent for securities trading above one dollar, a rule designed to enhance price transparency but one that fails to account for variation in stock price levels and liquidity needs. Our results suggest that inverted exchanges emerge as a partial market-driven solution to these regulatory constraints but also introduce new forms of liquidity fragmentation, arising from the discrepancy between gross and net prices. We elaborate on this in section 4.

The remainder of this paper is as follows. Section 3.2 discusses relative literature and empirical motivation for our model. In section 3.4, we present the data and further stylized facts. In section 2.4, we lay out the agents' optimization problems and present the model. In section 2.5, we characterize liquidity under different market environments. In section 2.6, we empirically test the predictions from the theory. In section 2.7, we present some policy implications. Section 3.8 concludes. The Appendix includes proofs of all propositions and additional tables and figures.

2.2 Related Literature and Empirical Motivation

This study contributes to multiple strands of literature, particularly research on exchange fee structures, market microstructure, and high-frequency trading. The pricing models adopted by exchanges have significant implications for liquidity provision, price efficiency, and market fragmentation. Harris (2013) discusses how make-take fee structures can distort displayed prices, complicating best execution practices, while Battalio et al. (2016) show that order routing decisions often maximize fee rebates rather than execution quality. Our theoretical model extends this discussion by examining how inverted exchanges affect liquidity dynamics in tick-constrained markets. Further, this paper complements the literature on collusion in dealer markets (Dutta and Madhavan (1997),Huang and Stoll (2001)). With the emergence of new technologies, this phenomenon has evolved to algorithmic collusion in market making (Cartea et al., 2022), whereby trading algorithms tacitly agree to collude on the optimal price grid and extract a rent from the tick size constraint.

Our findings also relate to the literature on market-making and HFT behavior. The classic work of Glosten and Milgrom (1985) models market-making under asymmetric information. Subsequently, a large body of research emerged, focusing on HFTs as market makers (Budish et al. (2015), Aït-Sahalia and Sağlam (2023)). Consistent with these studies, we assume that HFTs dominate liquidity provision due to their speed advantage. However, we depart from prior work by explicitly modeling how non-HFT participants respond to exchange fee incentives, particularly through strategic order placement on inverted venues. By distinguishing between retail traders and mutual funds, our framework highlights the role of order type choice in shaping market quality.

Our paper is most closely related to Li and Ye (2021) and Li et al. (2021). In a model of competition for liquidity provision, they show that execution algorithms may benefit from large tick sizes. Execution algorithms are unable to compete at the best bid-ask when the tick size is binding. When it is not binding, they alternate between limit and market orders depending on which side of the order book the price is leaning on. In contrast, our focus is on exchange fees. With a coarse pricing grid, the fundamental value of the asset will lean closer to one tick than the other. Hence, without fees, the mid-price does not reflect the true value. Take-make fees reduces this asymmetry by effectively shrinking the net spread, which has important implications on trading volume and top of book depth.

Additionally, to the best of our knowledge, this is the first paper to provide a theoretical framework for understanding inverted exchanges. Our model is able to match stylized facts consistent with the existing empirical literature. Notably, our paper provides a theoretical foundation to understand Comerton-Forde et al. (2019), whereby increased trading activity on inverted venues improves pricing efficiency when the minimum tick size is binding. As in Malinova and Park (2015), lowering fees to remove liquidity increases trading activity more substantially than raising rebates to providers. Corroborating Cardella et al. (2015), fee structure in our model determines the exchange's market share.

Finally, our work ties into policy discussions on optimal tick size regulation. The SEC's Rule 612 mandates a uniform minimum tick size of one cent for stocks priced above one dollar, a policy designed to enhance order book transparency. However, as demonstrated by Dayri and Rosenbaum (2015), Dyhrberg et al. (2023) and Fleming et al. (2024) rigid tick size regimes can lead to unintended consequences, such as excessive queuing and reduced price competition. In this line of literature, tick size design is centered around the trade-off between HFT intermediation and excessive undercutting and queuing. Our findings suggest that inverted exchanges partially offset these inefficiencies but also introduce new challenges related to liquidity fragmentation. This underscores the need for a more flexible approach to tick size regulation that accounts for differences in stock characteristics and market conditions.

2.3 Data and Stylized facts

2.3.1 Data Description

In this paper, we use four types of data to study the impact of exchange fees on market microstructure. First, we use the NYSE Trade and Quote (TAQ) database for tick-by-tick trade and quote data of all activity within the U.S. National Market System. Second, we merge this dataset with the NASDAQ HFT dataset, which identifies HFTs according to their trading and quoting behavior. Third, we use the SEC Market Information Data Analytics System (MIDAS) for their data on market depth. MIDAS collects data from consolidated tapes as well as proprietary feeds. These are typically used by high-frequency traders who have more sophisticated data needs. Finally, we use CRSP for fundamental variables such as industry sector and the number of shares outstanding to derive market capitalization. Our sample period runs from February 22, 2010 to February 26, 2010.

The NASDAQ OMX made the HFT dataset available to academics under a Non Disclosure Agreement. Previous papers using this dataset include Carrion (2013), Brogaard et al. (2014), O'Hara et al. (2014) and Brogaard et al. (2014). It contains all lit trading and quoting activity on a sample of 120 randomly selected stocks. The NASDAQ categorizes market participants as HFT based on their trading behavior, such as inventory management, end-of-day positioning, order cancellation rate and trading frequency. This dataset allows us to directly observe the HFT liquidity provision. Although it is restricted to the NASDAQ only, it accounts for more than a third of total trading activity and should be representative of general HFT behavior. In particular, we can rule out that HFTs may prefer to concentrate in make-take exchanges for external reasons. To illustrate stylized facts on intraday patterns, we draw charts using all of the SP500 constituents. For the empirical analysis, we separately used both the HFT dataset on the 120 stocks sample and the TAQ dataset on the SP500.

2.3.2 Overview of exchanges

The make-take model is the most popular type of fee structure: the exchange provides a rebate to liquidity makers while charging a trading fee to liquidity takers. In our sample, they include NYSE, NASDAQ, AMEX, ARCA, BATSZ, EDGX, PHLX and represent over 85% of total trading activity on lit venues. Conversely, inverted exchanges include EDGA, BX, BATSY and NSX. Finally, CHX is the only exchange using a fee-fee structure, a model charging positive fees to both liquidity takers and makers.

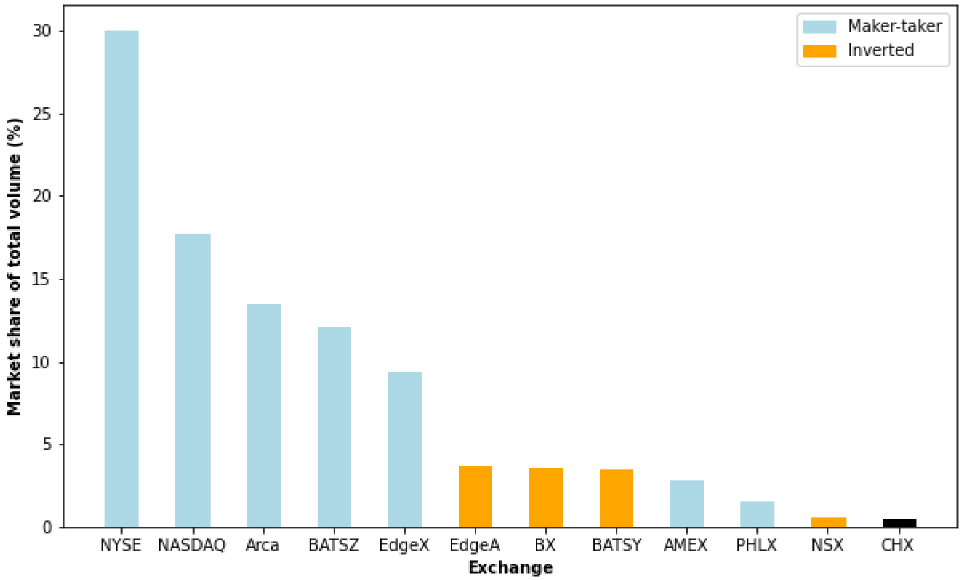


Figure 2.1: Exchange market share by fee structure

The US equity trading landscape operates within the confines of Regulation National Market System (NMS), a regulatory framework designed to ensure trades are executed at optimal prices. Its primary objective is to shield market participants from engaging in suboptimal transactions solely due to lack of awareness or accessibility to superior pricing elsewhere. Regulation NMS introduces the notion of a protected quotation, denoting a quote exhibited on an exchange that is immediately and broadly accessible, thereby establishing a clear rule to uphold its intent.

Furthermore, in adherence to Regulation NMS, trading centers are obligated to prevent trade-through, instances where transactions are settled at prices inferior to protected quotations. However, market participants are permitted to utilize inter-market sweep orders (ISOs) to execute orders on other exchanges, provided additional ISOs are routed, as necessary, to execute against the full displayed size of any superior-priced protected quotes. The regulation, by decreeing that trades must not occur at prices inferior to the best available price, can be construed as reinforcing brokers' obligation to route orders for best execution—a mandate that serves both as an investor protection measure and a requirement to secure the most advantageous terms for clients. Parameters for assessing best execution encompass factors such as speed, probability of execution, and price improvement, which denotes the opportunity for an order to execute at a superior price to the prevailing quote. These protected quotations are aggregated into a consolidated format known as the Security Information Processor (SIP), a real-time data feed encompassing all updates to protected quotes. At any given juncture, market participants can ascertain the optimal price available among protected quotations via reference to the SIP. Despite furnishing a common reference point, the SIP does not eradicate asynchronous market data dissemination. The identical information available through the SIP is also accessible via exchanges' proprietary direct feeds, with participants presented various options for accessing these disparate data sources. Hence, the TAQ data is not error-proof. Following the existing literature (Falkenberry (2001), Wah et al. (2017)), we filter the data to only select trades and quotes that reflect normal market activity. We restrict our focus to regular market hours from 09:30AM to 04:00PM ET. For quotes, we ignore locked and crossed markets, i.e. quotes that would indicate a negative spread. We use the National Best Bid Offer (NBBO), i.e. the highest bid and lowest ask in a security, sourced from among all available exchanges or trading venues. We further exclude abnormal quotes, as defined by quotations where the National Best Offer (NBO) is outside of the $[\frac{1}{3}NBB, 3NBO]$ range. For trades, we remove trades with correction indicators with sale condition codes B, C, G, H, L, M, N, O, P, Q, R, T, U, V, W, Z, 4, 5, 6, 7, 8, or 9, corresponding to abnormal trades. Finally, we consider trades occuring far from the NBBO to be aberrant when they do not lie within $[0.9NBB, 1.1NBO]$ Pricing structures may vary depending on a participant's volume on a given venue, with attainment of the most favorable, top-tier rebates contingent upon exceeding specified volume thresholds.

2.3.3 Intraday patterns and the tick size constraint

Volume, volatility, and spread exhibit a discernible intraday pattern that underpins market behavior. For example, the U-shaped pattern of trading volume is widely documented (Wood et al. (1985)). Figure 2.3 clearly exhibits the downward trend of volatility during the day. Trading volume experiences pronounced concentration during the initial and final hours of the trading session, with the first hour notably characterized by heightened volatility compared to subsequent periods. We observe a gradual narrowing of the bid-ask spread and an increase in the limit order book (LOB) depth throughout the day. Admati and Pfleiderer (1988) show that liquidity traders concentrate their activity and this effect is amplified when there are informed insiders.

The contracting spread pattern [1] has important ramifications for venue selection. In order

[1] As showcased in figure 2.11, a discernible trend emerges throughout the trading day, wherein an increas-

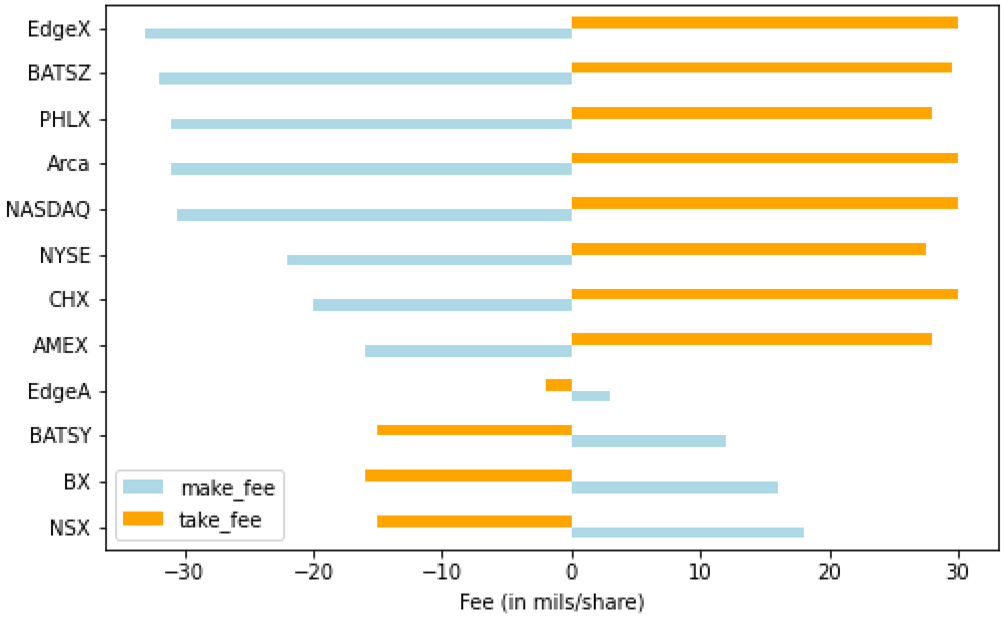


Figure 2.2: Fee schedule by exchange

to study venue switching, we introduce the concept of a "tick-constrained" stock—a security whose bid-ask spread consistently remains within a few ticks.
With these institutional details in mind, we illustrate in the next section the theoretical mechanism underlying HFT liquidity provision, before proceeding to a more complete model with price discreteness and inverted fees.

2.4 Model Setup

Agents. We consider a continuous-time trading environment where liquidity providers and liquidity demanders interact strategically. We are interested in determining what order types agents are sending and the associated equilibrium bid-ask spread. There is a single risky asset with fundamental value V_t which is common knowledge. There are three types of traders:

- **High Frequency Traders (HFTs)** are risk-neutral traders maximizing expected profits. They have a speed advantage, enabling them to place and cancel orders faster than other traders. By placing resting limit orders, they profit from the bid-ask

ing number of stocks transition towards becoming tick-constrained.

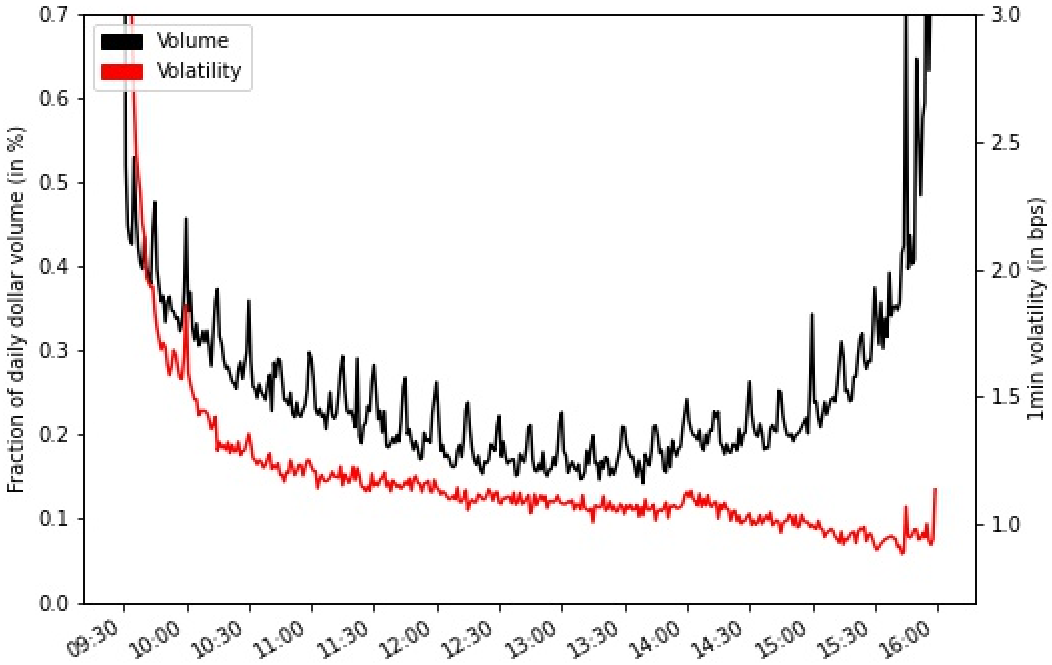


Figure 2.3: Intraday pattern of volume and volatility

spread.

- **Retail Traders (RT)** are exogenous liquidity demanders and can only submit market orders, accepting prevailing bid-ask prices. They do not compete with HFTs for liquidity provision.
- **Mutual Funds (MF)** are strategic liquidity demanders and have the flexibility to submit either limit or market orders to minimize transaction costs.

HFTs are present at every time period, providing liquidity. RTs and MFs only live for a single period and arrive randomly. We refer to them as non-HFTs. The non-HFT arrival is a Retail Trader with probability β and a Mutual Fund with probability $1-\beta$.

Trading motivation. Contrary to HFTs, non-HFTs are liquidity traders and have an inelastic demand. For now, we focus on the simple case where one needs to buy or sell one unit of asset. HFTs provides quotes to buy or sell, enabling gains from trade in this economy.

Modelling spread. Conditional on being able to fill their order, non-HFTs minimize their expected execution cost. HFTs are perfectly competitive and set the bid-ask spread to maximize their expected profits.

Limit Order Book. We model the Limit Order Book (LOB) as a pair of best bid and best offer. For limit orders resting at the same price, orders arriving earlier execute before later orders. HFTs are equally fast and their order messages are processed in random order if they arrive at the same time. The fundamental value V_t is common knowledge but when liquidity providers are unable to update their quotes following value jumps, they may get sniped by HFTs. For example, if there is a positive value jump, the previous sell order is now underpriced and HFTs race to exploit this arbitrage. Mutual funds are always subject to sniping risk because they are slower than HFTs.

Events and asset dynamics. An event is either a non-HFT coming to the market or a change in the value of the asset. As in Budish et al. (2015), we model the arrival of events as a compound Poisson process. Non-HFTs arrive at the market with Poisson intensity λ_L while the asset's fundamental value V_t has random jumps with arrival rate λ_V and i.i.d. jump values $\sigma_i \sim \mathcal{U}[-\Sigma, +\Sigma]$. Therefore, with probability $\pi = \frac{\lambda_V}{\lambda_L + \lambda_V}$, the next event is a value jump and the asset value increases or decreases with equal probability by an amount σ_i. With probability $1 - \pi$, the event is a non-HFT trader arrival. Among these non-HFTs, recall that there are β MFs and $1 - \beta$ RTs. Assume that RTs use buy or sell market orders with equal probability. Thus, the next event is an RT buy market order with probability $\frac{1}{2}(1 - \beta)(1 - \pi)$.

Exchange fees. Traditionally, make-take exchanges operate by charging take fees higher than make fees, incurring a profit of $f_t - f_m$ for each transaction. Hence, we assume $0 < f_m < f_t$. When there is no exchange fees, our model degenerates into Li et al. (2021). Fees are charged based on execution, irrespective of order type. As an example, a limit order lifting the ask removes liquidity. On the other hand, a limit order that does not get executed immediately provides liquidity and pays a make fee when filled.

Tick size constraint. Financial markets are quoted with a discrete price grid: a tick is the minimum price increment. This constraint does not exist under continuous pricing.

2.4.1 HFTs

As a benchmark, we first study how the bid-ask spread is set under continuous pricing without fees. In later sections, we show that price discreteness introduces frictions that HFTs are able to exploit. We model a financial market where N high-frequency traders (HFTs) act as liquidity providers, with $N \in [2, \infty]$. HFTs have a speed advantage, allowing them to post and cancel orders faster than other participants. HFTs maximize their expected profits and can choose between posting liquidity or removing it. Let h be the HFTs' quoted half bid-ask spread, i.e. the distance from the fundamental value V_t to the best bid or ask. HFTs strategically choose between liquidity provision and quote sniping. If an HFT provides liquidity, HFTs post a sell limit order at $V_t + h$ and a buy limit order at $V_t - h$. It earns a spread h when a non-HFT lifts the order. However, if the fundamental price moves before execution, the HFT's quote becomes stale, exposing it to the risk of being sniped. Without loss of generality, we study the expected payoff of the HFT's sell limit order. When a buy event occurs, the expected payoff of the HFT is:

$$(1 - \pi) \cdot h - \pi \frac{N-1}{N} \cdot (\sigma - h) \tag{2.1}$$

Conditional on a buy event, with probability $1 - \pi$, non-HFTs lift the order and the HFT receives h. With probability π, there is a value jump and the liquidity provider either manages to cancel their order with probability $\frac{1}{N}$ or fails to do so and gets sniped with probability $\frac{N-1}{N}$. The loss on stale quotes is $\sigma - h$. An HFT's outside option is to snipe stale quotes and does so successfully with probability $\frac{1}{N}$. Thus, the expected payoff for sniping is:

$$\frac{1}{N}\pi(\sigma - h) \tag{2.2}$$

HFTs post limit orders in the order book and earn profits from the bid-ask spread, but they also face the risk of adverse selection (i.e., being picked off by rivals). The decision to post liquidity depends on the expected profit from doing so, relative to the alternative of removing stale quotes posted by other HFTs.

2.4.2 Mutual Funds

Mutual Funds want to minimize their execution cost C by choosing between Market Orders (MO) and Limit Orders (LO). Their action space is defined by the set {MO, LO}. When using market orders, they buy at the available quoted ask and pay a half-spread h over the fundamental value. However, when the MF submits a buy limit order at $V_t + \xi$ with $\xi < h$, i.e. at a price lower than the price quoted by HFTs, HFTs immediately remove that liquidity and earn a profit of ξ by selling above the fundamental value. The MF pays

ξ with a limit order, instead of h with a market order. Hence, it benefits from a negative opportunity cost $h - \xi$. As $\xi \to h$, effectively, MFs execute at the fundamental value and HFTs do not make any profits trading against them. HFTs only earn profits from Retail Traders. Thus, MFs would always prefer to set limit orders quoted the fundamental value.

2.4.3 Equilibrium bid-ask spread

There is perfect competition among the HFTs. In this baseline model, an equilibrium consists of a Mutual Fund's order choice (Market Order or Limit Order) and the half-spread h set by HFTs. In equilibrium, an HFT should be indifferent between liquidity provision and stale quote sniping. Thus, the equilibrium bid-ask spread is the value $h = h_\beta$ which solves:

$$\frac{(1-\beta)(1-\pi)}{(1-\beta)(1-\pi)+\pi}h - \frac{N-1}{N}\frac{\pi(\sigma-h)}{(1-\beta)(1-\pi)+\pi} = \frac{1}{N}\frac{\pi(\sigma-h)}{(1-\beta)(1-\pi)+\pi} \tag{2.3}$$

The denominator $(1-\beta)(1-\pi)+\pi$ reflects the probability of either a value jump or a RT arrival. Because MFs always get executed at the fundamental value, HFTs do not earn any profits from them. They adjust the bid-ask spread to break-even solely based on the RT flow. The left-hand side is the value of providing liquidity to RTs while being subject to sniping risk. The right-hand side is the value of sniping stale quotes. The equilibrium half bid-ask spread is

$$h_\beta = \frac{\pi\sigma}{1-\beta(1-\pi)} \tag{2.4}$$

HFTs always maintain one unit in the LOB at the ask price $V_t + h_\beta$ and one unit at the bid price $V_t - h_\beta$. MFs submit limit orders at V_t when they arrive and HFTs immediately demand liquidity at V_t. When there is a positive (negative) value jump, HFTs race to snipe stale ask (bid) quotes.

In the following section, we present different market microstructures. We first start by deriving the equilibrium under fees. We then add the tick size friction and explain the role of inverted exchanges in discrete pricing. Under this framework, we subsequently derive results for the behavior of the order book, such as depth and trading volume.

2.5 Main Model

In the main model, we introduce exchange fees and the tick size constraint. Exchange fees alter net spreads. When traders provide liquidity, they earn a positive rebate f_m on their transactions. When traders remove liquidity, they incur a take fee f_t on their transactions.

In the presence of fees, canceling their own stale quote is not equivalent to sniping it. The HFT seller can cancel its order without incurring round-trip fees, because fees are only charged on transactions, not orders. In this setting, the sell order has an expected payoff given by:

$$\frac{(1-\beta)(1-\pi)}{(1-\beta)(1-\pi)+\pi}(h+f_m) - \frac{N-1}{N}\frac{\pi(\sigma-(h+f_m))}{(1-\beta)(1-\pi)+\pi} \tag{2.5}$$

The first term corresponds to liquidity provision to RTs while the second term represents the loss from being sniped in a value jump.

In equilibrium, HFTs should be indifferent between liquidity provision and stale quote sniping. Thus, the equilibrium bid-ask spread is the value h_β which solves:

$$\frac{(1-\beta)(1-\pi)}{(1-\beta)(1-\pi)+\pi}(h+f_m) - \frac{N-1}{N}\frac{\pi(\sigma-h-f_m)}{(1-\beta)(1-\pi)+\pi} = \frac{1}{N}\frac{\pi(\sigma-h-f_t)}{(1-\beta)(1-\pi)+\pi} \tag{2.6}$$

The right-hand side is the sniper's profit which now includes a cost f_t for removing liquidity. The equilibrium half bid-ask spread becomes

$$h_{\beta,N} = \frac{\pi\sigma}{1-\beta(1-\pi)} - f_m - \frac{\pi}{1-\beta(1-\pi)}\frac{1}{N}(f_t - f_m) \tag{2.7}$$

When an HFT wants to update its resting limit order, it does so by canceling her quote instead of sniping it for a cost of $f_t - f_m$. Hence, it benefits from a negative opportunity cost when providing liquidity. In equilibrium, the bid-ask spread is narrower by $\frac{\pi}{N}(f_t - f_m)$. Hence, compared with the benchmark case $h = h_\beta$, the tighter spread is not single-handedly explained by rebates.

Assumption 1. The number of HFTs is infinite, i.e. $N = \infty$.

As N increases, the value of sniping decreases as the sniping race becomes more competitive. An increase in N also decreases the value of liquidity provision as it decreases the probability of successful canceling on value jumps. Effectively, N affects the opportunity cost of canceling versus round-trip transactions given by $\frac{\pi}{N(1-\beta(1-\pi))}(f_t - f_m)$.

Proposition 6. As $N \to +\infty$, the equilibrium bid-ask spread becomes:

$$h_\beta = \frac{\pi\sigma}{1-\beta(1-\pi)} - f_m \tag{2.8}$$

Subsequently, we denote by $H_\beta = h_\beta + f_m$ the break-even spread excluding fees.

2.5.1 Tick constraint and the HFT rent

In an idealized market with continuous pricing, competition among HFTs would ensure that bid-ask spreads are as tight as possible, aligning prices closely with fundamental values.

However, in reality, exchanges enforce a minimum tick size constraint: prices can only be quoted in discrete increments of size Δ. In the United States,the minimum price increment is rigid and set at $\Delta = 0.01\$$.
This discretization introduces frictions, particularly when the tick size is large relative to the natural equilibrium spread. When the tick size constraint is binding:

- HFTs cannot tighten the spread further, even if competition would otherwise drive it lower.
- Non-HFT participants are unable to undercut HFTs, as they are constrained to quoting at fixed increments.
- HFTs can extract rents because their quotes maintain priority in the order book.

Formally, we express asset volatility in ticks such that $\sigma = L\Delta$. To put L in perspective, stocks like Ford (F) or General Motors (GM) have fundamental value jumps of about $L = 3$ ticks. Consider the extreme case when $h_\beta \to 0$. The bid-ask spread is binding at one tick. For illustrative purposes, we always use $V_0 = 0$. Assume that the bid-ask spread is one-tick wide and that V_0 sits in between these two price increments, i.e. $V_0 \in [-\frac{\Delta}{2}, +\frac{\Delta}{2}]$. We distinguish two cases:

- When $-\frac{\Delta}{2} \leq V_0 \leq -\frac{\Delta}{2} + f_m$, the HFT quotes a sell order at $-\frac{\Delta}{2}$. The HFT is willing to quote a price lower than the fundamental value because they will earn a positive profit with the maker fee.
- When $-\frac{\Delta}{2} + f_m \leq V_0 \leq \frac{\Delta}{2}$, the HFT quotes a sell order at $+\frac{\Delta}{2}$. If the order is executed, the HFT will receive $+\frac{\Delta}{2} + f_m - V_0 > 0$ as a rent for providing liquidity.

Remark 1. By definition, $\sigma = L\Delta$. Hence, the fundamental value will always sit in the middle of two ticks.

Assumption 2. The maker fee $f_m < \frac{\Delta}{4}$ is small enough such that sell orders are always quoted above fundamental value. This assumption is used to match reality and prevent crossed orders.

In the benchmark case where tick size is not binding, HFTs break even at the half bid-ask spread h_β and thus, it does not make sense to quote a second limit order. This is no longer the case when HFTs can extract rent from the tick size constraint. The HFT seller always quote at the tick above the breakeven price, i.e. $h_\beta < +\frac{\Delta}{2}$. Because the tick size is binding, MFs cannot undercut HFTs by posting a more aggressive quote. They are forced to use market orders and act effectively like the RTs.

Proposition 7. HFTs extract a rent from liquidity provision when the tick size constraint is binding. The expected profit to providing liquidity in an empty LOB is

$$(1-\pi)(\frac{\Delta}{2}+f_m)-\pi(\sigma-\frac{\Delta}{2}-f_m)=f_m-\pi\sigma+\frac{\Delta}{2}>0 \tag{2.9}$$

HFTs extract rent through two channels. First, they benefit from the incremental profits of $\beta(1-\pi)\frac{\Delta}{2}$ from the market orders of MFs. Under continuous pricing, when $\beta=0$, the HFTs would set the prices at the break-even spread $h_0=\pi\sigma-f_m<h_\beta$. If the break-even spread h_β is exactly at the tick, HFTs extract a profit $\frac{\Delta}{2}$ with probability $\beta(1-\pi)$ from the market orders of the MFs. Secondly, HFTs can no longer compete over prices and tighten the bid-ask spread. Therefore, they extract an additional profit of $\frac{\Delta}{2}-h_0$ from both RTs and MFs. We can decompose HFT's rent into three terms:

$$\underbrace{\frac{\Delta}{2}-h_0}_{\text{HFT rent}}=\underbrace{\beta(1-\pi)\frac{\Delta}{2}}_{\substack{\text{Rent on forced}\\ \text{MF market orders}}}+\underbrace{(1-\beta)(1-\pi)(\frac{\Delta}{2}-h_\beta)}_{\substack{\text{Incremental profits}\\ \text{from RT market orders}}}-\underbrace{\pi(h_\beta-\frac{\Delta}{2})}_{\substack{\text{Lower cost}\\ \text{from sniping}}} \tag{2.10}$$

Therefore, in partial equilibrium, there is a positive expected payoff to providing liquidity. HFTs can further compete on prices, with lower net spreads thanks to inverted fees (see section 2.5.2), or on speed, with more depth in the LOB (see section 2.5.3).

2.5.2 Inverted exchanges for reducing the net spread

Previously, make-take exchange fees widened the net cost of a one-tick spread. On the contrary, inverted exchanges charge a fee to liquidity providers while offering a rebate to liquidity takers. In tick-constrained markets, price competition is limited because traders cannot improve quotes in increments smaller than the tick size. This restriction allows high-frequency traders (HFTs) to dominate order queues by securing price priority at the best bid or ask. Inverted exchanges counteract this rigidity by enabling further price competition. Denote their make and take fees by $f_m^I<f_t^I<0$. The total exchange fee respects the participation constraint $T_I=f_t^I-f_m^I>0$.

Assumption 3. $f_m\leq -f_t^I<0<-f_m^I\leq f_t$

Essentially, this assumption means that getting a limit order filled on an inverted exchange is cheaper than executing a market order on a make-take exchange at the same price. This is consistent with real world exchange fees: inverted exchanges tend to have smaller fees and rebates.

Mutual Funds will post limit orders when at the front of the LOB queue because it reduces their execution cost in expectation by $f_t^i-f_m^j$, where i denotes the exchange populated by HFTs and j the exchange on which MFs provide liquidity. We further characterize the negative opportunity cost below. Denote by $h_0^I = H_0 - f_m^I$ the equilibrium half-spread under continuous pricing on inverted exchanges. By design, $h_0^I > h_0$.

Proposition 8.

We distinguish three cases:

- **Case I:** When $h_0 < h_0^I < \frac{\Delta}{2}$, HFTs can undercut prices on the make-take exchange by quoting on the inverted venue and earn $\frac{\Delta}{2} - h_0^I$. Non-HFTs are forced to use market orders and will always pick the inverted venue to obtain an execution cost lower by $f_t^I - f_t$.

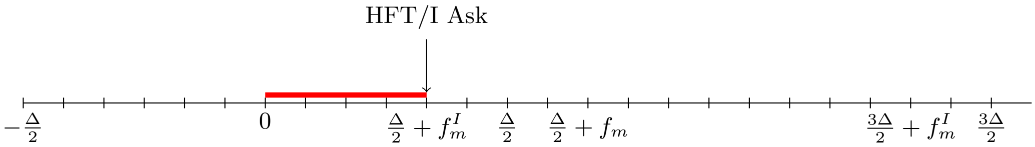


Figure 2.4: case I: $h_0 < h_0^I < \frac{\Delta}{2}$

Figure 2.4 shows the equilibrium quote of the HFT. The red region corresponds to the break-even spread range such that HFTs always quote at the nearest tick on inverted exchanges. We denote inverted exchanges by I and make-take exchanges by MT. There are four possible combinations for Bid and Ask: HFT/I, HFT/MT, MF/I, MF/MT.

When $h_0 < \frac{\Delta}{2} < h_0^I$, the break-even spread for HFTs under inverted fee structure is above the half-tick. They provide liquidity on make-take exchanges and set the ask quote at the nearest tick above at $\frac{\Delta}{2}$. It is not profitable to quote on inverted exchanges at the NBBO. They do not provide liquidity at the second tick level either as Non-HFTs will pay $+\frac{\Delta}{2} + f_t$ on make-take exchanges instead of $\frac{3\Delta}{2} + f_t^I$. Hence, the LOB on the inverted exchange is empty.

In this case, MFs can choose between using market orders at the make-take exchange, paying $\frac{\Delta}{2} + f_t$, and providing liquidity at inverted exchanges, paying $\frac{\Delta}{2} - f_m^I$. By providing liquidity in an empty LOB on inverted venues, MFs offer the best net price and effectively gain priority over the overall order queue. We assume that when an MF arrives after another MF and the inverted exchange LOB is not empty, MFs simply use market orders.

If MFs choose to post limit orders on inverted venues, HFTs only break even at the half-spread h_β. Therefore, we distinguish two sub-cases:

- **Case II:** When $h_0 < h_\beta < \frac{\Delta}{2} < h_0^I$, HFT cannot break even when providing liquidity in an inverted exchange at the nearest tick. Therefore, they provide liquidity on make-take exchanges only. MFs always chooses to undercut HFTs when they arrive on the market by posting orders on the inverted exchange at the tick $\frac{\Delta}{2}$, paying $\frac{\Delta}{2} - f_m^I$ instead of $\frac{\Delta}{2} + f_t$.

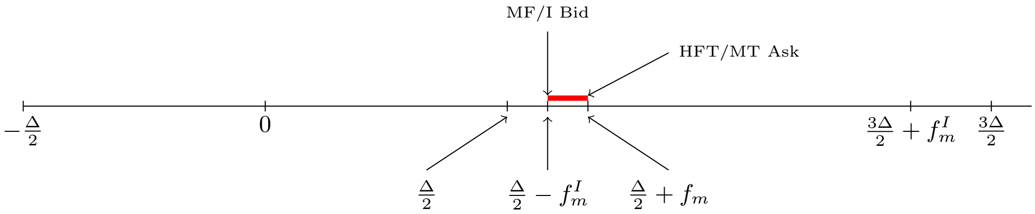


Figure 2.5: Case II: $h_0 < h_\beta < \frac{\Delta}{2} < h_0^I$

The MF puts a limit bid order at $\frac{\Delta}{2}$ on the inverted exchange which gets immediately accepted by the HFT. The HFT sells and makes a net profit of $\frac{\Delta}{2} - f_t^I$ while the MF pays $\frac{\Delta}{2} - f_m^I$.

- **Case III:** When $h_0 < \frac{\Delta}{2} < h_\beta < h_0^I$, HFT cannot break even when providing liquidity in a make-take exchange at the nearest tick. Therefore, they provide liquidity two ticks above on inverted exchanges. MFs always chooses to undercut HFTs when they arrive on the market by posting orders on the make-take exchange at the tick $\frac{\Delta}{2}$ → better execution (but not at fundamental value).

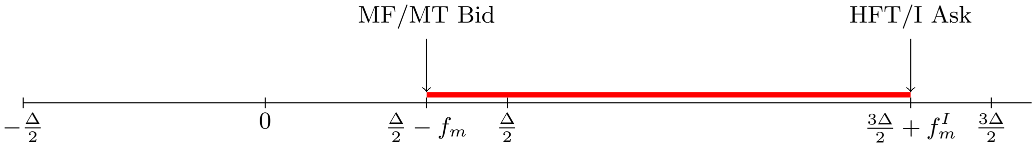


Figure 2.6: Case III: $h_0 < \frac{\Delta}{2} < h_\beta < h_0^I$

The MF puts a limit bid order at $\frac{\Delta}{2}$ on the make-take exchange which gets immediately accepted by the HFT. The HFT sells and makes a net profit of $\frac{\Delta}{2} - f_t$ while the MF pays $\frac{\Delta}{2} - f_m$.

We summarize these results in the following proposition:

Proposition 9. In the presence of inverted exchanges, there are three equilibrium regions for the Mutual Fund's decision:

- a pooling equilibrium when $h_0 < h_0^I < \frac{\Delta}{2}$: Mutual Funds are forced to used market orders and behave like Retail Traders. Their limit orders cannot be priced better than existing HFT orders, and market orders are the only order type guaranteeing them execution.
- an inverted equilibrium when $h_0 < h_\beta < \frac{\Delta}{2} < h_0^I$: Mutual Funds always post on inverted exchanges to undercut orders on traditional venues. Asset risk is low but not low enough for HFTs to out-compete the Mutual Funds.
- a make-take equilibrium when $h_0 < \frac{\Delta}{2} < h_\beta < h_0^I$: Mutual Funds always post on the traditional exchange and HFTs post on the inverted exchange.

Note that these three scenarii do not happen sequentially as the break-even spread increases. Recall that when HFTs cannot quote at the nearest tick on inverted exchanges, MFs can undercut them and HFTs do not earn profits from them. Assume that $h_0^I = \frac{\Delta}{2}$. For β sufficiently high, the difference in break-even spread is $h_\beta - h_0^I = \pi\sigma\frac{\beta(1-\pi)}{1-\beta(1-\pi)} + f_m^I - f_m$. Therefore, for sufficiently high β given by

$$\beta > \frac{1}{1-\pi} \cdot \frac{f_m - f_m^I}{\pi\sigma + f_m - f_m^I} \tag{2.11}$$

the second case never happens and HFTs switch between providing liquidity on make-take and inverted exchanges depending on where the fundamentals lie after the value jump.

2.5.3 Order book depth

A key takeaway from the previous sections is the positive profits to marginal liquidity provision in partial equilibrium, when there is a single unit of liquidity in the LOB. A natural solution is to allow the order book to have depth. HFTs provide liquidity at deeper levels because the effective spread is higher than the break-even spread.

Denote by D the depth of the order book. Assume the first $D-1$ limit orders in the LOB have a positive expected payoff. An HFT would post the marginal D^{th} order in the LOB at prevailing prices only if the rent is sufficiently high. Denote $p = \frac{(1-\beta)(1-\pi)}{1-\beta(1-\pi)}$ the probability of RT arrival instead of a value jump. With probability p^D, there are D Retail Traders arrivals in a row and the last HFT earns a profit of $h + f_m$ from liquidity provision. For $d < D$, with probability $p^d(1-p)$, there are d RT arrivals in a row. Conditional on this sequence, with probability $\frac{1}{N}$, the last HFT successfully wins the sniping race on $D-d+1$ orders while also canceling their own and make a profit of $(d-1)(\sigma - h - f_t)$, or gets sniped on his stale quote and makes a loss of $(\sigma - h - f_m)$. Therefore, the expected payoff would be $\frac{1}{N}(d-1)(\sigma - h - f_t) - \frac{N-1}{N}(\sigma - h - f_m)$.

Proposition 10. The expected payoff of marginal liquidity provision at depth D is:

$$g(D) = p^D(h + f_m) - (\sigma - h)(1 - p^D) + (1 - p^D)f_m \tag{2.12}$$

and the equilibrium gross half bid-ask spread at depth D is $H_D = \sigma(1 - p^D)$.

Hence, under discrete pricing, order book depth matters. HFTs are incentivized to compete for order time priority. Note that $\frac{\partial g}{\partial d} < 0$, i.e. the marginal gain from liquidity provision decreases the farther in the queue. As a consequence, the incentives for competing in the latency race are highest among top players.

2.5.4 Market fragmentation

When the break-even half-spread is sufficiently small, i.e. $H_0 < \pi\sigma + f_m^I$, inverted exchanges should dominate all trading activity. In reality, investors do not always obtain the optimal execution. Some orders may get filled at the NBBO but with worse outcomes after including fees. This may arise because of the brokerage's agency problem or simply, convenience. This can sustain fragmented liquidity even when inverted exchanges offer finer price improvement. We subsequently incorporate market fragmentation into our model. Non-HFTs are indifferent between exchanges that offer the same quotes, exclusive of fees. Therefore, they choose the make-take and inverted venues with equal probability $\frac{1}{2}(1 - \pi)$. Denote by D_{MT} and D_I the equilibrium depths on the make-take and inverted venues respectively such that the marginal HFTs providing liquidity break even. Formally, they are defined by:

$$\begin{aligned} D_{MT} &= \max\{d \geq 0 \mid (1 - (1 - \pi)^d)\sigma \leq \frac{\Delta}{2} + f_m\} \\ D_I &= \max\{d \geq 0 \mid (1 - (1 - \pi)^d)\sigma \leq \frac{\Delta}{2} + f_m^I\} \end{aligned} \tag{2.13}$$

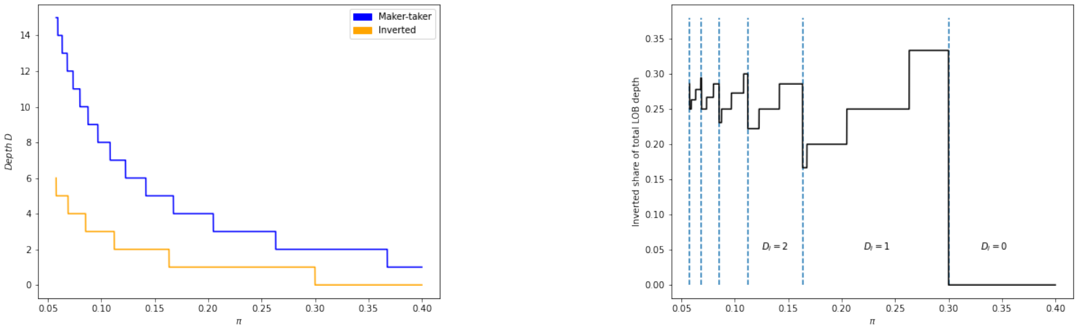
As the relative tick size increases, the equilibrium order book depth increases.

Proposition 11. Given D_I, the depth on the make-take exchange is determined by:

$$D_{MT} = \lfloor \frac{\log(1 - \frac{\frac{\Delta}{2} + f_m}{\sigma})}{\log(1 - \pi)} \rfloor \tag{2.14}$$

where $1 - \left(1 - \frac{\frac{\Delta}{2} + f_m^I}{\sigma}\right)^{\frac{1}{D_I + 1}} < \pi \leq 1 - \left(1 - \frac{\frac{\Delta}{2} + f_m^I}{\sigma}\right)^{\frac{1}{D_I}}$

We display the equilibrium depth in each venue in Figure 2.7a. Depth on the make-take exchange increases faster than on the inverted venue as the probability of a value jump π decreases. The make-take exchange is able to capture a higher proportion of the tick size rent compared to the inverted venue. Hence, the inverted share displays a staircase pattern as a function of π.



(a) Equilibrium depth

(b) Inverted share of total depth

Figure 2.7: Comparative depths. Chart drawn using following the parameters: $\sigma = 1, f_m = 0.1, f_m^I = -0.2$.

The fraction of total quoted depth is further shown in Figure 2.7b. When there is sufficiently more rent to be extracted and posting an additional layer of liquidity is profitable, inverted usage jumps as the make-take exchange LOB quoted depth does not increase as much relatively. For high values of π, liquidity provision in inverted exchanges is not profitable, and only the maker-taker model subsists. Further, the share of market depth has important implications on trading volume.

Proposition 12. The inverted market share in the fragmented market is given by:

$$\frac{\mathbb{E}[n_I]}{\mathbb{E}[n_I] + \mathbb{E}[n_{MT}]} = \frac{\frac{1}{2(1-\pi)} + D_I}{\frac{1}{(1-\pi)} + D_{MT} + D_I} \tag{2.15}$$

While the number of trades for each venue before the arrival of a news is the same and equal to $\frac{1}{2(1-\pi)}$, they differ in the number of stale quotes sniped. On average, after a trading session of length $\frac{1}{\lambda_V}$, the numbers of trades on the inverted and make-take exchanges are given by $\mathbb{E}[n_I] = \frac{1}{2(1-\pi)} + D_I$ and $\mathbb{E}[n_{MT}] = \frac{1}{2(1-\pi)} + D_{MT}$ respectively.

2.5.5 Market share under volatility regimes

Importantly, markets exhibit heterogeneous levels of volatility, which HFTs quickly incorporate into their spreads. For instance, volatility spikes in the market open and close, as shown in Figure 2.3. To match this stylized fact, we add volatility regimes (volatility of volatility) to our model. Assume the value jumps σ_t are i.i.d. and uniformly distributed. At time t, a value jump occurs and the new volatility regime for time $t+1$ is common knowledge. HFTs adjust their spread accordingly. For simplicity, we assume that the volatility regime follows a uniform distribution such that value jumps remain relatively small. The latter assumption is only for clarity of exposure and results remain similar without restrictions on the magnitude of the volatility regime.

Assumption 4.

$$\sigma_t \sim \mathcal{U}[-\Sigma_V, +\Sigma_V], \quad \forall t \in [0,\infty] \quad \textit{where} \quad \Sigma_V \in [\frac{\Delta}{2} + f_m, \frac{3\Delta}{2} + f_m^I] \tag{2.16}$$

This can be interpreted as important news causing a value jump as well as higher (short-term) volatility. When $\Sigma_V > \frac{3\Delta}{2} + f_m^I$, the equilibrium alternate between Case II and Case III, while offering little economic insight. As the ratio $\frac{\Sigma_V}{\Delta}$ increases, the tick size is less and less likely to be binding. When $\frac{\Sigma_V}{\Delta} \to \infty$, the tick size is infinitesimally small and the equilibrium reverts to that of the continuous benchmark. Thus, we exclude that range in our assumption and focus on tick size constrained environments. We compute trading volumes for each scenario in Section 2.5.2. Without loss of generality, we study the behavior of the LOB when HFTs provide liquidity on the ask side. Let us denote by n_{MT} and n_I the trading intensity on make-take and inverted venues respectively. Here, we are only interested in the trading volume of inverted exchanges relative to traditional exchanges. Hence, we define n_{MT} and n_I as the number of trades between two value jumps. Again, as in Section 2.5.2, we distinguish three cases:

- **Case I:** When $H_0 < \frac{\Delta}{2} + f_m^I$, the nearest tick is constraining, HFTs provide liquidity on inverted exchanges and all trades are market orders. With probability $\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V}$, the expected number of trades on inverted venues before the value jump is $\frac{1}{1-\pi}$. The total number of trades includes the stale quotes sniped during the value jumps and is given by: $\mathbb{E}[n_I|\sigma_t] = \frac{1}{1-\pi} + 1$

- **Case II:** When $H_0 > \frac{\Delta}{2} + f_m^I$ and $H_\beta \leq \frac{\Delta}{2} + f_m$, MFs undercut the HFTs by providing liquidity on the inverted exchange while HFTs provide liquidity on the make-take venue. This is true when $\sigma \leq \frac{1-\beta(1-\pi)}{\pi}(f_m - f_m^I)$. Therefore, with probability $\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m - f_m^I) - \frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V}$, the expected number of inverted trades is $\mathbb{E}[n_I|\sigma_t] = \frac{\beta}{1-\pi}$ and the expected number of make-take trades is $\mathbb{E}[n_{MT}|\sigma_t] = \frac{1-\beta}{1-\pi} + 1$.

- **Case III:** When $H_\beta > \frac{\Delta}{2} + f_m$, MFs undercut the HFTs by providing liquidity at the nearest tick on the make-take exchange while HFTs provide liquidity at the tick above on the inverted venue. This is true when $\sigma > \frac{1-\beta(1-\pi)}{\pi}(f_m - f_m^I)$. Therefore, with probability $1-\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m-f_m^I)-\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V}$, the expected number of make-take trades (inside NBBO) is $\mathbb{E}[n_{MT}|\sigma_t]=\frac{\beta}{1-\pi}$ and the expected number of inverted trades is $\mathbb{E}[n_I|\sigma_t]=\frac{1-\beta}{1-\pi}+1$.

Note that all trading sessions last the same amount of time on average because they only depend on the probability of a value jump arrival. In the last case, MFs post a limit order inside the NBBO and gets immediately accepted. This explains the footprint of hidden liquidity and price improvements in unconstrained markets. Given the average duration of a volatility regime is $\frac{1}{\lambda_V}$, the average number of trades during that period is the sum across all volatility states.

Proposition 13. The market share of the inverted exchange is determined by:

$$\frac{\mathbb{E}[n_I]}{\mathbb{E}[n_I]+\mathbb{E}[n_{MT}]} \tag{2.17}$$

where the trading volume on each exchange is given by:

$$\begin{aligned}\mathbb{E}[n_I] = {} & \mathbb{1}_{\beta<\frac{1}{1-\pi}\cdot\frac{f_m-f_m^I}{\pi\sigma+f_m-f_m^I}}\Bigg[\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V}\cdot(\frac{1}{1-\pi}+1)\\ & +(\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m-f_m^I)-\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{\beta}{1-\pi})\\ & +(1-\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m-f_m^I)+\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{1-\beta}{1-\pi}+1)\Bigg]\\ & +\mathbb{1}_{\beta\geq\frac{1}{1-\pi}\cdot\frac{f_m-f_m^I}{\pi\sigma+f_m-f_m^I}}\Bigg[\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V}\cdot(\frac{1}{1-\pi}+1)\\ & +(1-\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{1-\beta}{1-\pi}+1)\Bigg]\end{aligned} \tag{2.18}$$

$$\begin{aligned}\mathbb{E}[n_{MT}] = {} & \mathbb{1}_{\beta<\frac{1}{1-\pi}\cdot\frac{f_m-f_m^I}{\pi\sigma+f_m-f_m^I}}\Bigg[(\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m-f_m^I)-\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{1-\beta}{1-\pi}+1)\\ & +(1-\frac{1}{\Sigma_V}\frac{1-\beta(1-\pi)}{\pi}(f_m-f_m^I)+\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{\beta}{1-\pi})\Bigg]\\ & +\mathbb{1}_{\beta\geq\frac{1}{1-\pi}\cdot\frac{f_m-f_m^I}{\pi\sigma+f_m-f_m^I}}\Bigg[(1-\frac{\frac{\Delta}{2}+f_m^I}{\Sigma_V})\cdot(\frac{1-\beta}{1-\pi}+1)\Bigg]\end{aligned} \tag{2.19}$$

For small values of β, inverted market share increases with the fraction of mutual funds. When β crosses the threshold $\frac{1}{1-\pi}\frac{f_m-f_m^I}{\pi\sigma+f_m-f_m^I}$, the market never encounters Case II. The inverted volume is increasing in β in Case II while it is decreasing in Case III.

Consistent with Malinova and Park (2015), trading activity is largely driven by the demand of non-HFTs. In our model, market makers break even while non-HFTs have inelastic demand when they arrive. A reduction in take fee makes the exchange more attractive to liquidity demanders while an increase in the make fee has no impact in our model. Thus, make fees are irrelevant here. However, higher rebates affects the profitability of marketable orders and thus, the equilibrium depth of the LOB.

Contrasting the results of Section 2.5.4, the inverted market share is generally lower in fragmented markets than in the model with volatility regimes. The amount of non-HFT trading is the same on both exchanges. But the make-take exchange has a much deeper book and generates high volume through the stale quote sniping channel. In reality, brokers are incentivized by fees, not timely execution. Thus, more orders would be routed to the make-take exchanges.

In sum, the model highlights that inverted exchanges serve as a mechanism for mitigating tick-size inefficiencies by enabling finer price competition. However, the presence of HFTs and order queue dynamics create additional strategic frictions.

2.6 Empirical Results

We now turn to empirical evidence to assess whether these predictions hold in real-world trading data. In this section, we empirically test the predictions derived from our theoretical model using high-frequency trading and quote data. We aim to assess how inverted fee structures and tick-size constraints affect market dynamics, particularly focusing on High-Frequency Trader (HFT) activity, the usage of inverted exchanges, and the role of order imbalances.

Our theoretical model identifies that the rents extracted by High-Frequency Traders (HFTs) increase substantially under binding tick-size constraints. The economic intuition behind this prediction is straightforward. When tick sizes are binding, non-HFT traders—particularly slower liquidity providers such as mutual funds—are unable to competitively price-improve HFT quotes in finer increments. Consequently, HFTs enjoy greater intermediation profits, as competition from slower market participants diminishes. We summarize this economic intuition formally as our first empirical prediction.

Prediction 1 (HFT Activity). *HFT activity is higher for tick-constrained stocks characterized by lower spreads and lower volatility.*

Beyond their impact on HFT activity, the tick-size constraint also directly affects the competitive landscape of exchanges. A core theoretical insight of our model pertains to the role inverted fee structures play in alleviating tick-size constraints. In contrast to make-take exchanges, inverted venues offer rebates to liquidity takers and charge fees to liquidity providers. This structure effectively enables market participants, particularly institutional investors and mutual funds, to economically offer and achieve better net execution prices without violating the regulatory tick size. Thus, inverted venues become especially valuable under conditions where tick-size constraints significantly limit finer price increments, providing a practical avenue for market participants disadvantaged in traditional venues. Based on this mechanism, we explicitly formulate the following empirical prediction.

Prediction 2 (Inverted Exchange Usage). *The relative market share of inverted exchanges is positively associated with tighter tick-size constraints, specifically increasing in stocks for which the regulatory tick size represents a large fraction of the effective spread.*

Predictions 1 and 2 together imply a nuanced interaction between HFT behavior, market structure, and exchange competition. While binding tick-size constraints enable HFTs to extract increased intermediation rents, they simultaneously incentivize non-HFT traders to migrate toward inverted exchanges as alternative venues that mitigate these constraints through economically favorable execution terms. Thus, our theoretical model predicts a simultaneous increase in HFT participation and inverted exchange usage precisely in environments characterized by tighter effective spreads and binding tick increments.

Furthermore, our theoretical analysis indicates that liquidity provision incentives vary systematically with the relative positioning of fundamental asset values within discrete tick increments. Intuitively, when the fundamental asset price resides closer to one tick increment boundary (e.g., closer to the bid price), the other side (e.g., the ask side) becomes relatively crowded with liquidity providers. Thus, slower traders face substantial queue disadvantages when attempting to place limit orders on this crowded side. Consequently, these traders strategically turn to inverted exchanges, where they can economically undercut existing liquidity providers through effective net price improvements afforded by taker rebates.

This mechanism implies a clear, directional empirical relationship between order book imbalances and side-specific trading volume on inverted exchanges, summarized explicitly in our next empirical prediction.

Prediction 3 (Inverted Exchange Usage and Order Imbalance). *Trading volume on inverted exchanges disproportionately concentrates on the side of the order book experiencing larger liquidity imbalances. Specifically, mutual fund traders strategically utilize inverted venues primarily on the less competitive side to effectively mitigate queue disadvantages.*

Taken together, Predictions 2 and 3 establish a coherent narrative about how market participants adapt their trading venue choices in response to binding tick-size constraints and associated liquidity imbalances. Under such constraints, inverted exchanges not only provide overall improved net execution prices, but also enable strategic positioning within the order book.

In the subsequent sections, we systematically test each of these predictions using rigorous empirical methodologies and high-frequency data, aiming to provide clear and comprehensive evidence supporting our theoretical framework.

2.6.1 Market Quality Measures

To test the model's predictions, we first define several market quality metrics. Our primary variable, the *relative tick size*, is defined as the ratio of the minimum tick increment to the nominal stock price. This metric quantifies the economic significance of the tick constraint. The *quoted spread* measures liquidity ex ante and is computed as the difference between the displayed best bid and best ask prices. In contrast, the *effective spread* reflects ex post execution costs and captures realized transaction costs as the absolute deviation of the transaction price from the midpoint. We express these spreads as percentages relative to the midpoint price to allow cross-stock comparisons.

To measure the spread[2], we compute the dollar value-weighted relative effective half-spread (DVWRES), which is defined as:

$$DVWRES_t = \sum_{i \in TX_{t-1,t}} \frac{(Price_i \cdot Size_i)}{\sum_{i \in TX_{t-1,t}} (Price_i \cdot Size_i)} \frac{|Price_i - Midpoint_i|}{Midpoint_i}. \tag{2.20}$$

where $TX_{t-1,t}$ is the set of all transactions between minute $t-1$ and t. DVWRES accurately reflects realized transaction costs, capturing price improvement from hidden liquidity or sub-penny executions. Although Hagströmer (2021) indicates that midpoint-based effective spreads can slightly overstate transaction costs in discrete tick-size environments, we retain these standard definitions for comparability to the extant literature.

Finally, we measure *relative depth* as the time-weighted average quoted volume at the best bid and ask levels, relative to the concurrent 1-hour trading volume:

$$RelDepth_t = \frac{1}{Vol_{h_t}} \sum_{i \in U_{t-1,t}} \frac{(t_i - t_{i-1})(AskSize_{i-1} + BidSize_{i-1})}{\sum_{i \in U_{t-1,t}} (t_i - t_{i-1})}. \tag{2.21}$$

where Vol_{h_t} is the volume in the hour preceding minute t, $U_{t-1,t}$ is the set of all NBBO

[2] We also run the regressions with other spread measures (time-weighted vs dollar-weighted, gross vs effective) and do not find materially different results.

updates (new bid or new ask) between minute $t-1$ and t, and t_i is the time of the i^{th} update. This measure captures liquidity available at the top of the book, directly relevant for assessing execution probabilities and market impact.

The subsequent empirical analyses present regression results and visual illustrations that validate these theoretical predictions, providing robust evidence on how tick-size constraints and inverted fee structures jointly influence liquidity provision, venue market shares, and trading dynamics.

The NASDAQ HFT dataset identifies high-frequency traders (HFTs) solely at the group level, without distinguishing individual firms or explicitly identifying whether these traders predominantly supply or take liquidity. Nevertheless, the dataset provides valuable granularity through its detailed trade classification, specifying liquidity taker and maker roles for each transaction. Specifically, each transaction is classified into one of four categories: NN (non-HFT taker, non-HFT maker), NH (non-HFT taker, HFT maker), HN (HFT taker, non-HFT maker), and HH (HFT taker and maker). Consistent with the findings of Brogaard et al. (2018), we observe heightened HFT activity during significant price jumps, typically acting as contrarians to large price movements, thus mitigating excessive volatility.

To study HFTs' market behavior, we follow Carrion (2013) and construct three primary measures to quantify HFT participation in the market. The first measure concerns *HFT Activity*, i.e. the participation rate of HFTs in all transactions.

$$\text{HFT Activity} = \frac{HH + HN + NH}{HH + HN + NH + NN}, \tag{2.22}$$

The other two measures characterize the liquidity-induced states of HFT transactions, that is, whether their transactions are liquidity-taking or liquidity-making[3].

$$HFT^{Taker} = \frac{HH + HN}{HH + HN + NH + NN}, \tag{2.23}$$

$$HFT^{Maker} = \frac{HH + NH}{HH + HN + NH + NN}. \tag{2.24}$$

Moreover, to explicitly analyze HFTs' propensity to supply liquidity relative to their overall trading activity, we define the following variable:

$$HFT^{LP} = \frac{HH + NH}{HH + NH + HN} \tag{2.25}$$

which measures the fraction of HFT activity that contributes to liquidity provision.

[3] To be more accurate, HFT^{Taker} (respectively, HFT^{Maker}) is the fraction of trading in which there is at least one HFT taking liquidity (respectively, making liquidity). We double count HH transactions, defining them as both HFT liquidity-taking and liquidity-making

Aggregated across daily observations, HFTs neither exhibit a systematic net liquidity-taking nor liquidity-providing stance. This neutrality aligns with the interpretation that HFTs typically manage inventories aggressively and avoid prolonged directional exposures. However, our theoretical framework suggests that HFTs preferentially provide liquidity for stocks subject to significant tick size constraints. Due to their technological advantage, HFTs are more adept at capturing the economic rents arising from discrete price grids, exploiting the limited price improvement opportunities faced by slower market participants.

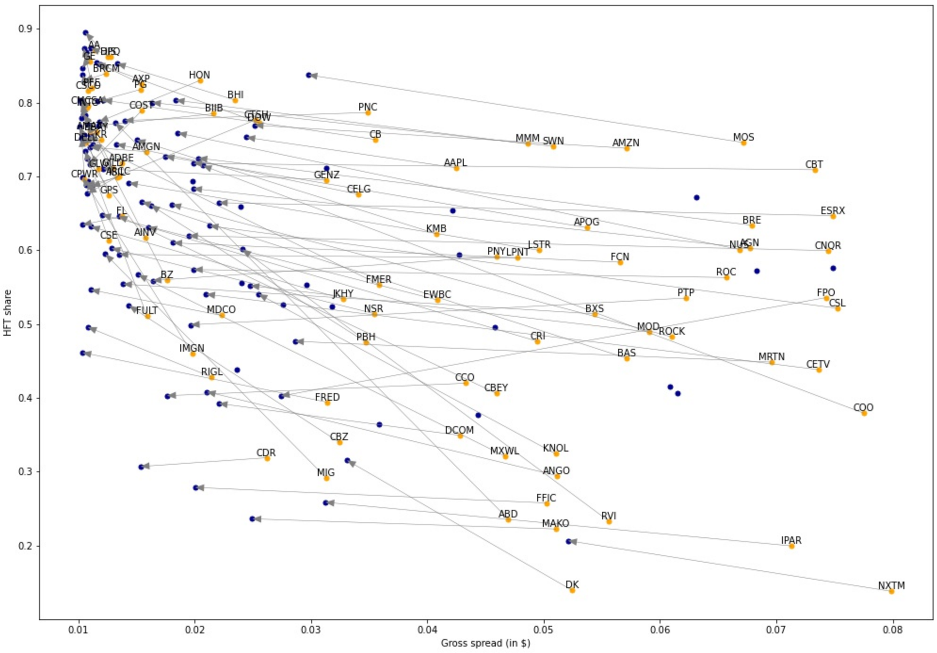


Figure 2.8: HFT share and spread through the day. The orange (blue) dots represent the average spread in the first (last) thirty minutes of the trading day. For clarity of exposure, the chart does not include stocks which had spreads above 0.10$ at the beginning of the day.

Figure 2.8 illustrates the intraday evolution of HFT trading activity alongside quoted spreads. We note a clear pattern: as volatility declines and spreads narrow over the trading session, HFTs become increasingly active. This observation reinforces our hypothesis that tighter tick constraints create greater incentives for HFT liquidity provision. To sharpen the exposition, we exclude stocks with extremely wide initial spreads (greater than $0.10), thereby concentrating the analysis on stocks meaningfully constrained by the tick size. Table 2.4 in the Appendix shows that, conditional on the NBBO being provided by only one type of trader, the HFT is more likely to be the liquidity provider as tick size and market capitalization increase.

Table 2.1 further corroborates these results through panel regression analysis. The dependent variable is the share of total trading attributable to HFTs. Independent variables include an indicator for stocks with spreads narrower than 1.5 cents—aligning with recent SEC amendments—and the relative tick size measured as the tick size divided by nominal stock price. In these regressions, we control for lagged volatility, relative depth of the order book, and fixed effects at stock, day, industry, and hourly levels. The results consistently show a positive and statistically significant association between HFT participation and tick-constrained environments. Controlling for other variables, a binding tick size is associated with a 2.02% higher HFT activity. More broadly, lower spreads and larger relative tick sizes predict heightened HFT activity, consistent with theoretical predictions that HFTs capture larger intermediation rents under constrained pricing environments.

Table 2.1: HFT Activity on tick size and spread

This table presents coefficients from regressions of HFT activity on microstructure measures. All variables are computed at the minute-level. The first variable indicates whether the average bid-ask spread is below 1.5 cents. This measures whether the tick size is binding or not. Overall results do not change with the type of spread or the threshold number. We chose 1.5 cents as this is the threshold used by the SEC in the Reg NMS Amendments as of September 2024. The relative tick size is the ratio of tick size to nominal price, expressed in %. Lagged volatility is the volatility of the 1-second return after aggregating transaction-level data (in bps). Relative depth is the ratio of order book depth to concurrent 1-hour trading volume (in %). Market capitalization is measured once at the market close each day. We run separate panel regressions at the stock level controlling for day and industry fixed effects, and stock and hour fixed effects.

Dependent Variable:		HFT ACTIVITY	
Model:	(1)	(2)	(3)
Variables			
$\mathbb{1}_{spread \leq 0.015\$}$	0.0651***	0.0290*	0.0202***
	(0.0092)	(0.0125)	(0.0056)
RelTickSize	-3.225***	-0.9329***	-2.287*
	(0.1897)	(0.1553)	(1.138)
lagged Volatility	-48.84***	-23.97***	-11.31
	(4.737)	(3.833)	(10.13)
lagged RelDepth	0.0006**	0.0003	-6.6×10^{-5}
	(0.0002)	(0.0001)	(0.0001)
$\mathbb{1}_{spread \leq 0.015\$}$ × RelTickSize	2.399***	1.396***	0.7380**
	(0.2362)	(0.2150)	(0.2417)
log(MarketCap)		0.0550***	
		(0.0051)	
Fixed-effects			
date	Yes	Yes	
industry	Yes	Yes	
hours			Yes
symbol			Yes
Fit statistics			
Observations	129,870	129,870	129,870
R^2	0.09528	0.14244	0.19834

*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

2.6.2 Market Share

Inverted exchanges facilitate trading within the quoted bid-ask spread, effectively offering liquidity demanders opportunities for finer incremental price improvement than traditional make-take exchanges, especially when the tick size constraint is binding. We define a stock as *tick-constrained* if its time-weighted average quoted spread remains less than 2.5 ticks. According to our model, the potential intermediation rents extracted by liquidity providers, especially high-frequency traders (HFTs), are highest when the minimum tick increment represents a significant fraction of the stock's nominal price. Therefore, inverted exchange usage is predicted to be greatest among stocks characterized by relatively tight nominal spreads and large relative tick sizes.

We empirically test this prediction using the following regression specification:

$$\text{InvShare}_{i,t} = \alpha + \beta_1 \cdot \text{RelTickSize}_{i,t} + \beta_2 \cdot \text{NBBOSpread}_{i,t} + X_{i,t}\gamma + \epsilon_{i,t},$$

where i indexes stocks and t indexes hourly intervals. Here, RelTickSizei, t is defined as the ratio of tick size to stock price, reflecting the economic significance of price discreteness. NBBOSpreadi, t represents the National Best Bid and Offer (NBBO) spread measured in absolute dollar terms. The vector $X_{i,t}$ contains control variables including relative order book depth, lagged absolute order imbalance, lagged HFT volume share, and lagged volatility. We employ stock and hour fixed effects to control for unobserved heterogeneity across securities and intraday variation in market conditions.

Inverted venue usage is positively correlated with the relative tick size, suggesting that traders actively seek price improvement opportunities when constrained by tick size rigidity. The effect is strongest for low-spread stocks, where the inability to quote within the bid-ask spread significantly impacts liquidity demanders.

First, *broker-driven market segmentation* may limit inverted venue usage. Brokers, incentivized by higher fees offered by make-take exchanges, may not consistently route orders to venues that minimize net trading costs for clients, as long as execution remains within the NBBO. Consequently, client orders may receive inferior net execution prices despite nominal execution at the best quoted prices.

Second, *liquidity concentration effects* could suppress inverted venue trading volumes. Traders, particularly institutional investors, prefer markets with higher liquidity concentration to minimize price impact and execution uncertainty (Admati and Pfleiderer (1988)). Inverted exchanges, being relatively new entrants, typically display thinner liquidity, discouraging significant order flow migration despite potential price improvements. Additionally, the presence of thinner order books on inverted exchanges intensifies adverse selection concerns, as liquidity providers face increased exposure to stale-quote sniping when fundamental values shift, further exacerbating liquidity concentration in traditional make-take venues.

Table 2.2: Inverted market share on tick size and relative depth

This table presents coefficients from regressions of inverted market share on relative tick size, NBBO spread (in \$), relative depth (in %), lagged absolute order book imbalance (in %), lagged HFT volume share (in %) and lagged 1-min volatility (in bps). All variables are computed at the minute-level. We run panel regressions at the stock level controlling for stock and hour fixed effects.

Dependent Variable:	InvShare			
Model:	ALL	ALL	TC only	TC only
Variables				
RelTickSize	1.191***	1.110***	1.662***	1.530**
	(0.2491)	(0.3029)	(0.4327)	(0.4696)
NBBO Spread		0.0177	-2.561***	-2.263***
		(0.0125)	(0.3267)	(0.2559)
RelDepth		0.0002*	0.0002**	0.0002*
		(9.06×10^{-5})	(8.63×10^{-5})	(8.74×10^{-5})
lagged AbsImbalance		-6.95×10^{-5}	-0.0048*	-0.0024
		(0.0015)	(0.0024)	(0.0020)
lagged HFT Activity		0.0042*		0.0065***
		(0.0018)		(0.0018)
lagged Vol				5.235
				(2.845)
Fixed-effects				
hours	Yes	Yes	Yes	Yes
symbol	Yes	Yes	Yes	Yes
Fit statistics				
Observations	180,681	141,718	112,051	89,626
R^2	0.07325	0.07736	0.08363	0.08358

Clustered (hours) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

This phenomenon can be exacerbated by the stale quote sniping channel. Because inverted venues are less profitable than make-take venues, HFTs supply fewer resting orders to the LOB. Thus, the resting depth at NBBO in inverted venues is smaller than that of regular venues. Upon a fundamental news arrival, the existing stale quotes are sniped. This naturally creates more volume at the make-take venue despite having the same amount of fundamental demand categorized by non-HFTs orders.

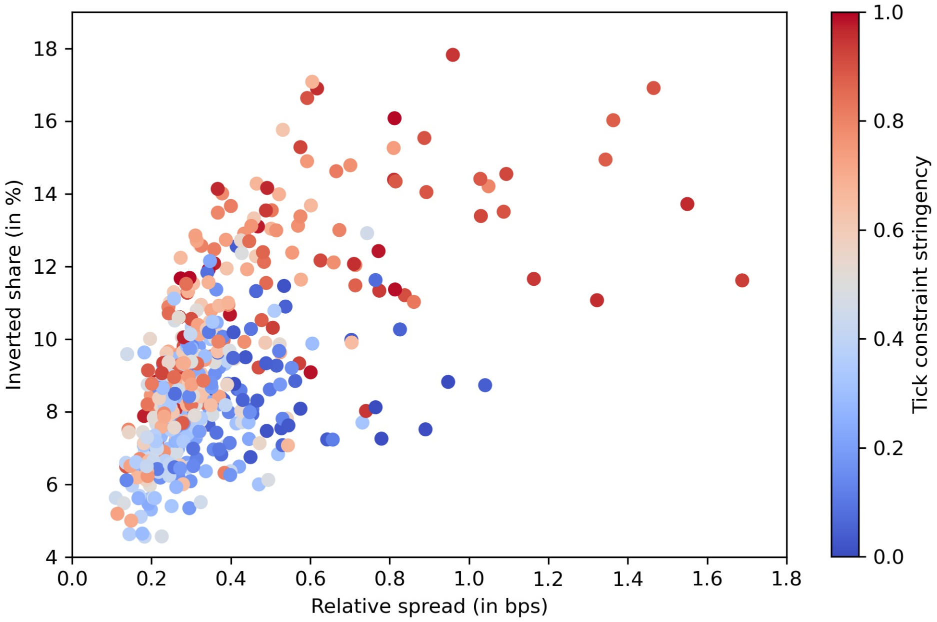


Figure 2.9: Inverted market share by spread

Figure 2.9 visualizes the empirical relationship between inverted market share and relative spread, revealing a concave pattern. While inverted venues capture a higher share of trading volume as the tick constraint becomes more severe, their market share is notably capped at approximately 20%, even for the most constrained stocks. We propose two potential explanations for this observed ceiling on inverted exchange volume.

2.6.3 Book depth and order imbalance

Thus far, we have analyzed aggregate market shares and overall HFT activity, linking these factors to tick size constraints and exchange fee structures. In this subsection, we extend our empirical analysis to examine how order imbalance affects liquidity dynamics, specifically through the channel of inverted venue usage.

In practice, a stock's fundamental value seldom resides exactly at the midpoint between two ticks; it typically leans closer to either the bid or the ask side. Following Hagströmer (2021), we leverage order imbalance as an indicator of the likely direction of this bias. When one side of the order book displays disproportionately higher depth, it signals the direction toward which the fundamental value is likely skewed. For instance, a surplus of sell orders at the best ask relative to buy orders at the best bid suggests the fundamental value resides closer to the bid tick.

Our theoretical predictions suggest that when the tick constraint is binding and the break-even spread falls within an intermediate range—not excessively tight nor excessively wide—inverted exchanges will attract more liquidity on the side of the order book opposite to the fundamental value. Specifically, if the fundamental value leans closer to the bid tick, we expect increased inverted venue liquidity provision at the ask side.

To empirically test whether inverted exchange usage systematically responds to order imbalance, we compute the time-weighted order book imbalance at the NBBO:

$$\text{Imbalance}_t = \sum_{i \in E_{t-1,t}} \frac{(t_i - t_{i-1})}{\sum_{\tau \in E_{t-1,t}} (t_i - t_{i-1})} \frac{(\text{BidSize}_i - \text{AskSize}_i)}{(\text{AskSize}_i + \text{BidSize}_i)}. \tag{2.26}$$

where $E_{t-1,t}$ is the set of all events (new top of book bid/ask or transaction) between minute $t - 1$ and t, and t_i is the time of the i^{th} update. We then estimate panel regressions to analyze whether inverted exchange trading volume, at the minute level, increases in the direction of the observed imbalance. Our primary regression specification is:

$$\frac{\text{InvVol}^B_{i,t}}{\text{InvVol}^B_{i,t} + \text{InvVol}^S_{i,t}} = \alpha + \beta_1 \text{Imbalance}_{i,t} + X_{i,t}\gamma + \epsilon_{i,t}, \tag{2.27}$$

where $\text{InvVol}^B_{i,t}$ and $\text{InvVol}^S_{i,t}$ denote inverted exchange trading volumes on the buy and sell sides, respectively, for stock i in time interval t. The vector of controls $(X_{i,t})$ includes lagged volatility, relative order book depth, and lagged order imbalance. Fixed effects at the stock and hourly levels control for unobservable differences across stocks and within-day variation.

Table 2.3 presents the results. We find a statistically significant positive relationship between order imbalance and the share of inverted buy volume, consistent with the theoretical prediction. Specifically, the coefficient on imbalance indicates that a 1% increase in order imbalance is associated with an approximate 0.12% increase in inverted buy-side volume share. This finding supports the hypothesis that traders strategically utilize inverted exchanges to gain price and queue priority, especially when the order book imbalance signals directional biases in fundamental value.

In the second regression specification, we include additional control variables such as lagged volatility and relative depth. The coefficient on lagged imbalance is negative, albeit only marginally statistically significant, suggesting some intraday reversion effects. These results indicate that inverted venue liquidity responds systematically to order imbalance signals, reflecting sophisticated trader strategies and highlighting the nuanced interplay between market structure, trading incentives, and information flow.

Table 2.3: Inverted share increases on the side of the order imbalance

This table presents coefficients from regressions of the inverted buy volume share on order book imbalance, lagged volatility (in bps), relative depth (in %), and lagged imbalance. All variables are computed at the minute-level. We run panel regressions at the stock level controlling for stock and hour fixed effects.

Dependent Variable:	$\frac{\text{InvVol}^B_{i,t}}{\text{InvVol}^B_{i,t}+\text{InvVol}^S_{i,t}}$	
Model:	(1)	(2)
Variables		
Imbalance	0.1215***	0.1301***
	(0.0177)	(0.0195)
lagged Vol		2.103
		(6.370)
RelDepth		-1.74×10^{-5}
		(5.11×10^{-5})
lagged Imbalance		-0.0092*
		(0.0047)
Fixed-effects		
hours	Yes	Yes
symbol	Yes	Yes
Fit statistics		
Observations	74,980	73,067
R^2	0.32458	0.32198

Clustered (hours) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

2.7 Policy Implications

The findings presented in this paper carry several important implications for regulatory policy regarding market design, especially with respect to tick size constraints and exchange fee structures. While a sufficiently large tick size enhances transparency by clearly delineating prices, it simultaneously restricts market competition by preventing liquidity providers from offering price improvements smaller than the minimum tick increment. As demonstrated by our theoretical and empirical analyses, excessively large tick sizes lead to increased market-maker rents, disproportionately benefiting high-frequency traders (HFTs) who leverage their speed advantage to dominate the limit order queues.

Conversely, an excessively small tick size may introduce detrimental market dynamics. As the cost of price undercutting diminishes, liquidity providers, particularly HFTs, frequently submit orders just inside existing quotes. While finer increments theoretically foster tighter spreads and improved market liquidity, they also increase execution uncertainty for marketable orders. Traders submitting marketable orders at the National Best Bid and Offer (NBBO) face heightened risks of unfilled executions or adverse price movements. Additionally, smaller tick increments could exacerbate latency arbitrage opportunities, as HFTs can swiftly adjust their quotes ahead of slower market participants, amplifying potential adverse selection costs for institutional investors.

Currently, U.S. equity markets adopt a binary tick size regime dictated by Regulation NMS, where the minimum increment is set at one penny for stocks priced above one dollar and one hundredth of a penny for stocks priced below one dollar. As illustrated in Figure 2.10, this structure results in substantial variation in the relative tick size across different price levels, inadvertently generating market inefficiencies at both ends of the price spectrum. Specifically, for low-priced stocks, the tick size becomes economically trivial, leading to aggressive undercutting and increased uncertainty regarding order execution. For higher-priced stocks, on the other hand, the tick size can become excessively large relative to price, leading to restricted price competition and elevated transaction costs.

International experience offers valuable insights for alternative approaches to tick size regulation. For example, the tick size regime employed by the Hong Kong Stock Exchange (HKEX) dynamically adjusts according to stock price tiers, maintaining relative tick sizes within a more consistent range across varying price levels. Such a tiered tick structure mitigates the economic distortions stemming from a fixed nominal tick size, allowing finer pricing increments for higher-priced securities and broader increments for lower-priced ones. This structure reduces the opportunity for HFT-driven rent extraction and promotes a more balanced competitive environment among liquidity providers.

Beyond tick size adjustments, our analysis underscores potential issues arising from fee-driven order routing incentives. Brokers often prioritize fee rebates over execution quality when routing client orders, potentially leading to suboptimal executions for end investors. Regulatory intervention aimed at transparency and enforcement of best-execution standards that explicitly account for net execution cost (inclusive of fees) could mitigate these agency conflicts. Strengthening disclosure requirements regarding routing practices and mandating greater transparency in fee structures might help align broker incentives with client interests, fostering competition based on execution quality rather than rebates.

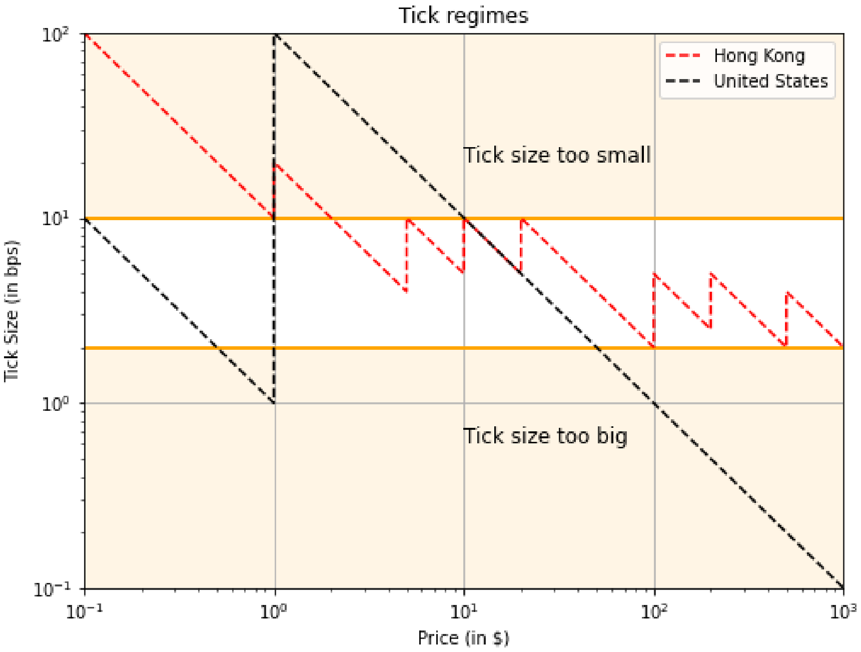


Figure 2.10: Tick regime in the USA and Hong Kong

The tick size as a fraction of price is displayed in Figure 2.10. The binary rule used in the U.S. creates inefficiencies in the market, making the tick too high for low prices and too low for high prices. In contrast, HKEX uses different tick sizes based on the price of the security, keeping the relative tick size fairly consistent across different price ranges.

Notably, liquidity disparities intensify further away from the midpoint, suggesting that slower liquidity providers may refrain from frequent price adjustments to avoid queue disadvantages relative to HFTs. This underscores the complexity inherent in liquidity provision incentives, reinforcing the need for nuanced regulatory frameworks that address not only tick size but also broader market design factors—including venue competition, order type diversity, and the technological landscape.

In summary, our results advocate for a more flexible tick size regime tailored to the price and liquidity characteristics of individual securities. Moreover, addressing broker incentives and improving transparency in routing practices could further enhance market efficiency and investor welfare.

2.8 Conclusion

This paper develops a model analyzing how exchange fee structures and minimum tick size constraints impact liquidity provision incentives, order book depth, trading volume, and relative venue market shares. We examine two primary fee models: the traditional make-take approach providing rebates to liquidity suppliers, and the inverted fee structure providing rebates to liquidity demanders.

Our model shows that inverted exchanges enable more granular effective pricing when the minimum tick size is sufficiently small and binding. The finer pricing grid afforded by rebating takers on the inverted venues allows institutional traders to effectively undercut the HFT top of book quotes. We derive equilibrium expressions for queue lengths, trading volumes and relative market shares across the different fee models and market regimes. The model yields several empirical predictions, which we confirm using market data from the U.S. NMS exchanges. In particular, we show that inverted venue usage increases for stocks facing more binding tick size constraints. It is especially more prevalent on the side with more supply as implied by the order book imbalance. However, inverted venues' usage is capped at around 25% market share due to factors such as agency conflicts in broker routing practices and liquidity concentration in primary exchanges. We also find evidence supporting the model's predictions on volume dynamics between make-take and inverted venues.

Finally, we discuss policy implications on optimal tick size levels to reduce intermediation rents. More granular tick size regimes scaling with price level may enhance market quality relative to the current U.S. binary approach.

Overall, this paper contributes to the literature on market microstructure by modeling the interactions of exchange fee structures and price discreteness. The analysis offers insights for better tick size regulation and routing practices.

2.9 Appendix

2.9.1 Additional tables and figures

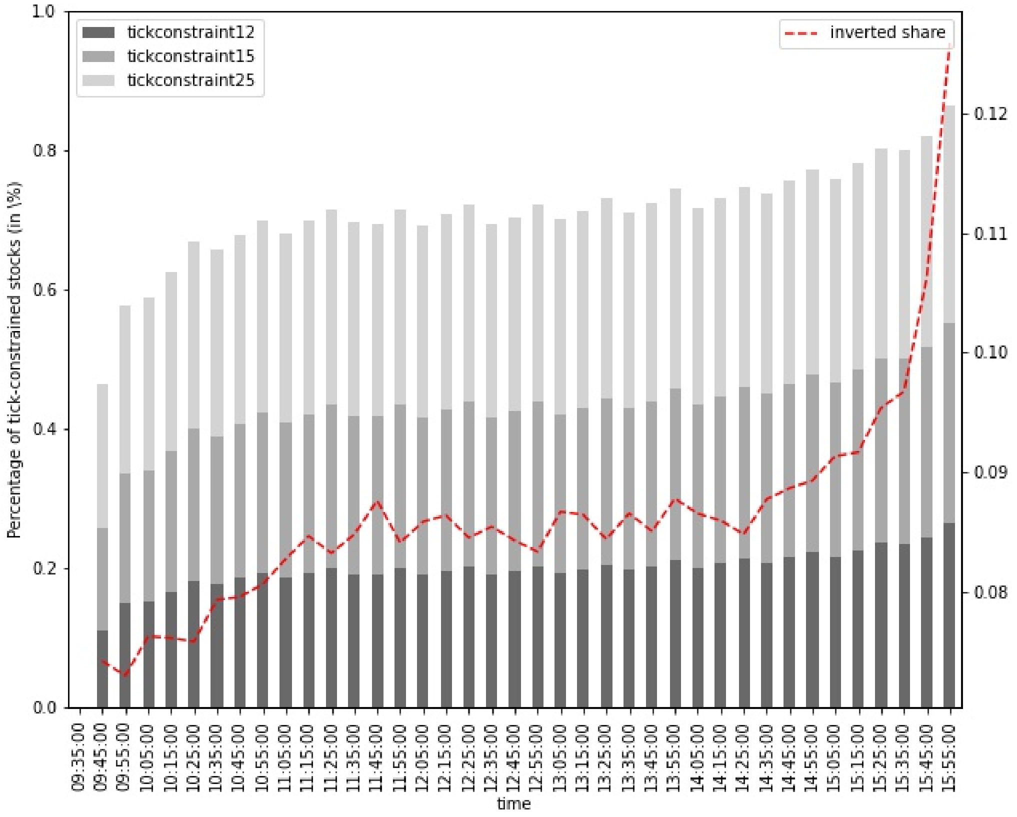
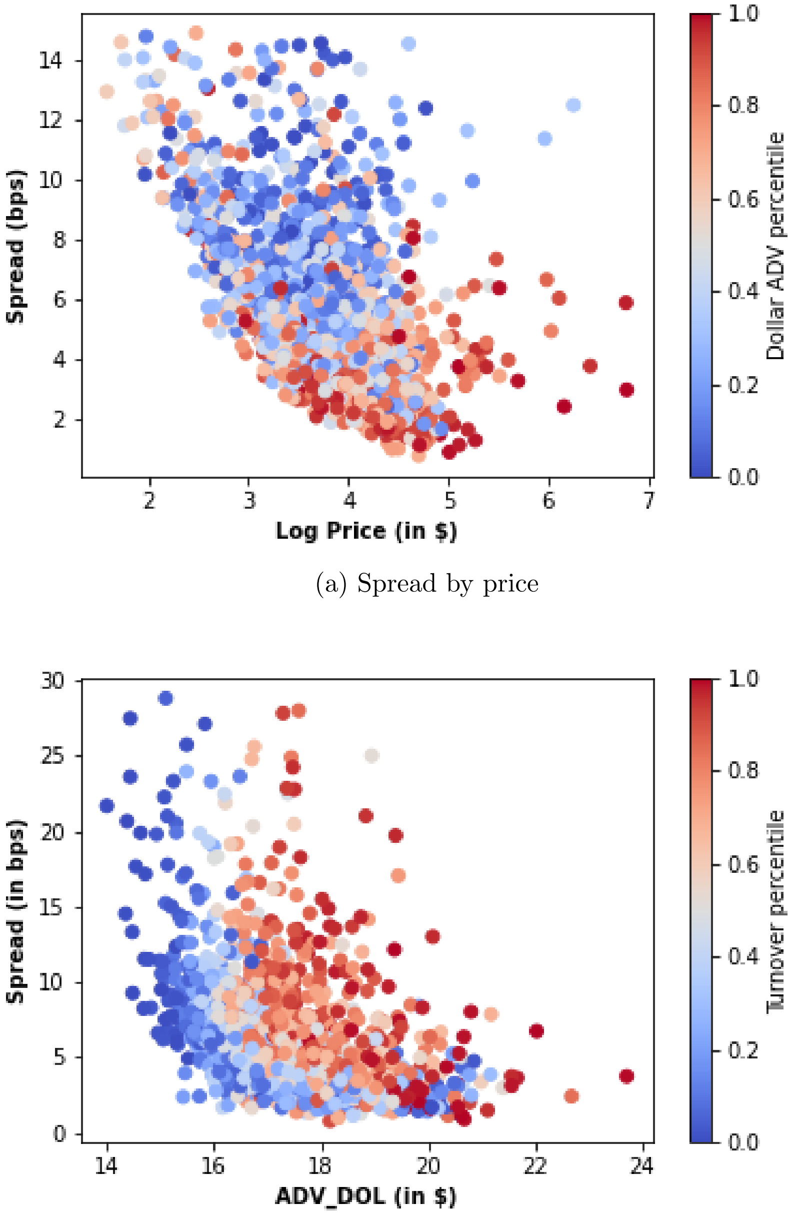


Figure 2.11: Tick constraint over time

		REL TICK	SPREAD	NHFT ONLY	HFT ONLY	BOTH NBBO	HFT LP
MKT CAP	REL TICK						
SMALL	1	4.1bps	0.156$	46.7%	8.7%	44.5%	48.9%
	2	7.9bps	0.033$	24.2%	7.6%	68.2%	55.4%
	3	13.8bps	0.025$	21.5%	5.5%	73.0%	63.3%
MEDIUM	1	2.1bps	0.063$	31.8%	8.7%	59.4%	45.2%
	2	3.2bps	0.033$	16.6%	8.1%	75.4%	48.1%
	3	8.4bps	0.016$	4.3%	4.4%	91.3%	69.9%
LARGE	1	1.3bps	0.067$	15.1%	6.0%	78.9%	55.9%
	2	2.4bps	0.015$	1.4%	6.7%	91.9%	62.7%
	3	5.6bps	0.011$	0.1%	0.9%	98.9%	77.5%

Table 2.4: Percentage of NBBO liquidity provided by each market participant

This table explains who provides liquidity across market capitalization and relative tick sizes. Market capitalization and relative tick sizes are both ranked into terciles, from smallest to largest. As expected, average relative spreads decrease with market capitalization. Importantly, HFT's share of the liquidity provision increase with both firm size and relative tick size. HFTs become more likely to be the liquidity provider when there is only one type of trader at the NBBO.



(a) Spread by price

(b) Spread by ADV

Figure 2.12: Spread on nominal price and ADV

Chapter 3

The Making of Dominant Currencies: Evidence in DeFi

We examine dominant currencies in decentralized finance (DeFi), their properties and determinants. We collect data from the largest DeFi exchange, Uniswap. By analyzing the swapping route between currency pairs of the top 50 currencies, we measure the daily dominance of each currency by its share of total trading volume, eigenvector and betweenness centralities of the trading network. We find that there are 3-5 currencies vying for dominance status in DeFi. Safety is a leading dominance attribute during the bust, while liquidity is more important during the boom. In line with findings from the dominant currency paradigm (DCP) literature, we find that utility coins are generally preferred as a dominant currency, whereas the dominance of stablecoins rises rapidly during market turmoil. We also find that an active money market, market size, and a currency's correlation with transaction costs are important determinants for dominance, suggesting essential design choices for future central bank digital currencies (CBDCs).

3.1 Introduction

The concept of a "dominant currency" has been a longstanding topic of interest in the field of international trade and finance. Historically, one currency—such as the Dutch Florin throughout the 18th century, followed by the British Pound (GBP) and, since the 1950s, the US Dollar (USD)—has facilitated the majority of global trade and financial flows for an extended period of time. The global dominant currency appears to be single and sticky despite competition from currencies such as Euro since its introduction in 1999 and more recently the Chinese Renminbi (RMB) (Ilzetzki et al., 2020; Bahaj and Reis, 2020; Gopinath and Itskhoki, 2022). The rise of dominant currencies has significant implications for monetary policy spillovers, financing, transaction costs, and financial market development (Gopinath et al., 2020). However, understanding what makes a currency dominant in the international financial system has been challenging due to the lack of historical observations and the slow pace of change in the macro environment.

We propose a different setting to examine the making of dominant currencies—decentralized finance (DeFi), a distributed financial system enabled by the blockchain technology. As an alternative to the traditional financial system, DeFi offers fast growing decentralized financial intermediary services ranging from payment, lending, trading, insurance, asset management, to derivative transactions and risk management, currently valued at over $50 billion in USD, about half of which is exchange trading.[1] Unlike the international financial system where there is one single well-established dominant currency, there is on-going competition for attaining the dominant currency status in DeFi among many cryptocurrencies including stablecoins, utilities coins and cryptoassets, offering a natural experiment to examine the properties and the making of dominant currencies, if there are any, in this environment some of which can be generalized to the real economy.

Moreover, there is a rich amount of data in DeFi. DeFi transactions are transparent and all historical data are publicly accessible, allowing us to examine currency usage choices in greater detail. In contrast, international payments and transactions are notoriously difficult to track and opaque. Furthermore, technology shocks in DeFi are rapid. New protocols and associated cryptocurrencies emerge and die out every day. The switching cost between protocols and currencies in DeFi is much lower thanks to the permissionless nature of the ecosystem. Consequently, values of crypto assets are extremely volatile, potentially causing frequent regime shifts of dominant currencies. Therefore, DeFi offers a unique laboratory setup to study the properties of dominant currencies and to test whether any of the mechanisms for the making of dominant currency proposed in the international money and finance literature can be corroborated by the emergence of dominant currencies for DeFi on blockchain.

Specifically, we examine dominant currencies in DeFi based on data from the Ethereum blockchain, which has the largest DeFi ecosystem with smart contract compatibility and a transaction fee (gas) that exists in proportion to computational complexity. We look at three protocols on the Ethereum blockchain: Uniswap, Aave and Compound. Uniswap is one of the most popular decentralized exchanges (DEXs). Aave and Compound are two of the largest protocols for loanable funds (PLFs) on the Ethereum blockchain. DEXs are asset exchanges with no centralized market makers. All asset swap transactions on

[1] The blockchain technology that underlies DeFi development has the potential to reduce contractual friction in financial transactions across geographic locations and to lower transaction costs through a distributed, censorship-resistant network with a tamper-proof and auditable record-keeping system.

DEXs are based on exchange rates set by a pre-programmed bonding curve (via smart contracts) between traders and a pool of passive liquidity providers. Cross-pool trades are possible to minimize price impacts of swap trades. Both Aave and Compound are PLF on the Ethereum blockchain. PLFs refer to protocols that feature overcollateralized lending. Depositors supply their assets to a PLF liquidity pool and receive supply interest based on a deposit contract. The deposited asset can be lent to PLF users and can serve as collateral for the depositors themselves to borrow other crypto assets.[2]

We begin our study by defining the dominance of a currency in DeFi utilizing unique properties of DEX or PLF and link the dominance status with one of the three roles of money, as a medium of exchange, a unit of account, or a store of value. The DEX protocols are different from traditional order-book based exchanges in several aspects. First, instead of order-books, a DEX protocol uses liquidity pools to facilitate trading and set exchange rates. Each liquidity pool services the swap trades of an asset pair. Liquidity providers deposit their holdings of an asset pair into the corresponding liquidity pool. Instead of trading against each other as in the traditional exchanges, buyers and sellers in DEX are trading against liquidity providers who are effectively passive market makers for the corresponding asset pair. The exchange rate between the asset pair is computed automatically according to a pre-programmed function that takes into account of the amount of each of the two assets left in the liquidity pool – the asset that makes up a larger proportion of the liquidity pool has relatively lower demand and hence commands a cheaper price and the price impact is less if liquidity providers deposit more of both assets in the pool. This feature of liquidity pools allows us to observe both the direction and the value of asset swap trades, enabling us to map out trading networks among cryptocurrencies dynamically. Second, for each pair-wise asset swap request, a DEX such as Uniswap uses an optimizing router to execute trades along a chain of asset swaps via corresponding liquidity pools that incurs the lowest total price impact. Utilizing this unique feature, we collect daily data on all executed trade routes and categorize asset swap trades into two types: those associated with the source or the target cryptocurrencies, which we call 'ultimate' trades, and those associated with cryptocurrencies in-between the source and the target tokens, which we call 'betweenness' trades. This categorization helps us to isolate the role of a medium of exchange from other roles performed by a crypto token. The 'betweenness' trades are driven by the desire to lower the transaction costs of the 'ultimate' trades. Therefore, a cryptocurrency featured more in the 'betweenness' trades is more liquid, and hence a better medium of exchange in the cryptocurrency market. It is *excessively* used in exchange transactions (since it is neither

[2] The lending contract on PLFs specifies the terms of the loan, such as a haircut based on the characteristics of the collateral, and the lending and borrowing interest rates, which are computed from a pre-specified formula used to clear the market.

the target nor source of any trades), a typical feature of a vehicle currency. In contrast, the 'ultimate' trades are motivated by reasons other than lowering transaction cost. The source or the target tokens are likely to be held for either investment, speculative, unit of account, store of value or any other purposes.

Based on this trade categorization, we compute several cryptocurrency dominance metrics. The first set are market-based. We measure a cryptocurrency's dominance in the role of medium of exchange as its percentage of (daily or weekly) total 'betweenness' trades, and its dominance in other potential roles of monetary assets as its percentage of total 'ultimate' trade volume, equal or value weighted. We also compute a cryptocurrency's dominance based on its percentage of total trade volume including both 'ultimate' and 'betweenness' trades. Though simple and intuitive, these market-based dominance metrics, equal or value weighted, however, are possibly affected by the size of ecosystem associated with each cryptocurrency, thus likely biased against or towards cryptocurrencies with larger market capitalizations. Alternatively, we also provide the second set of metrics utilizing the additional information in the dynamic trade networks. The dominance measures in this case are based (daily and weekly) network centrality metrics – the betweenness and the eigenvector centrality of a cryptocurrency. The betweenness centrality of a cryptocurrency, for example ETH, measures the fraction of those trades involving ETH in the trading route among all the trades in which ETH is neither the ultimate source nor the ultimate target asset. The eigenvector centrality of a cryptocurrency on the trading network is computed based on either only 'ultimate' trades or both 'ultimate' and 'betweenness' trades. The former eigenvector centrality of a cryptocurrency, for example ETH, is based only on trades where ETH is either the target or the source coin. The latter eigenvector centrality of ETH is based on all trades, reflecting the combined dominance of ETH in the trading network. These network-based metrics capture the dominance not purely based on the size but also the influence of a crypto token.

We find 3,301,933 betweenness trades, which accounts for 2.84% of total trades. We also find that betweenness centrality of a coin leads its eigenvector centrality but only during the boom. This suggests that liquidity (that is, the role of medium of exchange) drives dominance in the boom. We also find that betweenness centrality explains about 60% of the trade volume share and eigenvector centrality contributes to about 30% additionally in the sample. Finally, we measure the safety of a cryptocurrency. Stablecoins offer safety. We capture the difference among stablecoins by how successful they raise funds (its stablecoin market share). The safety measure is, hence, proportional to the market share of a coin in the stablecoin universe.

We graph time series dominant metrics for the top five cryptocurrencies in our sample and find that utility coins are generally preferred as a dominant currency, whereas the dominance of stablecoins rises rapidly during market turmoils. We also compute the daily Herfindahl indices of dominance metrics and find there are 3-5 cryptocurrencies battling for dominance status in DeFi, more dominant currencies during the boom period. Analysis of cross-autocorrelations among dominance metrics shows that safety is a leading dominance attribute, especially during the bust, while liquidity is more important during the boom.

Finally, we draw insights from DCP literature to form testable hypotheses on the determinants of these dominance metrics. We focus on the following asset-level characteristics: volatility, market size, safety, correlation with the gas fee (i.e. transaction cost) on the blockchain, and money market activities. We measure money market activities of a cryptocurrency by the amount of associated deposit in PLF. Results from cross-sectional time-series regression indicate that an active money market, market size, currency's correlation with (and hence ability to hedge against) transaction costs are important determinants for dominance. We also utilize the institutional features of DeFi to design difference-in-difference test of the importance of money market for currency dominance. We find again that when a money market is introduced to a cryptocurrency, its trade volume share, trade network centrality, and equally weighted betweenness centrality improve. Currently, as cryptocurrencies and stablecoins have become more popular, many central banks, especially those in the developing countries are considering to provide CBDCs as an alternative and competing for dominance in CBDC space.[3] Our findings on the determinants of dominant currency might be useful for governments to consider essential design choices for future CBDCs.

The remainder of this paper is organized as follows. In Section 3.2, we review the related literature. In Section 3.3, we give a quick overview of various DeFi protocols we draw data from in the paper. In Section 3.4, we describe the data collection process. In Section 3.5, we present the computation methodology of dominance metrics. In Section 3.6, we analyze the properties of various dominance metrics, concentration of dominance, as well as lead-lag relationships among them. In Section 3.7, we examine the determinants of dominance metrics at the cryptocurrency level by conducting time-series cross-sectional regressions. We present and discuss the estimation results. Section 3.8 concludes.

3.2 Related Literature on Dominant Currencies

To understand how a dominant currency is chosen by the majority of participants in cross-border transactions, the existing macro and finance literature analyzes how a dominant currency performs in each of the three traditional roles of money, namely as a unit of

[3] See https://www.atlanticcouncil.org/cbdctracker/ for countries that are active in launching and experimenting CBDCs.

account, a store of value, and a medium of exchange.

Dominant currency paradigm literature on international trade emphasizes the role of dominant currency as a unit of account to explain the choice of currency in trade invoicing (Gopinath et al., 2020). As a unit of account, the dominant currency is used for the majority of international trade and financial transactions. According to this line of inquiry, the interaction of nominal price stickiness with pricing complementarities and input-output linkages across firms generates complementarities in currency choice (Gopinath, 2015; Doepke and Schneider, 2017; Mukhin, 2022; Eren and Malamud, 2022); that is, exporters coordinate on the same currency of invoicing for the following two reasons: to be competitive in output pricing; and to be able to hedge their balance sheet against exchange rate shocks with the denominated currency of imported intermediate (real and financial) inputs. Financial intermediate inputs can be thought as working capital, trade credit, or any form of financial borrowing.[4] Some DeFi services specify invoice currencies but invoice currency choices are not directly observable other than in a few (anecdotal) cases where contracts are publicly available. However, indirect variables that capture reasons to trade that relate to invoicing or balance-sheet hedging: such as a network effect on the currency denomination of working capital/financial borrowings/transaction might explain the dominance of a currency from this perspective.

A large body of international finance literature emphasizes the role of the dominant currency as a store of value. A dominant currency such as the US Dollar (USD) preserves its value during global market crises and is therefore widely used as an international reserve/safe asset. This safety feature offered by assets denominated in the dominant currency means that the currency preserves value added in exchange transactions, leading to its wide use in the global financial market. Differences in financial development (hence the differences in access to safe assets) (Maggiori, 2017) or risk aversion of participants (Gourinchas and Rey, 2022) may drive the demand for an international safe asset. Gopinath and Stein (2021) argue that assets denominated in the dominant currency can be used as a savings device for export producers to hedge against invoicing risk. Chahrour and Valchev (2022) additionally suggest that safe assets are used as collateral to overcome contractual frictions in cross-border transactions. The dominant currency regime may persist due to the feedback between safe asset returns and the use of collateral. This literature also highlights several

[4] Gopinath et al. (2010) and Goldberg and Tille (2016) find that dollar pricing is more common in sectors classified as producing goods that are homogeneous and hence likely substitutes. The theoretical result in Mukhin (2022) indicates the size of market is important in determining the dominant currency. Amiti et al. (2022) find that Belgium firms that import their inputs are likely to choose dollar pricing. BIS (2014) documents that traded finance contracts are mostly dollarized even though they are sourced via local banks, indicating most trades are financed via dollars. Bahaj and Reis (2020) find that when the cost of financing working capital in RMB is lower due to the swap arrangement through central banks, trades are more likely to denominated in RMB.

key properties of safe assets denominated in the dominant currency: they have low returns, high convenience yield, and are widely held in savings accounts. In the DeFi setting, safety of a cryptoassets can be measured relatively easily due to the existence of stablecoin as an asset class.

Finally, there is also abundant literature emphasizing the role of dominant currency as a medium of exchange. A dominant currency guarantees the lowest transaction costs and maximizes the room for mutually beneficial trade when used as a medium of exchange. These benefits are achieved through high volume, liquidity, and network effects. Krugman (1980) characterizes currencies with such dominance as vehicle currencies. The aforementioned literature on dominant currencies as an invoicing currency for international trade also highlights the importance of a currency's network effect on its ability to serve as a medium of exchange in global trade. Interestingly, the money-search literature also focuses on the liquidity aspect of the exchange and how dominant currencies as vehicle currencies of trade lower transaction in payments and emphasizes the coexistence of multiple currencies (Devereux and Shi, 2013; Zhang, 2014; Wright and Trejos, 2001). Recently, Coppola et al. (2023) propose that market liquidity is an important determinant for firms to choose a currency to denominate their global debt contracts due to the feedback between the market and funding liquidity. In the DeFi setting, this would imply that dominant currencies tend to have higher market liquidity and, relatedly, lower price impacts, and an active money market.

The question of whether there should be a single or a basket of dominant currencies has been a topic of debate in international trade and finance. The theory of optimal currency areas in economics suggests that larger and more integrated markets are more likely to adopt a common currency to reduce transaction costs and increase economic efficiency (Mundell, 1957). The theory of globalization of production suggests that intensive reliance of production on global supply chains would lead to a single dominant currency (Gopinath et al., 2020). Conversely, the theory of fragmentation and localization of supply chains suggests that a basket of dominant currencies would be more appropriate (Bahaj and Reis, 2020). Furthermore, the money search literature also points out the possibility of multiple dominant currencies. Since the technology for payment and contractual financial service is improving rapidly which lowers transaction costs and contractual frictions and potentially reduces the switching cost between dominant currencies, it is possible that dominant currencies can be multiple and vary over time to serve different purposes.

In spite of its name, the classification of cryptocurrencies as currencies remains an ongoing debate among regulators, market participants, and academics. Given the distinctive characteristics of these digital assets—including extreme volatility, limited historical data, and evolving market structures—recent work has adapted classical factor models to better capture their pricing dynamics.

The notion that currencies may themselves exhibit risk factors is not new: Lustig et al. (2011) document a currency carry factor built by a long-short portfolio between high and low interest rate currencies. High interest rate currencies are negatively correlated to global FX volatility and thus deliver low returns in periods of unexpected high volatility (Menkhoff et al., 2012a). In a parallel work, Menkhoff et al. (2012b) show that momentum can be extended to international currency portfolios. Focusing on G-10 currencies, Aloosh and Bekaert (2021) argue that certain currency factors persist across different monetary regimes, reflecting both global financial linkages and underlying macroeconomic fundamentals. In particular, currency co-movements display a two-block structure, similar to a currency clustering factor within the dominant currency paradigm.

In the context of cryptocurrencies, a central question is whether established asset pricing frameworks, such as the Fama and French (1992) three-factor model, continue to offer meaningful insights. Value and momentum strategies extend beyond traditional equities to encompass currencies, commodities, and bonds (Asness et al., 2013), suggesting broader applicability to digital assets as well. Indeed, Liu et al. (2022) find that momentum trading appears to persist in crypto markets, reflecting both limited arbitrage capital and strong speculative flows. While size and momentum remain important factors in explaining cryptocurrency returns, the traditional value factor appears to lack a clear analogue in digital assets. In another related work, Liu and Tsyvinski (2021) find alternative factors such as investor attention appear to play a more prominent role. Liquidity in cryptocurrency markets is likely to play an important role, as shown by extensive work on commonality in stock market liquidity (Pástor and Stambaugh, 2003; Chordia et al., 2000; Korajczyk and Sadka, 2008).

Finally, the market structure and trading behaviors in crypto markets differ fundamentally from those in more established financial markets. Borri (2019) highlights conditional tail risk in cryptocurrency asset pricing, showing the existence of a crypto-specific systematic risk while exhibiting little exposure to broader equity and macro factors. Makarov and Schoar (2020) analyze arbitrage across exchanges and uncover persistent price dislocations tied to fragmented liquidity and regulatory constraints, suggesting that standard models, which assume integrated markets, may understate the cost of trading frictions.

3.3 A Primer on DeFi

DeFi is a financial ecosystem built on top of blockchains. DeFi protocols operate according to a set of rules pre-programmed into so-called "smart contracts" deployed on blockchains. Given the transparent nature of public blockchains, all transactions are publicly available.

This provides a perfect environment for us to conduct natural experiments on dominant currencies. In particular, our experiment is linked to DEXs and lending protocols.

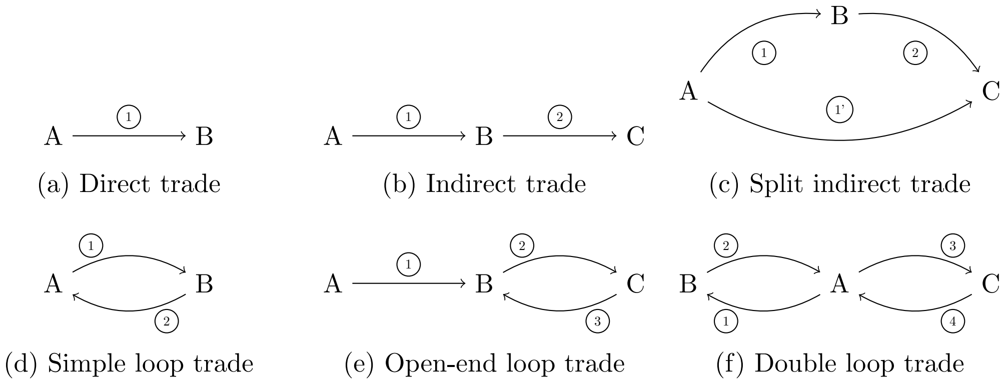
3.3.1 Ethereum

Blockchains that focus solely on simple payments such as the Bitcoin blockchain are not able to support complex algorithms needed for running a financial application. To be DeFi-compatible, a blockchain needs to be able to support smart contracts, which are essentially a piece of software deployed on chain. As the oldest blockchain with the embedded smart contract layer, Ethereum is also the most popular chain in terms of DeFi activities. Due to the distributed nature of blockchain, a state update (e.g. crediting a recipient's account balance) triggered by interacting with a DeFi protocol requires computational resources from all the validating nodes in the network. To encourage efficient usage of resources and discourage spamming, every transaction is subject to a fee. On the Ethereum network in particular, transaction fee is measured in gas: the higher the computational complexity of the transaction, the higher the gas consumption. The transaction fee amount is the gas price set by the transaction initiator multiplied by the gas consumption. Due to limited computational resources and network capacity, users are incentivized to set up a sufficiently high gas price for their transactions to be processed and committed to the blockchain. To this end, the more congested the network is—meaning a higher demand for transaction processing, the higher the gas price will generally be.

3.3.2 DEX

DEXs on blockchain usually refer to exchange protocols using automated market marking (AMM) algorithms. Unlike traditional order-book based exchanges that execute trades by matching buyers and sellers, an AMM-based DEX guarantees immediate trade execution by computing the exchange rate between the asset pair traded automatically according to a pre-programmed conservation function. Instead of trading against each other, DEX traders are effectively trading against liquidity providers, who contribute funds to liquidity pools. Each liquidity pool supports swap transactions of a specific asset/token pair. The conservation function of a liquidity pool stipulates that the coordinates of the token reserves in pool must follow a predetermined curve (or hyperplane, in case of more than two tokens in a pool) after each trade. The simplest and most commonly adopted conservation function is a constant product function $C = r_1 \times r_2$, where the product of reserves of token$_1$ and token$_2$—r_1 and r_2 respectively—must remain constant after each trade (Xu et al., 2023). At the time of writing, Uniswap is the largest DEX in terms of both total value locked (TVL) and trading volume. The protocol has witnessed three iterations—v1, v2, v3, with the latter two most widely adopted and hence being the source of empirical data for this study. Uniswap features token-pair pools; the v2 protocol applies the vanilla constant product function as its conservation function as discussed above, while the v3 protocol applies a derived form of it with an additional liquidity concentration factor to enhance capital efficiency of pool liquidity.

Not every possible combination of asset pairs would have a pool. For capital efficiency, typically in a liquidity pool, at least one of the two assets would be stablecoin, ETH or some other widely held token. Therefore, when a trade between two exotic tokens occurs, say token$_A$ and token$_B$, usually at least two liquidity pools would be interacted with, for example token$_A$-USDC and token$_B$-USDC. As such, the trade can be routed through token$_A$ → USDC → token$_B$, with USDC being the intermediary settlement token, all within one transaction. We term an unsplittable trade that involves interacting with one liquidity pool as an "atomic trade".



(a) Direct trade (b) Indirect trade (c) Split indirect trade

(d) Simple loop trade (e) Open-end loop trade (f) Double loop trade

Figure 3.1: Typical forms of transactions with an AMM-based DEX. Capital letters A, B, ... denote different cryptocurrencies; an arrow represents the direction of one atomic trade (e.g. A → B means to sell A and buy B); a number in circle indicates the order of trade within one transaction.

Figure 3.1 illustrates some typical forms of transactions with an AMM-based DEX. Note that this is an inexhaustive list of transaction forms: different trade routing methods can be combined and permutated to form a new route. Transaction (3.1a) represents a simple trade that consists of just one *atomic* trade which takes place at the token$_A$-token$_B$ liquidity pool. Transaction (3.1b) is an indirect trade that consists of two *atomic* trades: token$_A$ → token$_B$ and token$_B$ → token$_C$, with the former through the token$_A$-token$_B$ liquidity pool while the latter the token$_B$-token$_C$ liquidity pool. Transaction (3.1c) consists of three atomic trades: the sell volume of token$_A$ is first split into two parts, with one part flowing through an indirect trade interacting with token$_A$-token$_B$ and token$_B$-token$_C$ pools, while the other part traded directly with the token$_A$-token$_C$ pool. As the Uniswap router seeks to maximize the output token for users, the split can occur if going through one swap route is not optimal. This can be due to the fact that, routing a trade through different pools can optimize the use of combined liquidity, hence reduce the price impact and increase the output quantity. Transactions (3.1d), (3.1e) and (3.1f) involve loops: i.e. at least one token is both an source token for one atomic trade and an target token for another atomic trade in the same transaction. Those transactions are typically conducted for arbitrage purposes, with or without the same trading pair. If with the same trading pair, the transaction usually exploits price gaps between pools from different protocols (e.g. pools from Uniswap V2 and V3 with the same trading pair); if with different trading pairs, the transaction can result in profit by forming e.g. a triangular arbitrage. Transactions involving loops account for approximately 1.5% of the total number of transactions; they are typically performed by bots and do not represent genuine trading demand. We thus eliminate those transactions in our analysis.

3.3.3 PLF

PLFs on blockchain usually refer to protocols that offer overcollateralized lending. To be able to borrow from a PLF, deposits must first be supplied to the PLF as collateral. The utilization ratio of an asset, calculated as the total borrowed divided by total supplied amount, decides the supply and borrow interest rates of the asset based on a pre-programmed interest rate model. At the time of writing, Aave and Compound are the two largest PLFs in terms of TVL and trading volume on Ethereum. Lending and borrowing are the predominant service products on DeFi.

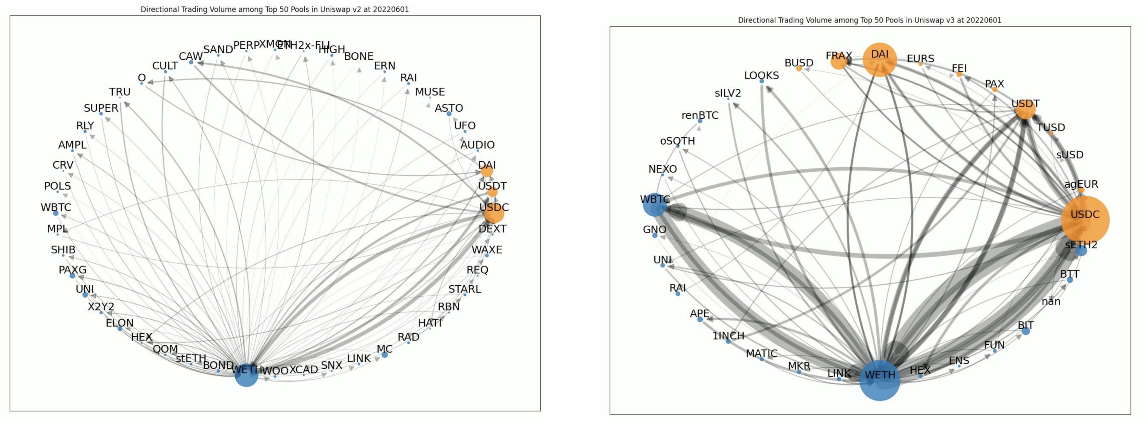
3.4 Data

We rely on transaction record from the Uniswap protocol, lending and borrowing data from the Compound protocol, and market-level data such as transaction cost in the Ethereum network, as well as token-specific data such as token market cap for this study.

3.4.1 DEX data

We fetch on-chain transaction data associated with Uniswap V2 and V3 from the Subgraph API. The data sample comprises 130,894 individual liquidity pools, including 122,458 V2 pools and 8,436 V3 pools until January 31, 2023. In total, the sample contains 28,608,912 transactions on Uniswap, including 18,635,768 transactions in V2 and 9,973,144 transactions in V3, from the protocols' deployment (on May 5, 2020 for V2 and May 4, 2021 for V3) until January 31, 2023. We observe 876,929 liquidity addition, 712,934 withdrawal, 43,100,698 exchange trades from top 50 pools on Uniswap V2 and the top 50 pools on Uniswap V3. We focus on transactions interacting with the top 50 pools on Uniswap V2 and the top 50 pools on Uniswap V3 by monthly trading volume USD. Due to the high concentration of liquidity in the few top pools, we believe that omitting the remaining pools would not significantly affect the results.

The volume data can be directly fetched by the Subgraph APIs as USD derived via Ethereum (ETH) prices from the tokens within the whitelist. Figures 3.2a and 3.2b give a snapshot of the trading network for Uniswap V2 and V3 on June 1 2022, respectively.



(a) Uniswap V2

(b) Uniswap V3

Figure 3.2: A Snapshot of the Trading Network on June 1 2022. Nodes denote different cryptocurrencies; the size of the node denotes the liquidity of that cryptocurrency; the node is yellow if the cryptocurrency is the stablecoin or blue otherwise. Directed edges denote the flow of volume between two cryptocurrencies; the weight of the edge indicates the amount of volume.

3.4.2 PLF data

For data in the lending and borrowing markets, we focus on the Compound protocol[5]. We use Compound's own API to collect data on the total supply, borrow volume in USD, as well as supply and borrow interest rates.

3.4.3 Ethereum network data

We fetch average daily gas fee in Wei and ETH price in USD from the Ethereum network from Etherscan. Wei is the smallest unit of ETH: 1 ETH $= 10^{18}$ Wei. The gas fee can

[5] We have also gathered data from Aave–another top PLF, and the data from Aave and the Compound protocol are highly correlated. For that reason and for the ease of continuous data collection, we only use data from the Compound protocol.

hence be easily converted to USD using the USD-denominated ETH price.

3.4.4 Token-specific data

We collect token-specific data—including price, price volatility, and market capitalization—from CoinGecko. We also collect token categories—e.g. governance tokens, stablecoins, utility coins—from Coinmarketcap.

3.4.5 Boom and Bust in the Crypto Market

To facilitate our analysis, we define the boom and bust periods of the crypto market based on the S&P crypto index value. We borrow the definition method from Aramonte et al. (2022). Specifically, we define a boom period as the period between a price trough and a peak with an increase of over 30%, and define a bust period as the period between a price peak and a trough with a decrease of over 30%. There also exist periods with below-30% price change which are neither a boom nor a bust. Figure 3.3 graphs the boom and bust delineation of our sample period.

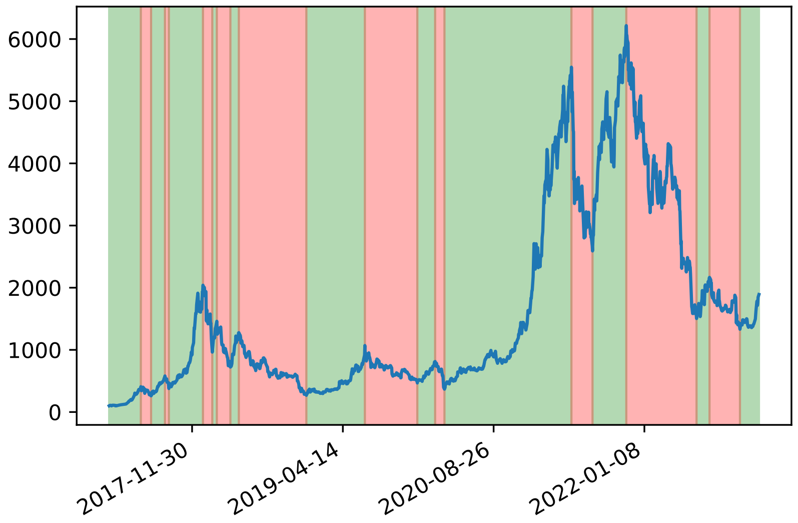


Figure 3.3: S&P Crypto index with boom (green) bust (red) periods. The blue line denotes the S&P Cryptocurrency Broad Digital Market Index; the green span denotes the boom period, while the red span denotes the bust period.

Table 3.1: Notations and Descriptions of Variables

The table reports a list of variables and their descriptions used in the study of decentralized finance (DeFi) markets. The variables are related to the trading volume, market capitalization, liquidity, and network centrality of cryptocurrencies traded on the Ethereum blockchain. The table includes variables such as total volume of atomic trades, in- and out-eigenvector centrality, count- and volume-weighted betweenness centrality, market capitalization, liquidity, and log return of token price. Additionally, some variables represent shares and ratios of the total volume or liquidity of all tokens, and some are dummy variables that indicate whether a token is a USD-pegged stable coin. The notations used for the variables are described in the table.

Variable	Descriptions
$V_{i,t}^{In}$	Total volume of atomic trades on day t where token$_i$ is the target token
$V_{i,t}^{Out}$	Total volume of atomic trades on day t where token$_i$ is the source token
$V_{i,t}$	Total volume of atomic trades on day t where token$_i$ is either the source or the target token, $V_{i,t} = V_{i,t}^{In} + V_{i,t}^{Out}$
$\mathit{VShare}_{i,t}^{In}$	Total volume of atomic trades on day t where token$_i$ is the target token, as a fraction of summed total in-volume of all tokens, $\mathit{VShare}_{i,t}^{In} = \frac{V_{i,t}^{In}}{\sum_k V_{k,t}^{In}}$
$\mathit{VShare}_{i,t}^{Out}$	Total volume of atomic trades on day t where token$_i$ is the source token, as a fraction of summed total out-volume of all tokens, $\mathit{VShare}_{i,t}^{Out} = \frac{V_{i,t}^{Out}}{\sum_k V_{k,t}^{Out}}$
$\mathit{VShare}_{i,t}$	Total volume of atomic trades on day t where token$_i$ is either the source or the target token, as a fraction of summed total volume of all tokens, $\mathit{VShare}_{i,t} = \frac{V_{i,t}}{\sum_k V_{k,t}}$
$\mathit{EigenCent}_{i,t}^{In}$	In-eigenvector centrality of token$_i$ on day t
$\mathit{EigenCent}_{i,t}^{Out}$	Out-eigenvector centrality of token$_i$ on day t
$\mathit{BetwCent}_{i,t}^{C}$	Count-weighted betweenness centrality of token$_i$ on day t
$\mathit{BetwCent}_{i,t}^{V}$	Volume-weighted betweenness centrality of token$_i$ on day t
$MCap_{i,t}$	Market capitalization of token$_i$ on day t
$\mathit{Liquidity}_{i,t}$	Total amount of token$_i$ in Uniswap pools at the end of day t
$\mathit{LiquidityShare}_{i,t}$	Total value of token$_i$ in Uniswap pools at the end of day t as a fraction of summed value of all tokens in Uniswap, $\mathit{LiquidityShare}_{i,t} = \frac{\mathit{Liquidity}_{i,t}}{\sum_k \mathit{Liquidity}_{k,t}}$
$\mathit{SupplyShare}_{i,t}$	Supply amount of token$_i$ as a fraction of total borrow amount from Compound at the end of day t, $\mathit{SupplyShare}_{i,t} = \frac{\mathit{Supply}_{i,t}}{\sum_k \mathit{Supply}_{k,t}}$
$R_{i,t}^{USD}$	Log return of token$_i$ price in USD on day t
$\mathit{IsStable}_i$	Dummy variable $\mathit{Stable}_i = 1$ if token$_i$ is a USD-pegged stable coin
$\mathit{StableShare}_{i,t}$	Market cap of token$_i$ as a fraction of total market cap of all stablecoins at the end of day t times $\mathit{IsStable}_i$
$\sigma_{i,t}^{USD}$	Past 30-day standard deviation of daily log return of token$_i$ price in USD on day t
$\mathit{GasPrice}_t$	Daily average gas price in USD of Ethereum network
$\sigma_{\mathrm{gas},t}^{USD}$	Past 30-day standard deviation of daily log return of gas price
$R_{\mathrm{SP},t}^{USD}$	Log return of the S&P Crpyto composite index price in USD on day t
$\sigma_{\mathrm{SP},t}^{USD}$	Past 30-day standard deviation of daily log return of S&P price
$\mathit{CorSP}_{i,t}$	Correlation between past 30-day log return of token$_i$ price in USD and that of S&P Crypto index price
$\mathit{CorETH}_{i,t}$	Correlation between past 30-day log return of token$_i$ price in USD and that of ETH price in USD
$\mathit{CorGas}_{i,t}$	Correlation between past 30-day log return of token$_i$ price in USD and that of gas price in USD
IsBoom_t	Dummy variable $\mathit{IsBoom}_t = 1$ if t is a boom period according to Figure 3.3

3.5 Measuring Currency Dominance in DeFi

In this section, we introduce a series of metrics to proxy currency dominance in DeFi. We refer the reader to Table 3.1 for the variable notations and definitions used in the rest of the paper.

3.5.1 Market share measurement

It is straightforward to measure currency dominance in DeFi transactions using market data. We first measure the dominance using simple percentages of total value traded for each currency: denoted as *VShare*.

As discussed in Section 3.4.1, we focus on tokens in Uniswap's top 50 pools by monthly trading volume. For each day t, we compute the sum of directional volume from token$_i$ to token$_j$ on day t, denoted by $V_{<i,j>,t}$ where $V_{<i,j>,t}$ equals total volume in USD of atomic trades on day t where token$_i$ is the source token and token$_j$ the target token. We denote $V_{i,t}^{in}$ or $V_{i,t}^{out}$ as the total inflow or outflow volume of atomic trades on day t where token$_i$ is either the source or the target token, and $V_{i,t}$ as the total traded volume on day t involving token$_i$ where $V_{i,t} = V_{i,t}^{In} + V_{i,t}^{Out}$. Hence, the dominance in inflow, outflow, or total trading volume for coin i at date t is captured by $VShare_{i,t}^{direction} = \frac{V_{i,t}^{direction}}{\sum_k V_{k,t}^{direction}}$ or $VShare_{i,t}$ where $direction = \{in, out\}$. We omit the direction superscript when we use the total volume to define $VShare$. Note that this measure of dominance can be partly explained by the size of ecosystem relating to each currency and may or may not reflect excessive use of these currencies as vehicles for invoicing or exchange transactions.

In Figure 3.4, we plot the time series of in-/outflow *VShare* of five major cryptocurrencies. We additionally mark four major events in the timeline: (i) 26 November 2020, the price oracle attack with DAI that caused liquidation on Compound, (ii) 5 May 2021, the introduction of Uniswap V3, (iii) 10 May 2022, the LUNA/Terra collapse, (iv) 11 November 2022, the FTX collapse.

Figure 3.4a, Figure 3.4d show that WETH has consistently high volume share—both directional and non-directional, throughout the observation period, partially explainable by the fact that most Uniswap V2 pools have a token pair with WETH on one side. This is because a large amount of Uniswap V2 liquidity was migrated from Uniswap V1[6], whose design stipulates that each pool must have WETH on one side. While any arbitrary token-pair can form a pool with Uniswap V2, the protocol inherited the feature of its deprecated V1 version due to liquidity migration. Despite this inherent mechanism, WETH's top position in volume share on Uniswap V2 was challenged by USDC around May 2022, coinciding with the

[6] `https://twitter.com/Uniswap/status/1262435668539715587`

Luna/Terra incident which caused a huge market turmoil. Notably, the price oracle attack incident around November 2020 caused a surge in trading volume of USDC and DAI, and squeezed the volume share of ETH. The effect of the attack is rather short-lived, as ETH quickly regained its top position. This is to contrast regime-switch-triggering events such as the introduction of Uniswap V3, which has a long-lasting effect of shaking the dominance of ETH.

Figure 3.4b, Figure 3.4e illustrate tokens' in-/outflow and total volume share in the Uniswap V3 market. Uniswap V3 has a concentrated liquidity provision feature that allows liquidity to be only provided for a certain price range defined by the liquidity provider. Uniswap V2 and V3's designs serve slightly different purposes: V2 is simple and gas-saving due to its low computational complexity, while V3 is more customizable to the expense of higher gas consumption. Therefore, the introduction of V3 does not deprecate V2. Thus, liquidity migration from V2 to V3 is rather limited, and V3 pool token-pair combinations quickly become much more diverse than V2. Clearly, absent mechanical constraints, multiple dominant currencies emerge; and in the case of Uniswap V3, they are WETH and USDC.

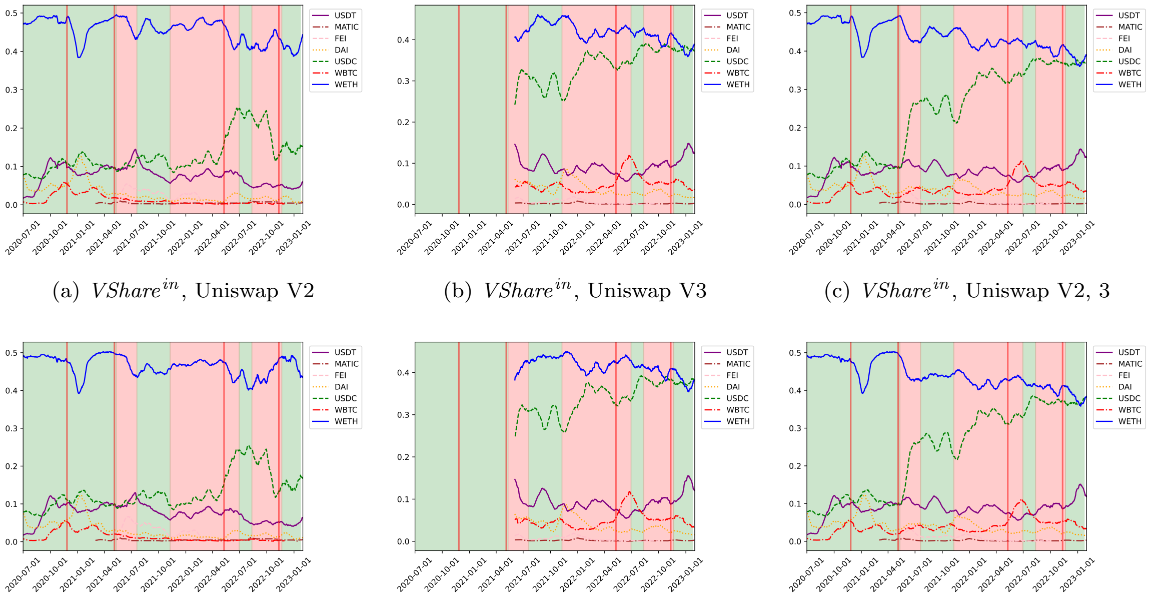
Figure 3.4c and Figure 3.4f present the change in tokens' volume share in the combined Uniswap V2 and V3 markets. Again, we observe USDC quickly takes over WETH's dominant status since the introduction of Uniswap V3, when token-pair combination is much less influenced by the legacy design of compulsory inclusion on WETH. Notably, the decline of WETH's top position in volume share appears to accelerate since the FTX collapse in November 2022. This is also reflected by the fact that there might be a sell-out of WETH in the Uniswap V2 market since there is an increase (a decrease) in WETH's outflow (inflow) volume share during the same period shown in Figure 3.4a and Figure 3.4d.

3.5.2 Eigenvector centrality

Figure 3.4c shows that WETH and USDC combined capture 80 to 90% of the market. However, that does not mean that all trades are between these two cryptocurrencies. In fact, we show next, either of these two cryptocurrencies are highly likely to be a counterparty of all other cryptocurrencies in Uniswap market when we utilize the trading network to compute eigenvector centrality, a measurement introduced by Bonacich (1987). It is computed by solving for $\boldsymbol{x_t}$ with the eigenvector equation below:

$$\boldsymbol{A_t x_t} = \lambda_t \boldsymbol{x_t}\,, \tag{3.1}$$

where $\boldsymbol{A_t} = (A_{ij,t})_{1\leq i,j,\leq n}$ is the aggregate-volume-weighed adjacency matrix of the directed network of token transactions on day t with eigenvalue λ_t, and $\boldsymbol{x_t} = (x_{i,t})_{1\leq i\leq n}$ is the centrality vector of tokens on day t. There is a unique positive solution if λ_t is the highest



(a) $VShare^{in}$, Uniswap V2

(b) $VShare^{in}$, Uniswap V3

(c) $VShare^{in}$, Uniswap V2, 3

(d) $VShare^{Out}$, Uniswap V2

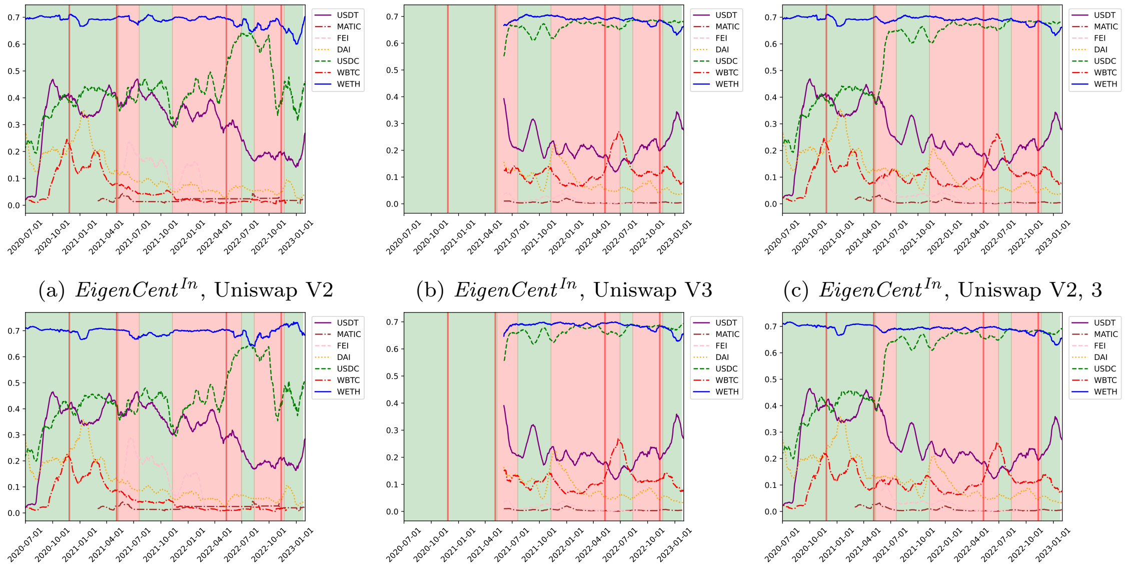
(e) $VShare^{Out}$, Uniswap V3

(f) $VShare^{Out}$, Uniswap V2, 3

Figure 3.4: 30-day moving average of volume of key cryptocurrencies. These figures plot the 30-day moving average of change in tokens' volume share for seven key cryptocurrencies in only Uniswap V2, only Uniswap V3, and both Uniswap V2 and V3; the green span denotes the boom period, while the red span denotes the bust period.

value by virtue of the Perron–Frobenius theorem (Mark EJ Newman, 2010). Since our networks do not have self-edges, the diagonal matrix elements are all zeros, i.e. $A_{kk,t} = 0$, $\forall k \in \{1, 2, ..., n\}$.

Let $EigenCent^{In}_{i,t}$ and $EigenCent^{Out}_{i,t}$ denote in-eigenvector centrality and out-eigenvector centrality, respectively, of token$_i$ on day t. Thus, $x_{i,t}$ represents $EigenCent^{Out}_{i,t}$ when $A_{ij,t} = V_{<j,i>,t}$, and $EigenCent^{Out}_{i,t}$ when $A_{ij,t} = V_{<i,j>,t}$.



(a) $EigenCent^{In}$, Uniswap V2 (b) $EigenCent^{In}$, Uniswap V3 (c) $EigenCent^{In}$, Uniswap V2, 3

(d) $EigenCent^{Out}$, Uniswap V2 only (e) $EigenCent^{Out}$, Uniswap V3 only (f) $EigenCent^{Out}$, Uniswap V2, 3

Figure 3.5: 30-day Moving Average of eigenvector centrality of key coins. These figures plot the 30-day moving average of eigenvector centrality for seven key cryptocurrencies in only Uniswap V2, only Uniswap V3, and both Uniswap V2 and V3; the green span denotes the boom period, while the red span denotes the bust period.

In Figure 3.5, we plot the eigenvector centrality of top five key tokens calculated by inflow trade volume and outflow trade volume. Overall, Figure 3.5 shows that a token's dominance in terms of eigenvector centrality is similar to those when measured with inflow transaction volume or outflow transaction volume. Similar to Figure 3.4, Figure 3.5a and Figure 3.5d show that WETH has consistently high eigenvector centrality throughout the observation period, partially explainable by the fact that most Uniswap V2 pools have a token pair with WETH on one side. Figure 3.5b and Figure 3.5e illustrate tokens' eigenvector centrality in the Uniswap V3 market. Clearly, absent mechanical constraints in the Uniswap V3 market, both WETH and USDC exhibit dominance.

Figure 3.5b and Figure 3.5e present the change in tokens' eigenvector centrality in the combined Uniswap V2 and V3 markets. Again, we observe the phenomenon of duo dominant currencies since the introduction Uniswap V3, when token-pair combination is much less influenced by the legacy design of compulsory inclusion on WETH. While only WETH's dominance is apparent before Q2 2021, centrality of stablecoins combined–particularly with USDT and USDC—is also quite high, but the centrality of each individual stablecoin falls far behind WETH. Since Q2 2021, USDT's centrality appears to steadily shift to USDC, which coincides with USDC's integration into mainstream payment network[7] and USDT's scandal concerning the reserves backing its value[8].

3.5.3 Betweenness centrality

Next, we turn to a unique measure of liquidity specific to each cryptocurrency in the Uniswap market, capturing how effective each crypto coin performs the function of medium of exchange. This measure leverages the fact that the Uniswap router optimally finds a path of liquidity pools that maximizes the number of output tokens for a given number of input tokens. Therefore, each executed swap transaction implies a trade route with the lowest price impact, the highest liquidity, and the cheapest slippage. Since liquidity is an important feature of a medium of exchange, we use a token's betweenness centrality (denoted by $BetwCent$) to capture its dominance as a medium of exchange. Specifically, a weighted betweenness centrality of token$_k$ on day t is calculated as:

$$BetwCent_{k,t} = \frac{\sum\limits_{i \neq j, i \neq k, j \neq k} A_{<i,j|k>,t}}{\sum\limits_{i \neq j, i \neq k, j \neq k} A_{<i,j>,t}}, \tag{3.2}$$

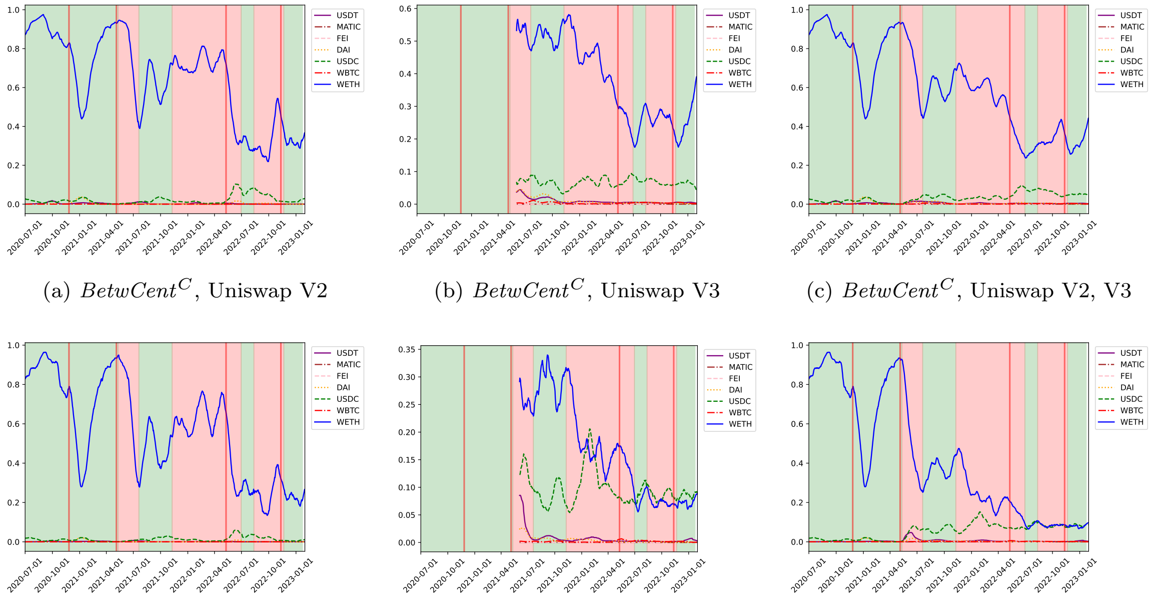
where $A_{<i,j>,t}$ represents all the transactions on day t where token$_i$ is the ultimate source token, and token$_j$ the ultimate target token; $A_{<i,j|k>,t}$ represents the subset of the abovementioned transactions where token$_k$ is an in-between node in the trade route. Specifically, when A represents the number of all the transactions, the betweenness centrality is count-weighted (or equal-weighted) denoted by $BetwCent^C_{i,t}$; when A represents transaction volume, the betweenness centrality is volume-weighted (or value-weighted), denoted by $BetwCent^V_{i,t}$. Since one transaction can contain multiple atomic trades (see Figure 3.1 for examples), we simply compute the transaction volume as the average of the volumes of the component trades.

The computation of betweenness centrality is based on the assumption that all the trading routes for the occurred transactions are the shortest (or the cheapest). The basic definition of the betweenness centrality by Brandes (2001) has to iterate the graph to find the shortest

[7] https://techcrunch.com/2021/03/29/visa-supports-transaction-settlement-with-usdc-stablecoin

[8] https://www.coindesk.com/markets/2021/05/13/tethers-first-reserve-breakdown-shows-token-49-backed-by-unspecified-commercial-paper/

path. By comparison, here the shortest route is given by the Uniswap router. Hence, we can compute the betweenness centrality directly by counting trading routes or accumulating trading volume along these routes. In addition, parallel edges between nodes are common in the graph that each edge represents a single transaction between two tokens which has the attribute of trading volume (in the case of value-weighted), unlike the unique relationship in the social networks. Note that betweenness centrality captures excessive use of a certain cryptocurrency since it is neither source or target of the trade. Furthermore, betweenness centrality is a measure of active liquidity provision of a cryptocurrency since the measure is computed based on realized trades.



(a) $BetwCent^C$, Uniswap V2

(b) $BetwCent^C$, Uniswap V3

(c) $BetwCent^C$, Uniswap V2, V3

(d) $BetwCent^V$, Uniswap V2 only

(e) $BetwCent^V$, Uniswap V3 only

(f) $BetwCent^V$, Uniswap V2, V3

Figure 3.6: Betweenness centrality of key coins. These figures plot the 30-day moving average of betweenness centrality for seven key cryptocurrencies in only Uniswap V2, only Uniswap V3, and both Uniswap V2 and V3; the green span denotes the boom period, while the red span denotes the bust period.

Figure 3.6 show that WETH has the highest betweenness centrality in V2 and V3 in terms of trade counts, indicating WETH is the preferred medium of exchange for small value transactions. When the trades are value weighted, we observe that USDC arises in betweenness centrality from November 2021 and overtakes WETH briefly early 2022, implying that USDC might be having less price impact for large value transactions during this period. However, value weighted betweenness centrality measures for both WETH and USDC drop after the Luna/Terra event, which implies, more trades were direct swap rather than intermediated by either WETH or USDC.

Table 3.2: Explaining trading volume with betweenness and eigenvector centralities

This table reports the decomposition of the contribution of *BetwCent* and *EignCent* to total trade volume by regression *VShare* on *BetwCent* and *EignCent*.

	(1)	(2)	(3)	(4)
Dependent Var	*VShare*	*VShare*	*VShare*	*VShare*
$BetwCent^C$	0.6724^{***}	0.2563^{***}		
	(0.0008)	(0.0003)		
AvgEigenCent		0.3970^{***}		0.4261^{***}
		(0.0002)		(0.0002)
$BetwCent^V$			0.7557^{***}	0.2581^{***}
			(0.0012)	(0.0005)
N	312,832	312,832	312,832	312,832
R^2	0.674	0.970	0.565	0.958

3.5.4 Relationship between Betweenness Centrality, Eigenvector Centrality and Volume Share

Dominance metrics *BetwCent* and *EignCent* capture the importance of a cryptocurrency in the trading network in different ways. Trades depicted in Figure 3.1b and Figure 3.1c increase both the betweenness centrality and eigenvector centrality of the cryptocurrencies between the source and the target cryptocurrency (that is, coin B in these graphs). Coin B is involved in these trades as a vehicle currency due to its liquidity and low price impact. By contrast, trades where a cryptocurrency is either the source or the target – such as direct swaps pictured in Figure 3.1a or in a chain – do not affect betweenness centrality of either the source (A) or the target coin (B) but increase the eigenvector centrality of both. These are the trades where investors either trade to hold the target coin or sell the source coin in their inventory. Investors trade these coins for investment or speculative purposes rather than for liquidity reasons. These investment or speculative purposes might relate to a currency's role being a unit of account or store of value. Therefore, the eigenvector centrality of a cryptocurrency reveals these additional functions of a coin beyond the medium of exchange captured in its betweeness centrality.

To highlight these differences in the dominance, we decompose the contribution of *BetwCent* and *EignCent* to total trade volume by regression *VShare* on *BetwCent* and *EignCent*. The results are reported in 3.2. We observe that liquidity dominance contributes 56.5% to 67.4% of the variation in trading volume share while eigenvector centrality contributes about 30%.

3.5.5 Store of value measurement

Typically safe assets offer the service of store of value. In the DeFi setup, there are multiple types of stablecoins that aim to offer this service. We measure the dominance of this service in several ways. One is the simple classification of stablecoin or not. The other is the market capitalization share of each stablecoin among the stablecoins in the sample reflecting the popularity of each stablecoin, which we term *StableShare*.

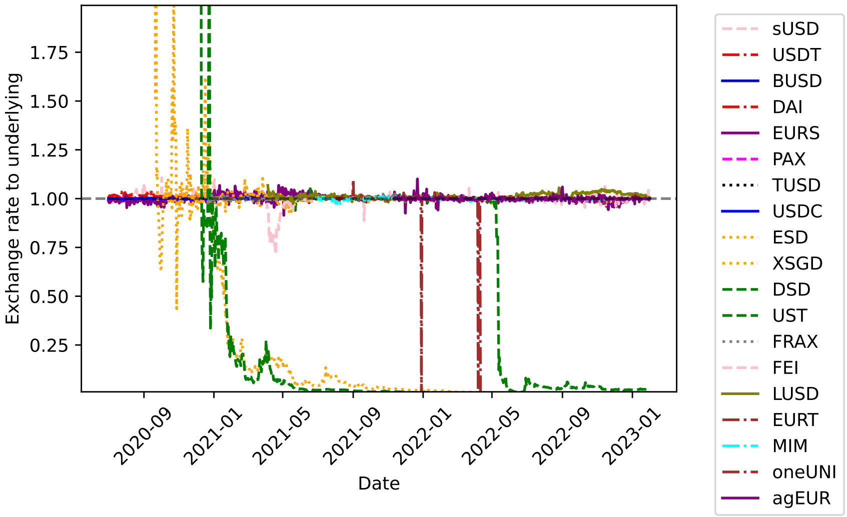


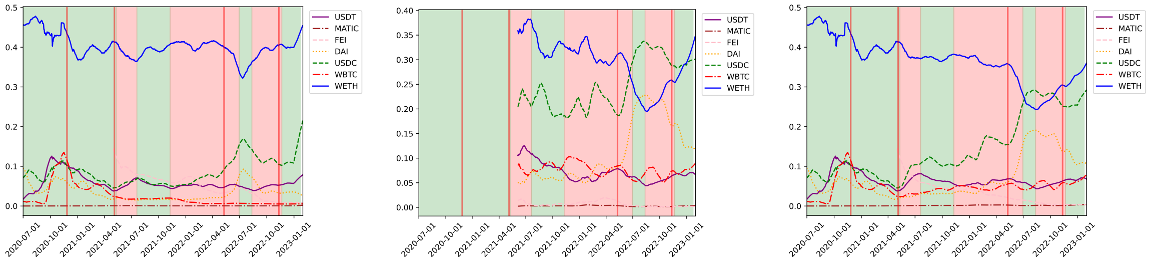
Figure 3.7: Stablecoin exchange rate to their underlying This figure plots the deviations from the peg for all the stablecoins.

Figure 3.8: Comparison of Stablecoin Market Dynamics and Peg Stability

Interestingly, not all stablecoins are successful in maintaining the peg to dollar. Figure 3.7 graphs daily deviations from the peg for all the stablecoins in our sample. It is clear from Figure 3.7 that some stablecoins are stable only in name.

3.5.6 Liquidity share

Figure 3.9 illustrates the share of liquidity provision of key cryptocurrencies in the Uniswap market. We define liquidity provision share as the share of the total value locked at Uniswap liquidity pools, denoted as *LiquidityShare*. This measure represents passive liquidity provision in the Uniswap market and potentially proxies for the status of a crypto asset both as a medium of exchange and as a store of value. It can be the former because larger the total value locked, the lower is the potential price impact of a trade; it can be the latter because the liquidity providers who stake a coin in its liquidity pools passively can expect lower impermanent loss relative to the price appreciation of the coin. Similarly as *VShare*, *LiquidityShare* might reflect the liquidity provision in proportion to the size of the ecosystem generated by each cryptocurrency and may not be indicative of any excessive of usage for liquidity provision.



(a) *LiquidityShare* Uniswap V2

(b) *LiqShare* Uniswap V3

(c) *LiquidityShare* Uniswap V2, V3

Figure 3.9: 30-day Moving average of liquidity provision on Uniswap of key cryptocurrencies. These figures plot the 30-day moving average of liquidity share for seven key cryptocurrencies in only Uniswap V2, only Uniswap V3, and both Uniswap V2 and V3; the green span denotes the boom period, while the red span denotes the bust period.

The general time series pattern of *LiquidityShare* dominance is similar to that of $VShare$. Wrapped ETH's initial dominance was taken over by stablecoins, in particular USDC after the Luna/Terra episode in early 2022. Slightly different from the $Vshare$ dominant metrics, we observe that in Figure 3.9c where we aggregate both Uniswap V2 and V3 liquidity provisions that there are three cryptocurrencies battling for dominance in passive liquidity provision: WETH, USDC and DAI, especially in the latter part of the sample.

Table 3.3 describes the summary statistics for all the above dominance metrics in the sample. It shows that for all dominance metrics the standard deviation is much larger relative to its mean and the distribution is highly skewed.

Table 3.3: Summary statistics

This table reports the summary statistics for the dominance metrics in the sample. The metrics include *VShare*, $VShare^{In}$, $VShare^{Out}$, *LiquidityShare*, $EigenCent^{In}$, $EigenCent^{Out}$, $BetwCent^{C}$, and $BetwCent^{V}$. The table reports the number of observations, the mean, the standard deviation, the minimum, the 25th percentile, the 50th percentile (median), the 75th percentile, and the maximum for each metric. The table highlights that the standard deviation is much larger relative to the mean for all metrics, indicating high variability in the data. Additionally, the distribution for all metrics is highly skewed.

	Obs	mean	std	min	25%	50%	75%	max
VShare	58511.0	1.637299e-02	0.066159	0.000000e+00	0.000254	0.000921	0.003284	5.000000e-01
$VShare^{In}$	58436.0	1.637689e-02	0.066194	0.000000e+00	0.000238	0.000905	0.003283	6.060238e-01
$VShare^{Out}$	58436.0	1.637689e-02	0.066381	0.000000e+00	0.000256	0.000932	0.003279	7.352508e-01
LiquidityShare	58436.0	1.637689e-02	0.059630	0.000000e+00	0.000456	0.001219	0.003593	5.000000e-01
$EigenCent^{In}$	58511.0	3.784705e-02	0.122233	-1.665335e-16	0.000568	0.002517	0.011358	9.365245e-01
$EigenCent^{Out}$	58511.0	3.778501e-02	0.122252	-1.665335e-16	0.000616	0.002565	0.011342	8.835949e-01
$BetwCent^{C}$	44852.0	1.223823e-02	0.091905	0.000000e+00	0.000000	0.000000	0.000000	9.878716e-01
$BetwCent^{V}$	44852.0	9.613459e-03	0.076991	0.000000e+00	0.000000	0.000000	0.000000	9.939918e-01

3.5.7 Dominance Concentration

To gauge the number of dominant currencies, we examine dominance concentration and in particular Herfindahl index of dominance measures across 50 cryptocurrency at each day. The inverse of the Herfindahl index can be regarded as the daily “effective” number of currencies. Except for the oracle price period, the overall high level of Herfindahl index indicates the existence of a few dominant currencies: as high as five (in January 2021—the earlier period of the sample) and as low as two (in August 2022—the later period of the sample).

We also compute the Herfindahl index for all other dominance measures except for eigenvector centralities[9]. The Herfindahl index for passive liquidity provision (TVL) shows that passive liquidity provision is less concentrated over time—suggesting that price impact is more evenly distributed in the latter part of the sample. The Herfindahl indices for both equally and value weighted betweenness centrality – the active liquidity dominance measure also show less concentration over time, albeit with volatile movements, indicating frequent regime changes.

3.6 Properties of Dominance

3.6.1 Lead and Lag Relationships

In this section, we examine how these dominance metrics relate to each other and how the relationships vary over the boom and bust cycle.

We first examine the lead-lag relationship of the two of our dominance metrics, betweenness centrality and eigenvector centrality by running a vector auto-regression of the two series. The estimation results are reported in Table 3.4. These results show clearly that betweenness centrality leads network centrality and this lead relationship comes from the boom but does not exist during the bust. This finding suggests that liquidity is a valuable attribute of dominant currencies during booms. Vehicle currencies captured by betweenness dominance are also the preferred investment/speculative currencies during booms.

To understand the contemporaneous relationship between all of the dominance metrics, we also compute the correlation matrix of dominance measures, including: $VShare$, $EigenCen^{Avg}$, $BetwCent^{V}$, $BetwCent^{C}$, $Stableshare$, and other coin-level characteristics such as volatility σ^{USD} and correlation with WETH and S&P Market crypto index, as well as associated lending and borrowing statistics in the lending platforms etc, which is presented in Figure 3.10. The heatmap shows that most dominance metrics are closely related and their correlation with risk variables are low.

[9] The Herfindahl index for the eigenvector centrality by definition is scaled to 1.

Table 3.4: Panel vector autoregression results with centrality measures

This table reports the estimation results of the vector auto-regression of two dominance metrics, betweenness centrality and eigenvector centrality. The table consists of three separate panels, and each panel reports four columns of results. The dependent variables are either ultimate eigenvector centrality or betweenness centrality. The results indicate that betweenness centrality leads network centrality during the boom but not during the bust. This finding suggests that liquidity is a valuable attribute of dominant currencies during booms. The preferred investment/speculative currencies during booms are captured by betweenness dominance. The fixed effects and time effects are included in the estimation, and the R^2 values are reported in the last row of each panel.

Panel 1: full sample

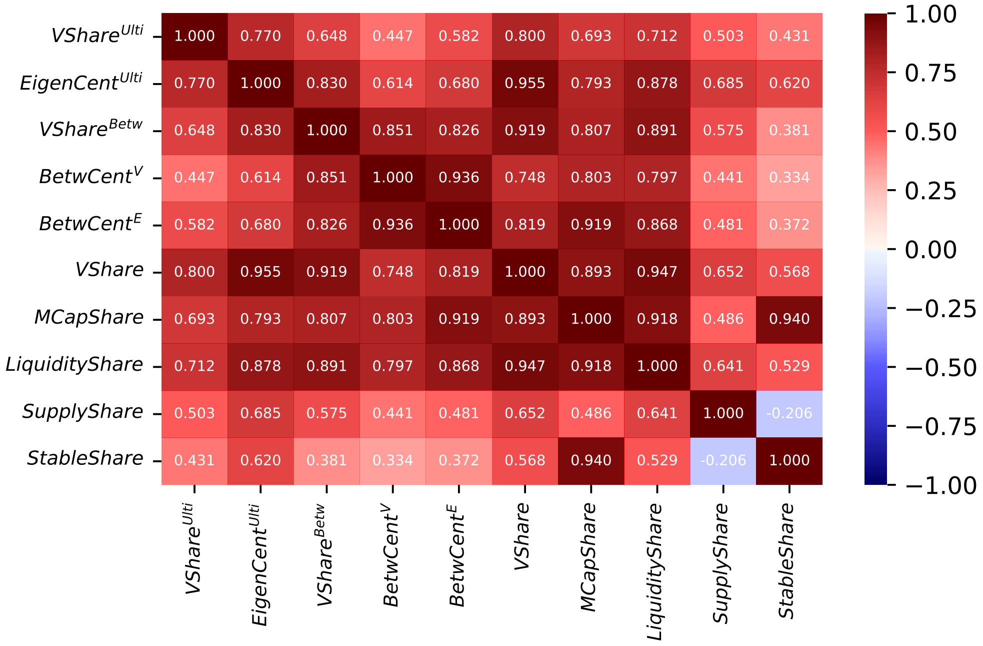
	(1)	(2)	(3)	(4)
Dependent Var	$EigenCent^{Ulti}$	$BetwCent^{V}$	$EigenCent^{Ulti}$	$BetwCent^{E}$
${EigenCent^{Ulti}}_{t-1}$	0.8430***	−0.0012*	0.8431***	−0.0003
	(0.0152)	(0.0007)	(0.0152)	(0.0005)
$BetwCent^{V}_{t-1}$	0.0080***	0.9741***		
	(0.0015)	(0.0044)		
$BetwCent^{E}_{t-1}$			0.0102***	0.9702***
			(0.0022)	(0.0054)
Fixed Effect	yes	yes	yes	yes
Time Effect	no	no	no	no
N	270,450	270,450	270,450	270,450
R^2	0.716	0.951	0.716	0.943

Panel 2: boom

	(1)	(2)	(3)	(4)
Dependent Var	$EigenCent^{Ulti}$	$BetwCent^{V}$	$EigenCent^{Ulti}$	$BetwCent^{E}$
${EigenCent^{Ulti}}_{t-1}$	0.7203***	−0.0054	0.7210***	−0.0056***
	(0.0274)	(0.0043)	(0.0273)	(0.0020)
$BetwCent^{V}_{t-1}$	0.0124***	0.8770***		
	(0.0043)	(0.0187)		
$BetwCent^{E}_{t-1}$			0.0094**	0.9499***
			(0.0037)	(0.0130)
Fixed Effect	yes	yes	yes	yes
Time Effect	no	no	no	no
N	143,520	143,520	143,520	143,520
R^2	0.557	0.808	0.557	0.910

Panel 3: bust

	(1)	(2)	(3)	(4)
Dependent Var	$EigenCent^{Ulti}$	$BetwCent^{V}$	$EigenCent^{Ulti}$	$BetwCent^{E}$
${EigenCent^{Ulti}}_{t-1}$	0.7203***	−0.0054	0.7210***	−0.0056***
	(0.0274)	(0.0043)	(0.0273)	(0.0020)
$BetwCent^{V}_{t-1}$	0.0124***	0.8770***		
	(0.0043)	(0.0187)		
$BetwCent^{E}_{t-1}$			0.0094**	0.9499***
			(0.0037)	(0.0130)
Fixed Effect	yes	yes	yes	yes
Time Effect	no	no	no	no
N	143,520	143,520	143,520	143,520
R^2	0.557	0.808	0.557	0.910



	$VShare^{Ulti}$	$EigenCent^{Ulti}$	$VShare^{Betw}$	$BetwCent^{V}$	$BetwCent^{E}$	VShare	MCapShare	LiquidityShare	SupplyShare	StableShare
$VShare^{Ulti}$	1.000	0.770	0.648	0.447	0.582	0.800	0.693	0.712	0.503	0.431
$EigenCent^{Ulti}$	0.770	1.000	0.830	0.614	0.680	0.955	0.793	0.878	0.685	0.620
$VShare^{Betw}$	0.648	0.830	1.000	0.851	0.826	0.919	0.807	0.891	0.575	0.381
$BetwCent^{V}$	0.447	0.614	0.851	1.000	0.936	0.748	0.803	0.797	0.441	0.334
$BetwCent^{E}$	0.582	0.680	0.826	0.936	1.000	0.819	0.919	0.868	0.481	0.372
VShare	0.800	0.955	0.919	0.748	0.819	1.000	0.893	0.947	0.652	0.568
MCapShare	0.693	0.793	0.807	0.803	0.919	0.893	1.000	0.918	0.486	0.940
LiquidityShare	0.712	0.878	0.891	0.797	0.868	0.947	0.918	1.000	0.641	0.529
SupplyShare	0.503	0.685	0.575	0.441	0.481	0.652	0.486	0.641	1.000	-0.206
StableShare	0.431	0.620	0.381	0.334	0.372	0.568	0.940	0.529	-0.206	1.000

Figure 3.10: Correlation matrix of key variables. This figure plots the correlation matrix of $VShare$, $EigenCen^{Avg}$, $BetwCent^{V}$, $BetwCent^{C}$, $Stableshare$, and other coin-level characteristics such as volatility σ^{USD} and correlation with WETH and S&P Market crypto index, as well as associated lending and borrowing statistics in the lending platforms etc.

.

To gauge the lead-lag relationship among all the dominance metrics, we then compute (cross-)auto-correlation of these variables with 7 lags, 14 lags, 21 lags, 28 lags following the methodology in Campbell et al. (1997). We examine whether the entries in the lower diagonal are larger than the counterparts in the upper diagonal of the cross-autocorrelation matrix. Over the full sample, we observe large auto-correlations among all dominance indices and we do not observe significant differences between these entries, indicating that there are no significant lead/lag relationships for these dominance metrics.

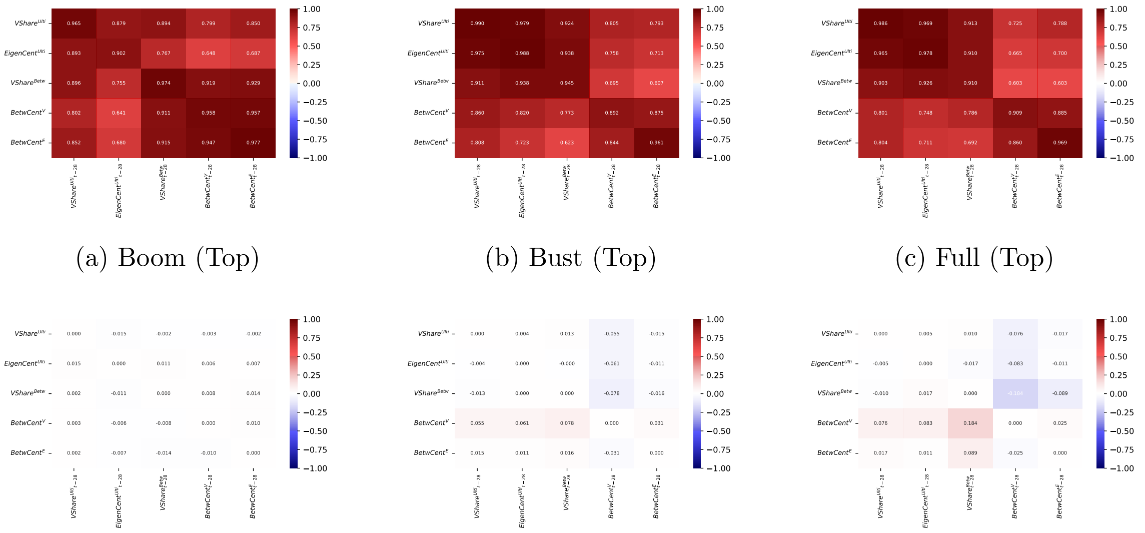
However, when we examine the boom and bust periods separately, we find some interesting differences between the lower and upper diagonal entries. The cross-autocorrelation matrix for 28 day lags for the boom subsample is shown in Figure 3.11d. We find that the correlations of lagged equal-weighted betweenness centrality $BetwCent^C$ (an active liquidity dominance measure) with all other metrics are larger than the corresponding correlation between $BetwCent^C$ and the lagged value of other metrics. This pattern is followed by the value-weighted $BetwCent^V$. Therefore, this indicates that during the boom, active liquidity provision is a main driver of currency dominance.

By contrast, Figure 3.11e shows a different pattern during the bust. First of all, we find that the store of value dominance measure ($StableShare$) leads (both passive and active) liquidity dominance metrics, indicating safety is preferred over liquidity in terms of currency dominance attributes. Interestingly, both market share dominance ($VShare$ and network centrality ($EigenCent$) lead $StableShare$, indicating that market demand or supply (in terms of market share and network centrality) affects how safety is valued during the bust. In summary, we find some evidence that during the market crash, cryptocurrency that is regarded safe moves up in liquidity dominance, i.e. used more to intermediate transactions; while during the market boom, cryptocurrencies with a high score in terms of active liquidity provision gain higher network centrality and larger trading volume shares, that is, liquidity is a key driver for currency dominance in a boom. This is consistent with the result in Gourinchas and Rey (2022) where they argue that investors' risk aversion leads some to prefer safe assets for settlement.

3.6.2 Time-Varying Market Concentration: Determinants

Next, we investigate how the Herfindahl index of relevant dominance metrics (or the inverse of the number of effective currencies) relates to market conditions by running the following market-wide regressions:

$$\text{Herfindahl}_t \sim IsBoom_t + GasPrice + \ln MarketVolume + \sigma_{\text{SP}}^{USD} + \sigma_{\text{gas}}^{USD} + \text{Herfindahl}_{t-1} \qquad (3.3)$$



(a) Boom (Top) (b) Bust (Top) (c) Full (Top)

(d) Boom (Bottom) (e) Bust (Bottom) (f) Full (Bottom)

Figure 3.11: Cross-autocorrelation with 28 days lag. These figures plot the cross-autocorrelation with 28 days lag over the boom, bust, and full period.

where *IsBoom* is a dummy for the boom period, $\sigma_{\mathrm{SP}}^{USD}$ is the volatility of the crypto market index, and the naming of the rest of the variables is self-explanatory. All independent variables are lagged.

The results are presented in Table 3.5. We observe that both the active and the passive liquidity provisions are more concentrated during the boom or when gas price or gas volatility is high, indicating that fewer currencies serve as vehicle currencies and hence are dominant under these market conditions. We find that higher gas price or gas volatility lead to a less concentrated trading volume share, indicating that there might be more direct asset swaps rather than cross-pool tradings. Hence, there is less use of vehicle currencies to minimize the transaction costs under these circumstances.

3.7 Drivers of Dominant Currencies

3.7.1 Testable Hypotheses

In this section, we develop testable hypotheses based on the findings from the existing financial, international monetary, and macroeconomic literature and construct specifications for these tests.

The DCP literature has identified at least four sets of factors that might affect a currency's dominant status. The first group relates to price stability. The second group concerns

Table 3.5: Herfindahl and Market Condition

The table shows the results of market-wide regressions investigating the relationship between Herfindahl index of relevant dominance metrics and market conditions. The dependent variables in each column represent different metrics of market concentration. The independent variables include the dummy for the boom period, gas price, natural logarithm of market volume, USD volatility of the crypto market index, and the previous period's Herfindahl index. The results are reported in terms of the estimated coefficients, and their standard errors are shown in parentheses. The number of observations is the same across all columns. The table suggests that both active and passive liquidity provisions are more concentrated during the boom or when gas price or gas volatility is high, indicating that fewer currencies serve as vehicle currencies and hence dominant under these market conditions. Higher gas price or gas volatility leads to less concentrated trading volume share, indicating that there might be more direct asset swaps rather than cross-pool tradings. Hence there is less use of vehicle currencies to minimize the transaction cost under these circumstances.

	(1)	(2)	(3)	(4)	(5)
Dependent Var	$HHI_{VShare^{Ulti}}$	$HHI_{BetwCent^{E}}$	$HHI_{BetwCent^{V}}$	$HHI_{VShare^{Betw}}$	HHI_{VShare}
HHI_{t-1}	0.4546***	0.6299***	0.6447***	0.5648***	0.4449***
	(0.0465)	(0.0381)	(0.0460)	(0.0525)	(0.0481)
$IsBoom$	−0.0018	0.0345***	0.0428***	0.0420***	0.0036
	(0.0026)	(0.0093)	(0.0105)	(0.0101)	(0.0023)
$MarketVolume$	0.0018	−0.0005	−0.0071	0.0021	0.0018
	(0.0020)	(0.0056)	(0.0063)	(0.0067)	(0.0018)
σ_{SP}^{USD}	0.1267	0.3516	−0.6414**	−0.6178	0.0185
	(0.0995)	(0.2920)	(0.3184)	(0.4076)	(0.0893)
$GasPrice^{USD}$	−16.6760***	46.9715**	49.7317*	−37.0482	−13.1549**
	(6.0647)	(20.8932)	(27.7102)	(28.3063)	(5.1528)
σ_{Gas}^{USD}	−0.0344***	0.0823***	0.0909***	0.0871***	−0.0140*
	(0.0115)	(0.0273)	(0.0320)	(0.0330)	(0.0085)
$AvgClustCoef$	0.0210	0.0485	0.0826	0.1033	0.0272
	(0.0228)	(0.0760)	(0.0754)	(0.0838)	(0.0198)
$NumClique/NumTxn$	0.0000	−0.0603***	−0.0525***	−0.0282*	−0.0044
	(0.0045)	(0.0158)	(0.0166)	(0.0159)	(0.0037)
Year-Month Dummies	yes	yes	yes	yes	yes
N	944	944	944	944	944
R^2	0.885	0.960	0.961	0.931	0.782

safety characteristics. The third group relates to financial product offerings, especially money market products such as deposits. The final group concerns characteristics of the currency related to the role of unit of account: the size of the market associated with each currency and the ability of the currency to hedge working capital needs.

The first group of factors are closely linked to the role of money as a store of value. Intuitively, if the real purchasing power of a cryptocurrency is stable, it is more likely to be used as a store of value. Indeed, in the trade literature, this (in)stability is captured by the bilateral exchange rate volatility between the invoicing currency for exports and the currency used to purchase consumption goods (which can be either the local currency or the currency used to pay for imported consumption goods). Since real consumption is denominated in dollars, a measure of purchasing power volatility is the volatility of the cryptocurrency in dollars. This leads to the following testable hypothesis.

Hypothesis 1 (Dollar volatility)**.** A cryptocurrency is more likely to be used for transactions when the volatility of its exchange rate against dollar is lower.

Another variation of Hypothesis 1 is that a cryptocurrency is viewed safe if a cryptocurrency can hedge again the crash risk. We use return correlation with crypto market index or ETH to capture the market risk embedded in each cryptocurrency.

Hypothesis 1a (Crash Hedge)**.** A cryptocurrency is more likely to be used for transactions if its return is negatively correlated with crypto market return during downturns.

In the trade literature, Gopinath and Stein (2021) have pinpointed that the special roles played by the safe asset—the currency that is regarded as "stable" in maintaining its real value and hence used as the invoice asset. This indicates that, in addition to lower volatility, traders might prefer the price stability of stablecoins when choosing it to invoice/settle a transaction, which leads to the second group of hypothesis.

Hypothesis 2 (Stablecoins)**.** Stablecoins are more likely to be used for transactions, especially during the volatile periods.

However, not all stablecoins are created equal. We use *StableShare* to capture the differential impacts among stablecoins.

Hypothesis 2a (Stablecoin Market Share)**.** Stablecoins that have larger market shares are more likely to be used for transactions, especially during the volatile periods.

The third group of factors relate to financial service offerings associated with a currency. Maggiori (2017), Gourinchas et al. (2019) and Gopinath and Stein (2021) suggest that exporting and importing firms have an incentive to invoice/transact in currencies associated with higher levels of financial services. If there is a well-developed debt/credit/derivatives market denominated in a certain currency, then trade finance, working capital finance, banking services and currency risk hedging in that particular currency would be cheaper and readily available. Financial services applications such as deposit, lending, investing, hedging, insurance, etc. in the crypto universe have only been developed in the last three years as blockchain developers have begun to tap into smart contract functionalities. These applications are characterized as decentralized finance (DeFi) as they are algorithmically based and without any centralized authority. We use *SupplyShare*, the share of crypto assets deposited in the DeFi lending protocol, as a proxy for the level of money market services— a most basic level of financial services—associated with each cryptocurrency. This leads to the following hypothesis.

Hypothesis 3 (Financial Service)**.** A currency with a larger money market is more likely to be used for trade transactions.

Finally, the literature shows that firms are more concerned about the currency mismatch between export sales denominated in the invoicing currency and working capital denominated in the local currency, than about the purchasing power of export sales for consumption goods directly. This leads to the fourth group of factors. In these cases, the bilateral exchange rate volatility with the currency used to pay for working capital may be an important consideration for firms to choose the currency for export invoices. In the blockchain environment, working capital is denominated in the utility coin, commonly the chain's native currency, associated with a specific chain. On the Ethereum blockchain—the world's largest DeFi network, each transaction is subject to a gas fee (akin to working capital) priced in ETH. This means that investors might also be concerned with the variation of gas fee. They would like to denominate their trade in a cryptocurrency which hedges this risk.

Hypothesis 4 (Gas Hedge)**.** A cryptocurrency is more likely to be used for transactions when its return covariance with gas price change, is lower.

Lastly, Mukhin (2022) find that the size of the market might explain why certain currencies are dominant. In the crypto setting, the size of a market associated with each cryptocurrency can be proxied by the market capitalization. For example, Bitcoin, one of the oldest and the most widely held crypto assets, has a market capitalization of $411.41 billion as of August 2022, double that of Ethereum in the second place and five times that of USDC, a stablecoin in the third place. This leads to the following hypothesis.

Hypothesis 5 (Market Size)**.** A cryptocurrency with a large market capitalization is more likely to be used for transactions.

In the next section, we develop specifications to test these hypotheses utilizing both cross-sectional and time-series variations.

3.7.2 Regression Specifications and Results

We now examine how each groups of factors identified in the hypotheses section explain each of dominant metrics (*AvgEigenCent*, *BetwCent*, *VShare*) by running the following fixed-effect regression for each coin at each day in our sample period:

$$\begin{aligned}\text{Dominance}_{i,t} \sim & \mathit{Dominance}_{i,t-1} + \mathit{CorETH}_{i,t-1} + \sigma^{USD}_{i,t-1} + \mathit{StableShare}_{i,t-1} \\ & + \mathit{CorGas}_{t-1} + \mathit{MCapShare}_{i,t-1} + \mathit{SupplyShare}_{i,t-1}\end{aligned} \tag{3.4}$$

We run the regression with the lagged dominance proxy to control for autocorrelation and report the results in Table 3.6. We observe that a coin's market capitalization (that is, a proxy for the size of ecosystem around one coin) and its supply share (that is, the amount

Table 3.6: Dominance regression: full sample with boom interaction

This table presents the results of fixed-effect regressions for each coin at each day in the sample period to examine how groups of factors explain each of the dominant metrics. The table displays the coefficients of the regression model and their respective standard errors. The dependent variable in the first column is the volume share of a cryptocurrency, and the dependent variable in the second column is the equally weighted betweenness centrality of a cryptocurrency. The independent variables include the lagged dominance proxy to control for autocorrelation and factors such as market capitalization, supply share, stablecoin, and correlation with gas prices. The coefficients for market capitalization and supply share are positive and statistically significant across all dominance metrics, supporting Hypothesis 5 and Hypothesis 3, respectively. Stablecoin contributes positively to all dominance metrics except for equally weighted betweenness centrality, supporting Hypothesis 6. Correlation with gas prices contributes positively to average eigenvector centrality during the boom period, weakly supporting Hypothesis 4. The table also indicates the impact of the market boom on the coefficients of some factors.

	(1)	(2)
Dependent Var	$VShare^{ulti}$	$VShare^{betw}$
$Dominance_{t-1}$	0.8681*** (0.0128)	0.8704*** (0.0137)
σ^{USD}_{t-1}	−0.0000* (0.0000)	0.0000*** (0.0000)
σ^{USD}_{t-1} : *IsBoom*	0.0000*** (0.0000)	−0.0000*** (0.0000)
$MCapShare_{t-1}$	0.0113 (0.0089)	−0.0031 (0.0290)
$MCapShare_{t-1}$: *IsBoom*	−0.0003 (0.0047)	0.0828*** (0.0176)
$LiquidityShare_{t-1}$	0.0480*** (0.0073)	0.1099*** (0.0177)
$LiquidityShare_{t-1}$: *IsBoom*	−0.0012 (0.0085)	−0.0020 (0.0190)
$SupplyShare_{t-1}$	0.0195*** (0.0043)	0.0428*** (0.0107)
$SupplyShare_{t-1}$: *IsBoom*	−0.0069* (0.0038)	−0.0198** (0.0089)
$CorGas_{t-1}$	0.0000 (0.0000)	0.0001 (0.0001)
$CorGas_{t-1}$: *IsBoom*	0.0001 (0.0001)	−0.0000 (0.0002)
$StableShare_{t-1}$	0.0464*** (0.0073)	0.0951*** (0.0143)
$StableShare_{t-1}$: *IsBoom*	−0.0121*** (0.0025)	−0.0496*** (0.0077)
$CorETH_{t-1}$	−0.0000 (0.0000)	0.0003*** (0.0001)
$CorETH_{t-1}$: *IsBoom*	0.0001 (0.0001)	−0.0001 (0.0001)
Fixed Effect	yes	yes
Time Effect	no	no
N	273,331	273,331
R^2	0.855	0.886

of deposit) positively contribute to its dominance measures consistently , supporting Hypothesis 5 and Hypothesis 3, respectively. In fact, the former effect is even stronger during the market boom. The latter effect, in contrast, is not uniform across all dominance measures during the boom, even negative for *VShare* and *LiquidityShare* but stronger for $BetwnCent^C$, that is, an active money market benefits active liquidity provision but do not contribute to volume share or passive liquidity provision during the boom.

We find that *StableShare* contributes positively to all dominance metrics except for equally weighted betweenness centrality, indicating that stablecoins are less likely to intermediate small value trades. In fact, we observe that stablecoins are much less likely to be dominant during the boom, indicating that they play more important roles during the bust.

We also observe that the coefficients *CorGas* are positive for network centrality metric $AvgEigenCent$ during the boom period, but statistically insignificant for all other metrics. This finding weakly supports Hypothesis 4, showing that currencies used to hedge price volatility of working capital asset are likely to be used as dominant currencies during the boom. We do not find strong evidence that either idiosyncratic volatility or market risk affect any dominant metrics, which may reflect the volatile nature of this market.

Finally, to examine the importance of money market to the currency dominance, we design a difference-in-differences test by utilizing the fact that PLFs such as Compound accept crypto-assets as deposits/collaterals only occasionally due to the decentralized nature of governance. The treatment group consists of coins accepted to the PLF and the control group is made of coins that have already been accepted prior to the beginning of the observation window. We choose the control group to deal with the endogeneity concern that PLFs might choose to accept tokens based on their predicted dominance. Our identification strategy exploits ex-ante differences in being accepted by a PLF across tokens, but it does not require the initial presence at a PLF to be random. It only requires that outcomes of treated and control tokens would have evolved similarly absent the treatment. We compare whether there are any significant differences in the change of dominance metrics after the treatment date for the coins that are treated with those in the control group. We have conducted this analysis for 14, 30 and 60 days before-and-after windows and reported the 60-day results in Table 3.7 since estimation outcomes are similar. In Table 3.7, $IsTreatedToken$ is a dummy for the treated coin(s), $AfterTreatedDate$ is a dummy for the PLF inclusion date, and the last regressor is the interaction between the two dummies. We include the token fixed effect in the difference-in-differences regressions and hence the dummy $IsTreatedToken$ is absorbed by the fixed effects. We find that the coefficients for the interaction variable is positive for all dependent variables except for *LiquidityShare* and statistically significant when the dependent variable is eigenvector and value-weight betweenness centrality. These results suggest that being included in the PLF (that is, having a money market) is an important driver for currency dominance.

Table 3.7: Difference-in-differences regression in money market

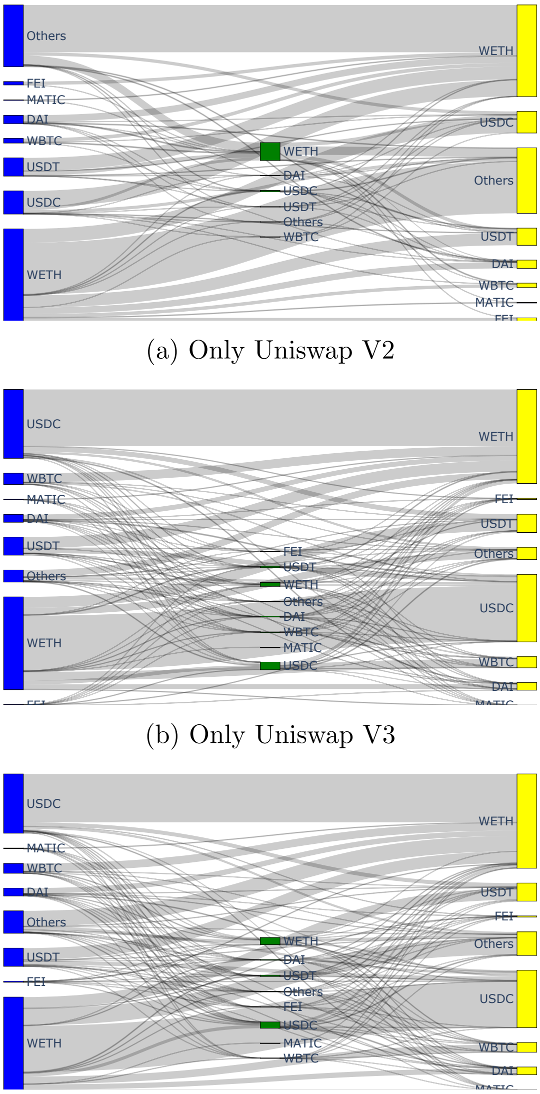
This table reports the results of the difference-in-differences test to examine the impact of a money market on currency dominance. The treatment group consists of tokens accepted to the PLF, while the control group comprises coins that had been accepted before the observation window began. The dependent variables include VShare^{Ulti}, $\mathit{EigenCent}^{Ulti}$, VShare^{Betw}, $\mathit{BetwCent}^{V}$, $\mathit{BetwCent}^{E}$, *EigenCent*, and *VShare*. The table shows the estimates of the coefficients for the interaction between *IsTreatedToken* and *AfterTreatedDate*, which indicates the treatment effect. The regressions control for *MCapShare*, *StableShare*, and σ^{USD}, and include token fixed effects. The results demonstrate that being included in the PLF (having access to a money market) is an important driver for currency dominance, as indicated by the significant and positive coefficients for the interaction variable for all dependent variables, except for *LiquidityShare*.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Var	VShare^{Ulti}	$\mathit{EigenCent}^{Ulti}$	VShare^{Betw}	$\mathit{BetwCent}^{V}$	$\mathit{BetwCent}^{E}$	*EigenCent*	*VShare*
Treatment : Post	0.0035^{***}	0.0149^{***}	-0.0013^{***}	-0.0001^{***}	-0.0001^{***}	0.0117^{***}	0.0027^{***}
	(0.0004)	(0.0016)	(0.0002)	(0.0000)	(0.0000)	(0.0013)	(0.0003)
MCapShare	0.6586^{***}	2.6941^{***}	0.0477	0.0040	0.0029^{**}	2.2211^{***}	0.5207^{***}
	(0.0473)	(0.1991)	(0.0313)	(0.0025)	(0.0012)	(0.1656)	(0.0381)
StableShare	−0.1914	−0.9332	0.4176^{***}	0.0330^{***}	0.0095^{**}	−0.5940	−0.0621
	(0.1713)	(0.7217)	(0.1135)	(0.0091)	(0.0043)	(0.6001)	(0.1380)
σ^{USD}	−0.0001	−0.0003	0.0000	0.0000	0.0000	-0.0003^{*}	−0.0000
	(0.0000)	(0.0002)	(0.0000)	(0.0000)	(0.0000)	(0.0002)	(0.0000)
Fixed Effect	yes	yes	yes	yes	yes	yes	yes
Time Effect	yes	yes	yes	yes	yes	yes	yes
N	5,351	5,351	5,351	5,351	5,351	5,351	5,351
R^2	0.055	0.053	0.008	0.008	0.013	0.051	0.053

It is intuitive that we do not find any positive or statistically significant results for liquidity share. In fact, *LiquidityShare* is an imperfect substitute for PLF since it represents passive liquidity provision and is an alternative way to deposit crypto-assets but without any associated borrowing or lending facilities. It is a less effective money market instrument since depositors to the liquidity pools share volatile trading fees rather than interest payments.

3.8 Conclusion

In this paper, we examine the properties of dominant currencies in the fast changing DeFi world to test the theories of dominant currency paradigm (DCP) in the international trade and finance literature. We find that unlike in the traditional finance, there are multiple dominant currencies in DeFi, including utility coins such as WETH and stablecoins such as USDC and DAI. We find that during the boom, liquid coins are more dominant while during market turmoils, stablecoins play a more dominant role. We find that an active money market is an important driver for currency dominance, suggesting an essential design choice of future CBDCs.



(a) Only Uniswap V2

(b) Only Uniswap V3

(c) Both Uniswap V2 and V3

Figure 3.12: Sankey plots for inflow, outflow, and intermediary volume of Uniswap V2, V3, and both V2 and V3 These figures plot the inflow, outflow, and intermediary volume of only Uniswap V2, only Uniswap V3, and both Uniswap V2 and V3. The blue rectangle denotes the source token, the green rectangle denotes the intermediary token, and the yellow rectangle denotes the target token. The thickness of bands denote the magnitude of volume.

3.9 Appendix

Table 3.8: Tokens acceptable by Compound and their time added to the protocol

The table shows the lending pool smart contract addresses and the time added to Compound in UTC for 13 different tokens: ETH, USDC, USDT, WBTC, DAI, UNI, SAI, REP, MKR, YFI, USDP, ZRX, and SUSHI. The lending pool smart contract is the address where the token is stored and from which it can be lent and borrowed. The time added to Compound is the date and time when the token was added to the Compound protocol, which is a decentralized finance (DeFi) platform that allows users to lend and borrow cryptocurrencies.

Token	Lending pool smart contract	Time added to Compound [UTC]
ETH	`0x4ddc2d193948926d02f9b1fe9e1daa0718270ed5`	2019-05-07 01:25:18
USDC	`0x39aa39c021dfbae8fac545936693ac917d5e7563`	2019-05-07 01:25:31
USDT	`0xf650c3d88d12db855b8bf7d11be6c55a4e07dcc9`	2020-04-15 21:13:06
WBTC	`0xc11b1268c1a384e55c48c2391d8d480264a3a7f4`	2019-07-16 19:47:37
DAI	`0x5d3a536e4d6dbd6114cc1ead35777bab948e3643`	2019-11-23 01:03:33
UNI	`0x35a18000230da775cac24873d00ff85bccded550`	2020-09-23 22:05:47
SAI	`0xf5dce57282a584d2746faf1593d3121fcac444dc`	2019-05-07 01:24:12
REP	`0x158079ee67fce2f58472a96584a73c7ab9ac95c1`	2019-05-07 01:24:48
MKR	`0x95b4ef2869ebd94beb4eee400a99824bf5dc325b`	2021-07-16 05:30:17
YFI	`0x80a2ae356fc9ef4305676f7a3e2ed04e12c33946`	2021-07-18 03:19:05
USDP	`0x041171993284df560249b57358f931d9eb7b925d`	2021-09-19 19:42:57
ZRX	`0xb3319f5d18bc0d84dd1b4825dcde5d5f7266d407`	2019-05-07 01:20:54
SUSHI	`0x4b0181102a0112a2ef11abee5563bb4a3176c9d7`	2021-07-18 03:12:59
FEI	`0x7713dd9ca933848f6819f38b8352d9a15ea73f67`	2021-09-15 02:26:35
BAT	`0x6c8c6b02e7b2be14d4fa6022dfd6d75921d90e4e`	2019-05-07 01:21:25
COMP	`0x70e36f6bf80a52b3b46b3af8e106cc0ed743e8e4`	2020-09-29 10:41:05
TUSD	`0x12392f67bdf24fae0af363c24ac620a2f67dad86`	2020-10-07 11:45:29
AAVE	`0xe65cdb6479bac1e22340e4e755fae7e509ecd06c`	2021-07-18 03:19:05
LINK	`0xface851a4921ce59e912d19329929ce6da6eb0c7`	2021-04-21 21:38:22
WBTC	`0xc11b1268c1a384e55c48c2391d8d480264a3a7f4`	2019-07-16 19:47:37

Table 3.9: Tokens acceptable by Aave and their time added to the protocol

This table provides a list of tokens and their corresponding lending pool smart contracts that have been added to the Aave protocol. Aave is a decentralized lending platform on the Ethereum blockchain that allows users to borrow and lend various cryptocurrencies. The time that each token was added to the platform is also provided in Coordinated Universal Time (UTC).

Token	Lending pool smart contract	Time added to Aave [UTC]
ETH	0x030ba81f1c18d280636f32af80b9aad02cf0854e	2020-11-30 22:20:30
USDC	0xbcca60bb61934080951369a648fb03df4f96263c	2020-12-01 14:23:56
USDT	0x3ed3b47dd13ec9a98b44e6204a523e766b225811	2020-11-30 22:20:30
WBTC	0x9ff58f4ffb29fa2266ab25e75e2a8b3503311656	2020-11-30 22:20:30
DAI	0x028171bca77440897b824ca71d1c56cac55b68a3	2020-12-01 14:22:02
UNI	0xb9d7cb55f463405cdfbe4e90a6d2df01c2b92bf1	2020-11-30 22:20:58
LINK	0xa06bc25b5805d5f8d82847d191cb4af5a3e873e0	2020-12-01 14:23:08
FRAX	0xd4937682df3c8aef4fe912a96a74121c0829e664	2021-09-11 23:42:46
GUSD	0xd37ee7e4f452c6638c96536e68090de8cbcdb583	2021-01-02 19:16:42
LUSD	0xce1871f791548600cb59efbeffc9c38719142079	2022-08-29 19:06:59
sUSD	0x6c5024cd4f8a59110119c56f8933403a539555eb	2020-12-01 14:23:43
TUSD	0x101cc05f4a51c0319f570d5e146a8c625198e636	2020-12-01 14:23:56
USDP	0x2e8f4bdbe3d47d7d7de490437aea9915d930f1a3	2021-07-25 12:17:36
1INCH	0xb29130cbcc3f791f077eade0266168e808e5151e	2022-07-30 17:30:33
AAVE	0xffc97d72e13e01096502cb8eb52dee56f74dad7b	2020-12-01 14:22:02
CRV	0x8dae6cb04688c62d939ed9b68d32bc62e49970b1	2020-12-27 21:46:55
DPI	0x6f634c6135d2ebd550000ac92f494f9cb8183dae	2021-08-21 17:42:40
ENS	0x9a14e23a58edf4efdcb360f68cd1b95ce2081a2f	2022-03-07 06:02:56
MKR	0xc713e5e149d5d0715dcd1c156a020976e7e56b88	2020-12-01 14:23:43
SNX	0x35f6b052c598d933d69a4eec4d04c73a191fe6c2	2020-12-01 14:23:43
stETH	0x1982b2f5814301d4e9a8b0201555376e62f82428	2022-02-27 16:22:12
WETH	0x030ba81f1c18d280636f32af80b9aad02cf0854e	2020-11-30 22:20:30

Bibliography

Admati, A. R. and Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *The Review of Financial Studies*, 1(1):3–40.

Aït-Sahalia, Y. and Sağlam, M. (2023). High frequency market making: The role of speed. *Journal of Econometrics*.

Alfonsi, A. and Acevedo, J. I. (2014). Optimal Execution and Price Manipulations in Time-varying Limit Order Books. *Applied Mathematical Finance*, 21(3):201–237.

Almgren, R. and Chriss, N. (2001). Optimal execution of portfolio transactions. *The Journal of Risk*, 3(2):5–39.

Almgren, R. and Chriss, N. A. (1997). Optimal Liquidation. *SSRN Electronic Journal*.

Aloosh, A. and Bekaert, G. (2021). Currency Factors. *https://doi.org/10.1287/mnsc.2021.4023*, 68(6):4042–4064.

Amiti, M., Itskhoki, O., and Konings, J. (2022). Dominant Currencies: How Firms Choose Currency Invoicing and Why it Matters. *The Quarterly Journal of Economics*, 137(3):1435–1493.

Aramonte, S., Doerr, S., Huang, W., and Schrimpf, A. (2022). DeFi lending: intermediation without information?

Asness, C. S., Moskowitz, T. J., and Pedersen, L. H. (2013). Value and Momentum Everywhere. *The Journal of Finance*, 68(3):929–985.

Avellaneda, M. and Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224.

Bahaj, S. and Reis, R. F. (2020). Jumpstarting an International Currency.

Baldauf, M., Frei, C., and Mollner, J. (2021). Principal Trading Arrangements: When Are Common Contracts Optimal? *Management Science*, 68(4):3112–3128.

Baldauf, M., Frei, C., and Mollner, J. (2024). Block trade contracting. *Journal of Financial Economics*, 160:103901.

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

Bernhardt, D. and Taub, B. (2008). Front-running dynamics. *Journal of Economic Theory*, 138(1):288–296.

Bertsimas, D. and Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50.

BIS (2014). Trade finance: developments and issues.

Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Borri, N. (2019). Conditional tail-risk in cryptocurrency markets. *Journal of Empirical Finance*, 50:1–19.

Brandes, U. (2001). A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177.

Brogaard, J., Carrion, A., Moyaert, T., Riordan, R., Shkilko, A., and Sokolov, K. (2018). High frequency trading and extreme price movements. *Journal of Financial Economics*, 128(2):253–265.

Brogaard, J., Hendershott, T., and Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies*, 27(8):2267–2306.

Brokmann, X., Sérié, E., Kockelkoren, J., and Bouchaud, J.-P. (2016). Slow Decay of Impact in Equity Markets. *https://doi.org/10.1142/S2382626615500070*, 01(02):1550007.

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

Campbell, J. Y., Lo, A. W., and MacKinlay, A. (1997). The Predictability of Asset Returns. In *The Econometrics of Financial Markets*, chapter 2, pages 27–82. Princeton University Press.

Cardella, L., Hao, J., and Kalcheva, I. (2015). Make and take fees in the us equity market. Technical report, Working Paper.

Carrion, A. (2013). Very fast money: High-frequency trading on the nasdaq. *Journal of Financial Markets*, 16(4):680–711.

Cartea, Á., Chang, P., and Penalva, J. (2022). Algorithmic collusion in electronic markets: The impact of tick size. *Working paper. Available at SSRN 4105954*.

Cartea, and Jaimungal, S. (2015). Incorporating Order-Flow into Optimal Execution. *SSRN Electronic Journal*.

Chahrour, R. and Valchev, R. (2022). Trade Finance and the Durability of the Dollar. *The Review of Economic Studies*, 89(4):1873–1910.

Chordia, T., Roll, R., and Subrahmanyam, A. (2000). Commonality in liquidity. *Journal of Financial Economics*, 56(1):3–28.

Comerton-Forde, C., Grégoire, V., and Zhong, Z. (2019). Inverted fee structures, tick size, and market quality. *Journal of Financial Economics*, 134(1):141–164.

Coppola, A., Krishnamurthy, A., and Xu, C. (2023). Liquidity, Debt Denomination, and Currency Dominance.

Dayri, K. and Rosenbaum, M. (2015). Large tick assets: implicit spread and optimal tick size. *Market Microstructure and Liquidity*, 1(01):1550003.

Devereux, M. B. and Shi, S. (2013). VEHICLE CURRENCY. *International Economic Review*, 54(1):97–133.

Doepke, M. and Schneider, M. (2017). Money as a Unit of Account. *Econometrica*, 85(5):1537–1574.

Duffie, D. and Dworczak, P. (2021). Robust benchmark design. *Journal of Financial Economics*, 142(2):775–802.

Dutta, P. K. and Madhavan, A. (1997). Competition and collusion in dealer markets. *The Journal of Finance*, 52(1):245–276.

Dyhrberg, A. H., Foley, S., and Svec, J. (2023). When bigger is better: The impact of a tiny tick size on undercutting behavior. *Journal of Financial and Quantitative Analysis*, 58(6):2387–2416.

Émery, M. and Yor, M. (2004). A parallel between Brownian bridges and Gamma bridges. *Publications of the Research Institute for Mathematical Sciences*, 40, num. 3(3):669–688.

Eren, E. and Malamud, S. (2022). Dominant currency debt. *Journal of Financial Economics*, 144(2):571–589.

Falkenberry, T. (2001). High frequency data filtering: a review of the issues associatedwith maintaining and cleaning a high frequency financial. Technical report, Sl]: Technical Report Tick Data.

Fama, E. F. and French, K. R. (1992). The Cross-Section of Expected Stock Returns. *The Journal of Finance*, 47(2):427–465.

Fleming, M., Nguyen, G., and Ruela, F. (2024). Tick size, competition for liquidity provision, and price discovery: Evidence from the us treasury market. *Management Science*, 70(1):332–354.

Frei, C. and Mitra, J. (2020). Optimal Closing Benchmarks. *SSRN Electronic Journal*.

Frei, C. and Westray, N. (2015). Optimal execution of a vwap order: A stochastic control approach. *Mathematical Finance*, 25(3):612–639.

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

Goldberg, L. S. and Tille, C. (2016). Micro, macro, and strategic forces in international trade invoicing: Synthesis and novel patterns. *Journal of International Economics*, 102:173–187.

Gopinath, G. (2015). The International price system. *Jackson Hole Symposium*, 1:1–7.

Gopinath, G., Boz, E., Casas, C., Díez, F. J., Gourinchas, P. O., and Plagborg-Møller, M. (2020). Dominant currency paradigm. *American Economic Review*, 110(3):677–719.

Gopinath, G. and Itskhoki, O. (2022). *Dominant Currency Paradigm: a review*, volume 6. Elsevier, Amsterdam.

Gopinath, G., Itskhoki, O., and Rigobon, R. (2010). Currency Choice and Exchange Rate Pass-Through. *American Economic Review*, 100(1):304–336.

Gopinath, G. and Stein, J. C. (2021). Banking, Trade, and the Making of a Dominant Currency. *The Quarterly Journal of Economics*, 136(2):783–830.

Gourinchas, P.-O. and Rey, H. (2022). Exorbitant Privilege and Exorbitant Duty.

Gourinchas, P.-O., Rey, H., and Sauzet, M. (2019). The International Monetary and Financial System. *Annual Review of Economics*, 11(1):859–893.

Guéant, O., Lehalle, C. A., and Fernandez-Tapia, J. (2013). Dealing with the inventory risk: A solution to the market making problem. *Mathematics and Financial Economics*, 7(4):477–507.

Hagströmer, B. (2021). Bias in the effective bid-ask spread. *Journal of Financial Economics*, 142(1):314–337.

Harris, L. (2013). Maker-taker pricing effects on market quotations. *USC Marshall School of Business Working Paper*.

Huang, R. D. and Stoll, H. R. (2001). Tick size, bid-ask spreads, and market structure. *Journal of Financial and Quantitative Analysis*, 36(4):503–522.

Huberman, G. and Stanzl, W. (2004). Price Manipulation and Quasi-Arbitrage. *Econometrica*, 72(4):1247–1275.

Ilzetzki, E., Reinhart, C. M., and Rogoff, K. S. (2020). Why is the euro punching below its weight? *Economic Policy*, 35(103):405–460.

Klöck, F., Schied, A., and Sun, Y. S. (2011). Price Manipulation in a Market Impact Model with Dark Pool. *SSRN Electronic Journal*.

Konishi, H. (2002). Optimal slice of a VWAP trade. *Journal of Financial Markets*, 5(2):197–221.

Korajczyk, R. A. and Sadka, R. (2008). Pricing the commonality across alternative measures of liquidity. *Journal of Financial Economics*, 87(1):45–72.

Krugman, P. (1980). Vehicle Currencies and the Structure of International Exchange. *Journal of Money, Credit and Banking*, 12(3):513.

Li, S., Wang, X., and Ye, M. (2021). Who provides liquidity, and when? *Journal of Financial Economics*, 141(3):968–980.

Li, S. and Ye, M. (2021). The optimal price of a stock: A tale of two discretenesses. *Working paper. Available at SSRN 3763516*.

Liu, Y. and Tsyvinski, A. (2021). Risks and Returns of Cryptocurrency. *The Review of Financial Studies*, 34(6):2689–2727.

Liu, Y., Tsyvinski, A., and Wu, X. (2022). Common Risk Factors in Cryptocurrency. *The Journal of Finance*, 77(2):1133–1177.

Lustig, H., Roussanov, N., and Verdelhan, A. (2011). Common risk factors in currency markets. *Review of Financial Studies*, 24(11):3731–3777.

Maggiori, M. (2017). Financial intermediation, international risk sharing, and reserve currencies. *American Economic Review*, 107(10):3038–3071.

Makarov, I. and Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics*, 135(2):293–319.

Malinova, K. and Park, A. (2015). Subsidizing liquidity: The impact of make/take fees on market quality. *The Journal of Finance*, 70(2):509–536.

Mark EJ Newman (2010). *Networks—An Introduction.* Oxford University Press.

McCulloch, J. and Kazakov, V. (2012). Mean Variance Optimal VWAP Trading. *SSRN Electronic Journal.*

Menkhoff, L., Sarno, L., Schmeling, M., and Schrimpf, A. (2012a). Carry Trades and Global Foreign Exchange Volatility. *The Journal of Finance*, 67(2):681–718.

Menkhoff, L., Sarno, L., Schmeling, M., and Schrimpf, A. (2012b). Currency momentum strategies. *Journal of Financial Economics*, 106(3):660–684.

Mukhin, D. (2022). An Equilibrium Model of the International Price System. *American Economic Review*, 112(2):650–688.

Mundell, R. A. (1957). A Theory of Optimum Currency Areas. *American economic review*, 377(1775):657–665.

O'Hara, M., Yao, C., and Ye, M. (2014). What's not there: Odd lots and market data. *The Journal of Finance*, 69(5):2199–2236.

Pástor, and Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685.

Röell, A. (1990). Dual-capacity trading and the quality of the market. *Journal of Financial Intermediation*, 1(2):105–124.

Stoikov, S. and Sağlam, M. (2009). Option market making under inventory risk. *Review of Derivatives Research*, 12(1):55–79.

Tóth, B., Lempérière, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., and Bouchaud, J. P. (2011). Anomalous Price Impact and the Critical Nature of Liquidity in Financial Markets. *Physical Review X*, 1(2):1–11.

Wah, E., Feldman, S., Chung, F., Bishop, A., Aisen, D., and Exchange, I. (2017). A comparison of execution quality across us stock exchanges. *Global Algorithmic Capital Markets: High Frequency Trading, Dark Pools, and Regulatory Challenges*, pages 91–146.

Wood, R. A., McInish, T. H., and Ord, J. K. (1985). An investigation of transactions data for nyse stocks. *The Journal of Finance*, 40(3):723–739.

Wright, R. and Trejos, A. (2001). International currency. *B.E. Journal of Macroeconomics*, 1(1).

Xu, J., Paruch, K., Cousaert, S., and Feng, Y. (2023). SoK: Decentralized Exchanges (DEX) with Automated Market Maker (AMM) Protocols. *ACM Computing Surveys*, 55(11):1–50.

Zhang, C. (2014). An information-based theory of international currency. *Journal of International Economics*, 93(2):286–301.