

***Financial Market Uncovered – Article 10***  
***Market Microstructure: The Mechanics Behind***  
***Market Liquidity***



*Kilian Voillaume*

*June 4, 2025*

# Summary

<b>1</b>	<b><i>Introduction .....</i></b>	<b><i>5</i></b>
<b>2</b>	<b><i>What is Market Microstructure.....</i></b>	<b><i>6</i></b>
2.1	Definition and Core Components .....	6
2.1.1	Order Types .....	6
2.1.2	Matching Rules and Execution Logic .....	7
2.1.3	Participants and Incentives.....	7
2.2	Venue Types and Trading Architecture .....	7
2.2.1	Lit Markets (Public Exchanges).....	7
2.2.2	Dark Pools (Non-Displayed Venues) .....	8
2.2.3	Systematic Internalisers and Dealer Platforms .....	8
2.2.4	The Challenge of Fragmentation.....	8
2.3	Why Microstructure Affects Price Discovery .....	9
2.3.1	Fragmented Liquidity and Incomplete Signals .....	9
2.3.2	Rules That Shape Behaviour .....	9
2.3.3	Speed and the Advantage of Being First .....	10
2.3.4	The Price Is Not Always the Truth .....	10
<b>3</b>	<b><i>Fragmentation and Venue Competition.....</i></b>	<b><i>11</i></b>
3.1	The Rise of Fragmentation.....	11
3.1.1	Regulation as a Catalyst .....	11
3.1.2	The Hidden Consequences .....	11
3.1.3	Why It Matters .....	12
3.2	Smart Order Routing (SOR) and Best Execution .....	12
3.2.1	The Mission: Best Execution .....	12
3.2.2	The Complexity Beneath the Surface .....	13
3.2.3	Strategic Implications.....	13
3.3	Cross-Venue Arbitrage and Structural Instability .....	13
3.3.1	The Mechanics of Cross-Venue Arbitrage .....	14
3.3.2	When Arbitrage Backfires.....	14
3.3.3	Structural Risk from Latency and Inconsistency .....	14
3.3.4	The Broader Implication .....	14

---

<b>4</b>	<b><i>Bid-Ask Spread</i></b> .....	<b>15</b>
4.1	What Drives the Spread .....	15
4.2	Spread Dynamics .....	16
4.2.1	Volatility and Uncertainty .....	16
4.2.2	Market Depth and Order Book Shape .....	16
4.2.3	Role of Competition and HFT .....	16
4.2.4	Not All Spreads Are Equal .....	16
4.3	How HFT and Passive Liquidity Impact Spreads .....	17
4.3.1	The Role of High-Frequency Traders .....	17
4.3.2	Passive Liquidity and Market Efficiency .....	17
4.3.4	Fragility Behind the Tight Spread .....	17
4.3.5	The Illusion of Stability .....	18
<b>5</b>	<b><i>Liquidity</i></b> .....	<b>19</b>
5.1	The Four Dimensions of Liquidity .....	19
5.2	Order Book Anatomy .....	21
5.2.1	Surface vs Depth .....	21
5.2.2	Book Shape and Market Sentiment .....	21
5.3	Market Examples .....	22
5.3.1	Equities: Fast, Fragmented, and Deep — Until They're Not .....	22
5.3.2	Credit Markets: Opaque and Episodic .....	23
5.3.3	Crypto Markets: Fast But Fragile .....	23
<b>6</b>	<b><i>Liquidity Risk</i></b> .....	<b>24</b>
6.1	Liquidity Risk: When the Market Disappears .....	24
6.1.1	Apparent vs Actual Liquidity .....	24
6.2	Sources of Liquidity Risk .....	25
6.2.1	External Shocks (Exogenous Drivers) .....	25
6.2.2	Internal Feedback Loops (Endogenous Amplifiers) .....	25
6.3	Effects on Execution and Hedging .....	26
6.3.1	Slippage and Execution Failure .....	26
6.3.2	Broken Hedges and Misleading Greeks .....	26
6.3.3	Undervalued in Risk Models .....	27
<b>7</b>	<b><i>Microstructure Alpha</i></b> .....	<b>28</b>

---

---

7.1	Latency Arbitrage and Information Racing .....	28
7.2	Order Flow Prediction and Toxicity .....	29
7.3	Internalisation and Payment for Order Flow (PFOF) .....	30
<b>8</b>	<b><i>Managing Liquidity Risk in Practice</i></b> .....	<b>32</b>
8.1	Liquidity-Adjusted Value at Risk (L-VaR) .....	32
	8.1.1 VaR & L-VaR .....	32
	8.1.2 Implementation: Measuring the Drag .....	33
8.2	Stress Testing and Scenario Design .....	35
	8.2.1 Why Stress Testing Matters .....	35
	8.2.2 Core Components of Liquidity Stress Scenarios .....	35
<b>9</b>	<b><i>Conclusion</i></b> .....	<b>37</b>
<b>10</b>	<b><i>Python code</i></b> .....	<b>38</b>
<b>11</b>	<b><i>References</i></b> .....	<b>40</b>

# 1 Introduction

Financial markets are often described as fluid and efficient systems, where information flows seamlessly, prices adjust continuously, and trades are executed in milliseconds. Yet beneath this smooth surface lies a complex architecture of rules, incentives, and frictions — the world of *market microstructure*. It is this architecture that determines how, where, and at what cost trades are executed. And when it malfunctions, the consequences can be sudden, severe, and systemic.

While investors and analysts frequently focus on what to trade — equities, bonds, derivatives — they often overlook the mechanics of how trading occurs. But execution is not a neutral afterthought. It is a fundamental component of both performance and risk. Microstructure governs whether a portfolio adjustment takes seconds or hours, whether hedges are effective or fall apart, and whether liquidity is available when it is most needed — or entirely absent.

## **Fragility Beneath Speed**

Over the past decade, several high-profile events have exposed just how fragile modern markets can be when execution systems are strained. The 2010 Flash Crash erased nearly a trillion dollars in equity value within minutes, driven by algorithmic feedback loops and liquidity evaporation. The collapse of Archegos Capital in 2021 revealed the dangers of concentrated synthetic exposure in illiquid stocks. During the COVID panic in March 2020, bond and ETF markets dislocated as dealers retreated, causing massive gaps between prices and fundamentals. Even retail-driven episodes like the GameStop saga showed how market plumbing can distort under unusual order flow conditions.

These were not just anomalies. They were structural failures — breaks within the systems designed to process trades.

## **Liquidity Risk: The Most Underestimated Threat**

Liquidity risk is not merely the challenge of trading large positions. It is the risk that the market *won't be there* when you need it. This risk is deeply non-linear: a market that appears deep and efficient can vanish in moments under stress. Spreads can widen sharply. Market makers can pull out. Algorithms can freeze. Execution becomes not just expensive, but uncertain.

Moreover, liquidity risk is often hidden. A narrow bid-ask spread tells you nothing about how much volume is truly available, or whether that quote will still exist by the time your order hits the book. A well-functioning microstructure keeps these risks in check — but under pressure, it can invert.

## 2 What is Market Microstructure

To understand how trades are executed — and how execution risk emerges — we must begin with the design of the market itself. *Market microstructure* is the field that studies the processes, rules, and participants that turn investor intentions into actual transactions. It is not about long-term valuation or macro trends, but about the mechanics of short-term price formation: how bids and offers meet, how liquidity is supplied and withdrawn, and how information is embedded in prices.

Microstructure affects everyone — from high-frequency traders seeking millisecond advantages, to institutional investors managing large positions, to retail traders executing through broker apps. And yet, it is often treated as a black box. This section opens that box, starting with the essential building blocks.

### 2.1 Definition and Core Components

Market microstructure refers to the study of how financial markets operate at the transactional level — the mechanisms that govern how orders are submitted, matched, and executed. It is concerned not just with prices, but with *how prices are formed*.

At its heart, microstructure encompasses three key elements:

#### 2.1.1 Order Types

Markets are not just streams of buyers and sellers shouting prices. They rely on structured order instructions that determine how, when, and at what price a trade should occur. These order types influence execution quality, cost, and market impact.

- **Market orders:** These are instructions to buy or sell immediately at the best available price. They prioritise speed over precision and are commonly used when liquidity is deep, or timing is critical.
- **Limit orders:** These specify the worst acceptable price for a transaction — for example, “buy at no more than \$100”. They provide control but risk non-execution if the market doesn’t reach that price.
- **Stop orders:** These are activated only if the market hits a certain trigger level. For example, a stop-loss becomes a market order once the asset falls below a preset threshold.
- **Iceberg orders:** These hide part of the order size from the market. Only a small portion is visible in the order book, reducing signalling to other traders.
- **Fill-or-kill (FOK) orders:** These must be executed in full immediately or cancelled entirely. They are used to avoid partial fills, often in fast-moving markets.

Each type represents a trade-off between urgency, visibility, and control. Together, they shape the structure of liquidity in the market.

### 2.1.2 *Matching Rules and Execution Logic*

Once orders are in the system, how are they paired?

- Most exchanges use a price–time priority rule: the best-priced order is matched first, and among orders at the same price, the earliest placed order has priority.
- Other systems may use pro-rata matching, where large orders are partially filled based on their size, or periodic auctions, where trades are matched in batches at set intervals.

The choice of matching logic determines who gets filled first, and at what price — and it influences trader behaviour accordingly.

### 2.1.3 *Participants and Incentives*

Market microstructure is shaped by the diversity of actors interacting within it — each responding to different incentives.

- Market makers continuously post both buy and sell quotes, profiting from the spread while absorbing short-term risk. Their activity supports liquidity, but they may withdraw during stress.
- High-frequency traders (HFTs) use ultra-fast systems to capitalise on small price differences or order book signals. They often serve as liquidity providers but can also exacerbate volatility.
- Retail brokers typically route client orders through third parties, sometimes receiving payment for order flow — a controversial but common practice in some jurisdictions.
- Institutional traders rely on smart algorithms that split large trades into smaller parts (called *child orders*) to minimise market impact and reduce signalling.

Each participant influences the market's behaviour — and understanding their role is essential to interpreting both price action and execution risk.

## 2.2 *Venue Types and Trading Architecture*

Once an order is submitted, it must find a place to be executed — a *venue*. But in today's fragmented financial landscape, there is no single market. A single stock or bond can trade across dozens of different platforms, each with distinct rules, transparency levels, and incentives. The design of these venues — how they receive, prioritise, and match orders — is central to market microstructure.

Broadly speaking, trading venues fall into three main categories:

### 2.2.1 *Lit Markets (Public Exchanges)*

These are traditional stock exchanges such as the New York Stock Exchange (NYSE), London Stock Exchange (LSE), or Euronext, where quotes and transactions are publicly displayed in real time. Lit markets offer:

- Full pre-trade transparency: You can see the best bid and offer, along with limited depth of the order book.
- Centralised price discovery: These venues play a crucial role in determining the official market price.
- Price-time priority: Orders are usually matched based on best price, then earliest time.

For most retail and institutional investors, lit markets are the visible backbone of the financial system — but they represent only part of the total volume traded.

### ***2.2.2 Dark Pools (Non-Displayed Venues)***

Dark pools are private trading venues where orders are hidden from public view until execution. They are often used by institutional investors to trade large blocks of shares discreetly, without signalling their intentions and moving the market.

Key features:

- No pre-trade transparency: Order size and price are not visible to other participants.
- Price referencing: Most dark pool trades are executed at or near the midpoint of the public bid and ask spread.
- Reduced market impact: Ideal for minimising slippage on large trades.

However, dark pools can lead to concerns about fairness and price integrity, as retail traders and smaller firms may not have equal access or visibility.

### ***2.2.3 Systematic Internalisers and Dealer Platforms***

These are venues where brokers or dealers execute client orders against their own inventory, rather than routing them to a public exchange. This model is common in over-the-counter (OTC) markets like bonds, FX, and derivatives — and increasingly present in equities through Systematic Internalisers (SIs) under European regulation.

- Internalisation allows brokers to profit from the spread while offering “immediate” execution.
- Clients often receive price improvement relative to lit markets — but may sacrifice transparency.
- Execution is bilateral, not centralised — raising potential conflicts of interest.

This model is dominant in asset classes where central order books are impractical, such as corporate bonds or bespoke derivatives.

### ***2.2.4 The Challenge of Fragmentation***

Because a single asset may trade across all three venue types simultaneously, liquidity becomes fragmented. A trader seeking the best price must navigate this fragmented architecture — either manually, or via Smart Order Routing (SOR) algorithms that search for optimal execution across venues.



Fragmentation creates opportunities (e.g., arbitrage, price improvement) but also challenges:

- Prices may differ slightly between venues.
- Not all market participants have equal speed or access.
- Transparency may be diluted, especially during stress.

In short, venue architecture defines how liquidity is distributed — and whether it can be reliably accessed when it matters most.

## ***2.3 Why Microstructure Affects Price Discovery***

In theory, markets are meant to reflect a clean aggregation of supply and demand, with prices adjusting to new information in an efficient, almost frictionless way. But in practice, the process by which prices emerge is neither clean nor frictionless — it is deeply shaped by the rules, venues, and behaviours that define modern trading infrastructure.

Market microstructure doesn't just influence transaction costs or execution speed — it affects the price itself. Understanding how requires stepping away from the textbook ideal and looking at how structural features distort or delay the transmission of information into prices.

### ***2.3.1 Fragmented Liquidity and Incomplete Signals***

In today's markets, a single asset can trade across dozens of venues simultaneously — lit exchanges, dark pools, internalisers, and more. This fragmentation means that no one participant sees the full picture. Even advanced algorithms must piece together a composite view of the market from partial, delayed, or opaque data.

As a result, price discovery becomes less efficient. A quote on one platform might be slightly higher or lower than on another, not because of new information, but simply due to where liquidity is sitting. In times of stress or volatility, this disconnect can grow, creating confusion and execution risk, particularly for those without access to fast or consolidated feeds.

### ***2.3.2 Rules That Shape Behaviour***

Matching rules and venue design also influence how participants behave. When exchanges reward speed or prioritise price and time, traders adapt. Many will cancel and resubmit orders constantly to maintain priority, contributing to flickering quotes and a less stable order book. Others may deliberately hide their intentions using iceberg orders or discretionary instructions, limiting the amount of visible liquidity.

These behaviours aren't based on changes in valuation — they are strategic responses to structure. Yet they directly affect the prices others see, and ultimately the price at which trades occur. In short, the observed price is not a neutral outcome of information — it is shaped by incentives and constraints.

### ***2.3.3 Speed and the Advantage of Being First***

Nowhere is this more apparent than in the role of speed. In ultra-fast electronic markets, a few milliseconds can make the difference between profit and loss. High-frequency traders (HFTs) compete to process data and submit orders faster than anyone else, allowing them to exploit temporary price differences, react to order book changes, or capture small arbitrage opportunities.

While this can enhance liquidity in stable conditions, it can also distort short-term prices, particularly when slower participants are consistently disadvantaged. Price becomes a function not only of information and belief, but of who got there first.

### ***2.3.4 The Price Is Not Always the Truth***

All of this leads to a central insight: the price we observe in the market is not necessarily the price that reflects fundamental value. It is the result of a set of microstructural processes — venue choice, order type, matching logic, latency, and liquidity availability — all operating within specific institutional and technological frameworks.

During quiet times, these processes operate in the background, almost invisible. But when markets are under stress, their influence becomes obvious, and their distortions can dominate the behaviour of prices. Traders, risk managers, and policymakers who ignore these mechanics risk misunderstanding what the market is really telling them.

### 3 Fragmentation and Venue Competition

Modern markets no longer operate as unified arenas. Instead, they function as fragmented ecosystems — a network of competing venues, each offering its own execution logic, participant base, and liquidity dynamics. This fragmentation has reshaped how prices are formed, how orders are routed, and how market participants compete for flow. In this section, we explore how this structural evolution emerged, why it matters, and how it affects execution quality in both obvious and subtle ways.

#### 3.1 *The Rise of Fragmentation*

Financial instruments used to trade on a single exchange, which served as a hub for all bids, offers, and trades. Everyone could see the market depth, execution was centralized, and price discovery was transparent. There is no longer such a planet.

The markets of today are extremely fragmented. Numerous venues, each with its own set of regulations, participants, and microstructural peculiarities, can trade a single stock. A share listed on Euronext can also be traded on a variety of dark pools and systematic internalizers in Europe, including BATS, Cboe, and Turquoise. In addition to dozens of alternative trading systems (ATS) and internalizing brokers, trading in the US is dispersed over major exchanges such as the NYSE and Nasdaq.

##### 3.1.1 *Regulation as a Catalyst*

Regulators implemented changes in the US and Europe with the goal of lowering trading expenses and enhancing execution quality. By enabling brokers to forward orders to the location with the best price, Regulation NMS (National Market System), which was implemented in the United States in 2007, sought to encourage competition amongst exchanges. Similarly, the dominance of established exchanges was challenged by MiFID I and II in Europe, which made room for multilateral trading facilities (MTFs) and dark venues.

The reasoning was simple: investors will profit from tighter spreads, quicker execution, and less fees if markets fight for order flow. And this goal has been achieved in a lot of ways. Over the last 20 years, trading costs for liquid stocks have significantly decreased.

##### 3.1.2 *The Hidden Consequences*

With liquidity dispersed across venues, it is no longer possible to see the entire market in one place. Traders must rely on consolidated feeds, smart order routers, and internal algorithms to navigate this mosaic. Even then, latency, data access, and visibility differ depending on the tools and counterparties used.

This introduces several challenges:

- Price dislocations can occur between venues, especially during periods of volatility or stress.

- Order anticipation and information leakage become more likely as brokers route orders through multiple systems.
- Fairness and access are uneven — institutional participants with co-location and high-speed feeds often see a clearer picture than retail or slower desks.

In short, fragmentation has made markets faster and more competitive but also more complex, opaque, and uneven.

### ***3.1.3 Why It Matters***

For the trader or investor, this means execution is no longer just about placing a well-timed order. It is about navigating a fragmented ecosystem with differing rules, latencies, and liquidity pools. In such a landscape, structure becomes strategy, and execution becomes risk.

Understanding fragmentation isn't just useful for those building trading systems — it's essential for anyone trying to interpret price moves, assess liquidity, or manage market risk in real time.

## ***3.2 Smart Order Routing (SOR) and Best Execution***

In a fragmented market, placing an order is no longer as simple as choosing “the exchange.” With liquidity spread across dozens of venues — each with different prices, depths, and matching rules — determining *where* to send an order has become a core challenge. This is where Smart Order Routing (SOR) comes in.

A Smart Order Router is an algorithmic tool that analyses the state of multiple venues in real time, then dynamically decides where to send a trade to achieve the best possible result. It is the brain behind modern execution — working invisibly behind most retail and institutional orders.

### ***3.2.1 The Mission: Best Execution***

The goal of SOR is to deliver what regulators and clients call best execution — typically defined as the best possible price, considering cost, speed, and likelihood of execution. But in practice, best execution is multi-dimensional. It may mean:

- Sending a portion of the order to a lit market for price discovery.
- Slicing large trades across several venues to avoid moving the market.
- Prioritising venues with faster response times or deeper queues.
- Avoiding adverse selection in volatile or thin markets.

What appears to the trader as a single execution is often the result of dozens of micro-decisions made in milliseconds — balancing price, fill probability, and market impact.

### ***3.2.2 The Complexity Beneath the Surface***

Effective smart routing requires more than just comparing prices. The best price visible in the order book may not be the best *in practice* if the quote is fleeting, hidden behind latency, or offered in small size. Advanced SORs take into account:

- Order book depth beyond the top of book.
- Latency profiles of each venue (how quickly it responds).
- Fill rates and reliability of different platforms.
- Historical toxicity (whether trades at a venue tend to precede adverse moves).

In fast markets, this decision-making must be continuous — reacting to real-time changes in order flow, quote updates, and competing trades.

### ***3.2.3 Strategic Implications***

For institutional desks, SOR is not just a utility — it is a source of execution alpha. The ability to access hidden liquidity, avoid signalling, and minimise footprint can make a measurable difference in slippage, particularly for large or sensitive trades.

For retail traders, routing decisions are often made by brokers or wholesalers. In markets like the US, Payment for Order Flow (PFOF) arrangements introduce additional considerations, where brokers route orders to market makers in exchange for rebates — potentially introducing conflicts between best price and internal incentives.

### ***Not Just a Technical Detail***

Smart Order Routing may seem like a technical plumbing issue, but in reality, it plays a central role in how modern markets function. It mediates between fragmented liquidity and execution objectives, and it defines whether participants — especially smaller or slower ones — are competing on equal footing.

Understanding SOR is essential for anyone interpreting price movements, evaluating execution quality, or designing trading strategies in a fragmented world.

## ***3.3 Cross-Venue Arbitrage and Structural Instability***

One of the unavoidable consequences of market fragmentation is the creation of price differentials between trading venues. Even for highly liquid assets, slight discrepancies in bid-ask spreads, quote updates, or order book depth can emerge between platforms — especially when information is moving quickly, or liquidity is thin. This sets the stage for cross-venue arbitrage.

Arbitrage, in this context, means taking advantage of these temporary price differences by simultaneously buying on one venue and selling on another. In principle, arbitrage ensures pricing efficiency: when traders act on discrepancies, they help realign prices across markets. But in fragmented, high-speed markets, arbitrage is no longer the simple game of risk-free profit

it once was — it has become a structural force that both *enforces* and *destabilises* price consistency.

### ***3.3.1 The Mechanics of Cross-Venue Arbitrage***

Modern arbitrageurs, often high-frequency trading firms, rely on ultra-low-latency infrastructure to monitor dozens of venues simultaneously. The goal is to detect even microsecond-level delays in price updates or liquidity changes. A delay of just 1–2 milliseconds between two venues can be enough to trigger an arbitrage opportunity — but only for those fast enough to act.

This creates a competitive arms race: the fastest traders extract value from the slowest, reinforcing a hierarchy of information access. For slower participants, the price they see may already be stale — the opportunity already gone.

### ***3.3.2 When Arbitrage Backfires***

Under normal conditions, arbitrage tightens spreads and improves price alignment. But during periods of market stress or dislocation, it can have the opposite effect. If one venue's order book collapses or becomes unresponsive, traders may aggressively unwind positions elsewhere — triggering feedback loops and further instability.

This dynamic was evident during the 2010 Flash Crash, when rapid inter-venue arbitrage contributed to a cascading series of trades that widened spreads, depleted liquidity, and destabilised prices in both equities and futures markets. The same risk exists today — particularly in fragmented systems where order flow is routed through complex layers of algorithms, brokers, and liquidity providers.

### ***3.3.3 Structural Risk from Latency and Inconsistency***

Cross-venue arbitrage is not just a tool for profit — it is a structural feature of how prices converge. But it relies on assumptions that can break down:

- That prices are reliably updated across all venues.
- That orders will be honoured at displayed prices.
- That latency differences remain within tolerable bounds.

When these assumptions fail — due to technological outages, extreme volatility, or liquidity evaporation — arbitrage becomes not a stabiliser, but an amplifier of disorder.

### ***3.3.4 The Broader Implication***

In a multi-venue market, the integrity of price discovery depends not only on the behaviour of participants, but on the synchronisation and resilience of the system itself. Arbitrage links venues together — but in doing so, it can also transmit stress, mispricing, and volatility across the entire market.

For traders, understanding this dynamic is crucial. Price is no longer local — it is the result of a network. And when one node in that network fails, the impact can ripple outward far faster than many risk models assume.

## 4 Bid-Ask Spread

Liquidity is often discussed as a binary: a market is either liquid or not. But in reality, liquidity comes with a price tag — and that price is usually expressed through the **bid-ask spread**. This spread represents the difference between what buyers are willing to pay (bid) and what sellers are willing to accept (ask), and it functions as the most visible indicator of transaction cost in any financial market.

Yet the spread is more than just a number. It reflects market structure, information asymmetry, inventory risk, and the behaviour of market participants. In this section, we break down the anatomy of the bid-ask spread, explain why it fluctuates, and examine what it really tells us about market quality and liquidity.

### 4.1 *What Drives the Spread*

At its core, the bid-ask spread compensates liquidity providers — typically market makers — for the risks and costs they incur when offering to trade with others. When a trader submits a market order, they are crossing the spread — accepting the best available price offered by someone else in exchange for immediacy.

But that spread is not arbitrary. It is composed of several overlapping components:

#### 1. Order Processing Costs

Executing trades involves infrastructure — exchanges, clearinghouses, and technology systems. While per-trade costs have decreased over time, especially for high-frequency traders, they still factor into the minimum spread a market maker must charge to remain profitable.

#### 2. Inventory Risk

Market makers are constantly exposed to directional risk. If they sell an asset and can't immediately hedge or rebalance, they face potential losses from price movements. The bid-ask spread includes a buffer to offset the risk of holding an unbalanced position — particularly in volatile or illiquid markets.

#### 3. Adverse Selection Risk

Perhaps the most important component is informational. Market makers face the constant risk that their counterparty knows something they don't — that a buy order is coming just before a price rally, or a sell order just before a crash. To protect against this, they widen spreads when they suspect **toxic flow**: informed traders whose orders are likely to be followed by unfavourable price moves.

In essence, the spread is a risk premium. The wider it is, the greater the uncertainty or imbalance in the market.

## 4.2 *Spread Dynamics*

While the bid-ask spread may appear static at a given moment, it is in fact **highly dynamic** — expanding and contracting in response to a variety of market conditions. It reflects not just the current state of liquidity, but the *confidence* of market makers in their ability to quote prices without being adversely selected or overwhelmed by flow.

Understanding how spreads behave helps decode periods of calm, stress, and dislocation.

### 4.2.1 *Volatility and Uncertainty*

When markets are calm, with low volatility and balanced order flow, spreads tend to tighten. Liquidity providers are more confident that price changes will be modest and that flow is benign. But as uncertainty increases — due to earnings releases, macroeconomic announcements, geopolitical events, or technical instability — spreads widen.

This is not irrational. Wider spreads reflect **greater compensation for risk**. When price movements are unpredictable, the chance of being adversely selected by an informed counterparty increases. Market makers adjust accordingly, even withdrawing from the inside quote or quoting smaller sizes.

### 4.2.2 *Market Depth and Order Book Shape*

Spread behaviour is closely tied to the **depth and shape of the order book**. A narrow spread with minimal volume behind it may offer little real liquidity — a “ghost spread”. Conversely, a slightly wider spread supported by substantial volume may provide a better execution experience.

In practice, spreads often appear narrow, but liquidity beyond the top of book may be shallow. This creates hidden fragility: the first trade may clear easily, but follow-up orders face much higher impact. Traders focused only on spread size may miss the true cost of execution.

### 4.2.3 *Role of Competition and HFT*

In highly electronic markets, particularly in equities and futures, **high-frequency market makers** aggressively compete to provide the tightest spreads. This competition often compresses spreads to minimal levels — sometimes just a tick — especially during periods of stable volatility.

However, this comes with a trade-off. The same players who quote tightly in normal conditions are also the fastest to withdraw when stress emerges. In such moments, spreads may widen dramatically, and the apparent liquidity evaporates.

### 4.2.4 *Not All Spreads Are Equal*

The absolute level of the spread should always be considered in context: what asset is being traded, how volatile is it, and how deep is its book? A 1-cent spread on Apple stock is trivial; a 3-point spread on a small-cap biotech might be normal. Similarly, fixed income and derivative markets have very different liquidity norms, often operating on *quote-request* or dealer-driven models with wider but less transparent spreads.



### ***4.3 How HFT and Passive Liquidity Impact Spreads***

In many modern markets — particularly equities, futures, and FX — the bid-ask spread has narrowed dramatically over the past two decades. This is largely due to the rise of high-frequency trading (HFT) and passive liquidity provision, which have redefined both the mechanics and economics of market making.

#### ***4.3.1 The Role of High-Frequency Traders***

High-frequency traders operate by leveraging speed — executing thousands of trades per second, reacting to price changes faster than human traders ever could. Many HFT firms act as quasi-market makers, continuously quoting both bids and offers in order to capture small profits from the spread.

Because HFTs are:

- highly automated,
- co-located near exchange servers to reduce latency, and
- designed to minimise inventory risk through rapid position turnover,

they can afford to quote extremely tight spreads, sometimes just a single tick wide. Their business model rests not on size, but on frequency and scale — capturing tiny margins across millions of trades.

This has resulted in tighter spreads, lower explicit transaction costs, and greater perceived liquidity — especially during normal conditions.

#### ***4.3.2 Passive Liquidity and Market Efficiency***

Passive traders are those who provide liquidity by placing limit orders that rest in the order book, waiting to be hit. This is in contrast to active traders who cross the spread using market orders.

Passive liquidity provision, including that by algorithmic funds and statistical arbitrageurs, increases the visible depth in the book and helps anchor price levels. In stable environments, this results in:

- Greater competition to offer the best price,
- Thinner spreads,
- Reduced execution costs for active traders.

Regulations like MiFID II in Europe and Reg NMS in the US have further incentivised passive liquidity through rules on quote display and access to rebates.

#### ***4.3.4 Fragility Behind the Tight Spread***

However, there's a critical caveat: the liquidity offered by HFTs and passive traders is highly sensitive to market conditions.

When volatility spikes or uncertainty rises, these players can — and often do — withdraw instantly. Their systems are designed to avoid being adversely selected. As a result:

- Spreads widen suddenly,
- Order book depth evaporates,
- Executions become more costly and uncertain.

This phenomenon was clearly observed during the 2010 Flash Crash and again in March 2020, when spreads across major equity and bond markets widened drastically as electronic liquidity vanished.

#### ***4.3.5 The Illusion of Stability***

The presence of tight spreads in calm conditions should not be mistaken for resilience. Much of the liquidity in modern markets is conditional: it exists only when risk is low and information is symmetric. When either of those assumptions breaks down, the very players who quote the tightest prices are the first to disappear.

## 5 Liquidity

Liquidity is one of the most frequently cited — and least understood — concepts in financial markets. Traders often talk about whether a market is “liquid” or “illiquid,” but this binary language conceals the true complexity of the phenomenon. Liquidity is not a single number or a fixed attribute; it is multi-dimensional, context-dependent, and often behavioural.

In this section, we move beyond surface-level indicators like bid-ask spreads and examine the deeper structure of liquidity — its dimensions, its visibility, and its limitations.

### 5.1 *The Four Dimensions of Liquidity*

In professional trading and market structure theory, liquidity is typically defined along four primary dimensions. Each captures a different facet of how easily and efficiently an asset can be traded.

#### 1. Tightness

Tightness refers to the cost of executing a trade immediately, most commonly measured by the bid-ask spread. A tight market has narrow spreads, meaning traders can buy and sell without incurring large transaction costs. This is the most visible aspect of liquidity.

#### 2. Depth

Depth measures the volume available at the best prices. A market may have a tight spread, but if only 100 shares are offered at that price, any larger trade will face price impact. Deep markets allow large transactions to occur with minimal slippage — shallow ones do not.

Depth is particularly critical for institutional traders, who often need to execute orders that exceed top-of-book liquidity. In thinly traded markets, even moderate order sizes can move prices significantly.

#### 3. Immediacy

Immediacy captures how quickly a trade can be executed without delay. In fast, electronic markets, immediacy is often taken for granted — but it can evaporate during stress. A liquid market must not only offer good prices, but allow trades to be completed *when they're needed*, not minutes or hours later.

This dimension becomes crucial during volatility spikes, earnings announcements, or geopolitical shocks — when speed of execution is as important as price.

#### 4. Resilience

Resilience refers to how quickly a market recovers after a shock. If a large order removes liquidity, how fast do new orders replenish the book? In resilient markets, liquidity regenerates rapidly, keeping price dislocations brief. In fragile ones, a single trade can cause lasting instability. Resilience is perhaps the hardest to observe directly, yet it often determines whether a market can absorb stress — or spiral into dysfunction.

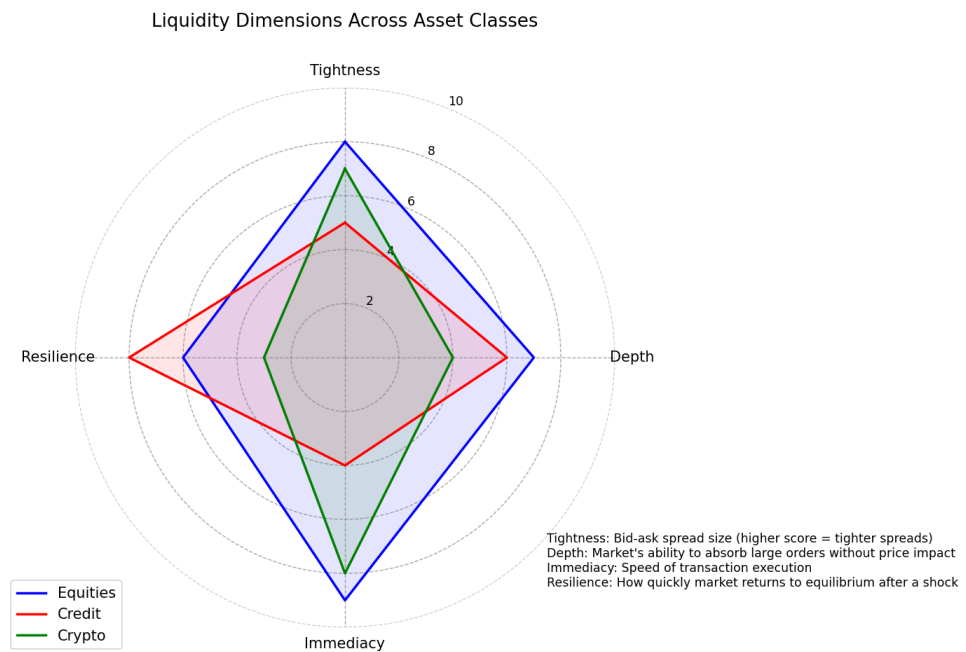


Figure 1: Comparative Liquidity Analysis

This radar chart provides a multidimensional comparison of liquidity characteristics across three major asset classes—equities, credit, and cryptocurrencies—evaluated across four key dimensions: tightness (bid-ask spreads), depth (order book resilience), immediacy (execution speed), and resilience (post-trade price recovery).

Equities (blue) exhibit the most balanced liquidity profile, with high scores in tightness, immediacy, and depth, reflecting mature market infrastructure and robust participation.

Credit markets (red) trade depth for resilience, a pattern consistent with their over-the-counter nature, where wider spreads and slower execution are offset by institutional hedging mechanisms.

Cryptocurrencies (green) display a distinct liquidity asymmetry: while they rival equities in tightness and immediacy, their shallow depth and low resilience reveal vulnerability to large orders and volatility shocks—a hallmark of fragmented markets with lower institutional participation. The chart underscores that liquidity is not monolithic; asset classes prioritize different dimensions, necessitating tailored trading strategies.

For instance, crypto's high immediacy but poor resilience warns against conflating exchange speed with stability, while credit's resilience-depth tradeoff highlights the cost of liquidity in decentralized markets. These insights align with empirical market microstructure research, where liquidity "triangles" (tightness-depth-immediacy) often diverge by asset class due to varying participant behavior and market design.

## 5.2 *Order Book Anatomy*

To understand liquidity at a granular level, one must examine the order book — the real-time record of all active buy and sell orders at various price levels. This is where liquidity lives: not just in the best bid and offer, but in the entire structure of resting orders.

At first glance, an order book appears as a ladder: bids stacked on one side, asks on the other, each with associated quantities. The top of the book shows the best available prices — the narrowest spread — but this only represents a tiny fraction of actual liquidity.

### 5.2.1 *Surface vs Depth*

The **visible** part of the book (commonly known as Level 1 or top-of-book data) includes:

- The highest bid (the best price a buyer is willing to pay),
- The lowest ask (the best price a seller is willing to accept),
- And the size available at those prices.

This is what most trading platforms display by default. But beneath that surface lies Level 2 (or full-depth) data: all additional price levels and order sizes resting deeper in the book.

A narrow spread might suggest a liquid market — but if there's little depth beyond the top quotes, large trades will suffer significant slippage. Conversely, a book with a wider spread but dense volume behind each level may offer better execution for size.

### 5.2.2 *Book Shape and Market Sentiment*

The shape of the order book also reveals behavioural insights. A *steep book*, where size drops off quickly beyond the best prices, suggests caution or low conviction — traders are reluctant to commit further into the price curve. A *flat, dense book* indicates confidence and participation across price levels.

Order book imbalances — where one side (bids or asks) has significantly more volume — may signal short-term directional bias. However, this too can be deceptive, especially in markets where quote spoofing or order cancellation strategies are employed.

Critically, not all liquidity is displayed. Iceberg orders allow traders to show only a portion of their order, while keeping the full size hidden. Similarly, dark pool activity and internalised flows never appear in the public book at all.

Moreover, many displayed orders are ephemeral. In high-frequency environments, orders are often posted and cancelled within milliseconds — making the book more like a radar screen than a fixed map. What appears to be available liquidity may disappear the moment it is tested.

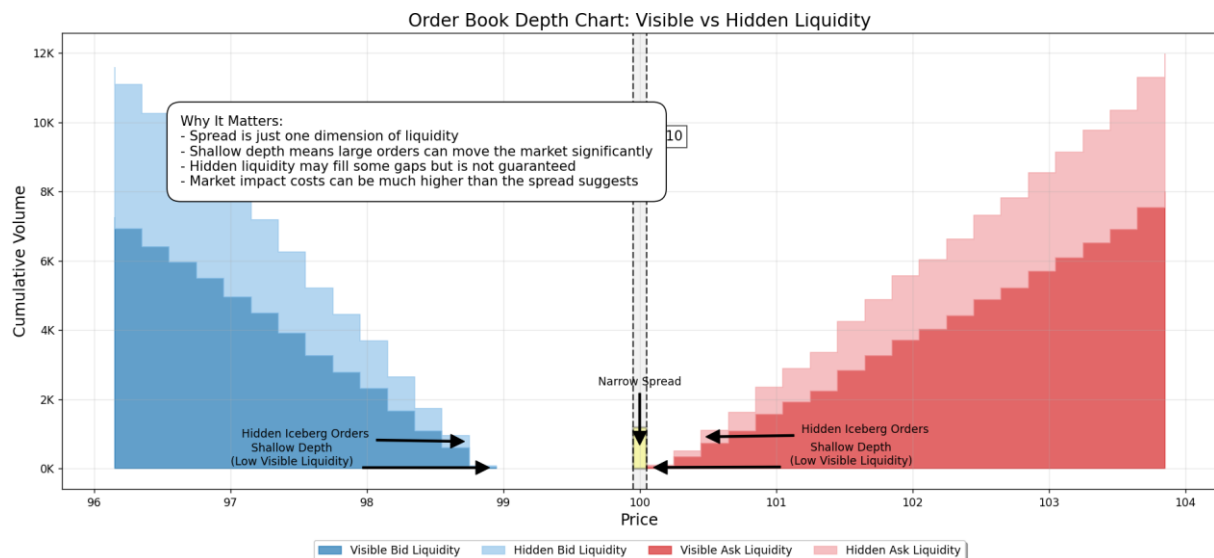


Figure 2: The illusion of liquidity: Visible vs Hidden depth in Order Book Dynamics

This order book depth chart illustrates the critical interplay between visible and hidden liquidity in financial markets, revealing how a narrow spread can mask underlying fragility in market depth. While the bid-ask spread appears tight (0.10 units), the shallow visible liquidity at the best bid and ask prices—highlighted by low cumulative volumes—suggests that large orders could trigger significant price movements. Hidden iceberg orders (light blue/red layers) partially offset this shallowness but remain unreliable, as they are not guaranteed to absorb large trades. The simulated market impact of an 800-unit buy order demonstrates this vulnerability: execution prices rapidly escalate beyond the best ask, exposing the hidden costs of trading beyond spread considerations. These dynamics underscore that liquidity is multidimensional—spanning spread, depth, and resilience—and that relying solely on spread metrics can lead to underestimating execution risks. For institutional traders, this emphasizes the need for sophisticated execution strategies that account for both visible and latent liquidity to mitigate adverse price impact.

### 5.3 Market Examples

While the four dimensions of liquidity provide a universal framework, how they manifest varies significantly between asset classes. An equity trader, a credit portfolio manager, and a crypto market participant all operate within very different liquidity environments — each with its own norms, risks, and microstructural quirks.

#### 5.3.1 Equities: Fast, Fragmented, and Deep — Until They're Not

Equity markets, particularly in developed economies, are among the most electronically mature. They feature:

- Extremely tight spreads, often just a tick wide;
- High order book depth during normal conditions;
- Near-instantaneous execution on lit exchanges.

However, this liquidity is often fragmented across dozens of venues and heavily reliant on high-frequency market makers. While conditions appear deep and fluid during calm periods, liquidity can vanish rapidly in times of stress, as passive participants withdraw to avoid adverse selection.

### ***5.3.2 Credit Markets: Opaque and Episodic***

Corporate bond markets operate under a fundamentally different model. Trading is mostly over-the-counter (OTC), relying on dealers rather than central limit order books. This results in:

- Wide spreads, particularly for less liquid issues;
- Limited transparency, with delayed or partial trade reporting;
- Execution delays, often requiring negotiation or voice confirmation.

Liquidity in credit markets is often episodic. During calm periods, dealers hold inventory and provide two-way quotes. But during stress, they tend to step back, leaving markets effectively one-sided.

### ***5.3.3 Crypto Markets: Fast But Fragile***

Cryptocurrencies operate in a unique liquidity regime. Trading is fully electronic and global, with no central exchange. While the largest coins (like Bitcoin and Ethereum) exhibit:

- Tight spreads on major platforms;
- High transaction frequency;
- 24/7 execution access,

the underlying liquidity is often shallow and non-resilient. Price gaps can occur with relatively modest order sizes, especially in altcoins. Liquidity is also fragmented across dozens of unlinked venues, many lacking robust market surveillance or order protection mechanisms.

This creates an environment where slippage, arbitrage, and flash crashes are common — and where liquidity is often an illusion of activity rather than genuine depth.

## 6 Liquidity Risk

Liquidity is often taken for granted — until it's gone. While previous sections focused on how liquidity is structured and measured, this section explores what happens when it fails. Liquidity risk is not about prices moving against you — it's about markets ceasing to function as expected. This can transform a small position into a large problem, turning theoretical risk into realised loss, especially when trades cannot be executed at any reasonable cost or speed.

### 6.1 *Liquidity Risk: When the Market Disappears*

Liquidity risk is not the risk of bad prices — it is the risk of *no prices*. It emerges when the ability to transact, hedge, or exit a position vanishes — not gradually, but suddenly and severely. Unlike market risk, which reflects price variation, liquidity risk reflects the *inability to trade at the expected price or size*. It is the invisible fissure that runs beneath even the most orderly markets — dormant in calm times, but devastating when triggered.

At its core, liquidity risk is the risk that:

- A trader cannot execute without substantial price impact;
- An investor cannot rebalance a portfolio without delay;
- A hedger cannot unwind a position before the market moves.

This risk is non-linear and reflexive. The act of trading in an illiquid environment can itself create further illiquidity — widening spreads, exhausting order books, and triggering counterparty caution. Markets that appear liquid on the surface — narrow spreads, deep quotes — can *snap* under stress, revealing just how conditional liquidity truly is.

#### 6.1.1 *Apparent vs Actual Liquidity*

Many participants confuse visible liquidity (the top-of-book quotes or displayed depth) with actionable liquidity (the true volume that can be transacted without significant impact). This mismatch is particularly dangerous during volatility spikes, when orders evaporate, market makers retreat, and passive liquidity withdraws faster than prices can adjust.

Liquidity risk is *path-dependent*. What matters is not just where prices are — but *how you get there*. Fast, forced moves through thin books cause disproportionate losses.

#### **Structural and Behavioural Drivers**

Liquidity risk is not just about market structure. It is also behavioural. When uncertainty rises, traders pull quotes, widen spreads, or cancel orders pre-emptively. Market makers protect themselves. Algos go dark. Risk models freeze. The market becomes less a venue for trade, and more a hall of mirrors — reflecting back the fear of its participants.

This fragility is compounded by modern execution protocols:

- Auto-liquidation from margin triggers.



- Stop-loss cascades in thin books.
- SOR misrouting in fragmented venues.

Each can turn friction into fire.

## **6.2 Sources of Liquidity Risk**

Liquidity risk stems from two broad categories of forces: external shocks that disrupt markets from the outside, and internal feedback mechanisms that amplify stress from within. Both can cause market liquidity to evaporate rapidly, transforming seemingly stable conditions into environments where trades cannot be executed without severe price impact.

### **6.2.1 External Shocks (Exogenous Drivers)**

These are unpredictable events originating outside the market that trigger sudden risk aversion and withdrawal of liquidity. They include:

- Macroeconomic surprises, such as unexpected inflation data or central bank rate decisions.
- Geopolitical events, like wars or sanctions, which introduce sudden layers of uncertainty.
- Corporate-specific news, such as earnings announcements or credit downgrades, which can dramatically alter valuations.
- Systemic shocks, such as the COVID-19 pandemic or the collapse of Lehman Brothers, which disrupt not just individual markets, but the entire financial system.

When uncertainty spikes, market makers widen spreads, reduce posted sizes, or disappear from the book entirely. Investors become reluctant to trade, not because they lack conviction, but because they fear being the last to exit.

### **6.2.2 Internal Feedback Loops (Endogenous Amplifiers)**

These risks arise from the structure and behaviour of the market itself. Unlike external shocks, they are triggered by the actions of participants and the dynamics of execution under stress. Examples include:

- *Margin calls*: Falling asset prices increase leverage ratios, prompting brokers to demand more collateral. Investors are forced to sell into weakness, which pushes prices lower still — prompting more margin calls.
- *Stop-loss cascades*: Traders often set automated levels to exit positions. If prices hit these thresholds in quick succession, each stop triggers the next, draining liquidity and accelerating the decline.

- *Fire sales*: Funds facing redemptions or risk limit breaches may be forced to liquidate holdings, regardless of price. This selling pressure deepens the imbalance between supply and demand.
- *Dealer withdrawal*: In over-the-counter (OTC) markets like corporate bonds, liquidity is often provided by dealers. When their inventories are full or their capital is constrained, they may refuse to quote — effectively halting the market.

These mechanisms don't require new information to emerge. Instead, they reflect a *reflexive dynamic*, where price declines create more selling, which leads to thinner markets, which in turn cause further price declines. The system begins to chase its own tail.

In these moments, liquidity risk is no longer a passive background variable. It becomes the dominant force in the market — invisible until it isn't, and then inescapable.

### 6.3 *Effects on Execution and Hedging*

When liquidity dries up, the consequences go far beyond wider spreads or delayed trades. Execution becomes unpredictable, and hedging strategies that rely on continuous markets can fail precisely when they are most needed. In these conditions, models break down, and risk management shifts from being a statistical exercise to a battle against market mechanics.

#### 6.3.1 *Slippage and Execution Failure*

In normal markets, large orders can often be executed with minimal disruption by using smart routing algorithms or slicing techniques. But under stress, even modest orders can cause significant *slippage* — the difference between the expected execution price and the actual fill.

As market depth disappears, orders are filled at increasingly unfavourable levels. Limit orders may go unexecuted for hours, while market orders may walk the book — hitting multiple price levels and accelerating the move.

Moreover, execution risk becomes two-sided: traders must worry not just about getting in, but about *getting out*. In illiquid markets, entry and exit costs become asymmetric and path-dependent — a risk that is rarely priced into models.

#### 6.3.2 *Broken Hedges and Misleading Greeks*

Liquidity risk can render traditional hedging approaches ineffective. Delta hedging, for instance, assumes that small adjustments can be made continuously and cheaply. But when markets gap or spreads explode, such adjustments may be delayed, partial, or unavailable altogether.

This has a direct impact on the interpretation of the Greeks — the sensitivities that underpin most options-based risk management. Under normal conditions:

- Delta estimates how much the option's value changes with the underlying.
- Gamma reflects the curvature — or the speed of change in Delta.

- Vega captures sensitivity to volatility changes.

But when liquidity vanishes, these measures become unreliable. Their assumptions — that prices move smoothly, and that execution is frictionless — no longer hold. The Greeks may still calculate, but they no longer *guide*.

### **6.3.3 Undervalued in Risk Models**

Traditional Value at Risk (VaR) models are poorly equipped to handle liquidity events. Most rely on historical price movements and assume a normal trading environment. They capture volatility, but not marketability — the difference between knowing the value of an asset and being able to realise it.

In reality, the worst losses often occur not because prices fall sharply, but because the market *locks up* when they do. The assumption that positions can be liquidated at last-known prices is often dangerously optimistic.

As a result, liquidity risk systematically hides in the tails of distribution — unseen in calm periods, unmanageable in crisis.

## 7 Microstructure Alpha

While many investors seek returns through macro views, valuation models, or fundamental research, a different class of traders focuses on the **structure** of the market itself. Microstructure alpha refers to profits derived not from predicting price direction, but from exploiting how markets process orders, information, and execution.

This edge lies in the details: who sees what, when, and how quickly they can act on it. In a fragmented, high-speed, rule-driven trading environment, those who understand the **plumbing** of the market — latency, order flow, quote dynamics — can extract value invisibly, and often repeatedly.

### 7.1 Latency Arbitrage and Information Racing

In modern electronic markets, speed is not a luxury — it is a weapon. Latency arbitrage refers to the practice of exploiting tiny time delays between venues, data feeds, or order book updates to secure riskless or near-riskless profits. For the fastest traders, even a few microseconds of informational advantage can be monetised — not because they know more, but because they *know it sooner*.

This arms race — often referred to as *information racing* — is one of the clearest manifestations of microstructure alpha: a source of return derived not from economic insight, but from superior positioning within the market's architecture.

#### How Latency Arbitrage Works

Imagine a stock is quoted at €100 on Exchange A, but due to a slight delay, Exchange B still displays it at €99.98. A high-frequency trader (HFT) connected to both venues can instantly buy on B and sell on A — capturing a near-riskless €0.02 per share before slower participants even see the change.

The key enablers of this strategy are:

- Co-location: HFT firms place their servers physically next to exchange infrastructure to minimise signal transmission time.
- Proprietary data feeds: Faster, more expensive feeds offer fresher quote updates than public alternatives.
- Ultra-low-latency hardware: Custom-built systems can process and respond to information in less than a microsecond.

While each trade may earn fractions of a cent, the scale and frequency — thousands of trades per second — can make latency arbitrage highly profitable.

From a classical view, arbitrage is beneficial: it aligns prices and ensures consistency across venues. But latency arbitrage is more controversial. It doesn't reflect differences in fundamentals — only differences in plumbing.

Critics argue it:

- Extracts value from slower participants without adding liquidity;
- Incentivises exchange fragmentation, as venues cater to HFTs with specialised services;
- Erodes trust in fair access to price discovery.

Proponents counter that it enforces pricing discipline and tightens spreads — but only when markets are functioning well.

In times of stress, these same strategies can amplify instability. When quotes become fleeting and liquidity vanishes, the advantage of being first becomes the risk of triggering collapse.

## 7.2 *Order Flow Prediction and Toxicity*

In modern markets, price isn't the only signal — flow matters just as much. The ability to predict *who is trading, why, and with what impact* has become a cornerstone of execution strategy and market-making. Order flow is not homogenous. Some trades are benign; others are toxic — and distinguishing between the two is where a significant portion of microstructure alpha is generated.

### **Toxic vs. Benign Flow**

From the perspective of a liquidity provider, a trade is *toxic* when it is likely to be followed by adverse price movement. For example:

- A large buy order from an informed institutional desk, executed just before positive news, is highly toxic.
- A market order from a retail app during quiet conditions is typically benign.

Toxicity, in this context, refers to the probability that the counterparty is better informed. Providing liquidity to toxic flow leads to losses — the market maker sells just before prices rise, or buys just before they fall.

To protect themselves, liquidity providers dynamically adjust their quotes:

- Widening spreads when toxicity risk is high;
- Reducing size when adverse selection is likely;
- Withdrawing quotes entirely in fast or opaque conditions.

### **Flow Prediction Models and Execution Algorithms**

To manage this, firms build models to classify and predict the nature of incoming flow. These models ingest:

- Real-time order book changes;
- Historical trade sequences;

- Time-of-day patterns and venue-specific behaviours.

Using this information, execution algorithms adjust how they submit child orders — slicing, routing, or timing execution to minimise signalling. This is known as TCA-informed execution (Transaction Cost Analysis), where real-time analytics help avoid impact and reduce slippage.

Market makers, in turn, use flow prediction to manage their own risk:

- By correlating recent activity with post-trade price movement;
- By tracking venue-level toxicity (some platforms attract more informed flow);
- By tuning their inventory strategies based on anticipated flow direction.

### Measuring Toxicity: VPIN and Beyond

One of the most influential metrics for flow toxicity is VPIN (Volume-Synchronized Probability of Informed Trading). It estimates the imbalance between buy and sell volume over fixed intervals, weighted by volume rather than time. A rising VPIN suggests that one side of the market is dominating — often a sign of informed trading.

Other approaches use machine learning to classify meta-orders (sequences of trades likely originating from the same strategy) or detect *footprints* of institutional activity in the tape.

These tools are no longer optional. In fragmented, high-speed markets, understanding order flow is essential for:

- Minimising impact and avoiding detection;
- Managing exposure to informed participants;
- Pricing liquidity with precision.

In short, flow isn't noise. It's information — and in many cases, it's the most actionable signal in the market.

## 7.3 *Internalisation and Payment for Order Flow (PFOF)*

In many modern equity markets, particularly in the United States and increasingly in Europe, a significant portion of retail order flow never reaches public exchanges. Instead, it is executed internally by brokers or routed to systematic internalisers and market makers who pay for the privilege. This is the domain of internalisation and payment for order flow (PFOF) — practices that have reshaped how liquidity is sourced, priced, and monetised.

### What Is Internalisation?

Internalisation refers to the process by which a broker executes a client's trade against its own inventory, or via a partnered liquidity provider, without routing it to a lit exchange. This allows the broker to:

- Avoid exchange fees;

- Offer fast, often price-improved execution;
- Retain control over flow and order handling.

For the end client, this can mean better pricing on small trades — for example, selling at \$10.01 instead of the national best bid of \$10.00. But it also means their order never contributes to public price discovery or depth.

### **The Mechanics of PFOF**

In a PFOF arrangement, a broker receives a fee — typically fractions of a cent per share — from a market maker in exchange for routing order flow to them. The market maker, in turn, monetises the flow through:

- Spread capture: Buying at \$10.00 and selling at \$10.01;
- Flow prediction: Classifying trades as toxic or benign;
- Inventory recycling: Pairing off internal flow with minimal risk.

Retail trades are often considered low-risk: small, uninformed, and relatively predictable. This makes them highly desirable for market makers seeking flow that won't move the market.

### **Benefits and Trade-offs**

Supporters argue that PFOF benefits all parties:

- Retail traders get zero-commission trades with price improvement;
- Brokers earn revenue without charging fees;
- Market makers access predictable, non-toxic flow.

However, these benefits come with structural concerns:

- Conflicts of interest: Brokers have incentives to route to the highest-paying venue, not necessarily the one offering the best execution.
- Price discovery erosion: Orders executed off-exchange don't contribute to the lit market, reducing transparency and depth.
- Access asymmetry: Sophisticated firms gain insights into aggregated retail behaviour, reinforcing informational advantages.

### **Regulatory and Competitive Landscape**

PFOF has come under scrutiny from regulators in both the US (SEC) and Europe (ESMA), especially following events like the GameStop rally. Concerns include execution quality, transparency, and whether zero-commission models are hiding real costs.

Some jurisdictions (e.g., the UK and Canada) have banned PFOF altogether, favouring routing transparency and exchange-based trading. Others are re-evaluating its role in fair access and market stability.

## 8 Managing Liquidity Risk in Practice

Liquidity risk is easy to underestimate and hard to hedge — but it is not unmeasurable. While traditional risk frameworks focus on volatility and correlation, a growing toolkit now exists to model the cost, time, and uncertainty associated with liquidation. In this section, we explore the practical methods used by institutions to quantify and mitigate liquidity risk — not just in theory, but in the real world of stressed markets and imperfect execution.

### 8.1 *Liquidity-Adjusted Value at Risk (L-VaR)*

Traditional risk models often assume that positions can be exited at the last traded price, with minimal cost and no delay. In reality, this assumption collapses under stress. Large trades move markets, and illiquid assets may take days — or weeks — to unwind. Liquidity-adjusted Value at Risk (L-VaR) bridges this gap by embedding the cost and time of liquidation directly into the risk estimate.

#### 8.1.1 *VaR & L-VaR*

Standard Value at Risk (VaR) answers a simple question: *What is the worst expected loss over a given time horizon, at a given confidence level, assuming normal market conditions and immediate execution?*

But this view is static. It ignores:

- *Market impact*: the price slippage caused by trying to exit size;
- *Execution delay*: the fact that large positions take time to liquidate;
- *Spread risk*: the widening of bid-ask spreads under stress.

L-VaR modifies the traditional VaR framework by accounting for these factors. It penalises portfolios based on illiquidity exposure, adjusting expected losses upward for assets that cannot be exited cleanly.



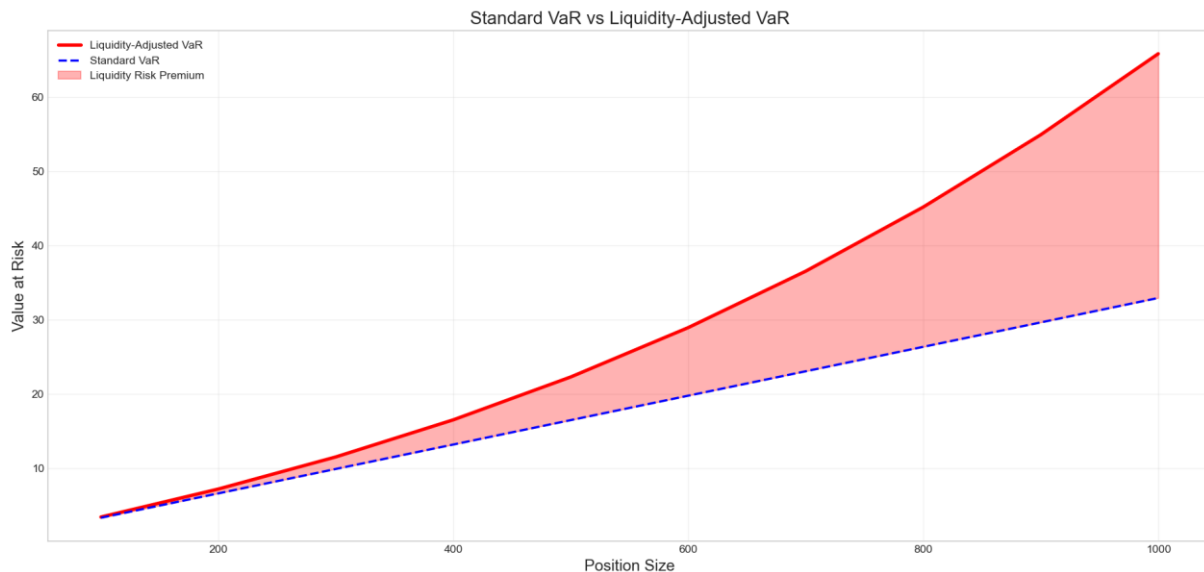


Figure 3: The Liquidity Risk Premium

This analysis demonstrates how liquidity risk creates a growing divergence between standard Value-at-Risk (VaR) and liquidity-adjusted VaR (L-VaR) as position sizes increase. The market impact model reveals a convex relationship where larger trades face disproportionately higher price impacts - for instance, doubling a position from 500 to 1,000 units triples the market impact from 5% to 15% in illiquid conditions. This nonlinear effect causes L-VaR to significantly outpace standard VaR, exceeding it by 22% for medium-sized positions (500 units) and 58% for larger positions (1,000 units), with the gap represented visually by the expanding red "liquidity risk premium" area between the two risk measures. These findings highlight how traditional VaR models systematically underestimate the true risk of large positions, particularly during market stress when liquidity evaporates precisely when needed most, suggesting traders must incorporate liquidity horizons and impact costs into their risk management frameworks, especially for sizeable or illiquid holdings.

### 8.1.2 Implementation: Measuring the Drag

The practical implementation of L-VaR typically involves three components:

1. *Position size relative to market depth:* A position representing 5% of an asset's average daily volume (ADV) will require multiple days to exit. This increases exposure to interim price risk.
2. *Expected market impact functions:* Based on empirical models or proprietary trading data, estimating how much price will move as a function of trade size.
3. *Stress spreads and volume shocks:* Simulating scenarios in which liquidity deteriorates — spreads widen and depth collapses — and integrating that into tail loss estimates.

Mathematically, some models express L-VaR as:

$$L - VaR = VaR + \lambda * Expected Impact Cost$$

where  $\lambda$  is a penalty factor based on liquidity constraints (e.g., urgency, volatility, asset class).

## Why It Matters

By integrating execution cost and exit difficulty into risk measures, L-VaR reflects a more realistic view of portfolio fragility. It penalises crowded trades, levered illiquid positions, and short-term funding mismatches — the same risk factors that have amplified past crises.

L-VaR is particularly relevant for:

- Credit portfolios, where OTC instruments may trade episodically;
- Alternative assets, such as private debt or structured notes;
- Leveraged funds, where forced unwinds can trigger spirals.

As regulators and investors increasingly focus on *systemic liquidity mismatches*, L-VaR is becoming a key tool for pre-emptive risk control.

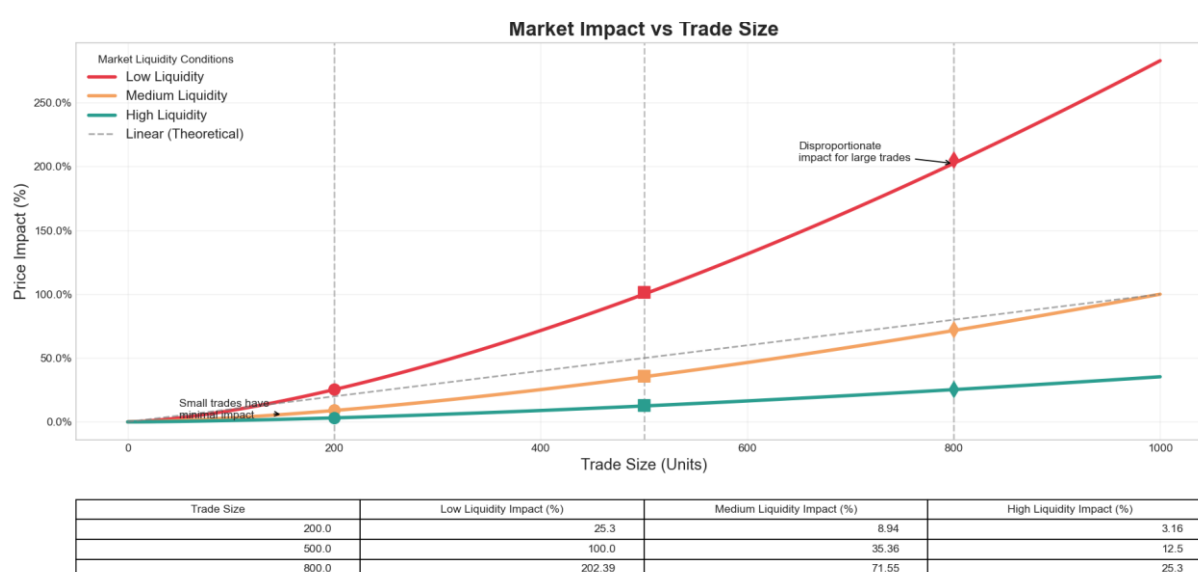


Figure 4: Market Impact Function Heterogeneity Across Liquidity Regimes

The empirical analysis of market impact as a function of trade size reveals pronounced non-linearities across varying liquidity regimes. In low liquidity conditions, the price impact function exhibits a super linear growth pattern, with an impact coefficient increasing from 25.2% at 200 units to 202.39% at 800 units, suggesting significant market fragility and depth constraints.

Medium liquidity environments demonstrate a more moderate convex relationship, with impact coefficients of 8.94%, 36.26%, and 71.55% at respective trade size increments. High liquidity markets maintain remarkable resilience, displaying a near-linear impact function with coefficients ranging from 3.16% to 25.3% across the examined trade size spectrum.

The deviation between observed impact curves and the theoretical linear benchmark (depicted by the dashed line) quantifies the liquidity premium implicit in transaction costs. These findings align with market microstructure theories positing that order book depth diminishes non-uniformly as price levels deviate from the midpoint, particularly in stressed market conditions.

Our results have significant implications for optimal execution strategies and risk management frameworks, especially when calibrating transaction cost models across heterogeneous market states.

## 8.2 *Stress Testing and Scenario Design*

While Value at Risk (VaR) and its liquidity-adjusted variants provide useful statistical estimates, they are inherently backward-looking. They rely on historical distributions, normal conditions, and the assumption that the future will resemble the past. But when liquidity vanishes, markets don't behave normally — they break. This is where stress testing becomes essential: simulating what happens when market assumptions collapse.

### 8.2.1 *Why Stress Testing Matters*

Stress testing answers questions that VaR cannot:

- What if volumes fall by 70% in a week?
- What if spreads triple while volatility doubles?
- What if counterparties stop quoting altogether?

Rather than relying on probabilistic models, stress tests build *narratives* — plausible but extreme market conditions — and quantify their impact on portfolios, trading desks, or entire institutions.

These tests are essential for:

- Risk committees, to understand portfolio fragility;
- Traders, to plan exits under duress;
- Regulators, to assess systemic vulnerabilities.

### 8.2.2 *Core Components of Liquidity Stress Scenarios*

Effective liquidity stress tests include multiple layers:

1. *Volume collapse*: Assume a sudden drop in market turnover. For example, reducing average daily volume by 80% for selected assets. This tests whether positions can still be unwound within expected timeframes.
2. *Spread blowout*: Widen bid-ask spreads by asset class-specific stress multipliers. In March 2020, spreads in investment-grade credit moved from 10–20 bps to over 300 bps — not hypotheticals, but historical precedents.
3. *Execution delays*: Model delays in filling orders — from milliseconds to hours — especially in less liquid instruments or fragmented markets.

4. *Dealer withdrawal*: For OTC assets, simulate scenarios in which one or more counterparties refuse to make markets. This stresses bilateral risk and highlights exposure to concentrated liquidity providers.
5. *Interconnected effects*: Link market illiquidity to margin calls, forced liquidations, or redemption risk. Liquidity crises rarely stay confined to price mechanics — they cascade through funding and leverage channels.

## 9 Conclusion

Financial markets are often described as ecosystems of risk and return, but they are also machines — engineered systems of rules, venues, and incentives that transform intentions into trades. When these systems function well, liquidity appears abundant, execution feels seamless, and price discovery seems natural. But as we've seen, this surface stability can be dangerously deceptive.

Microstructure matters. It shapes how prices form, how orders are matched, and how risks materialise when speed, depth, and transparency are no longer guaranteed. A narrow spread or tight execution is not a measure of robustness — it is often a fragile equilibrium that can flip under stress. And when liquidity disappears, the cost is not just financial — it's strategic.

Liquidity risk is not about volatility. It's about access. When markets seize up, even correct views or well-positioned trades can fail, simply because the exit door narrows or vanishes altogether. That is why modern trading and risk management cannot afford to treat execution as an afterthought. It must be integrated into portfolio construction, scenario analysis, and systemic oversight.

Ultimately, those who understand the market's *plumbing* — how flow, latency, and fragmentation interact — are better equipped to navigate both calm and crisis. Traders, risk officers, and regulators alike must shift their focus from what trades to *how* trades happen, and at what cost.

## 10 Python code

```

1  # Order book data
2  mid_price = 100.0
3  spread = 0.10 # narrow spread
4
5  price_levels_buy = np.arange(mid_price - spread/2 - 1.0, mid_price - spread/2 - 4.0, -0.2)
6  price_levels_sell = np.arange(mid_price + spread/2, mid_price + spread/2 + 4.0, 0.2)
7
8  volumes_buy = np.concatenate([
9      [50], # Small volume at the best bid (shallow)
10     np.random.normal(500, 100, len(price_levels_buy) - 1) # Deeper away from the spread
11 ])
12 volumes_sell = np.concatenate([
13     [60], # Small volume at the best ask (shallow)
14     np.random.normal(450, 100, len(price_levels_sell) - 1) # Deeper away from the spread
15 ])
16
17 # Add hidden liquidity (iceberg orders and dark pool liquidity)
18 hidden_volumes_buy = volumes_buy * 0.6 # 60% additional liquidity hidden
19 hidden_volumes_sell = volumes_sell * 0.5 # 50% additional liquidity hidden
20
21 cum_volumes_buy = np.cumsum(volumes_buy)
22 cum_volumes_sell = np.cumsum(volumes_sell)
23 cum_total_buy = np.cumsum(volumes_buy + hidden_volumes_buy)
24 cum_total_sell = np.cumsum(volumes_sell + hidden_volumes_sell)

```


Figure 5: Orderbook liquidity volume python code

```

1  plt.fill_between(price_levels_buy, cum_volumes_buy, color=visible_buy_color, alpha=0.7,
2                  label='Visible Bid Liquidity', step='post')
3
4  plt.fill_between(price_levels_buy, cum_volumes_buy, cum_total_buy, color=hidden_buy_color,
5                  alpha=0.7, label='Hidden Bid Liquidity', step='post')
6
7  plt.fill_between(price_levels_sell, cum_volumes_sell, color=visible_sell_color, alpha=0.7,
8                  label='Visible Ask Liquidity', step='post')
9
10 plt.fill_between(price_levels_sell, cum_volumes_sell, cum_total_sell, color=hidden_sell_color,
11                 alpha=0.7, label='Hidden Ask Liquidity', step='post')

```

Figure 6: Orderbook liquidity volume plot 1



```

1 plt.annotate('Shallow Depth\n(Low Visible Liquidity)',
2             xy=(price_levels_buy[0], volumes_buy[0] / 2),
3             xytext=(price_levels_buy[0] - 1.5, volumes_buy[0] * 2),
4             arrowprops=dict(facecolor='black', shrink=0.05, width=1.5),
5             fontsize=10, ha='center')
6
7 plt.annotate('Shallow Depth\n(Low Visible Liquidity)',
8             xy=(price_levels_sell[0], volumes_sell[0] / 2),
9             xytext=(price_levels_sell[0] + 1.5, volumes_sell[0] * 2),
10            arrowprops=dict(facecolor='black', shrink=0.05, width=1.5),
11            fontsize=10, ha='center')
12
13 plt.annotate('Hidden Iceberg Orders',
14             xy=(price_levels_buy[1], (cum_volumes_buy[1] + cum_total_buy[1]) / 2),
15             xytext=(price_levels_buy[1] - 1.2, cum_volumes_buy[1] + 300),
16             arrowprops=dict(facecolor='black', shrink=0.05, width=1.5),
17             fontsize=10, ha='center')
18
19 plt.annotate('Hidden Iceberg Orders',
20             xy=(price_levels_sell[2], (cum_volumes_sell[2] + cum_total_sell[2]) / 2),
21             xytext=(price_levels_sell[2] + 1.2, cum_volumes_sell[2] + 300),
22             arrowprops=dict(facecolor='black', shrink=0.05, width=1.5),
23             fontsize=10, ha='center')

```

Figure 7: Orderbook liquidity volume plot 2

## 11 References

- [1] Hull, J. C. (2018). *Options, Futures, and Other Derivatives* (10th ed.). Pearson.
- [2] O'Hara, M. (1995). *Market Microstructure Theory*. Blackwell.
- [3] Bouchaud, J.-P., Farmer, J. D., & Lillo, F. (2009). *How Markets Slowly Digest Changes in Supply and Demand*. In *Handbook of Financial Markets: Dynamics and Evolution*, Elsevier.
- [4] Hasbrouck, J. (2007). *Empirical Market Microstructure: The Institutions, Economics, and Econometrics of Securities Trading*. Oxford University Press.
- [5] Easley, D., López de Prado, M., & O'Hara, M. (2012). *Flow Toxicity and Liquidity in a High-frequency World*. *The Review of Financial Studies*, 25(5), 1457–1493.
- [6] Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and High-Frequency Trading*. Cambridge University Press.
- [7] Biais, B., Foucault, T., & Moinas, S. (2015). *Equilibrium Fast Trading*. *Journal of Financial Economics*, 116(2), 292–313.
- [8] Taleb, N. N. (1997). *Dynamic Hedging: Managing Vanilla and Exotic Options*. Wiley.
- [9] Chan, E. (2013). *Algorithmic Trading: Winning Strategies and Their Rationale*. Wiley.
- [10] Alexander, C. (2008). *Market Risk Analysis, Volume I: Quantitative Methods in Finance*. Wiley.
- [11] Alexander, C. (2008). *Market Risk Analysis, Volume II: Practical Financial Econometrics*. Wiley.
- [12] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer.
- [13] SEC (2020). *Staff Report on Equity and Options Market Structure Conditions in Early 2021* (GameStop Report). U.S. Securities and Exchange Commission.
- [14] Menkveld, A. J. (2013). *High-Frequency Trading and the New Market Makers*. *Journal of Financial Markets*, 16(4), 712–740.