Design of the U.S. Treasury Secondary Market: Venue Sorting, Competing Mechanisms, and Allocative Efficiency*

Jiaer Janet Liu†

October 30, 2026

Latest Version

Abstract

Is the current fragmented structure of the U.S. Treasury secondary market, in which centralized and decentralized venues coexist, optimal? This paper develops a venue-choice model to study allocative efficiency under current and alternative market structures as aggregate risk varies. Venues differ in their pricing and matching mechanisms, and agents sort across venues based on the tradeoff between execution certainty and transaction costs. When aggregate risk is low, gains from trade are primarily realized through trading in the centralized venue, the central limit order book, which mitigates inefficiencies associated with search frictions in the decentralized over-the-counter venue. However, as aggregate risk increases, adverse selection intensifies for limit orders, discouraging limit-order submission and reducing order book depth, while agents substitute toward decentralized over-the-counter trading and recover gains from trade that would otherwise be lost. Quantifying the model using transaction-level U.S. Treasury data, this paper finds that the current hybrid market structure achieves 87 percent of the first-best allocation, outperforming both over-the-counter-only and fully centralized market structures. These findings rationalize the current fragmented market.

Keywords: venue, central limit order book, over-the-counter, allocative efficiency, US Treasuries.

*I am greatly indebted to my dissertation chair, Pierre-Olivier Weill, for his invaluable guidance. I am deeply grateful to my dissertation committee, including John Asker, Andrew Atkeson, Martin Hackmann, and Lee Ohanian, for their constructive advice on this project and their continuous mentorship throughout different stages of my Ph.D. I am especially thankful to Darrell Duffie, Zhiguo He, David Baqaee, Saki Bigio, Alexander Bloedel, Simon Board, Joao Guerreiro, Oleg Itskhoki, Patrick Kennedy, and Will Rafey for their insightful comments. I thank the participants of the UCLA Macro-Finance Reading Group, UCLA Macro Proseminar, UCLA IO Proseminar, UCLA Theory Proseminar, and 2026 Macro-Finance Research (MFR) Summer Session at UChicago for their comments. Finally, I would like to express my gratitude to the UCLA U.S. Treasury TRACE Data Study Group, in particular Mahyar Kargar, Shuo Liu, and Lian Chen. Their insights have been invaluable in helping me understand the dataset, and I have greatly benefited from their generous willingness to share their knowledge. Any comments or suggestions would be greatly appreciated.

†Department of Economics, University of California, Los Angeles. E-mail: jiaerliu98@g.ucla.edu

1 Introduction

In modern financial markets, many assets trade through multiple trading mechanisms, raising concerns about the inefficiencies associated with such market fragmentation and generating considerable academic and policy debate over the optimal organization. A prominent example is government bonds, which can be traded through a centralized trading venue, the central limit order book (CLOB), as well as through decentralized over-the-counter (OTC) trading. This form of fragmentation is common across jurisdictions, including the United States, Europe, and China. This paper uses the U.S. Treasury secondary market as a setting to investigate this market-design concern. Is this fragmented market structure allocatively efficient? Today, the U.S. Treasury secondary market has about $1 trillion in daily trading volume, so even small improvements in market design can generate substantial welfare gains in dollar terms. Thus, if the current fragmented structure is not allocatively efficient, appropriate regulatory measures are worth considering. Conversely, if the current fragmented structure is allocatively efficient, why fragmentation arises and why the coexistence of centralized and decentralized trading constitutes an optimal market design deserve an economic explanation.

To answer the research question, I develop a venue-choice model in a competitive environment featuring heterogeneous agents and aggregate risk. The model formalizes the trading mechanism of each *venue* and explains how agents choose among venues. The model is populated by a continuum of heterogeneous agents who differ in their preferences for Treasury bonds. Aggregate risk is modeled as a common preference shock that shifts all agents' preferences by the same amount. In the model, the trading mechanisms of the two venues differ along two dimensions: the intensive margin, pricing, and the extensive margin, matching. In the central limit order book, limit orders are submitted first, quoting bid or ask prices on a discrete price grid determined by the tick size. These limit orders then rest on the book as standing commitments to trade at the quoted prices. Later, market orders arrive and execute immediately against these standing limit orders at the best available quotes. Thus, matching is centralized and governed solely by price priority. In contrast, in the decentralized over-the-counter venue, each agent randomly searches for a counterparty and, once they meet, trade occurs only if there are gains from trade between them, and transaction prices are determined through bilateral Nash bargaining. The centralized venue competes with the decentralized venue for agent participation.

The theoretical solution of the model exhibits endogenous venue sorting and consequently endogenous market fragmentation. The equilibrium sorting is determined by the speed–cost trade-off. Agents with high urgency to trade sort into the centralized venue, which offers greater execution certainty despite higher trading costs. In contrast, agents with low urgency to trade sort into the decentralized over-the-counter venue. Within the centralized venue, agents with the highest urgency

to trade submit market orders, while those with intermediate urgency to trade provide liquidity by submitting limit orders. However, as aggregate risk increases, fewer agents choose to submit limit orders due to the adverse-selection concern — the possibility that aggregate shocks move the asset's common value against standing limit orders, rendering their quotes stale and exposing them to being picked off by other agents using market orders, thereby creating a winner's curse.

This equilibrium venue-sorting pattern has theoretical implications for how allocative efficiency varies with aggregate risk. Allocative efficiency is defined as the fraction of the first-best (maximal attainable) gains from trade that is realized through trading in the venues. When aggregate risk is low, gains from trade are primarily realized through the centralized venue, as agents with high urgency to trade endogenously select into this venue and trade directly through it, thereby avoiding the inefficiency arising from search frictions associated with decentralized over-the-counter trading that would otherwise occur in the absence of the centralized venue. However, the ability of the centralized venue to realize these gains from trade depends on order book depth, i.e., the mass of limit orders. As aggregate risk increases, adverse selection discourages limit-order submission and reduces order book depth, limiting the volume of trade and thus the gains from trade realized through the centralized venue. In extreme cases, sufficiently high aggregate risk can cause the order book to become empty, preventing any trading through the centralized venue. In contrast, the over-the-counter venue allows agents to continue trading outside the order book, recovering gains from trade that would otherwise be lost in a purely centralized market due to a thin or empty order book under high aggregate risk. Over-the-counter trading thus sustains gains from trade in a high aggregate-risk market environment.

Building on these theoretical results, the quantitative analysis answers the research question and provides implications for market design aimed at maximizing allocative efficiency. First, in terms of allocative efficiency, the coexistence of centralized and decentralized trading venues constitutes the optimal market structure. This benchmark hybrid market structure achieves 87 percent of the first-best allocation, outperforming alternative market structures, including a fully centralized market. Second, the current tick size may be too large given the observed level of aggregate risk in the U.S. Treasury secondary market, and the model provides an estimate of the optimal tick size. That said, because aggregate risk varies over time, the optimal tick size also varies over time, making it difficult for a fixed tick size to remain optimal under all market conditions, regardless of whether the tick size is currently misspecified or correctly specified. The coexistence of the over-the-counter venue mitigates the inefficiencies arising from this inherent imperfection in the design of the central limit order book. The quantitative findings, together with the aforementioned theoretical implications for allocative efficiency, rationalize and support maintaining a fragmented market.

This paper uses transaction-level data from the U.S. Treasury secondary market in September 2022. I merge two datasets: the full order book data from BrokerTec, a central limit order

book platform operated by the Chicago Mercantile Exchange (CME), and the Trade Reporting and Compliance Engine (TRACE) transaction data. My empirical analysis focuses on the interdealer segment of the market, because both over-the-counter and central limit order book transactions are observed and dealers actively choose between the two venues. The merged dataset contains all on-the-run U.S. Treasury nominal coupon securities.

The setup of the venue-choice model is motivated by several stylized facts in the data. First, I document that the majority of interdealer over-the-counter transaction prices fall between the best bid and best ask quoted on the order book at the same second, suggesting that traders take transaction costs into account when choosing a venue. Second, cross-venue substitution patterns in the data suggest that execution certainty is another important determinant of venue choice. I document that, relative to non-banks, banks substitute away from over-the-counter trading toward the central limit order book, in particular market orders, as the end-of-day capital assessment approaches, in pursuit of execution certainty (speedy trade). Third, aggregate risk is quantitatively payoff-relevant at the trading horizon. On average, I observe four episodes per day in which the price of the 10-year U.S. Treasury moves monotonically by more than 10 basis points[1] within an interval shorter than two minutes, the average time that a limit order waits in the order book before execution.

These stylized facts motivate a model that incorporates both the pricing (transaction-cost) and matching (execution-probability) mechanisms of each venue in an environment with aggregate risk. Specifically, because a limit order represents a standing commitment to trade at a firm quote, new information may arrive before execution and induce all agents to revalue the asset, rendering the quote stale. Consequently, the ex post payoff from trading using a limit order depends on the aggregate shocks that arrive *between* limit-order submission and execution. Aggregate risk is therefore payoff-relevant and affects venue choice because different trading venues expose agents to this form of risk to different degrees. To capture this idea, the model features a sequential setup in which limit orders are submitted first, the aggregate state is subsequently realized, and then market orders are submitted and immediately executed against the existing limit orders. Over-the-counter trading takes place in the final stage because execution through this venue is slower than through the central limit order book, and also for tractability. Modeling the arrival of an aggregate shock between the limit-order stage and the market-order stage also mechanically generates fluctuations in venue participation across aggregate states, a pattern that is also empirically relevant.

The equilibrium outcomes of the model, including venue market shares, the bid–ask spread,

[1]The 10 bp threshold is heuristic but is intended to capture economically meaningful price movements. A substantially smaller threshold, such as 1 bp, would include bid–ask bounces since the CLOB tick size for the 10-year U.S. Treasury is 1.5625 bp, whereas a substantially larger threshold, such as 100 bp, would exclude most relevant market events. As a reference, consider the Federal Open Market Committee announcement on September 21, 2022. Immediately following the announcement of a 75 bp increase in the rate, the price of the 10-year U.S. Treasury declined by approximately 40 bp. This reference suggests that a 10 bp threshold is of the appropriate order of magnitude.

over-the-counter price dispersion, order-flow patterns, and limit-order fill rates, are broadly consistent with the data under reasonable parameter values. Specifically, the model implies that the majority of trading is executed through the central limit order book, with the over-the-counter venue serving as an outside option, consistent with the empirical evidence. The model also predicts that the bid–ask spread binds at one tick, in line with the empirical fact that the bid–ask spread equals one tick for the vast majority of the trading day. In addition, the model implies that many over-the-counter transactions occur at prices between the best bid and the best ask, a pattern that is also observed in the data. Importantly, the model highlights the quantity margin as the channel through which the market incorporates aggregate risk ex ante and absorbs aggregate shocks ex post: ex ante limit-order mass decreases with the level of risk, while ex post market-order flow becomes one-sided. This theoretical prediction is observed in the data around scheduled announcement events. This prediction that limit-order mass decreases in a high-risk environment is also empirically relevant during COVID-19 crisis period (Duffie (2025)).[2] The model also predicts low limit-order execution rates, with a majority of limit orders remaining unfilled and failing to trade. This pattern is also observed in the data, although the empirical fraction of unfilled limit orders is even larger.

I estimate the model parameters by minimizing the distance between the model-implied and empirical moments and evaluate and use the model through counterfactual analyses. I select moments that are informative about venue choice, pricing, and matching, including venue market shares, price dispersion, price skewness, and the asymmetry of limit-order-book depth. I conduct numerical comparative statics to shed light on parameter identification. I then conduct a quasi-experimental event study exploiting a change in tick size to examine its effects on market liquidity and assess the model's predictions. Specifically, on November 19, 2018, BrokerTec reduced the tick size for the 2-year U.S. Treasury by half, after which CLOB trading volume increased by 30%. The model predicts an increase of 14.16%, which has the same sign though a smaller magnitude. Importantly, I use the model to compute allocative efficiency under two counterfactual market structures: a fully centralized, CLOB-only market and a fully decentralized, OTC-only market. I implement the CLOB-only counterfactual by setting the OTC search parameter to its polar value of zero, thereby shutting down the OTC venue, and implement the OTC-only counterfactual by removing the CLOB trading from the model. Both counterfactual market structures yield lower allocative efficiency than the current hybrid structure.

Rather than fully centralizing the market, this paper points to two alternative policy tools that can improve allocative efficiency while preserving the current hybrid structure of the U.S. Treasury secondary market, one for normal times and the other for periods of market stress. First, to support market functioning during normal times, policymakers can optimize the tick size. Second, to sustain

[2] As documented by Duffie (2022) (Figure 5.8), Duffie (2025), during the COVID-19 crisis, "dealer-to-dealer market depth (measured using BrokerTec CLOB 10-year Treasury data) worsened by a factor of more than ten".

market functioning during periods of market stress, the Federal Reserve can intervene through direct asset purchases, as it did during the COVID-19 crisis. These two policy tools follow directly from the implications of the model's theoretical results for allocative efficiency. During normal times, i.e., when aggregate risk is zero or sufficiently small, an appropriately designed tick size would allow the market itself to achieve high allocative efficiency, reducing the need for more radical mandates, such as full centralization. However, when aggregate risk is high, allocative efficiency deteriorates substantially even with both venues operating, as the mass of limit orders becomes very small, as predicted by the theory, and search frictions in the over-the-counter venue presumably increase as well. Together, these two forces substantially reduce successful trade executions and the gains from trade realized. In this case, direct market intervention may be necessary, such as Fed asset purchases. Ex post Fed asset purchases can improve allocative efficiency by turning otherwise unsuccessful selling attempts into successful trade executions, thereby correcting the large-scale ex post misallocation arising from widespread failures of trade execution, i.e., a market failure ("market freeze"). This market-design-based rationale may be relevant to the COVID-19 crisis, although many other considerations also motivated the Fed's asset purchase program during this dash-to-cash episode.

To conclude, although this paper uses data from the U.S. Treasury secondary market, its insights may extend beyond this setting, such as to other asset markets or even beyond. In fact, benchmark government bond markets in the Euro area, UK, and China also exhibit a similar fragmented market structure, with a central limit order book (CLOB) coexisting alongside an over-the-counter (OTC) venue. Since the quantitative results may depend on market-specific institutional features and therefore should be extrapolated with caution, I summarize three theoretical insights into market design that may be applicable across countries and discuss their practical relevance.

First, the centralized matching *directs* assets from owners to non-owners who participate in the venue, thereby mitigating the inefficiency associated with search frictions and random matching in the over-the-counter venue. In practice, centralized venue aggregates and consolidates order information from a wide range of market participants into a common matching engine, thereby allowing the venue to reallocate assets across a very large population and direct them to those who value them most. This is fundamentally different from decentralized over-the-counter trading, even under modern electronic request-for-quote arrangements, where transactions are typically conducted within small groups or between pairs of traders rather than among a very broad set of market participants. Inefficiency instead arises primarily from rationing, that is, from an imbalance between the masses of limit and market orders that prevents some trades from being executed. The model shows that the only environment in which neither limit orders nor market orders are rationed is one with a zero tick size and no aggregate risk, which also coincides with the Walrasian equilibrium. However, such conditions are unlikely to hold at all times in practice.

Second, rather than focusing on private information, this paper builds on the existing insight that common-value risk can microfound the adverse-selection problem faced by limit orders, owing to the sequential trading structure and firm quotation system of the central limit order book, and further recognizes that flexible price negotiation in the over-the-counter venue immunizes against this problem. To illustrate, consider a case in which negative news arrives. In the model, excess sell market orders will follow the negative news ("selling pressure"), increasing the likelihood that buy limit orders are executed, relative to the case in which positive news arrives. As a result, limit-order execution is adversely selected. That is, in this example, the common value moves against buy limit-order traders: they purchase the asset at a quote based on stale information that reflects an erroneously high common value and subsequently hold the asset when it becomes less valuable. This is reminiscent of the winner's curse in common-value auctions studied by Milgrom and Weber (1982).[3] Over-the-counter trading avoids this adverse selection because counterparties negotiate using symmetric information. For example, in practice, an agreement to trade is reached simultaneously through a phone call, rather than through an asynchronous trading process that mechanically makes the quoting party (limit-order trader), who moves first, less informed than the executing party (market-order trader), who moves second, thus creating asymmetric information and adverse selection. This common-value channel of adverse selection may be particularly relevant for government bond secondary markets, where public macroeconomic and monetary news is important for asset prices.

Third, the model suggests two possible explanations for why in some fixed-income markets an exchange-style central limit order book technology is available yet remains inactive in equilibrium, with all trades occurring instead over-the-counter. This phenomenon has remained puzzling in the literature, as discussed in Hugonnier, Lester, and Weill (2025) Chapter 12. In this model, the first explanation is a market-design problem, whereby the tick size is mis-specified, and the second is an asset-specific characteristic, whereby aggregate risk is sufficiently large. These two explanations are also interrelated, since the optimal tick size is endogenous to the level of aggregate risk. Furthermore, the model suggests that, in the absence of aggregate risk, a large tick size acts only as a wedge that distorts prices away from the efficient price and thus distorts the allocation. However, in the presence of aggregate risk, a large tick size also acts as an incentive device that discourages informed market orders, mitigates the adverse-selection problem faced by limit-order traders, and thereby encourages limit-order submission, allowing book depth to recover toward its no-risk level and increasing the amount of trade that can be realized through the central limit order book. The optimal tick size thus balances ex post distortion against ex ante incentives.

The results should, however, be interpreted in light of several limitations. First, the trading

[3] In common-value auctions, winning is itself informative, and the value of the asset must therefore be evaluated conditional on winning. Here, the execution of a limit order plays the analogous role of winning in an auction.

mechanisms considered here may represent only a subset of those used in practice, and the implications for allocative efficiency may change when additional mechanisms coexist. Second, even within the two venues studied here, several potentially important features are abstracted from, including time priority in limit order book matching and the role of order size. Centralization versus decentralization is a long-standing question in the organization of industries, and the increasing availability of detailed data may allow future research to study this question in greater depth.

Related Literature

This paper builds on previous theoretical work on venue choice between centralized and decentralized markets but differs in at least one of the following two aspects (e.g., Liu, Vogel, and Zhang (2018), Glode and Opp (2020), Colliard, Foucault, and Hoffmann (2021), Dugast, Üslü, and Weill (2022), Lee and Wang (2025), Yoon (2026)). First, the centralized market is not modeled as a frictionless market in which a unique price clears the market,[4] but instead as the trading mechanism that operates in practice: the central limit order book. This distinction is crucial for carefully answering counterfactual questions such as full centralization and is also necessary to match the granular transaction-level data.[5] Modeling frictions in the centralized market is also theoretically exciting because it now allows one to study how the inefficiencies in each market interact. Second, within the central limit order book, the choice between limit orders and market orders is endogenous, in addition to the choice between centralized and decentralized markets. Endogenizing liquidity provision (i.e., agents *choose* to submit limit orders) is relevant for theoretically studying market liquidity and (dys)functioning, especially in environments with high aggregate risk. Empirical evidence also suggests that traders actively choose between these two order types, and in modern CLOBs, no trader is forced to quote (i.e., no trader is exogenously assigned to be a market maker who must submit quotes).

Accordingly, this paper is grounded in the market microstructure literature and seeks to bridge these two strands of the literature: over-the-counter and limit order book. On the over-the-counter side, this paper draws on Duffie, Gârleanu, and Pedersen (2005), Weill (2020), and Hugonnier, Lester, and Weill (2025) for the heterogeneous-agent setup and random matching. On the limit order book side, I draw on and combine two ideas. The first comes from Foucault (1999), whose setup features a common value of the asset that can change over time and thus generates pick-off risk (i.e., adverse selection) for limit orders whose quotes are fixed over time. In his setting, this creates a winner's curse problem for limit-order traders because limit orders are more likely to be

[4]There are several variations of this approach, such as demand schedules or a Walrasian market. However, these variations in previous works share the same structure in which a unique frictionless price clears the centralized market, which is not the case in practice.

[5]For example, whether the trade is executed at the ask price or the bid price.

executed at a loss when their quotes become mispriced.[6] This is also an equilibrium result in my setting. The second idea comes from Dugast, Marta, and Riva (2026), which emphasizes the impact of the quantity margin, i.e., the mass of limit orders, on inefficiency. This is also a focus of this paper, and I similarly view the tick size as a design parameter that affects the depth of the book. When the model combines these two ideas, the interpretation of the tick size as a design parameter can be further framed as an ex ante incentive device for addressing the adverse-selection risk faced by limit-order traders due to informed market orders.

Once venue choice is modeled, the framework developed in this paper speaks to market fragmentation and its welfare implications, such as Chen and Duffie (2021) and Babus and Parlatore (2022). This paper differs from these works in that they study fragmentation within a single trading-mechanism arrangement,[7] rather than fragmentation across distinct trading mechanisms. A very large reduced-form finance literature[8] studying trading-mechanism choice and the effects of trading mechanisms on bond-market outcomes motivates a structural treatment of this topic, which appears to be underexplored in the literature.

The welfare implications contribute to the recent IO literature on financial markets (e.g., Brancaccio, Li, and Schürhoff (2017), Allen and Wittwer (2023), Brancaccio and Kang (2025), Wittwer and Allen (2026)) by incorporating aggregate risk into the environment. This allows one to study how allocative efficiency varies across aggregate market conditions, which speaks to market resilience, i.e., market (dys)function when aggregate risk is high. In this way, it speaks to the ongoing debate over the effective design of the U.S. Treasury secondary market for ensuring the resilience of this global safe asset, a concern raised by previous episodes of market stress (Duffie (2023), Duffie (2025)). In this paper, I show that the central limit order book can become dysfunctional when aggregate risk is high, suggesting that fully centralizing over-the-counter trading may not be a robust reform approach. This directly responds to the open question left by Allen and Wittwer (2023), where the authors argue that studying a market organized through a central limit order book is the next step toward understanding the effects of full centralization.

Another technique that this paper draws from the IO-spatial literature is the use of a matching function to study search frictions (e.g., Buchholz (2022), Brancaccio, Kalouptsidi, et al. (2023)). This approach is used not only for tractability when combining two marketplaces, but also because, with a matching function, one can immediately see that the total amount of trade, and thereby the gains from trade, realized through the centralized venue is capped by the total mass of limit orders submitted to the book. Therefore, the matching-function approach provides a straightforward way to introduce the aforementioned focus on the quantity margin into the equilibrium and study its

[6] This pick-off idea is similar to the stale-quote sniping described in Budish, Cramton, and Shim (2015).

[7] That is, exchange trading across multiple exchange platforms, or dealer-intermediated over-the-counter trading with a varying number of active dealers

[8] Please see Footnote 2 of Allen and Wittwer (2023) for a short list of papers using reduced-form approaches.

impact on allocative efficiency.

Finally, this paper also builds on the empirical literature on U.S. Treasury market liquidity in two ways. First, building on previous empirical work using BrokerTec limit order book data (e.g., M. Fleming, Nguyen, and Ruela (2024), Adrian, M. Fleming, and Nikolaou (2025)), I further combine this dataset with TRACE data to provide a more granular picture of the liquidity landscape of the U.S. Treasury secondary market. Second, because the equilibrium here features thin order book depth in high-risk environments, this paper also rationalizes some empirical facts related to dysfunction in the U.S. Treasury market during the COVID-19 crisis, as documented by Duffie, M. J. Fleming, et al. (2023), Duffie (2025), and references therein.

2 Data and Motivating Facts

2.1 Data

This paper uses transaction-level data from the U.S. Treasury secondary market in September 2022. I merge two datasets: the full order book data from BrokerTec, a central limit order book (CLOB) platform operated by the Chicago Mercantile Exchange (CME),[9] and transaction-level data from the Trade Reporting and Compliance Engine (TRACE). The merged dataset covers all on-the-run nominal U.S. Treasury securities across the 2-, 3-, 5-, 7-, 10-, 20-, and 30-year maturities. For reference, the yield to maturity of the on-the-run 10-year U.S. Treasury averaged approximately 3.5-3.7% during September 2022, and a 10 basis point change in its transaction price corresponded to approximately a 1.3 basis point change in yield.

The empirical analysis focuses on the ATS & Interdealer segment of the market,[10] where both over-the-counter and central limit order book transactions are observed and traders actively choose between the two venues. By contrast, the data indicate that customers rarely, if ever, transact through the central limit order book, instead trading almost exclusively with dealers over-the-counter. Given the paper's theoretical focus on endogenous venue choice, I restrict the analysis to the ATS & Interdealer segment, where the data suggest that traders face no apparent barriers to accessing the CLOB and can therefore endogenously choose between venues. I leave the dealer-to-customer segment and the absence of central limit order book trading in that segment for future research.

I construct the empirical sample as follows. First, I identify all direct transactions in TRACE that are not executed through an alternative trading system (ATS, including the BrokerTec CLOB) and classify them as over-the-counter transactions. Second, I combine these observations with the

[9] BrokerTec refers to its electronic trading platform as the central limit order book (CLOB). For additional information, see https://www.cmegroup.com/markets/brokertec/brokertec-clob.html.

[10] See https://www.finra.org/finra-data/browse-catalog/about-treasury for the definitions of the ATS & Interdealer and Dealer-to-Customer segments.

BrokerTec full order book data, which identifies transactions executed through the central limit order book. The BrokerTec CLOB accounts for only approximately 60% of ATS trading volume reported in TRACE, while the trading protocols underlying the remaining ATS trades cannot be determined with confidence; therefore, I exclude them from the analysis.

The TRACE data provide transaction-level information, including execution price, quantity, execution time, CUSIP, and the types of both counterparties. Unlike U.S. corporate bond TRACE, the data do not include anonymous dealer identifiers; only the counterparty type is observed. The types of trading entities in the ATS & Interdealer segment include bank dealers, non-bank dealers, and principal trading firms (PTFs). Throughout the paper, I analyze PTFs together with dealers. This is because matching on the BrokerTec CLOB is centralized and anonymous. In other words, an empirical analysis that excludes transactions involving PTFs from the ATS & Interdealer segment would be inconsistent with the centralized matching process of the central limit order book, in which traders cannot choose their counterparties or reveal their identities; instead, they can only direct their orders to the centralized platform and trade according to its rules.

The BrokerTec full order book data contain two components, both of which I use in the empirical analysis. The first is the message data, which record every order submission, modification, and cancellation. The second is the transaction data, which record every realized transaction, i.e., order execution. Every BrokerTec CLOB transaction consists of one market order (aggressive order) matched against one or more limit orders (passive orders): a sell market order executes against standing buy limit orders at the bids, whereas a buy market order executes against standing sell limit orders at the asks. The transaction data report the execution price, quantity, execution time, CUSIP, and trade direction, that is, whether the transaction is initiated by a sell market order or a buy market order. Using the full order book data, I reconstruct the complete limit order book and recover the best bid and best ask at every event timestamp. Because every order is assigned a unique identifier, I also recover the complete life cycle of each limit order, allowing me to measure limit-order waiting times and thereby discipline the empirically relevant time horizon of the one-period model developed in this paper,[11] as well as the limit-order fill rates.

The Treasury secondary market during the sample period covered by this paper is open for 22 hours per day, from 19:30 to 17:30 the following day. Following Duffie, M. J. Fleming, et al. 2023, I restrict the sample to observations between 7:30 and 17:30, the core trading hours.[12]

[11] In particular, only new information arriving while a limit order remains active constitutes aggregate risk that is payoff-relevant to the limit-order trader. Therefore, the observed waiting times of limit orders provide an empirical horizon over which to identify common-value shocks in the data in a manner consistent with the theory.

[12] Duffie, M. J. Fleming, et al. 2023 uses the interval from 7:30 to 17:00, but I extend the sample by 30 minutes to 17:30. This is because the empirical analysis in Section 2.2 exploits cross-venue variation within this time interval.

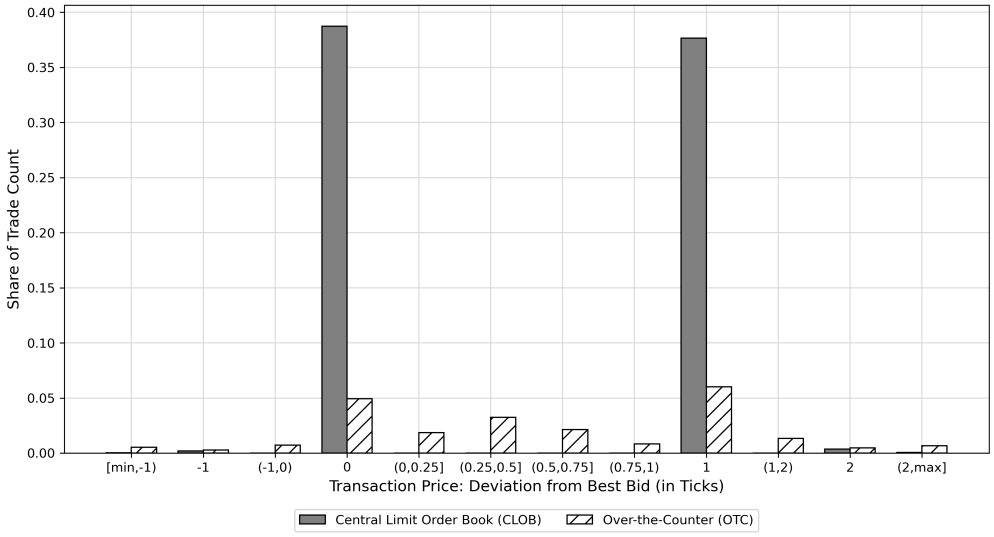


Figure 1: Distribution of Transaction Prices Relative to the Best Bid: CLOB versus OTC

Notes: This figure uses transaction data for the on-the-run 10-year U.S. Treasury. Overall, approximately 77% of transactions are executed on the CLOB (gray bars), while the remaining 23% are executed OTC (hatched bars).

2.2 Motivating Facts

Fact 1: Price Dispersion Trading venues shape the cross-sectional distribution of transaction prices. CLOB transaction prices are concentrated at the best bid and best ask, typically separated by one tick (the smallest allowable increment between quoted prices). In contrast, OTC transaction prices are more dispersed and continuously distributed, with a substantial fraction falling between the prevailing best bid and best ask. Figure 1 illustrates this difference by plotting the distribution of transaction prices separately for CLOB and OTC venues, measured by the share of transactions in each price bin. Transaction prices are expressed in ticks relative to the prevailing best bid.[13]

Three empirical observations from Figure 1 motivate the model's pricing mechanisms. First, almost all CLOB transactions occur at the top of the book, concentrating at the best bid (0 in the figure) and best ask (1 in the figure), with only a small fraction occurring at deeper quote levels when large market orders walk the book. Accordingly, the model focuses on the best bid and best ask, abstracting from deeper levels of the limit order book for simplicity. Second, the tick size acts as an effective constraint on CLOB pricing and plays an important role in determining the bid–ask spread. This figure, together with Figures A1 and A2 shows that the bid–ask spread equals one tick for approximately 87% of trading hours and immediately before 99.5% of CLOB transactions.

[13]For CLOB trades, this measure is exact because I reconstruct the complete limit order book and recover the best bid and best ask at every event timestamp. The realized transaction price is therefore measured relative to the best bid at the exact microsecond of execution. For OTC trades, this measure is only an approximation because TRACE reports timestamps only to the second level. I therefore use the state of the order book at the instant the second begins. Specifically, for each second *SS*, the reference best bid for an OTC trade is computed from all order-book events with timestamps strictly earlier than HH:MM.SS.000000; no events occurring during second *SS* are incorporated.

Accordingly, the model explicitly incorporates the tick size. Third, OTC trading allows for finer pricing than the discrete tick grid of the CLOB. This figure shows that 43% of OTC transactions occur between adjacent tick values,[14] with 35% of OTC transactions occurring between the prevailing best bid and best ask, where CLOB trading is not feasible when the bid–ask spread binds at one tick. Accordingly, the model adopts Nash bargaining for OTC pricing, which naturally generates continuous transaction prices in the presence of a continuum of heterogeneous agents.

Fact 2: Execution Probability (Speed) Execution probability is an important determinant of venue choice. To show that traders value timely execution, I exploit the end-of-day regulatory balance-sheet assessment faced by bank dealers as an exogenous source of variation in venue choice. As the assessment approaches, timely execution of U.S. Treasury transactions becomes particularly valuable for bank dealers relative to non-bank dealers, who are not subject to the same regulation. I estimate a multinomial logit model of trading-method choices among market order on the CLOB, limit order on the CLOB, and OTC, using OTC as the base outcome, with a difference-in-difference specification. This specification exploits differences in end-of-period balance-sheet regulation, following identification strategy similar to those in Du, Tepper, and Verdelhan (2018), Rapp and Waibel (2023), and Correa, Du, and Liao (2025), among others. Appendix Figure A3 plots the intraday trading-method shares for banks and non-banks, providing suggestive evidence of parallel trends. The regression specification is

$$\log\left(\frac{\Pr(Y_i = j)}{\Pr(Y_i = \text{OTC})}\right) = \alpha_j + \beta_{1j}\mathbf{1}\{\text{Bank}\}_i + \beta_{2j}\mathbf{1}\{\text{17:00-17:30}\}_i + \beta_{3j}\mathbf{1}\{\text{Bank}\}_i \times \mathbf{1}\{\text{17:00-17:30}\}_i + \Gamma_j'\mathbf{X}_i$$

where $j \in \{\text{Limit-Order}, \text{Market-Order}\}$. Here, Y_i denotes the trading method used for transaction i, $\mathbf{1}\{\text{Bank}\}_i$ indicates whether either counterparty is a bank, $\mathbf{1}\{\text{17:00–17:30}\}_i$ identifies transactions during the final 30 minutes before market close, and $\mathbf{X}_i$ denotes a vector of controls. The coefficient of interest is β_{3j}. Table 1 reports $\hat{\beta}_{3,\text{Market-Order}} > \hat{\beta}_{3,\text{Limit-Order}} > 0$, indicating that, relative to non-banks, banks exhibit a greater increase in the relative odds of choosing CLOB trading over OTC near market close, with a stronger shift toward using market orders than limit orders on the CLOB.

These estimates suggest that execution probability plays an important role in venue choice and motivate its explicit incorporation into the model through matching mechanism. This is because, in practice, OTC trading requires counterparty search and bargaining, making execution potentially slower and less certain; on the CLOB, limit orders must wait for execution and face execution uncertainty, whereas market orders execute immediately against available limit orders. Thus, as timely execution becomes more valuable, banks substitute toward trading methods offering faster and more certain execution.

[14] That is, OTC transaction prices fall into bins such as $(-1, 0)$, $(0, 1)$, $(1, 2)$, etc.

Table 1: Determinants of Venue Choice

	Specification (I)		Specification (II)		Specification (III)	
	(1) Limit	(2) Market	(3) Limit	(4) Market	(5) Limit	(6) Market
$\mathbf{1}\{\text{Bank}\}$	-2.189***	-2.473***	-2.190***	-2.473***	-2.231***	-2.513***
	(0.058)	(0.052)	(0.058)	(0.052)	(0.056)	(0.053)
$\mathbf{1}\{\text{17:00-17:30}\}$	-0.270*	-0.212	-0.264**	-0.204	-0.308**	-0.249*
	(0.139)	(0.146)	(0.134)	(0.142)	(0.130)	(0.136)
$\mathbf{1}\{\text{Bank}\} \times \mathbf{1}\{\text{17:00-17:30}\}$	0.585*	1.488***	0.570*	1.476***	0.575*	1.490***
	(0.324)	(0.262)	(0.336)	(0.271)	(0.326)	(0.268)
Price	0.060***	0.047***	0.065***	0.051***	-0.025	-0.042
	(0.007)	(0.007)	(0.006)	(0.006)	(0.028)	(0.026)
Trade Size ($ mn)	-0.057***	-0.054***	-0.058***	-0.054***	-0.059***	-0.055***
	(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Day Fixed Effect	No	No	Yes	Yes	Yes	Yes
CUSIP Fixed Effect	No	No	No	No	Yes	Yes
Observations	3,609,916	3,609,916	3,609,916	3,609,916	3,609,916	3,609,916
Pseudo R^2	0.0666		0.0670		0.0726	
$\beta_{3,\text{Market-Order}} - \beta_{3,\text{Limit-Order}}$	0.904		0.907		0.915	
p-value	0.000		0.000		0.000	

Notes: This table reports estimates from multinomial logit regressions of choice among trading methods: limit order on the CLOB, market order on the CLOB, or OTC. Specifications (I), (II), and (III) differ only in the fixed effects included. Columns (1), (3), and (5) report coefficients for limit orders relative to the OTC category, while columns (2), (4), and (6) report coefficients for market orders relative to the OTC category. Standard errors are clustered at the trading-day level and reported in parentheses. The final two rows report the difference between the $\mathbf{1}\{\text{Bank}\} \times \mathbf{1}\{\text{17:00-17:30}\}$ coefficients for market order and limit orders and the two-sided p-value for the null hypothesis that the two coefficients are equal. The data cover all tenors of on-the-run U.S. Treasuries in September 2022. For detailed sample construction specifically used only for this regression, please see the note beneath Figure A3. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Fact 3: Common-Value Shocks and Limit-Order Adverse Selection Common-value shocks generate one-sided market order flow following their arrival, exposing resting limit orders to adverse selection due to the sequential trading structure of the CLOB. To document this mechanism, I follow the literature in interpreting persistent price changes that are not immediately reversed as shocks to the common value of an asset (e.g., Hasbrouck (1991), Brogaard, Hendershott, and Riordan (2014), among others). Figure 2 illustrates this adverse-selection mechanism using a representative negative-shock episode. During the first five seconds following the onset of the shock, sell market orders account for 83% of trading volume, executing against resting buy limit orders at successive bid levels, compared with only 17% for buy market orders. Thus, limit orders are more likely to execute precisely when the common value moves against them: buy limit orders are more likely to execute following negative shocks, whereas sell limit orders are more likely to execute following positive shocks by symmetric reasoning. This directional market order flow is informative about the asset's common value and exposes limit-order traders to the winner's curse. In the episode shown, buyers whose resting buy limit orders execute at the onset of the shock pur-

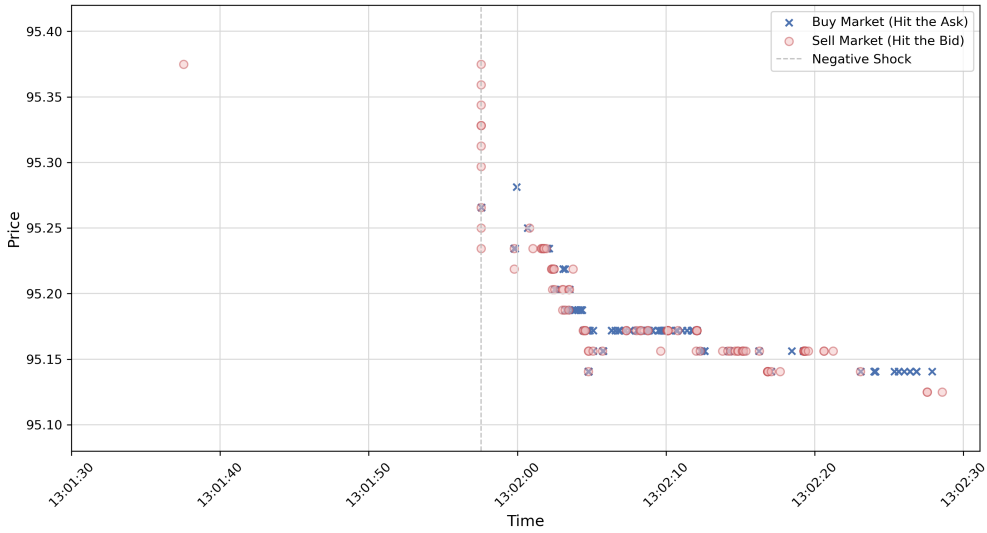


Figure 2: An Example Episode: One-sided Market Order Flow Following a Common-Value Shock

Notes: This figure uses CLOB transaction data for the on-the-run 10-year U.S. Treasury on September 12, 2022. Figure A4 shows this representative episode over a longer horizon, in which the transaction price moves sharply over a short period without an immediate reversal, which this paper interprets as a common-value shock.

chase the asset at stale bid prices that are too high relative to the updated common value. This relationship between common-value shocks and one-sided market order flow also holds systematically across the data. Appendix Figures A5, A6, and A7 provide additional examples around scheduled major macroeconomic announcements, and Figure A8 compares market-order direction across positive-shock, negative-shock, and non-shock periods using all identified common-value shocks in the sample (see Appendix A for the shock-identification procedure).

This adverse-selection mechanism motivates the model's sequential timing structure. In practice, quotation and execution on the CLOB are *asynchronous*: limit orders move first and market orders move second (e.g., as modeled in Glosten and Milgrom (1985), Glosten (1994), and Biais, Martimort, and Rochet (2000), among others). That is, limit orders are submitted before the arrival of market orders and remain executable as market conditions change, representing a commitment to trade. Accordingly, the model assumes that limit orders are submitted ex ante and remain executable at their quoted prices without re-optimization. The aggregate state is then realized, followed by market-order decisions made ex post. This sequential timing generates an information asymmetry consistent with the practical operation of the CLOB: market-order traders move second and can condition their trading decisions on newly arrived information that is not available to limit-order traders who submit quotes in the first place. As a result, market-order traders have more information about asset values (the informed party), and limit-order traders who submit quotes have less information (the uninformed party). Thus, market orders generate adverse selection for limit-order traders as an equilibrium outcome.

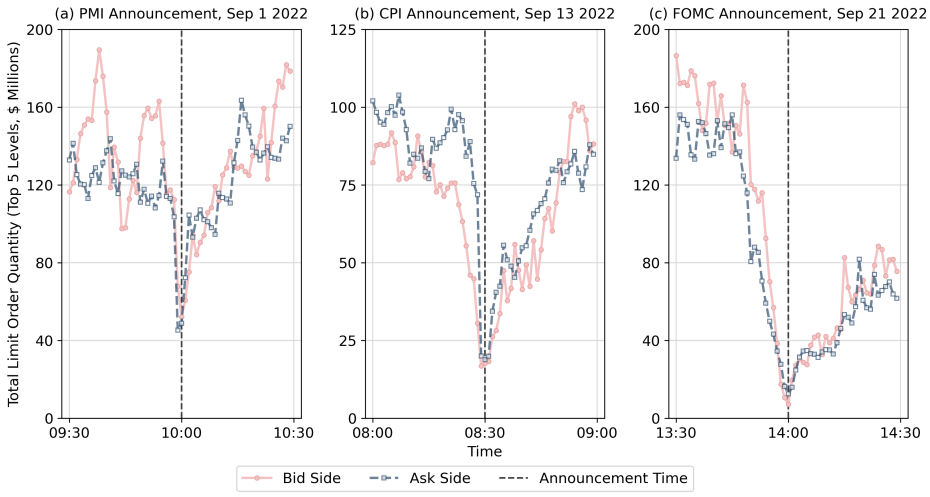


Figure 3: Quantity of Limit Orders Before and After Scheduled Announcements

Notes: This figure uses CLOB message data for the on-the-run 10-year U.S. Treasury on the corresponding trading days. Each dot represents the time-weighted average quantity of outstanding limit orders within a one-minute interval.

Fact 4: Book Depth Figure 3 documents a sharp decline in limit order book depth ahead of scheduled major macroeconomic announcements, measured by the outstanding limit order quantity at the top five quote levels on each side of the market. Following the literature (e.g., M. J. Fleming and Remolona (1999), Nakamura and Steinsson (2018), among others), I exploit narrow time windows around scheduled announcements as a natural setting for studying common-value uncertainty. This is because, while the timing of these announcements is known in advance, their informational content remains uncertain, and this content would affect market participants' assessments of monetary policy and economic fundamentals, and thus the common value of the Treasury. I therefore examine limit order book depth within 30-minute windows before and after scheduled PMI, CPI, and FOMC announcements. Across these events, limit order quantities decline sharply immediately before the announcements, consistent with limit-order traders reducing their exposure to common-value uncertainty, potentially due to the aforementioned adverse-selection risk.

This empirical pattern motivates the model's emphasis on the quantity margin in incorporating aggregate risk and determining allocative efficiency. As common-value uncertainty increases, the adverse-selection risk faced by limit-order traders discourages limit-order submission and reduces order book depth. Accordingly, the model explicitly incorporates endogenous limit-order submission decisions to study how aggregate risk affects CLOB depth and, thereby, allocative efficiency. The idea is that, with fewer outstanding limit orders, the CLOB has less capacity to accommodate incoming market orders, limiting the volume of trade and the gains from trade it can realize.

To summarize, I document four stylized facts that motivate the model. First, trading venues shape the cross-sectional distribution of transaction prices. Second, execution certainty is an important determinant of venue choice. Together, these two facts motivate a model that incorporates

both the pricing (transaction costs) and matching (execution probabilities) mechanisms of each venue. Third, common-value shocks expose limit-order traders to adverse selection through the sequential trading structure of the CLOB. Fourth, limit order book depth declines in the presence of common-value uncertainty. Together, the latter two facts motivate a model with a sequential setup featuring aggregate risk. I now proceed to outline the model.

3 Model

In this section, I develop a venue-choice model with heterogeneous agents and aggregate risk in a competitive environment. The model is populated by a unit mass of continuum agents with type $\delta \sim F(\delta)$, where F is a continuous cumulative distribution function, and initial inventory $q \in \{0, 1\}$, who trade an indivisible asset with total supply $k < 1$. Each aggregate state $s \in \{+, -\}$ corresponds to a common preference shock ϵ_s affecting all agents' values. ϵ_s takes the value $+\alpha$ in the positive state and $-\alpha$ in the negative state. The goal of the model is to study allocative efficiency: an efficient market should allow owners and non-owners who are initially misallocated to successfully sell and buy the asset through the venues into which they sort. The roadmap is as follows. The setup is presented in Section 3.1. The equilibrium solutions presented in Section 3.2 explain venue sorting, market fragmentation, and how aggregate risk is incorporated ex ante and absorbed ex post. Section 3.3 studies allocative efficiency and its implications for market structure and market design.

3.1 Setup

3.1.1 Timing

1. *Limit-order stage.* An owner $q = 1$ (non-owner $q = 0$) who chooses to submit sell (buy) limit order do so by posting at the quoted ask price A (the quoted bid price B).

2. *Shock stage.* A common preference shock realizes. The aggregate state is $s \in S = \{+, -\}$ with $\Pr(s = +) = \frac{1}{2}$, and the shock is

$$\epsilon_s = \begin{cases} +\alpha, & s = +, \\ -\alpha, & s = -. \end{cases}$$

3. *Market-order stage.* After observing s, agents who did not post a limit order may submit a market order, which execute immediately against resting limit orders, or opt out of the CLOB.

4. *Over-the-counter stage.* Following CLOB trading (Stages 1-3), all agents enter the OTC venue with the asset position q' carried over from the previous stage. Agents are randomly

matched and may trade bilaterally by Nash-bargaining.

5. *Payoff stage.* A type-δ agent holding q'' units in state s receives $(\delta + \epsilon_s)q''$, where δ captures the idiosyncratic-value component and ϵ_s captures the common-value component.

In short, the choice set $d \in D$ is as follows. For an owner ($q = 1$), the CLOB-stage (Stages 1–3) decision d consists of choosing whether to post a sell limit (SL) order, submit a sell market (SM) order, or opt out of the CLOB altogether (Opt-out). Likewise, for a non-owner ($q = 0$), the decision d consists of choosing whether to post a buy limit (BL) order, submit a buy market (BM) order, or opt out of the CLOB altogether (Opt-out). All agents then enter the OTC trading stage with the asset position resulting from the CLOB stage. For tractability, the model considers only the best quotes in the central limit order book, the so-called top of the book. The state of the book is therefore summarized by the best ask A and the best bid B ($A \geq B$), together with the total quantity posted at each side.

The model's timing is motivated by three considerations. First, the sequential timing reflects the practical operation of the CLOB: limit orders are submitted first at quoted prices, followed by market orders that execute against resting limit orders on the opposite side of the market at the best available quotes. For tractability, OTC trading takes place after CLOB trading. Second, the common preference shock ϵ_s is realized between limit-order and market-order submission, allowing market-order traders to condition their decisions on the realized aggregate state, while resting limit orders remain committed to trade at their quoted prices. This generates informative market order flow and mechanically induces fluctuations in venue demand across aggregate states. Third, the model features no principal, auctioneer, or exogenously designated market maker. Instead, all agents endogenously choose their trading methods d, consistent with modern financial markets. These modeling choices are broadly supported by the facts documented in Section 2.2.

Before presenting the remaining details of the model, I first introduce the Walrasian equilibrium to build intuition for the environment. Figure 4 illustrates the Walrasian version of the model through an aggregate demand–supply diagram (left panel) and the aggregate shock ϵ_s as a shift in aggregated demand (right panel). Walrasian equilibrium corresponds to the efficient allocation (first-best). Under this allocation, the asset is held by the highest-k fraction of agents ranked by type. The ex post Walrasian price is given by $P_s^* = F^{-1}(1-k) + \epsilon_s$, which induces the ex post efficient allocation. When the aggregate shock has zero mean, i.e., $\mathbf{E}_s[\epsilon_s] = 0$, the ex ante Walrasian price is $P^* = F^{-1}(1-k) = \mathbf{E}_s[P_s^*]$. The ex ante Walrasian allocation is also ex post efficient. This is because the aggregate shock is a common preference shock that shifts every agent's valuation by the same amount without changing the ranking of types. The Walrasian equilibrium serves as a benchmark for evaluating the (in)efficiency of the trading venues. I now turn to a detailed description of each venue's trading mechanism.

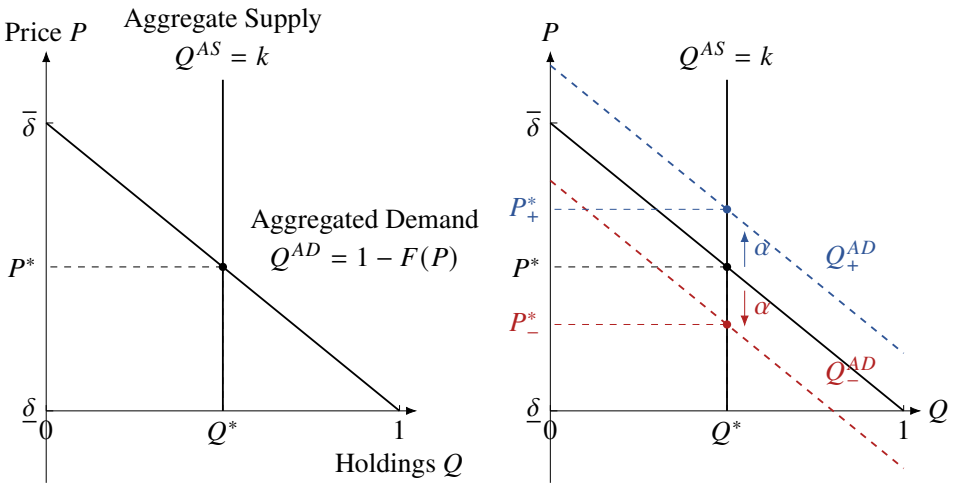


Figure 4: Walrasian Equilibrium with Demand Shifts

3.1.2 Mechanism: Pricing and Matching

The model treats each trading venue as a *mechanism* (P, ϕ), characterized by a trade price P, the intensive margin, and a trade probability ϕ, the extensive margin. The intensive margin is modeled through the pricing mechanism, where the central limit order book follows the canonical bid–ask structure, whereas the over-the-counter venue features classic Nash bargaining. The extensive margin is modeled through the matching mechanism. In the central limit order book, sell market orders directly match against buy limit orders, and buy market orders directly match against sell limit orders. In contrast, over-the-counter matching is random, and a trade occurs if and only if the non-owner's (buy-side) type exceeds the owner's (sell-side) type within a bilateral relationship. Both the pricing and matching mechanisms differ across the two trading venues. In this sense, the two venues are different but competing trading mechanisms. Both the intensive and extensive margins of each trading venue are endogenously determined.

I first present the pricing mechanism. In the CLOB, transactions are executed at the quoted bid and ask prices. A sell limit order posts the ask, and a buy market order submitted later matches against the sell limit order, so both orders execute at the ask price,

$$P^{SL} \equiv A = P^* + \frac{1}{2}cm = P^{BM}, \tag{1A}$$

and, symmetrically, a buy limit order posts the bid, and a sell market order submitted later matches against the buy limit order, so both orders execute at the bid price,

$$P^{BL} \equiv B = P^* - \frac{1}{2}cm = P^{SM}, \tag{1B}$$

where P^* is the ex ante Walrasian price, c is the tick size parameter, and m is an (equilibrium)

integer, capturing the fact that the bid and ask can only be posted at multiples of tick.[15] Therefore, by subtracting the two, integer $m \in \{1, 2, 3, \ldots\}$ denotes the equilibrium bid–ask spread measured in ticks, $m = \frac{A-B}{c}$. Equation 1 implies that the tick size c is a wedge on the efficient price P^* and assumes that the mid-quote $\frac{1}{2}(A + B)$ is an unbiased proxy for P^*.

In the OTC venue, transaction prices are determined through Nash bargaining with bargaining power parameter θ. Upon matching, if $\delta_{no} > \delta_o$, an owner with type δ_o and a non-owner with type δ_{no} negotiate the ex post price to share the joint gains from trade according to

$$P^{OTC} = \arg\max_p \big(p - (\delta_o + \epsilon_s)\big)^{1-\theta}\big((\delta_{no} + \epsilon_s) - p\big)^{\theta},$$

which yields

$$P^{OTC}(\delta_o, \delta_{no}, s) = \theta\delta_o + (1 - \theta)\delta_{no} + \epsilon_s. \tag{2}$$

The price-setting mechanisms in Equations 1 and 2 capture three differences between the two trading venues along the intensive margin, as summarized in Panel A of Table 2. First, they distinguish centralized and decentralized trading according to whether prices are uniform across agents or personalized. CLOB prices are independent of individual types (A and B are equilibrium scalars), whereas OTC prices are set bilaterally and depend on the types of the matched agents. Second, CLOB quotes are firm commitments to trade regardless of the state of the world, since they are state-invariant, whereas OTC prices adjust flexibly to the realized state and incorporate it into the negotiated price. Third, CLOB prices must be multiples of the tick size and are therefore discrete, whereas OTC prices vary continuously.

I now turn to the matching mechanism in the two trading venues. In the CLOB, execution occurs between market orders and previously posted limit orders on the opposite side of the book: buy market orders match against sell limit orders, while sell market orders match against buy limit orders. Matching follows a pro-rata rule: among limit orders posted at the same quoted price, incoming market-order flow is allocated pro rata. Since the model considers only the top of the book, execution occurs at a single quoted price level on each side of the book.[16] Let N_s^d denote the equilibrium mass of agents choosing d in state s. Execution probabilities satisfy

$$\phi_s^{SL} = \min\left\{\frac{N_s^{BM}}{N^{SL}}, 1\right\}, \qquad \phi_s^{BM} = \min\left\{\frac{N^{SL}}{N_s^{BM}}, 1\right\}, \tag{3a}$$

[15] I model the half tick following the convention in the literature.

[16] The price-priority rule directs market orders to execute first against the most aggressive limit orders (the highest bid or lowest ask), followed by the next-best quotes, and so forth. For simplicity, the model considers only the top of the book, with only one quoted price level on each side of the market in equilibrium. Nevertheless, price priority remains relevant for determining equilibrium quotes and is incorporated into the model through the no one-shot deviation condition introduced later in the equilibrium definition.

$$\phi_s^{BL} = \min\left\{\frac{N_s^{SM}}{N^{BL}}, 1\right\}, \qquad \phi_s^{SM} = \min\left\{\frac{N^{BL}}{N_s^{SM}}, 1\right\}. \tag{3b}$$

The intuition for this matching function is as follow. Since limit orders are submitted before the aggregate state is realized, whereas market orders are submitted afterward, N^{SL} and N^{BL} are state-invariant, while N_s^{BM} and N_s^{SM} are state-dependent. The term $\frac{N_s^{BM}}{N^{SL}}$ can be interpreted as the sell limit order fill rate. If there is excess buy market order flow in state s, such that $N_s^{BM} > N^{SL}$, then all sell limit orders are filled and are therefore assured of execution, so that $\phi_s^{SL} = 1$; otherwise, only a fraction of the sell limit orders are filled, $\phi_s^{SL} < 1$. Similar logic applies to other execution probability terms. The operator $\min\{\cdot, 1\}$ ensures that $\phi \in [0, 1]$, so that ϕ is a valid probability.

In the OTC venue, agents are randomly matched, and a matched owner–non-owner pair trades if and only if there are gains from trade

$$\delta_o + \epsilon_s \leq \delta_{no} + \epsilon_s.$$

Consequently,

$$\phi_s^{OTC-O}(\nu) = \lambda \min\left\{\frac{N_s^{OTC-NO}}{N_s^{OTC-O}},\ 1\right\}\left(1 - G_s^{OTC-NO}(\nu)\right), \tag{4A}$$

$$\phi_s^{OTC-NO}(\nu) = \lambda \min\left\{\frac{N_s^{OTC-O}}{N_s^{OTC-NO}},\ 1\right\} G_s^{OTC-O}(\nu). \tag{4B}$$

Here, G denotes the cumulative distribution of owners (OTC-O) and non-owners (OTC-NO) after CLOB trading, i.e., at the beginning of the OTC trading stage, and $\nu \equiv \delta + \epsilon_s$ denotes the ex post valuation. Since trade occurs if and only if $\delta_o \leq \delta_{no}$, low-valuation owners have a higher probability of executing a trade than high-valuation owners because they are more likely to meet a counterparty with a higher valuation. Accordingly, the execution probability of owners in the OTC venue, $\phi^{OTC-O}(\nu)$, decreases with ν. The same reasoning applies to non-owners, whose execution probability in the OTC venue, $\phi^{OTC-NO}(\nu)$, increases with ν. This matching function assumes that matching in the OTC venue is random. Specifically, the probability that an agent meets a counterparty of a given type on the other side of the market is proportional to the density of types on that side, g, up to the scaling factor $\lambda \min\{\cdot, 1\}$. In this scaling factor, the parameter $\lambda \in [0, 1]$ captures the search efficiency of the OTC venue relative to the CLOB, whose search efficiency is normalized to one, and $\min\{\cdot, 1\}$ keeps track of the tightness between the two sides of the market in the OTC venue.

The execution-probability Equations 3 and 4 capture three differences between the two trading venues along the extensive margin, as summarized in Panel B of Table 2. First, they formalize the distinction between centralized and decentralized trading. In the CLOB, execution depends only on the aggregate order imbalance and is independent of an agent's individual type once the

trading venue has been chosen.[17] In contrast, execution in the OTC venue depends on the types of the matched counterparties and is therefore bilateral, with personalized execution probabilities. Second, CLOB trading is sequential, with market orders executing against previously posted limit orders, whereas OTC trading is determined simultaneously upon matching. As a result, the information structure also differs across the two venues. In the CLOB, aggregate shocks arrive between limit-order submission and market-order submission, creating asymmetric information between the two counterparties. In contrast, OTC counterparties negotiate and agree on the trade simultaneously upon matching, so neither party has an informational advantage.

Table 2: Economic Intuition for the Parameterization of Trading Mechanisms

Panel A. Intensive-Margin: Pricing

	Central Limit Order Book (CLOB)	Over-the-counter (OTC)
Centralization	Centralized: δ-invariant	Bilateral: δ-specific
Firmness	Firm commitment: s-invariant	Flexible negotiation: s-specific
Grid	Discrete: quotes restricted to multiples of the tick size c	Continuous: $\theta \in [0, 1]$

Panel B. Extensive-Margin: Matching

	Central Limit Order Book (CLOB)	Over-the-counter (OTC)
Centralization	Centralized: δ-invariant	Bilateral: δ-specific
Timing of Decision	Sequential	Simultaneous
Information Structure	Asymmetric information: s is realized between the limit-order and market-order stages	Symmetric information: s is common knowledge

3.1.3 Value Functions

Since the timing is sequential, the value functions and decision rules are characterized by backward induction. The choice-specific value functions are constructed by weighting the values under execution and non-execution by the corresponding execution probabilities.

Stage 4 — Over-the-counter stage. Utility $u(q', \delta, s)$ denotes the continuation value upon entering the OTC venue (i.e., at the beginning of Stage 4) for an agent with post-CLOB asset position

[17] This aligns with the practice that all orders are pooled in a common matching engine and matched anonymously and centrally according to the same rule.

q', type parameter δ, in aggregate state s:

$$u(1,\delta,s) = \underbrace{\delta+\epsilon_s}_{\text{flow payoff}} + \underbrace{\lambda(1-\theta)\min\left\{\frac{N_s^{OTC-NO}}{N_s^{OTC-O}},1\right\}\int_{\nu>\delta+\epsilon_s}(\nu-(\delta+\epsilon_s))g_s^{OTC-NO}(\nu)\,d\nu}_{\text{expected surplus from OTC}}, \quad (5a)$$

$$u(0,\delta,s) = \lambda\theta\min\left\{\frac{N_s^{OTC-O}}{N_s^{OTC-NO}},1\right\}\int_{\nu<\delta+\epsilon_s}(\delta+\epsilon_s-\nu)g_s^{OTC-O}(\nu)\,d\nu. \quad (5b)$$

The derivation of Equation 5 is provided in Appendix B. $g_s^{OTC-NO}(\nu)$ and $g_s^{OTC-O}(\nu)$ denote the conditional probability density functions satisfying

$$\int_{\underline{\delta}+\epsilon_s}^{\bar{\delta}+\epsilon_s} g_s^{OTC-NO}(\nu)\,d\nu = 1, \qquad \int_{\underline{\delta}+\epsilon_s}^{\bar{\delta}+\epsilon_s} g_s^{OTC-O}(\nu)\,d\nu = 1. \quad (6)$$

Stage 3 — Market-order stage. After observing the aggregate state s, the owner's choice-specific state-contingent valuations $W(q,\delta,s \mid d)$ are

$$W(1,\delta,s|\text{SM}) = \underbrace{\phi_s^{SM}}_{\text{Prob. of execution}} \underbrace{\big(B+u(0,\delta,s)\big)}_{\text{Value if executed}} + \underbrace{(1-\phi_s^{SM})}_{\text{Prob. of failure}} \underbrace{u(1,\delta,s)}_{\text{Value if not executed}}, \quad (7a)$$

$$W(1,\delta,s|\text{Opt-out}) = u(1,\delta,s). \quad (7b)$$

That is, if the sell market order is executed, the agent receives the bid price B (since sell market orders execute at the bid, as described in Equation 1) and enters the OTC venue as a non-owner with continuation value $u(0,\delta,s)$. Otherwise, the agent retains the asset and enters the OTC venue as an owner with continuation value $u(1,\delta,s)$.

Symmetrically, the non-owner's choice-specific state-contingent valuations are

$$W(0,\delta,s|\text{BM}) = \phi_s^{BM}\big(u(1,\delta,s)-A\big) + (1-\phi_s^{BM})u(0,\delta,s), \quad (8a)$$

$$W(0,\delta,s|\text{Opt-out}) = u(0,\delta,s). \quad (8b)$$

Stage 1 — Limit-order stage. At the beginning of the game, the owner's ex ante choice-specific valuations $V(q,\delta \mid d)$ are

$$V(1,\delta|\text{SL}) = \mathbb{E}_s\Big[\phi_s^{SL}\big(A+u(0,\delta,s)\big) + (1-\phi_s^{SL})u(1,\delta,s)\Big], \quad (9a)$$

$$V(1,\delta|\text{NotSL}) = \mathbb{E}_s\Big[\max\Big\{W(1,\delta,s|\text{SM}), W(1,\delta,s|\text{Opt-out})\Big\}\Big]. \quad (9b)$$

That is, if the agent submits a sell limit order (Equation 9a), if the order is executed, the agent

receives the ask price A (since sell limit orders execute at the ask, as described in Equation 1) and enters the OTC venue as a non-owner, and if the order fails to execute, the agent retains the asset and enters the OTC venue as an owner. Alternatively, if the agent does not submit a sell limit order (Equation 9b), the valuation is the maximum attainable value after observing the aggregate state, since the agent retains the option to submit a market order or not.

Symmetrically, the non-owner's ex ante choice-specific valuations are

$$V(0,\delta|\text{BL}) = \mathbb{E}_s\Big[\phi_s^{BL}\big(u(1,\delta,s) - B\big) + (1-\phi_s^{BL})u(0,\delta,s)\Big], \tag{10a}$$

$$V(0,\delta|\text{NotBL}) = \mathbb{E}_s\Big[\max\Big\{W(0,\delta,s|\text{BM}), W(0,\delta,s|\text{Opt-out})\Big\}\Big]. \tag{10b}$$

3.1.4 Decision Rules

Agents' decisions d are characterized by the following indicator functions, where $\mathbf{1}_d\{\cdot\}$ equals 1 if the condition is satisfied and 0 otherwise.

Stage 3 — Market-order stage. After the aggregate shock ϵ_s is realized, the decision to submit a market order depends on the immediate payoff relative to the continuation value in the OTC:

$$\mathbf{1}_{SM}(\delta,s) = \mathbf{1}\Big\{B + u(0,\delta,s) > u(1,\delta,s)\Big\}, \tag{11a}$$

$$\mathbf{1}_{BM}(\delta,s) = \mathbf{1}\Big\{u(1,\delta,s) - A > u(0,\delta,s)\Big\}. \tag{11b}$$

That is, after observing the state s, the owner submits a market order only if $W(1,\delta,s|\text{SM}) > W(1,\delta,s|\text{Opt-out})$. Substituting the expressions for W and rearranging the inequality yields the condition above. The condition for the non-owner is derived analogously.

Stage 1 — Limit-order stage. Ex ante, the decision to submit a limit order depends on the relative valuations:

$$\mathbf{1}_{SL}(\delta) = \mathbf{1}\Big\{V(1,\delta \mid \text{SL}) > V(1,\delta \mid \text{NotSL})\Big\}, \tag{12a}$$

$$\mathbf{1}_{BL}(\delta) = \mathbf{1}\Big\{V(0,\delta \mid \text{BL}) > V(0,\delta \mid \text{NotBL})\Big\}. \tag{12b}$$

3.1.5 Aggregation

Here I present the order quantities N and execution probabilities ϕ, both of which are aggregate equilibrium objects obtained by aggregating agents' decisions. For tractability, I assume that the asset is initially allocated randomly across agents. That is, the initial joint distribution satisfies $\mu(q=1,\delta) = kf(\delta)$ and $\mu(q=0,\delta) = (1-k)f(\delta)$. The aggregation procedure, as well as the characterization of the law of motion, extends naturally to more general initial distributions.

Stage 1 — Limit-order stage. The total supply of limit orders is determined by aggregating all agents for whom submitting a limit order is optimal:

$$N^{SL} = \int_{\underline{\delta}}^{\bar{\delta}} \mathbf{1}_{SL}(\delta) \cdot k \, f(\delta) \, d\delta, \tag{13a}$$

$$N^{BL} = \int_{\underline{\delta}}^{\bar{\delta}} \mathbf{1}_{BL}(\delta) \cdot (1-k) \, f(\delta) \, d\delta. \tag{13b}$$

Stage 3 — Market-order stage. After the aggregate shock ϵ_s is realized, agents who did not choose the limit-order (NotSL and NotBL) strategy in Stage 1 re-optimize. In contrast, agents who submitted limit orders in Stage 1 cannot re-optimize because limit orders are commitments to trade. The mass of market orders in state $s \in \{+, -\}$ therefore consists of agents who did not submit limit orders and for whom an immediate execution attempt is optimal in state s:

$$N_s^{SM} = \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{SL}(\delta)) \cdot \mathbf{1}_{SM}(\delta, s) \cdot k \, f(\delta) \, d\delta, \tag{14a}$$

$$N_s^{BM} = \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{BL}(\delta)) \cdot \mathbf{1}_{BM}(\delta, s) \cdot (1-k) \, f(\delta) \, d\delta. \tag{14b}$$

These aggregate order quantities N, specified in Equations 13 and 14, determine the execution probabilities ϕ through the CLOB matching mechanism introduced previously in Equation 3. That is, the matching function satisfies the short-side matching rule in each aggregate state s. Orders on the short side are guaranteed execution, whereas the execution probability of orders on the long side equals the ratio of the aggregate order quantity on the short side to that on the long side.

3.1.6 Law of Motion

To characterize how $g_s^{OTC}(\cdot)$ evolves endogenously, I track both the endogenous change in the distribution resulting from ownership changes through CLOB trading and the exogenous shift in the distribution induced by the realization of the aggregate state. The transition follows ten mutually exclusive and collectively exhaustive paths, as specified in the following equation. Again, let $v \equiv \delta + \epsilon_s$ denote the ex post valuation. The density of agents entering the OTC stage is then given by

$$g_s^{OTC-O}(v) = \frac{f(v-\epsilon_s)}{N_s^{OTC-O}} \Bigg[\underbrace{(1-k)}_{\text{Non-Owner}} \times \Bigg(\underbrace{\mathbf{1}_{BL}(v-\epsilon_s) \cdot \phi_s^{BL}}_{\text{Path 1: Buy Limit Success}} + \underbrace{(1 - \mathbf{1}_{BL}(v-\epsilon_s)) \cdot \mathbf{1}_{BM}(v-\epsilon_s, s) \cdot \phi_s^{BM}}_{\text{Path 2: Buy Market Success}} \Bigg)$$
$$+ \underbrace{k}_{\text{Owner}} \times \Bigg(\underbrace{\mathbf{1}_{SL}(v-\epsilon_s) \cdot (1-\phi_s^{SL})}_{\text{Path 3: Sell Limit Fail}} + \underbrace{(1 - \mathbf{1}_{SL}(v-\epsilon_s)) \cdot \mathbf{1}_{SM}(v-\epsilon_s, s) \cdot (1-\phi_s^{SM})}_{\text{Path 4: Sell Market Fail}} \Bigg)$$

$$+ \underbrace{(1 - \mathbf{1}_{SL}(v - \epsilon_s)) \cdot (1 - \mathbf{1}_{SM}(v - \epsilon_s, s))}_{\text{Path 5: Opt-out CLOB}} \Bigg) \Bigg], \tag{15a}$$

$$g_s^{OTC-NO}(v) = \frac{f(v - \epsilon_s)}{N_s^{OTC-NO}} \Bigg[\underbrace{k}_{\text{Owner}} \times \Bigg(\underbrace{\mathbf{1}_{SL}(v - \epsilon_s) \cdot \phi_s^{SL}}_{\text{Path 6: Sell Limit Success}} + \underbrace{(1 - \mathbf{1}_{SL}(v - \epsilon_s)) \cdot \mathbf{1}_{SM}(v - \epsilon_s, s) \cdot \phi_s^{SM}}_{\text{Path 7: Sell Market Success}} \Bigg)$$
$$+ \underbrace{\Big(1 - k\Big)}_{\text{Non-Owner}} \times \Bigg(\underbrace{\mathbf{1}_{BL}(v - \epsilon_s) \cdot (1 - \phi_s^{BL})}_{\text{Path 8: Buy Limit Fail}} + \underbrace{(1 - \mathbf{1}_{BL}(v - \epsilon_s)) \cdot \mathbf{1}_{BM}(v - \epsilon_s, s) \cdot (1 - \phi_s^{BM})}_{\text{Path 9: Buy Market Fail}}$$
$$+ \underbrace{(1 - \mathbf{1}_{BL}(v - \epsilon_s)) \cdot (1 - \mathbf{1}_{BM}(v - \epsilon_s, s))}_{\text{Path 10: Opt-out CLOB}} \Bigg) \Bigg]. \tag{15b}$$

where the support of $g^{OTC-O}(v)$ and $g^{OTC-NO}(v)$ is $v \in [\underline{\delta} + \epsilon_s, \bar{\delta} + \epsilon_s]$. The indicator function $\mathbf{1}_d(v - \epsilon_s)$ denotes the decision rule of the agent whose ex post valuation is v. Equivalently, $\mathbf{1}_d(v - \epsilon_s)$ evaluates the decision of the agent whose original type is $\delta = v - \epsilon_s$. Thus the argument $(v - \epsilon_s)$ maps the post-shock valuation back to the corresponding original type in the initial distribution.

3.1.7 Equilibrium Definition

Definition 1 (Competitive Equilibrium)**.** *A competitive equilibrium consists of:*

- CLOB quotes, consisting of the ask A and the bid B (equivalently, $m \in \{1, 2, 3, \dots\}$), and OTC transaction prices $\{P^{OTC}\}$.
- Execution probabilities $\phi_s^{SL}, \phi_s^{BL}, \phi_s^{SM}, \phi_s^{BM}, \phi_s^{OTC-O}, \phi_s^{OTC-NO}$.
- Continuation values $u(q', \delta, s)$ and the implied value functions $V(q, \delta \mid d)$ and $W(q, \delta, s \mid d)$.
- Decision rules $\mathbf{1}_{SL}, \mathbf{1}_{BL}, \mathbf{1}_{SM,s}, \mathbf{1}_{BM,s}$.
- Aggregate order quantities in the CLOB, $N^{SL}, N^{BL}, N_s^{SM}, N_s^{BM}$.
- Distributions of agents in the OTC, $G_s^{OTC-O}(v)$, $G_s^{OTC-NO}(v)$.

such that for every aggregate state $s \in \{+, -\}$:

1. Optimality. Each agent's venue choice $d^*(\delta)$ is optimal. That is, the indicator satisfies

$$\mathbf{1}_d(\delta) = \begin{cases} 1, & d = d^*(\delta), \\ 0, & d \neq d^*(\delta). \end{cases}$$

2. Consistency. The aggregate objects N, ϕ, G are consistent with agents' optimal choices d^*.

3. No one-shot deviation. If the bid–ask spread exceeds one tick $A - B > c$ (equivalently, $m \geq 2$), no limit-order trader can increase her payoff by undercutting the prevailing ask or overbidding the prevailing bid, i.e., no profitable deviation. That is,

$$V(1, \delta^o \mid SL(A)) \geq V(1, \delta^o \mid SL(A - c)),$$
$$V(0, \delta^{no} \mid BL(B)) \geq V(0, \delta^{no} \mid BL(B + c)),$$

for every $\delta^o \in \{\delta \mid d^*(\delta) = SL\}$ and $\delta^{no} \in \{\delta \mid d^*(\delta) = BL\}$. Otherwise, when $A - B = c$ (equivalently, $m = 1$), deviation is infeasible.

The one-shot deviation condition can be understood as follows. If the bid–ask spread exceeds one tick, fixing all other agents' choices, a single limit-order trader can undercut the prevailing ask by one tick, from A to $A - c$, or overbid the prevailing bid by one tick, from B to $B + c$. By the price-priority rule of CLOB matching, this single deviating quote becomes the only best available quote and therefore receives priority in execution. Consequently, this single deviating trader executes with certainty, i.e., $\phi = 1$, in every aggregate state. This incentive to deviate is based on the Bertrand-style price-competition intuition underlying the price-priority rule of the limit order book: by sacrificing quoted price (reducing gains from trade conditional on an execution, i.e., the intensive margin), an agent gains execution priority (increasing execution probability, i.e., the extensive margin). Hence, define the gain from deviation as follows.

Definition 2. *The gain from deviation* Δ *is*

$$\begin{aligned}\Delta(1, \delta^o; A, c) &\equiv V(1, \delta^o \mid SL(A - c)) - V(1, \delta^o \mid SL(A)) \\ &= \mathbb{E}_s\left[A - c + u(0, \delta^o, s)\right] - V(1, \delta^o \mid SL(A)),\end{aligned}$$

$$\begin{aligned}\Delta(0, \delta^{no}; B, c) &\equiv V(0, \delta^{no} \mid BL(B + c)) - V(0, \delta^{no} \mid BL(B)) \\ &= \mathbb{E}_s\left[u(1, \delta^{no}, s) - (B + c)\right] - V(0, \delta^{no} \mid BL(B)).\end{aligned}$$

That is, by undercutting the prevailing ask, the seller guarantees execution. Upon execution, the seller receives the undercutting ask $A - c$, enters the OTC stage as a non-owner, and obtains the continuation value $u(0, \delta^o, s)$. Taking expectations over aggregate states gives the ex ante value, $\mathbb{E}_s\left[A - c + u(0, \delta^o, s)\right]$. The expression for a non-owner follows symmetrically.

However, when the bid–ask spread equals one tick, i.e., $A - B = c$, quote improvement is infeasible. This is because any limit order (passive order) quoted aggressively enough to cross the bid–ask spread becomes a market order (aggressive order), rather than entering the book as a resting limit order. That is, when $A - c = B$, there is no allowable tick between the bid and ask, so quoting one tick below the current ask (or one tick above the current bid) crosses the spread and becomes a

market order. Consequently, such a deviation is not feasible for sell (or buy) limit order traders.[18]

3.1.8 Equilibrium Characterization

The following definitions, lemmas, and corollaries are useful for the subsequent analysis.

Definition 3. *Define the OTC reservation value of the asset as* $R(\delta, s) \equiv u(1, \delta, s) - u(0, \delta, s)$.

Lemma 1. *For every* $s \in \{+, -\}$*, the OTC reservation value* $R(\delta, s)$ *is weakly increasing in* δ.

Corollary 1. *The market-order-stage decision rules in Equation 11 can then be rewritten as*

$$\mathbf{1}_{SM}(\delta, s) = \mathbf{1}\{B > R(\delta, s)\},$$
$$\mathbf{1}_{BM}(\delta, s) = \mathbf{1}\{A < R(\delta, s)\}.$$

Lemma 1 and Corollary 1 together provide intuition for why solving the optimality condition in Equilibrium 1 reduces to determining the cutoff types. This is because the type-invariant equilibrium objects of the CLOB, such as ask A and bid B, partition the type-specific equilibrium objects of the OTC, such as reservation value $R(\delta, s)$, and thereby partition the type space and gives rise to equilibrium sorting. In addition, Corollary 1 provides a natural economic interpretation of the market-order decision rules. An owner attempts to sell the asset immediately (i.e., using market order) after observing the aggregate state s only if the selling price (the bid price B) exceeds his reservation value, while a non-owner attempts to buy the asset immediately only if the buying price (the ask price A) is below her reservation value. The proof of Lemma 1 is provided in Appendix B.1, and Corollary 1 follows immediately from the lemma.

Now I turn to characterizing the deviation condition in Equilibrium 1, which pins down the quotes. Substituting $V(1, \delta \mid SL)$ and $V(0, \delta \mid BL)$ expressed in Equations 9 and 10, into the gain from deviation Δ defined in Definition 2, and collecting the continuation value terms and expressing them in terms of the reservation value $R(\delta, s)$, yields

$$\Delta(1, \delta^o; A, c) = \mathbb{E}_s\left[(1 - \phi_s^{SL})\big(A - R(\delta^o, s)\big)\right] - c, \tag{16a}$$
$$\Delta(0, \delta^{no}; B, c) = \mathbb{E}_s\left[(1 - \phi_s^{BL})\big(R(\delta^{no}, s) - B\big)\right] - c. \tag{16b}$$

Here, the first component represents the marginal benefit of improving the quote, increasing the execution probability from ϕ^{SL} to 1, while the second component represents the marginal cost of improving the quote, sacrificing a one-tick gain c.

[18] In practice, a trader can submit a sell limit order at the best bid only if the order quantity is sufficiently large to exhaust the entire buy-limit quantity at the best bid. The remaining unmatched quantity may then rest in the book as a sell limit order at the previous best bid, which subsequently becomes the best ask. However, in this model, a single-agent deviation can never exhaust the entire quantity at the best quote.

The terms $\Delta(1, \delta^o; A, c)$ and $\Delta(0, \delta^{no}; B, c)$ satisfy the following monotonicity property stated in Lemma 2. As will be shown, this is a useful property because it reduces the dimensionality of the problem. The proof is in Appendix B.2.

Lemma 2. *$\Delta(1, \delta; A, c)$ is weakly decreasing in δ, whereas $\Delta(0, \delta; B, c)$ is weakly increasing in δ.*

Given the monotonicity in Lemma 2, it is sufficient to verify the no deviation condition only for the lowest-type sell limit trader and the highest-type buy limit trader, rather than for all limit order traders. Thus, it reduces the dimensionality of the problem and facilitates the analytical solution. Lemma 2 also aligns with economic intuition. Among all sell limit traders, the lowest type has the greatest urgency to sell and therefore the strongest incentive to deviate by undercutting the ask to gain execution priority. Among all buy limit traders, the highest type has the greatest urgency to buy and therefore the strongest incentive to deviate by overbidding the bid to gain execution priority. The following definition and corollary formalize this simplified equilibrium condition.

Definition 4. *Define four cutoff types as*

$$\delta^{SL-1} \equiv \inf\{\delta \mid d^*(\delta) = SL\}, \qquad \delta^{SL-2} \equiv \sup\{\delta \mid d^*(\delta) = SL\},$$
$$\delta^{BL-1} \equiv \inf\{\delta \mid d^*(\delta) = BL\}, \qquad \delta^{BL-2} \equiv \sup\{\delta \mid d^*(\delta) = BL\}.$$

Corollary 2. *If $A - B > c$, the no one-shot deviation condition in Equilibrium 1 reduces to*

$$\Delta(1, \delta^{SL-1}; A, c) = \mathbb{E}_s\left[(1 - \phi_s^{SL})\big(A - R(\delta^{SL-1}, s)\big)\right] - c \le 0,$$
$$\Delta(0, \delta^{BL-2}; B, c) = \mathbb{E}_s\left[(1 - \phi_s^{BL})\big(R(\delta^{BL-2}, s) - B\big)\right] - c \le 0.$$

When $A - B = c$, quote improvement is infeasible. Therefore, the no one-shot deviation condition is automatically satisfied.

Corollary 2 also shows that the tick size c is a market-design parameter, because it affects agents' ex ante incentives to undercut the ask and overbid the bid through the term Δ. This incentive determines the equilibrium bid and ask, which in turn determine all other equilibrium objects, including the aggregate order quantities N.[19] The tick size c affects the incentive through the Bertrand-style price competition: it changes the tradeoff between execution certainty and transaction price. This is because, to gain execution priority, an agent must undercut the current ask or overbid the current bid by at least one tick. The larger the tick size, the more costly it becomes to acquire such execution priority, thereby reducing the incentive to undercut or overbid, in line with Corollary 2, which establishes that Δ decreases with c.

[19] This is closely related to the literature on the effects of tick size on order-book depth and provides one possible microfoundation for this relationship.

To summarize, the equilibrium is obtained in two steps: first, solving the venue-sorting problem for a given set of quotes, and second, determining which quotes can be sustained in equilibrium. Specifically, given the bid and ask, the optimal venue choices d^* and aggregate equilibrium objects (N, ϕ, G) are characterized by cutoff types, corresponding to marginal agents who are indifferent between two trading methods. The no one-shot deviation condition Δ then pins down the equilibrium bid and ask, with the condition needing to be imposed only at the relevant cutoffs. An equilibrium is obtained when such deviations are not profitable or when the constraint binds, i.e., at one tick because further quote improvement is no longer feasible.

To shed light on how risk (α), friction or fragmentation (λ), and the wedge (c) distort the equilibrium, I first present the frictionless economy obtained by setting these parameters to zero, which admits a fully closed-form solution while preserving the key economic forces that determine the equilibrium. The frictionless equilibrium coincides with the Walrasian equilibrium. This result is stated in the following Proposition 1 and the proof is provided in Appendix B.3.

Proposition 1 (Frictionless Economy)**.** *Suppose that there is no aggregate risk, $\alpha = 0$, and no over-the-counter venue fragmenting the market, $\lambda = 0$. As the tick size converges to zero, $c \to 0^+$, the central limit order book competitive equilibrium converges to the Walrasian equilibrium. Moreover, for any sufficiently small $c > 0$, limit orders are rationed, $0 < \phi^{SL} < 1$, $0 < \phi^{BL} < 1$, whereas market orders execute with certainty, $\phi^{SM} = \phi^{BM} = 1$.[20] As $c \to 0^+$, the limit-order execution probabilities converge monotonically to one from below, $\phi^{SL}, \phi^{BL} \uparrow 1$. Therefore, in the frictionless limit, $\alpha = \lambda = c = 0$, $\phi^{SL} = \phi^{BL} = \phi^{SM} = \phi^{BM} = 1$, so there is no failure to trade.*

Proposition 1 identifies an inefficiency in this market: the failure to trade ($\phi < 1$), which can be viewed as a specific form of market failure. Specifically, the failure to trade in the central limit order book arises from rationing, that is, from the imbalance between the two types of orders on the two sides of the market induced by endogenous venue sorting, summarized by the equilibrium object ϕ. This form of inefficiency is in addition to the canonical search friction in over-the-counter, which also generates inefficiency in this market (e.g., Duffie, Gârleanu, and Pedersen (2005), Weill (2020)). I now relax these parametric assumptions step by step and characterize the full equilibrium. I first open the over-the-counter venue ($\lambda > 0$) while maintaining the risk-free environment ($\alpha = 0$), and then introduce aggregate risk into the environment ($\alpha > 0$). The theoretical solutions for these two environments are presented separately in the next subsection.

[20] The economic intuition is that, in equilibrium, for a marginal owner to be indifferent between submitting a sell market order at the bid and a sell limit order at the ask, where the ask exceeds the bid, the sell limit order must face execution risk, i.e., $\phi^{SL} < 1$, to offset the benefit of a higher execution price. This is precisely the speed-cost tradeoff that sustains the equilibrium. The same logic applies to the non-owner side.

3.2 Theoretical Solution

3.2.1 Benchmark Case: $\alpha = 0$

For the benchmark case without aggregate risk, $\alpha = 0$, I first characterize the equilibrium analytically. I then present a numerical example and explain the equilibrium. Finally, I conduct comparative statics with respect to the tick size c. As discussed in the previous subsection, the tick size is a market design parameter, and therefore I examine how it affects the equilibrium outcomes.

The equilibrium characterization consists of two parts, venue sorting and quote deviation. Proposition 2 characterizes the equilibrium venue sorting by deriving the number and ordering of the cutoff types for given bid and ask quotes. Proposition 3 establishes the no profitable deviation condition that pins down the equilibrium bid and ask. In Proposition 2, the cutoff types correspond to agents who are indifferent between two trading-method options, where the indifference conditions compare the corresponding OTC reservation values with the CLOB equilibrium objects. In Proposition 3, the bid–ask spread equals one tick. The equilibrium is unique because the equilibrium bid–ask spread is unique, and the venue sorting induced by the unique bid–ask spread is also unique. Notice that when there is no aggregate risk, the state index s is redundant, so I omit it. The proofs of Propositions 2 and 3 are provided in Appendices B.4 and B.5, respectively.

Proposition 2 (Venue Sorting)**.** *When $c > 0$, given bid B and ask A, suppose limit orders are rationed on both sides of the CLOB, $0 < \phi^{SL} < 1$, $0 < \phi^{BL} < 1$. Then the optimal venue choice partitions the type space according to the four cutoff types defined in Definition 4 as follows:*

For owners ($q = 1$),

$$\begin{aligned} \delta < \delta^{SL-1} \quad &\Leftrightarrow \quad \textit{Sell Market}, \\ \delta^{SL-1} < \delta < \delta^{SL-2} \quad &\Leftrightarrow \quad \textit{Sell Limit}, \\ \delta > \delta^{SL-2} \quad &\Leftrightarrow \quad \textit{Opt Out CLOB}. \end{aligned}$$

At cutoff δ^{SL-1}, the agent is indifferent between Sell Market and Sell Limit, while at cutoff δ^{SL-2}, the agent is indifferent between Sell Limit and Opt Out CLOB.

For non-owners ($q = 0$),

$$\begin{aligned} \delta < \delta^{BL-1} \quad &\Leftrightarrow \quad \textit{Opt Out CLOB}, \\ \delta^{BL-1} < \delta < \delta^{BL-2} \quad &\Leftrightarrow \quad \textit{Buy Limit}, \\ \delta > \delta^{BL-2} \quad &\Leftrightarrow \quad \textit{Buy Market}. \end{aligned}$$

At cutoff δ^{BL-1}, the agent is indifferent between Opt Out CLOB and Buy Limit, while at cutoff δ^{BL-2}, the agent is indifferent between Buy Limit and Buy Market.

The cutoff types satisfy the following indifference conditions:

$$R(\delta^{SL-1}) = \frac{B - \phi^{SL} A}{1 - \phi^{SL}}, \qquad R(\delta^{SL-2}) = A,$$
$$R(\delta^{BL-1}) = B, \qquad R(\delta^{BL-2}) = \frac{A - \phi^{BL} B}{1 - \phi^{BL}}.$$

The ordering of the four cutoff types is

$$\delta^{SL-1} < \delta^{BL-1} < \delta^{SL-2} < \delta^{BL-2}.$$

Proposition 3 (Quote Deviation)**.** *The deviation gain presented in Corollary 2 satisfies*

$$\Delta(1, \delta^{SL-1}) = \Delta(1, \delta^{BL-2}) = (m-1)c, \quad \forall m \in \{1, 2, 3, \dots\}.$$

Therefore, for all $c > 0$, the unique equilibrium bid–ask spread that results in no profitable deviation is $m = 1$, i.e., the equilibrium bid–ask spread equals one tick, and the deviation gain is zero.

Corollary 3. *When the tick size converges to zero, $c \to 0^+$, the unique equilibrium bid–ask spread converges to zero, i.e., $\lim_{c\to 0^+}(A - B) = \lim_{c\to 0^+} c = 0$. In this limit, the competitive equilibrium replicates the Walrasian allocation, where $A = B = P^*$ and $\phi^{SL} = \phi^{BL} = \phi^{SM} = \phi^{BM} = 1$, so there is no failure to trade.*

Corollary 4. *When the tick size converges to zero, $c \to 0^+$, in this limit, no agent chooses to trade in over-the-counter. The central limit order book alone achieves the Walrasian allocation, even when the over-the-counter venue is available. That is, the market endogenously resists fragmentation.*

Taken together, Propositions 2, 3, and Corollaries 3, 4 characterize the equilibrium and its limiting properties within the limit-order rationing configuration. Specifically, the conditions $0 < \phi^{SL} < 1$, $0 < \phi^{BL} < 1$ select the equilibrium configuration in which limit orders constitute the long side of the CLOB and therefore face uncertain execution, while market orders constitute the short side and execute with certainty. The relevance of this equilibrium configuration is supported in two ways. First, in the analytically tractable frictionless economy of Proposition 1, with $\alpha = \lambda = 0$, the equilibrium exhibits exactly this execution pattern for a positive tick size: limit orders are rationed, whereas market orders execute with certainty. Second, I numerically verify that the limit-order rationing conditions hold for equilibrium reported in the risk-free environment. Thus, the equilibrium configuration underlying Proposition 2 – Corollary 4 is both analytically grounded in the frictionless economy and numerically relevant for the equilibrium studied here.

I now proceed to present one numerical example in Figure 5, which illustrates the equilibrium. Figure 5 shows that low-type owners ($[99, 99.44]$, the blue region) choose market orders to sell the

asset at the bid price B, middle-type owners ($[99.44, 100.1]$, the red region) choose limit orders to sell the asset at the ask price A, and high-type owners ($[100.1, 101]$, the black region) opt out and do not participate in the central limit order book. Vice versa for non-owners. All agents use their post-CLOB updated holdings to enter the over-the-counter venue, where they trade if and only if they meet a counterparty such that there are positive gains from trade. The ownership distribution evolves endogenously. I illustrate the support of the distributions of owners and non-owners entering the over-the-counter venue (the green region), that is, the subsets of the type space on which g^{OTC-O} and g^{OTC-NO} are strictly positive.

This equilibrium venue sorting can be interpreted as the classic speed–cost tradeoff. In this numerical example, market orders are on the short side, while limit orders are on the long side. Consequently, limit orders face execution risk ($\phi^{SL} = \phi^{BL} < 1$), whereas market orders are guaranteed execution ($\phi^{SM} = \phi^{BM} = 1$) but must cross the bid–ask spread, resulting in a worse execution price. Therefore, low-type owners and high-type non-owners, the most urgent agents, submit market orders because they substantially value guaranteed execution even though they must pay the bid–ask spread, i.e., the transaction cost. These agents demand liquidity, consistent with the insight of Demsetz (1968) that immediacy is costly. Middle-type owners and middle-type non-owners, the middle-urgency agents, submit limit orders and provide liquidity. High-type owners and low-type non-owners, patient agents, opt out of the centralized venue and may instead trade in the over-the-counter venue.

This speed–cost interpretation can be further validated using the data. In the data, market orders are executed immediately with essentially zero delay. Limit orders are executed with uncertainty (i.e., many are never filled) and, on average, wait about two minutes before execution. Although I do not have data on the execution time of over-the-counter trades, it is reasonable to assume that they take longer than two minutes because they often involve manual quote requests and bilateral negotiation. Therefore, the over-the-counter venue can be interpreted as the slowest trading venue. The interpretation of equilibrium sorting as a speed–cost trade-off is also broadly consistent with the substitution pattern presented in Table 1.

This venue sorting suggests that the centralized venue plays a first-order role in allocative efficiency. This is because owners whose types are below the Walrasian price (100 in the numerical example presented in Figure 5) and non-owners whose types are above the Walrasian price are initially misallocated, and they endogenously trade through the centralized venue. Moreover, the agents whose failure to trade would generate the largest inefficiency are low-type owners and high-type non-owners, and these agents choose to submit market orders, which are always executed in this benchmark case. Therefore, the most severe misallocation is resolved by the central limit order book. In this sense, the centralized venue improves allocative efficiency.

I now use comparative statics with respect to the tick size to examine how this market-design

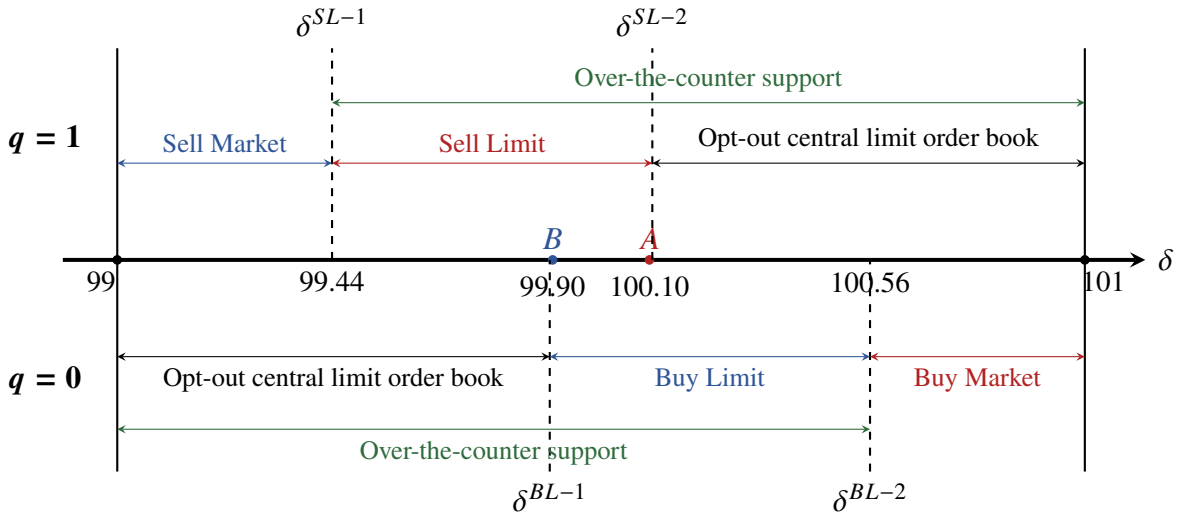


Figure 5: Optimal Venue Choice: A Numerical Example ($\alpha = 0$)

Parameters: tick-size $c = 0.2$, shock size $\alpha = 0$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = kf(\delta)$, asset supply $k = 0.5$.

parameter affects equilibrium. The parameter configuration in Figure 6 corresponds to a symmetric equilibrium, so I report only the owner-side venue choice and the sell limit order execution probability, and the non-owner-side venue choice and the buy limit order execution probability are symmetric. When the tick size converges to zero, the equilibrium converges to the Walrasian benchmark (the purple point), as stated in Corollary 3. Given Proposition 3, the equilibrium bid–ask spread is always one tick, regardless of the tick size. Therefore, in Panel (a), as the tick size increases, the bid–ask spread also increases. Since the bid–ask spread is the cost of using a market order to obtain execution certainty, the mass of agents choosing market orders decreases (the blue region shrinks) as the tick size increases. In contrast, limit orders avoid this transaction cost, so a larger tick size and the associated wider bid–ask spread induce more agents to submit limit orders instead (the red region expands). Although limit orders save transaction costs, they always face execution risk in this benchmark case, as shown in Panel (b). That is, $\frac{N^{BM}}{N^{SL}} \leq 1$, therefore execution probability $\phi^{SL} = \min\{\frac{N^{BM}}{N^{SL}}, 1\} \leq 1$. As the tick size increases, since the mass of limit orders rises while the mass of market orders falls, the execution probability of limit orders decreases, as shown in Panel (b). By contrast, market orders are always executed with certainty because $\phi^{BM} = \min\left\{\frac{N^{SL}}{N^{BM}}, 1\right\} = 1$. Therefore, the speed–cost trade-off illustrated in the previous numerical example is preserved throughout the parameter space.

Figure 6 suggests that an excessively large tick size may explain why, in some fixed-income markets, the central limit order book is available but little or no trading occurs through it. Panel (c) shows that the over-the-counter market share, measured by the share of executed volume, increases with the tick size. Beyond a threshold tick size, no trade is realized through the CLOB, and all

trading occurs OTC, as indicated by the gray region. This occurs because the tick size, and hence the implied bid–ask spread, becomes too large, making the transaction cost of obtaining execution certainty through market orders prohibitively high. As a result, no agents are willing to cross the spread by submitting market orders. Without market orders, limit orders cannot be filled, and hence no trade is realized through the CLOB. Some empirical evidence from equity markets supports the predicted negative relationship between tick size and centralized venue's market share (Kwan, Masulis, and McInish (2015), Werner et al. (2023), etc.).[21]

3.2.2 General Case: $\alpha > 0$

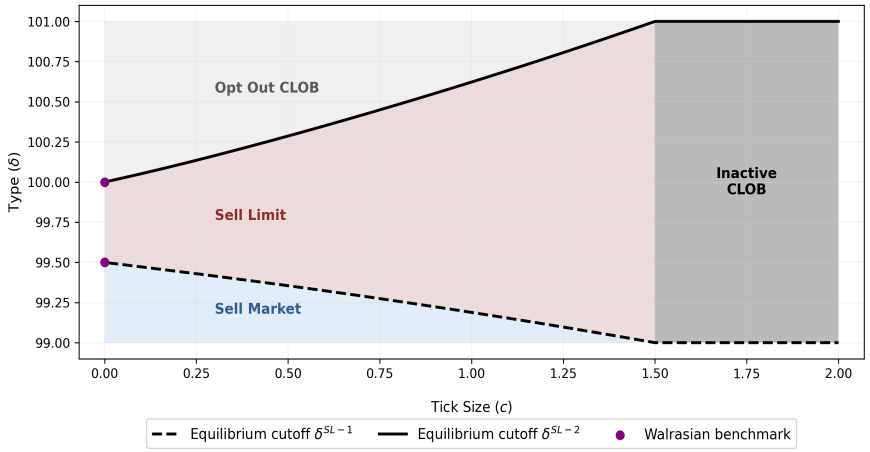
For the general case with aggregate risk, $\alpha > 0$, I first characterize the equilibrium, present a numerical example, and contrast it with the previous benchmark case. The key difference is that venue sorting features informed market orders, which generate adverse selection for limit orders, as stated in Proposition 4. I provide a simplified proof in Appendix B.6.[22] I then conduct comparative statics with respect to the shock size α and examine its effects on the equilibrium both ex ante and ex post, highlighting the quantity-margin channel. The parameter α can be interpreted as the asset's risk. To see this, one can simply calculate the variance of the ex post efficient price $\text{Var}(P_s^*) = \alpha^2$.

Proposition 4 (Market orders are informative about the aggregate state; therefore, limit orders are adversely selected)**.** *In equilibrium, sell-market order flow is weakly larger in the negative state, $N_-^{SM} \geq N_+^{SM}$, and buy-market order flow is weakly larger in the positive state, $N_+^{BM} \geq N_-^{BM}$. Therefore, sell limit orders execute weakly more often in the positive state, $\phi_+^{SL} \geq \phi_-^{SL}$, and buy limit orders execute weakly more often in the negative state, $\phi_-^{BL} \geq \phi_+^{BL}$. That is, limit orders are weakly more likely to execute when the common value of the asset moves against them.*

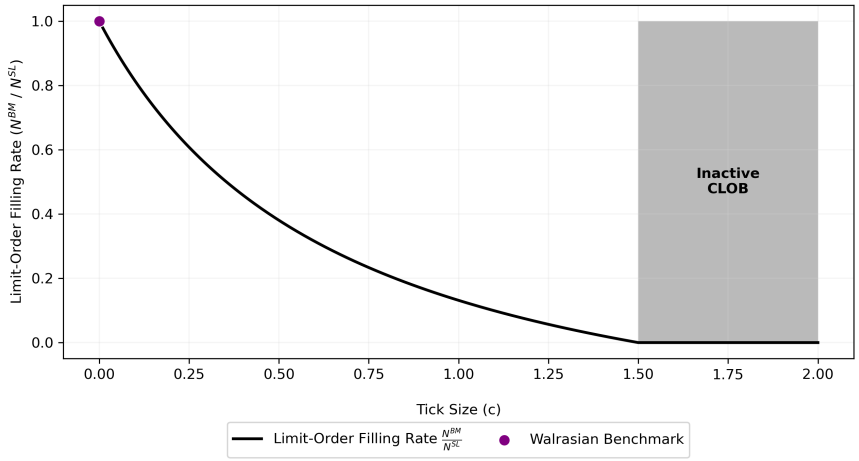
The quote deviation forces the bid–ask spread to the one-tick constraint. Unlike in the benchmark case, here the deviation gain Δ at the one-tick spread remains strictly positive. The numerical

[21] For example, Kwan, Masulis, and McInish (2015) exploit a regulatory discontinuity in the U.S. equity market, where the minimum tick size increases from 0.0001 for stocks priced below one dollar to 0.01 for stocks priced at or above one dollar. Using a regression discontinuity design, they find that crossing the one-dollar threshold discontinuously reduces the market share of centralized venue (see Table 2 and Figure 2). Similarly, Werner et al. (2023) study the reduction in tick size following the termination of the U.S. Tick Size Pilot in 2018 and find that smaller tick sizes increase the market share of centralized venue, particularly for stocks whose bid–ask spreads were previously constrained by the minimum tick size (see Table 6, Panel A). One should note that the empirical finance literature on the relationship between tick size and centralized venue market share is mixed, though.

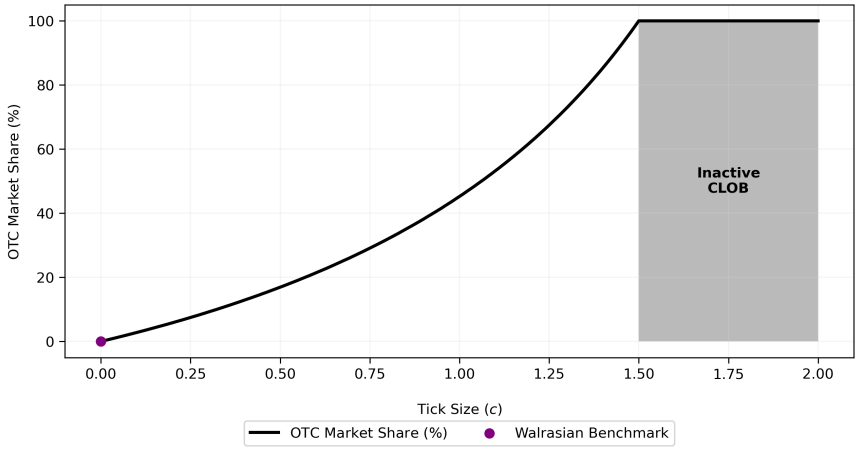
[22] Specifically, I prove the proposition only in an environment without over-the-counter trading ($\lambda = 0$) and with aggregate risk ($\alpha > 0$). This simplified proof nevertheless sheds light on the theoretical mechanism underlying the proposition: the aggregate state shifts agents' reservation values, generating excess sell market orders in the negative state and excess buy market orders in the positive state ("trading pressure"). Extending the proof analytically to the case with $\lambda > 0$ requires establishing the property $R(\delta, +) \geq R(\delta, -)$ for all δ, for which I have not found a general analytical argument. When $\lambda = 0$, the reservation value R simplifies to $R(\delta, s) = \delta + \epsilon_s$, making this property immediate. Nevertheless, all numerical analysis in this paper satisfy $R(\delta, +) \geq R(\delta, -)$ and support the proposition in the general setting with $\lambda > 0$.



(a) Owner-side Venue Choice



(b) Sell Limit Order Filling Rate



(c) Over-the-Counter Market Share

Figure 6: Comparative Statics with respect to the Tick Size

Parameters: shock size $\alpha = 0$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

solution suggests that, when $\alpha > 0$, the deviation gain Δ increases with the bid–ask spread and, importantly, is close to zero but remains slightly positive when the bid–ask spread is constrained at one tick (see Appendix C). Therefore, the one-tick-spread equilibrium arises not because undercutting or overbidding is strictly unprofitable (as in the benchmark case, where $\Delta = 0$ in equilibrium), but because further undercutting or overbidding becomes infeasible once the spread is constrained to one tick. Since Δ is already slightly positive in the one-tick-spread equilibrium, equilibria with larger bid–ask spreads cannot be sustained because they admit profitable deviations.

Contrasting the numerical example presented in Figure 7 with the benchmark case in Figure 5 shows that venue sorting responds to aggregate risk through the endogenous emergence of informed market orders. In this numerical example, informed market orders are sell market orders submitted by owners whose types lie between $\delta^{SL-2} = 99.66$ and $\delta^{SM}_{-} = 100.25$, and buy market orders submitted by non-owners whose types lie between $\delta^{BM}_{+} = 99.75$ and $\delta^{BL-1} = 100.34$. These are informed market orders because these agents submit market orders in one aggregate state but not in the other. Consequently, their order flow is informative about the aggregate state (i.e., information-driven). By contrast, low-type owners (owners whose types lie between $\underline{\delta} = 99$ and $\delta^{SL-1} = 99.17$) and high-type non-owners (non-owners whose types lie between $\delta^{BL-2} = 100.83$ and $\bar{\delta} = 101$) always submit market orders regardless of the aggregate state. Equivalently, their trading decisions do not respond to the arrival of aggregate shock, so their market orders are not informative about the aggregate state. Following the convention in the literature, these agents can instead be interpreted as liquidity-driven, since they have the greatest trading urgency.

The informed market orders generate adverse selection for limit orders. As shown in Figure 7, the mass of limit orders is state invariant because it is determined ex ante. However, $N^{SM}_{-} > N^{SM}_{+}$, where the difference is generated by the informed sell market orders. Consequently, $\phi^{BL}_{-} > \phi^{BL}_{+}$. In this numerical example, $\phi^{BL}_{+} = \min\left\{\frac{99.17-99}{100.83-100.34}, 1\right\} \approx 0.35 < 1$, whereas $\phi^{BL}_{-} = \min\left\{\frac{(99.17-99)+(100.25-99.66)}{100.83-100.34}, 1\right\} = \min\left\{\frac{0.76}{0.49}, 1\right\} = 1$. The same argument applies symmetrically to sell limit orders. This numerical example illustrates Proposition 4. That is, buy limit orders are more likely to be filled when the common value of the asset decreases, reflecting the winner's curse in this common-value setting. Since the execution probabilities ϕ are state dependent, the law of motion implies that the support of the over-the-counter distribution g^{OTC} is also state dependent, which in turn implies that the OTC price distribution is state dependent, in particular its skewness, as can be immediately seen from this figure's asymmetric over-the-counter support across the two states. Therefore, aggregate shocks are absorbed ex post through both the quantity and price margins. On the quantity margin, market order flow becomes one-sided, with an excess mass of sell market orders following a negative shock and an excess mass of buy market orders following a positive shock. On the price margin, the OTC transaction price distribution becomes right-skewed in the negative state and left-skewed in the positive state. In contrast, both market order flow and

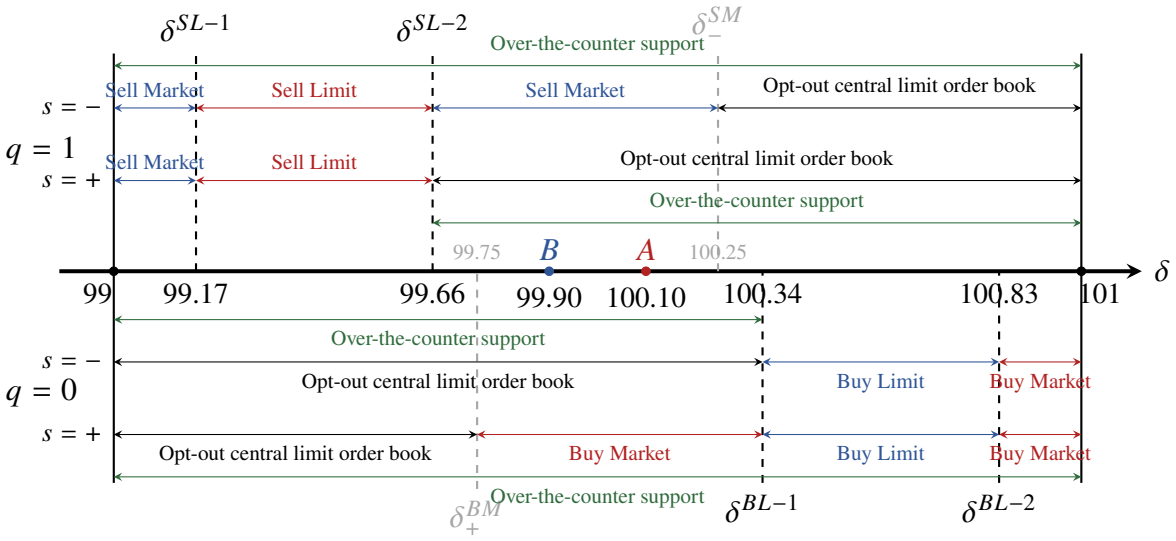


Figure 7: Optimal Venue Choice: A Numerical Example ($\alpha = 0.3$)

Parameters: tick-size $c = 0.2$, shock size $\alpha = 0.3$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = kf(\delta)$, asset supply $k = 0.5$.

the OTC transaction price distribution are symmetric in the benchmark case.

The comparative statics with respect to the shock size α are presented in Figure 8, which shows that agents incorporate aggregate risk ex ante through the *quantity* (extensive) margin, the mass of limit orders. As shown in Panel (a), as the shock size increases, the mass of owners choosing limit orders (the red region) shrinks. In other words, the mass of limit orders decreases with aggregate risk, consistent with the theoretical intuition that greater risk discourages liquidity provision. In addition, ex post, the mass of informed market orders (the green region) increases with aggregate risk, making the adverse selection faced by limit orders more severe. As shown in Panel (b), this is reflected in the widening gap in limit-order execution probabilities across states (the hatched area). As a result, Panel (c) shows that the over-the-counter market share increases with aggregate risk, suggesting that agents substitute away from the central limit order book toward the over-the-counter as aggregate risk increases.

Notice that, although the model parametrizes aggregate risk ϵ_s to have zero mean and all agents are risk neutral, equilibrium venue choice still responds to aggregate risk. The reason is that execution probabilities (extensive margin) enter the value functions and are thus payoff-relevant. As previously shown, when α increases, the adverse selection faced by limit orders becomes severe, so only lower type owners are still willing to continue providing liquidity at the same ask price.[23] Intuitively, more severe adverse selection faced by limit-order traders requires a larger gain from trade as compensation. Since lower-type owners derive a larger gain from submitting sell limit orders,[24]

[23] In Figure 8, the bid–ask spread remains fixed at one tick, and the tick size is held constant.

[24] The ex ante surplus from submitting a sell limit order over holding the asset is $\mathbb{E}_s[\phi_s^{SL}(A - R(\delta, s))]$, which increases as δ decreases.

the sell-limit region shifts downward along the type distribution as α increases to provide such compensation. Eventually, the sell-limit region reaches the lower bound of the type distribution, at which point no owner finds it optimal to submit a limit order and the order book is empty.[25]

These comparative statics also illustrate that informed market orders, and the associated adverse-selection problem, arise only when aggregate risk is sufficiently large relative to the tick size. In this numerical example, informed market orders emerge only when $\alpha > 0.2$. When aggregate risk is small relative to the tick size, it does not affect agents' trading decisions because the shock is not large enough to move the common value of the asset above the ask or below the bid. Consequently, the mass of limit orders remains the same as in the case without aggregate risk (the flat region following $\alpha = 0$ in Panel (a) of Figure 8), recovering the no-risk level of book depth.

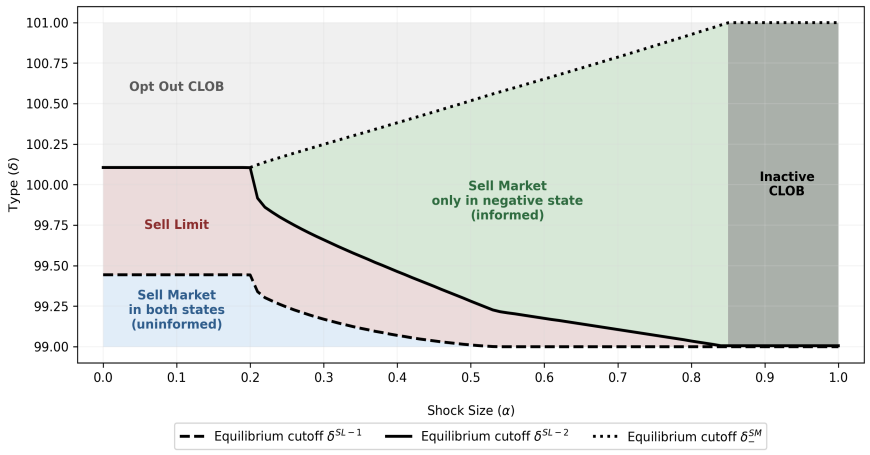
The analysis in this subsection suggests another reason why few or no trades are executed through the central limit order book in some fixed-income markets, even when the technology is available. Unlike the benchmark case, in which CLOB inactivity arises because no agents submit market orders due to an excessively large tick size, here it arises because no agents are willing to submit limit orders and commit to firm quotes when large aggregate shocks may occur afterward. This might rationalize why CLOB liquidity can collapse during periods of heightened aggregate uncertainty while over-the-counter venues remain active. More generally, CLOB trading requires the participation of both market-order and limit-order traders, introducing a coordination dimension to market design: if either side is absent, no trades can occur on the CLOB, and all trading takes place through the over-the-counter venue.

3.3 Allocative Efficiency

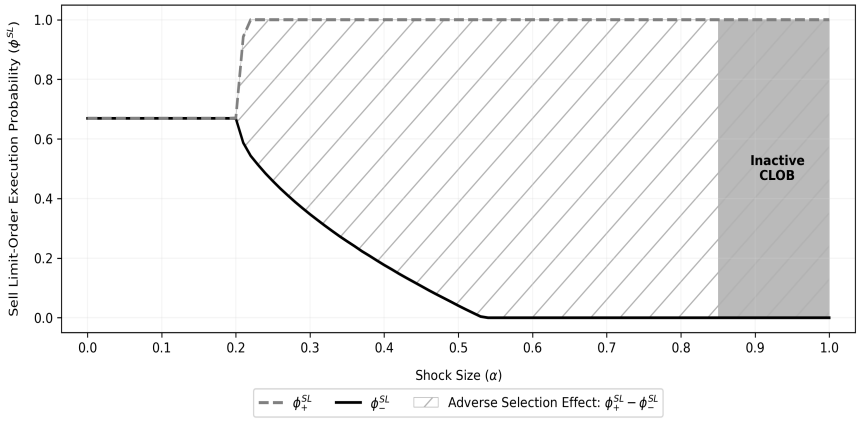
This section discusses the implications of the model for market structure and market design by analyzing allocative efficiency. I use the Walrasian allocation as the benchmark for evaluating allocative efficiency. To construct, I first define ex post allocative efficiency in state s as follows:

$$\begin{aligned}\text{Allocative Efficiency}_s &\equiv \frac{\text{Realized gains from trade}}{\text{First-best gains from trade}} \\ &= \frac{\displaystyle\int_{\underline{\delta}}^{\bar{\delta}} (\delta + \epsilon_s)\left[\tilde{\mu}_s(1,\delta) - \mu(1,\delta)\right] d\delta}{\displaystyle\int_{F^{-1}(1-k)}^{\bar{\delta}} (\delta + \epsilon_s) f(\delta) d\delta - \int_{\underline{\delta}}^{\bar{\delta}} (\delta + \epsilon_s)\mu(1,\delta) d\delta}. \end{aligned} \tag{17a}$$

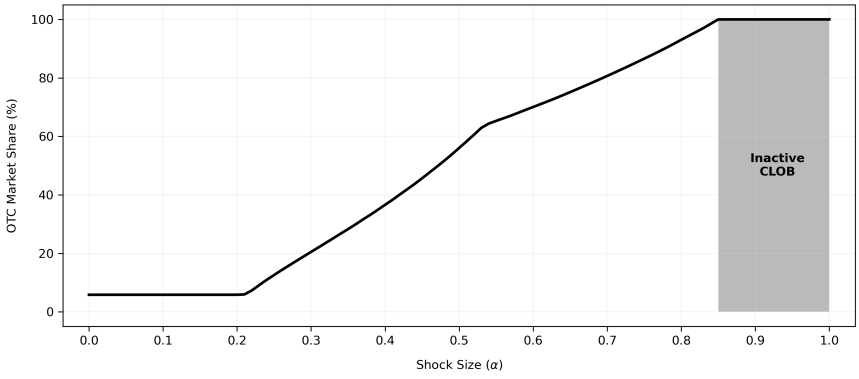
[25] Notice that the empty book is a special result of assuming that the type distribution has bounded support. Alternatively, with an unbounded type distribution, such as the normal distribution, all findings are qualitatively preserved, except that the book becomes very thin, but never empty, when risk is sufficiently large.



(a) Owner-side Venue Choice



(b) Sell Limit Order Execution Probability



(c) Over-the-Counter Market Share

Figure 8: Comparative Statics with respect to the Shock Size

Parameters: tick size $c = 0.2$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

The (ex ante) allocative efficiency is then given by

$$\text{Allocative Efficiency} = \mathbb{E}_s[\text{Allocative Efficiency}_s]. \tag{17b}$$

Here, $\mu(1,\delta)$ is the initial ownership density, $\tilde{\mu}_s(1,\delta)$ is the final ownership density, and the first-best allocation is the allocation in which the highest-k fraction of agents hold the asset. The density $\tilde{\mu}_s(1,\delta)$ in the notation of the model setup, with $v \equiv \delta + \epsilon_s$, is

$$\begin{aligned}\tilde{\mu}_s(1,\delta) =& N_s^{OTC-O} g_s^{OTC-O}(\delta+\epsilon_s)\left[1-\phi_s^{OTC-O}(\delta+\epsilon_s)\right] \\ &+ N_s^{OTC-NO} g_s^{OTC-NO}(\delta+\epsilon_s)\phi_s^{OTC-NO}(\delta+\epsilon_s),\end{aligned}$$

where the first component consists of post-CLOB owners who fail to sell in the OTC, and the second component consists of post-CLOB non-owners who successfully buy in the OTC.

Allocative efficiency is a percentage between 0% and 100%, and it is the welfare notion that this paper uses to evaluate market structure and market design. It measures the percentage of the maximal attainable welfare improvement that is realized through trading venues. Autarky ($\tilde{\mu} = \mu$) corresponds to 0%, while the first-best allocation corresponds to 100%. Viewing a venue as a mechanism (P, ϕ), it is the execution probability ϕ that determines allocative efficiency, whereas the transaction price P is purely a transfer. This can be seen from the fact that the model implies that the joint distribution over (q, δ) evolves as follows

$$\text{initial exogenous } \mu(\cdot,\delta) \xrightarrow{\phi_s^{\text{CLOB}}} g_s^{OTC}(\cdot,\delta) \xrightarrow{\phi_s^{\text{OTC}}} \text{final } \tilde{\mu}_s(\cdot,\delta). \tag{18}$$

3.3.1 Optimal Tick Size on the Central Limit Order Book

I use numerical methods to compute the optimal tick size for a given aggregate risk level α. The optimal tick size is defined as the tick size that maximizes allocative efficiency, given α and the other parameter values. Figure 9 presents one numerical result. As shown in the figure, when $\alpha = 0$, the optimal tick size (top panel) is zero, and allocative efficiency (bottom panel) is 100%. This corresponds to Corollaries 3 and 4. A numerical example illustrating the grid search for the optimal tick size for an arbitrary level of the shock size is presented in Appendix Figure A14.

The top panel of Figure 9 shows that the optimal tick size $c^*(\alpha)$ increases with aggregate risk α. This is because a larger tick size reduces the mass of informed market orders, as implied by Corollary 1. Specifically, as in Corollary 1, if the ask increases and the bid decreases (i.e., the bid–ask spread widens as the tick size increases; recall that the spread always binds at one tick), a smaller mass of agents sorts into market orders, since the OTC reservation value R is increasing in type. Importantly, a larger tick size first eliminates informed market orders while preserving uninformed

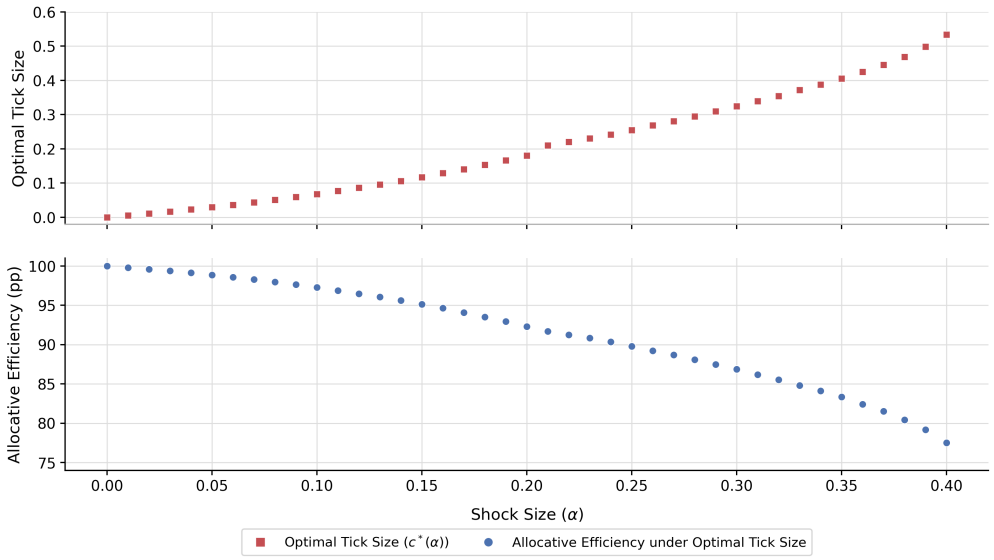


Figure 9: Allocative Efficiency under Optimal Tick Size

Parameters: search efficiency $\lambda = 0.5$, bargaining power $\theta = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

ones. The reason is that uninformed market-order traders are those with the most extreme types (lowest-type owner and highest-type non-owner), whose reservation values R are therefore extreme and lie far from the bid and ask, allowing them to continue submitting market orders despite the wider spread; in contrast, (marginal) market-order traders whose reservation values R lie close to the bid and ask are informed market-order traders, making them the first to switch away from market orders as the spread widens.

Nevertheless, the bottom panel of Figure 9 shows that even when the tick size is set at its optimal level, maximized allocative efficiency decreases with the level of aggregate risk. This is because, as previously illustrated in Figure 8, if the tick size is sufficiently large relative to the shock size, informed market orders disappear, adverse selection is eliminated, limit order book depth increases to the no-risk level, and the equilibrium is restored to the corresponding no-risk configuration. However, this no-risk equilibrium is itself distorted by the sufficiently large tick size, as can be seen from Figure 6, where the distortion from the Walrasian equilibrium increases with the tick size. The optimal tick size therefore balances this ex post distortion against the ex ante incentive effect.

Although the result in Figure 9 is based on a single numerical example, it offers a broader insight: the optimal tick size varies with aggregate risk. Even if the tick size is chosen optimally ex ante to maximize expected allocative efficiency, it need not remain optimal ex post because the realized level of aggregate risk may differ from its anticipated (e.g., average) level. This gives rise to tick-size misspecification, where a fixed tick size is no longer optimal for the prevailing level of aggregate risk. This is particularly relevant in practice because the tick size set by BrokerTec remains fixed over long periods, potentially spanning years or even decades, while aggregate risk in

the Treasury market can change substantially. Consequently, tick-size misspecification is an inherent possibility in a CLOB design with a fixed tick size, a specific form of pricing rigidity that can generate extra inefficiency. Appendix Figure A15 provides a numerical example illustrating the efficiency loss from such misspecification. One way to mitigate the efficiency loss from a misspecified tick size is through the operation of over-the-counter trading, which is one of the implications discussed in the next subsection.

3.3.2 Efficiency Gain from Over-the-Counter

Now I turn to the study of market structure by comparing allocative efficiency across CLOB-only, OTC-only, and hybrid market structures. The main welfare implication is that the hybrid market structure, in which an over-the-counter venue coexists with the central limit order book, is optimal.

I first examine the benchmark case in which there is no aggregate risk, and the allocative efficiency results are presented in Figure 10. When there is no aggregate risk, as shown previously in Corollary 3, $c = 0$ is optimal since it replicates the Walrasian allocation, and allocative efficiency is therefore maximized at 100% (purple). This result is independent of the level of search frictions in the over-the-counter venue (which is negatively correlated with the search-efficiency parameter λ). This follows from Corollary 4: as long as $c = 0$ and $\alpha = 0$, no agents use the over-the-counter venue, so the level of search frictions is welfare-irrelevant. As previously discussed, the tick size in this risk-free environment acts only as a wedge that distorts the allocation, and the larger the tick size, the larger the distortion. This can be seen from Figure 10, where the allocative efficiency under the wide-tick case ($c = 1.5$, dotted) is lower than that under the medium-tick case ($c = 1$, dash-dotted), which is in turn lower than that under the narrow-tick case ($c = 0.5$, dashed).

Figure 10 also suggests that, in the benchmark case with $\alpha = 0$, as long as the tick size is not set at its optimum, $c^*(0) = 0$, opening an over-the-counter venue strictly improves allocative efficiency. This can be seen by comparing allocative efficiency for $\lambda > 0$ with that for $\lambda = 0$, where $\lambda = 0$ corresponds to the absence of OTC trading. This corresponds to the aforementioned result that the over-the-counter venue mitigates the efficiency loss due to a misspecified tick size. That is, for a given tick size $c > 0$, allocative efficiency increases with search efficiency λ. Moreover, for a given level of search efficiency λ, comparing allocative efficiency with that under $\lambda = 0$ shows that the larger the tick size, the greater the distortion and, consequently, the larger the efficiency loss recovered through over-the-counter trading.

Second, I examine allocative efficiency in the general case in which aggregate risk is present. Which market structure is optimal depends on whether the tick size is optimal. Specifically, I present a numerical example in Figure 11. For most levels of aggregate risk α, the hybrid market structure (black) performs at least as well as either single-venue market structure, a CLOB-only market (red dashed) or an OTC-only market (blue dash-dotted). The only parameter configuration in which

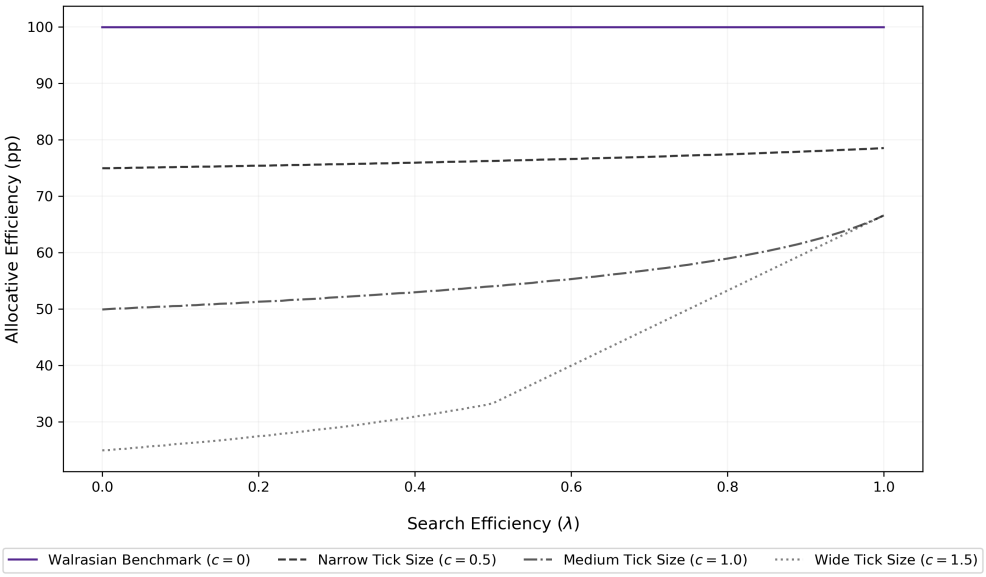


Figure 10: Allocative Efficiency: Benchmark Case ($\alpha = 0$)

Parameters: shock size $\alpha = 0$, bargaining power $\theta = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

fully centralizing the market dominates the hybrid market structure in terms of allocative efficiency is around the case in which the shock size α happens to coincide with the level at which the fixed tick size c is optimal (dashed vertical line, $\alpha \approx 0.23$). The main idea is therefore that centralization is desirable only if the central limit order book is designed with the optimal tick size; otherwise, allocative efficiency deteriorates. However, as suggested previously, there is no reason to expect the level of risk to remain permanently at this specific value, because aggregate conditions in the Treasury market change substantially over time. Therefore, a fully centralized market is unlikely to be an attractive organization. By contrast, introducing a central limit order book into a market organized solely through over-the-counter trading never reduces allocative efficiency, as can be seen from the gap between the blue and black curves.

Figure 11 shows that the allocative efficiency of OTC trading is invariant to risk and therefore robust, whereas the allocative efficiency of the CLOB is initially high but deteriorates sharply as risk increases, particularly beyond the level at which the fixed tick size is optimal. This formalizes the common intuition in the market-microstructure literature that CLOB ceases to function effectively during periods of market stress, whereas OTC is more resilient. In fact, in Figure 11, the regions in which the allocative efficiency of the hybrid market coincides with that of the OTC-only market are precisely those in which the shock is sufficiently large that the CLOB becomes inactive, as shown in Figure 8. Since no trades are executed through the CLOB in these regions, all gains from trade

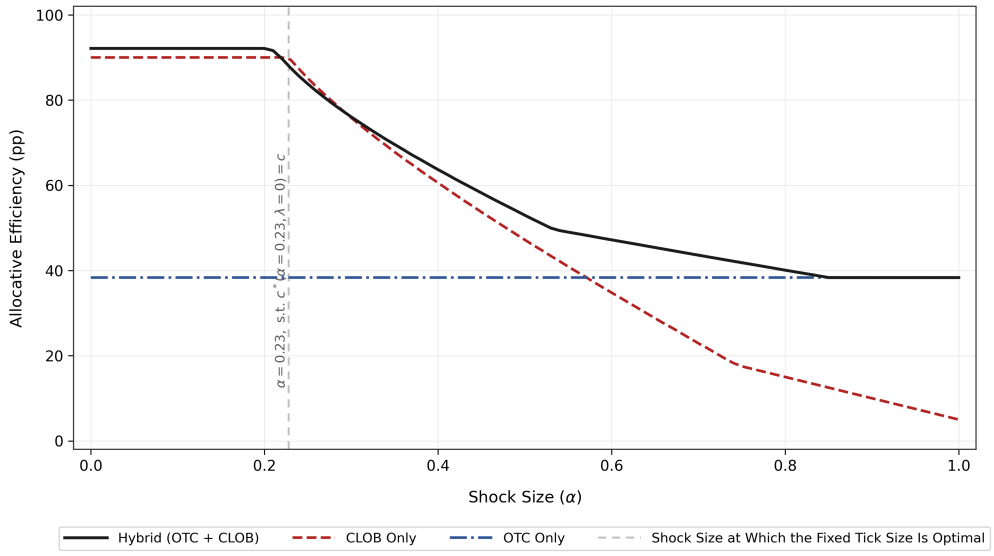


Figure 11: Allocative Efficiency: General Case ($\alpha > 0$)

Parameters: tick size $c = 0.2$, search efficiency $\lambda = 0.5$ for OTC only and Hybrid and $\lambda = 0$ for CLOB only, bargaining power $\theta = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = kf(\delta)$, asset supply $k = 0.5$.

are realized through OTC trading, and the two allocative-efficiency therefore coincide.[26]

To conclude the theoretical section, the key idea is that the trading mechanisms of the two venues are not perfect substitutes but instead exhibit a degree of complementarity, and the key welfare implication is that their coexistence constitutes the optimal market organization. The trading mechanisms formalized in Section 3.1.2 illustrate how the central limit order book complements the decentralized over-the-counter venue through their distinct pricing and matching mechanisms across different aggregate environments. The CLOB performs well when aggregate risk is low (no search inefficiencies) but deteriorates as aggregate risk increases (adverse selection concern), whereas OTC trading is robust to aggregate risk and therefore acts as a stabilizer that improves market resilience. In most cases, the allocative efficiency of the hybrid market structure exceeds the envelope of the two venues operating separately, i.e., max {OTC-only, CLOB-only}, providing direct evidence of this complementarity. The main welfare implication of the model is therefore that the coexistence of the two venues constitutes the optimal market structure and that policymakers should allow market participants to freely substitute across venues and determine the degree of fragmentation endogenously (supporting an intuition discussed in Chapter 7 of Foucault, Pagano, and Röell (2023)). I now turn to quantifying the model using the data.

[26]Notice that, under the CLOB-only market structure, allocative efficiency remains positive even for these values of α. This is because venue choices respond endogenously to the market structure. When the OTC venue does not compete with the CLOB, the CLOB is more active than under the hybrid structure. Therefore, for sufficiently large shocks, for example when $\alpha \to 1$, the CLOB remains active in the CLOB-only market, whereas the CLOB becomes inactive in the hybrid market because all agents optimally choose to opt out of the CLOB and trade in OTC instead.

4 Quantitative Analysis

4.1 Estimation

4.1.1 Minimum-Distance Estimator

The models with and without aggregate risk are estimated separately using different data samples. The model without aggregate risk uses the full sample, and the model with aggregate risk uses observations within the five minute windows before and after the shock event shown in Figure 2. I select this event because it is the cleanest identified shock event in the sample, exhibiting a sharp price discontinuity that is not immediately reversed. I estimate the model without aggregate risk using the subsample of observations whose pre-trade bid–ask spread equals one tick,[27] so that the data used for estimation are consistent with the model equilibrium. The pre-trade bid–ask spread equals one tick for 75% of the five-minute window preceding the shock event. This subsection presents the estimation results for using 10-year Treasury. The remaining maturities are estimated separately, and the corresponding results are reported in the Appendix D.

The parameters to be estimated are $\mathbf{\Gamma} \equiv \{\lambda, \sigma, \theta, k\}$. Notice that the theoretical results in Section 3.2 do not depend on parametric assumption for the type distribution, and a uniform distribution is used in those numerical examples solely for expositional simplicity.[28] For the estimation, I instead assume that the type δ follows a normal distribution, $\delta \sim N(100, \sigma^2)$, as a bell-shaped distribution is empirically more plausible than a uniform distribution. Because the location parameter is not identified, i.e., adding a common constant to all ex-ante types leaves the model-implied moments unchanged, I normalize the mean to 100, consistent with the conventional par value around which U.S. Treasury prices are quoted. The tick size is directly observed from the data and fixed at $c = 0.015625$. The aggregate shock size is set to $\alpha = 0.140625$, corresponding to a nine-tick sweep of the book, which is directly measured from the event shown in Figure 2 at the exact microsecond when the shock arrives.

I estimate the parameters $\mathbf{\Gamma}$ using a minimum-distance estimator (e.g., Gavazza (2016)) that matches model-implied moments to their empirical counterparts. Let M^{data} denote the vector of empirical moments and $M^{\text{model}}(\mathbf{\Gamma})$ the corresponding model-implied moments. The estimator solves

$$\widehat{\mathbf{\Gamma}} = \arg\min_{\mathbf{\Gamma}} Q(\mathbf{\Gamma}), \quad Q(\mathbf{\Gamma}) = g(\mathbf{\Gamma})' W g(\mathbf{\Gamma}), \quad g(\mathbf{\Gamma}) = M^{\text{model}}(\mathbf{\Gamma}) - M^{\text{data}},$$

where the weighting matrix is $W = \text{diag}\left(s_j^{-2}\right)$, $s_j = \left|M_j^{\text{data}}\right|$, the absolute value of the j-th empirical

[27] Using this restriction removes approximately 3% of the observations.

[28] This is because, under the uniform distribution, the limit-order filling rates $\frac{N^{BM}}{N^{SL}}$, $\frac{N^{SM}}{N^{BL}}$ are exactly equal to the fractions of the corresponding cutoff intervals. The only theoretical result that depends on the distributional assumption is stated in Footnote 25.

moment, so that the function Q measures squared proportional deviations from the data moments.

The targeted moments used for estimation are as follows. (i) The OTC market share, defined as the number of OTC trades divided by the total number of trades across the OTC and CLOB venues. (ii) OTC price dispersion, measured as the standard deviation of normalized OTC prices, where normalized OTC prices are defined as $P^{OTC} - P^*$.[29] (iii) OTC price skewness, defined as the share of OTC trades executed above the contemporaneous best ask minus the share executed below the contemporaneous best bid. (iv) Limit-order-book asymmetry, measured by the ratio of sell-side to buy-side depth at the best quotes, $\frac{N^{SL}}{N^{BL}}$, where the average sell-side and buy-side limit-order masses, N^{SL} and N^{BL}, are computed as time-weighted averages. For the model with aggregate risk, the sequential structure of the model determines the timing of the targeted moments. The first three moments are measured using the five-minute post-shock window because, in the model, both market-order submission (and hence CLOB trading) and OTC trading occur after the aggregate shock. In contrast, the limit-order-book asymmetry is measured using the five-minute pre-shock window because, in the model, limit orders are submitted before the shock. All moments are directly constructed from the merged TRACE and BrokerTec full order book data. Together, these moments identify the key dimensions of the model: the OTC market share disciplines endogenous venue choice, OTC price dispersion and skewness discipline OTC pricing relative to the CLOB, and limit-order-book asymmetry disciplines the equilibrium state of the order book.

The non-targeted moment used to evaluate the model fit is the limit-order execution probability. In the data, it is measured as the fraction of limit orders posted at the best available quote that are eventually executed. In the model, the corresponding moments are ϕ^{SL} and ϕ^{BL} for the benchmark model, and ϕ^{SL}_{-} and ϕ^{BL}_{-} for the model with aggregate risk, since the identified event corresponds to a negative shock (the positive-shock state serves as the counterfactual for this event). I choose this as the non-targeted moment because it implies both allocative efficiency and the adverse-selection risk faced by limit-order traders.

Although all parameters are jointly estimated using the full set of moments, each parameter has a primary source of identification. The search parameter λ is primarily identified by the OTC market share because, given the tick size c, λ determines the execution certainty of the OTC venue relative to the CLOB and therefore governs the speed–cost trade-off underlying equilibrium venue choice, so a greater λ gives a larger OTC market share. The standard deviation of the type distribution, σ, is primarily identified by OTC price dispersion, as greater heterogeneity in types generates greater dispersion in OTC transaction prices through the Nash-bargaining pricing-setting equation (Equation 2). The bargaining-power parameter θ is primarily identified by the skewness of the

[29] For the model without aggregate risk, P^* is proxy by the contemporaneous mid-quote, $\frac{A+B}{2}$. For the model with aggregate risk, OTC prices are normalized relative to the pre-shock (ex ante) mid-quote. The normalized prices are winsorized at the 2% level before computing the standard deviation to reduce the influence of outliers in observed OTC transaction prices.

OTC price distribution because it determines the relative weights placed on buyers' and sellers' valuations in the Nash-bargaining price. Finally, the asset supply k is primarily identified by the limit-order-book asymmetry, N^{SL}/N^{BL}, through the aggregation equations for equilibrium limit-order masses (Equation 13), because a larger asset supply increases the mass of agents who own the asset and can therefore provide more sell liquidity to the market. Figure A16 and Table A1 in the Appendix present a total of 16 (4×4) numerical comparative statics illustrating and explaining the parameter identification.

4.1.2 Estimation Results

The estimation results for the benchmark model without aggregate risk are reported in Table 3. The benchmark equilibrium achieves 87.06% allocative efficiency; equivalently, 12.94% of the first-best gains from trade remain unrealized, i.e., misallocation. Scaling the model-implied welfare measures per unit of trading volume by the $890.90 billion in observed trading volume in the estimation sample yields approximately $401.46 million in realized gains from trade and $59.66 million in misallocation. The benchmark model closely matches all targeted moments but overpredicts the non-targeted limit-order execution probability. One possible explanation is that, in the data, many limit orders are actively canceled or revised before execution, reducing the observed execution probability relative to the model.[30] Nevertheless, the model correctly predicts that the majority of limit orders (i.e., more than half) remain unfilled. Because the benchmark model overpredicts the execution probability of limit orders, the estimated allocative efficiency should be interpreted as an upper bound on the true allocative efficiency.

The estimation results for the model with aggregate risk are reported in Table 4. The model with aggregate risk matches both the targeted and non-targeted moments. In particular, it correctly predicts the asymmetric execution probabilities of buy and sell limit orders following a negative aggregate shock. After the negative shock, all buy limit orders resting at the best bid are executed, whereas none of the sell limit orders resting at the best ask are executed. This asymmetric execution pattern is consistent with adverse selection against resting limit orders following a common-value shock, supporting the model's sequential trading structure and Proposition 4.

Comparing the estimation results in Tables 3 and 4 provides further validation of the parameterization. First, agents are more heterogeneous in the risk environment than in the benchmark environment, with the estimated standard deviation $\hat{\sigma}$ approximately doubling. This widening of type dispersion can be rationalized by the microfoundation in Duffie, Gârleanu, and Pedersen (2007) as an increase in the scale of the risk being hedged. Second, the over-the-counter venue is estimated

[30] The model treats all limit orders as firm commitments to trade. In practice, however, some limit orders are actively revised or canceled and therefore may not represent firm trading intentions. For example, some limit-order quotes are revised a dozen times within a single second or canceled very shortly after submission.

Table 3: Benchmark Model without Risk: Parameter Estimates and Model Fit

Panel A. Targeted Moments	**Data**	**Model**
OTC market share	0.2112	0.2112
OTC price dispersion	0.0112	0.0112
OTC price skewness	0.0131	0.0131
Limit order book asymmetry	0.9763	0.9763
Panel B. Estimated Parameters	**Symbol**	**Value**
Search efficiency in OTC	λ	0.9339
Type standard deviation	σ	0.0322
Bargaining power in OTC	θ	0.5045
Asset supply	k	0.4978
Panel C. Non-targeted Moment	**Data**	**Model**
Sell limit order execution probability	11.64%	45.71%
Buy limit order execution probability	11.41%	46.69%
Panel D. Welfare Analysis		
Allocative efficiency		87.06%

Notes: Panel A compares the empirical moments with the model predicted moments. Panel B reports the estimated parameter values. Panel C reports the non-targeted moment as test of model fitness. Panel D reports the model-implied allocative efficiency at the estimated parameter values. The minimized criterion value $Q(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}) = 0$.

to be substantially more search-frictional in the risk environment, as reflected in the sharp decline in the estimated search efficiency parameter $\hat{\lambda}$. Third, in the benchmark environment, the estimated bargaining power parameter $\hat{\theta}$ is close to 0.5, consistent with approximately symmetric bargaining power between the two sides of the market. Following the negative shock, however, bargaining power becomes asymmetric. Although the smaller estimated value of $\hat{\theta}$ may appear counterintuitive following a negative shock, one possible interpretation is related to the adverse selection faced by buy limit orders submitted by initial non-owners. In the negative state, buy limit orders are adversely executed, converting initial non-owners into post-CLOB owners and making the subsequent OTC trading opportunity more valuable for rebalancing their asset holdings. Consequently, the smaller $\hat{\theta}$ implies a greater expected OTC surplus for post-CLOB owners, as reflected in Equation 5. Finally, the estimated asset supply parameter $\hat{k}$ varies little across the two environments, as expected, since secondary-market trading reallocates ownership but does not change the outstanding amount of a given Treasury CUSIP.

These two tables reveal that allocative efficiency differs substantially across aggregate states. Taken together, they suggest that allocative efficiency is high when the market is not subject to aggregate risk but deteriorates sharply during periods of aggregate risk. The two venues together achieve about 87% of the first-best allocation in the environment without aggregate risk (Table 3, Panel D), but only about 4% during the identified negative shock event (Table 4, Panel D). These

Table 4: General Model with Negative Shock: Parameter Estimates and Model Fit

Panel A. Targeted Moments	**Data**	**Model**
Post-shock OTC market share	0.0785	0.0785
Post-shock OTC price dispersion	0.0412	0.0412
Post-shock OTC price skewness	0.0438	0.0438
Pre-shock limit order book asymmetry	1.1574	1.1574
Panel B. Estimated Parameters	**Symbol**	**Value**
Search efficiency in OTC	λ	0.0020
Type standard deviation	σ	0.0604
Bargaining power in OTC	θ	0.4142
Asset supply	k	0.4850
Panel C. Non-targeted Moment	**Data**	**Model**
Sell limit order execution probability	0%	0%
Buy limit order execution probability	100%	100%
Panel D. Welfare Analysis		
Allocative efficiency		4.07%

Notes: Panel A compares the empirical moments with the model predicted moments. Panel B reports the estimated parameter values. Panel C reports the non-targeted moment as test of model fitness. Panel D reports the model-implied allocative efficiency at the estimated parameter values. The minimized criterion value $Q(\hat{\mathbf{\Gamma}}_{\text{NegativeShock}}) = 0$.

findings suggest that persistent aggregate risk can substantially impair market allocative efficiency, thereby creating scope for welfare-improving Federal Reserve intervention. Moreover, even in the benchmark environment, the market does not attain the first-best allocation because the tick size acts as a wedge that distorts the allocation, thereby creating scope for welfare-improving tick-size reduction. Section 5 analyzes both policies in detail.

Figure 12 plots the distribution of transaction prices relative to the efficient price P^* separately for each venue. The price deviation $P - P^*$ measures the markup (or markdown, if negative) relative to the efficient price. This figure illustrates that the design of trading mechanisms can generate and shape the distribution of markups, even in the absence of market power, as in this paper's competitive model. The figure also highlights a limitation of the model: OTC transaction prices exhibit bunching at focal price levels, such as the CLOB best bid, best ask, and midpoint, which the model cannot fully replicate.

4.2 Counterfactual: Alternative Market Structures

I calculate allocative efficiency under two alternative market structures, CLOB only and OTC only, at the estimated parameter values $\hat{\mathbf{\Gamma}}_{\text{benchmark}}$, and compare it with the current hybrid structure. The CLOB-only counterfactual is implemented by setting $\lambda = 0$, while the OTC-only counterfactual is

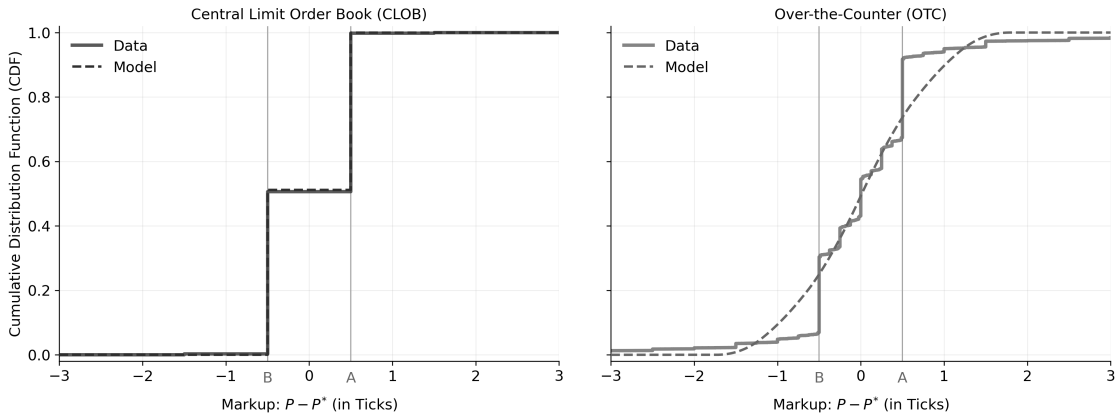


Figure 12: Distribution of Transaction Prices by Trading Venue

Notes: This figure plots the cumulative distribution functions (CDFs) of transaction prices relative to the efficient price P^*, expressed in ticks, separately for the Central Limit Order Book (CLOB, left panel) and Over-the-Counter (OTC, right panel) venues. Here, P^* denotes the mid-quote, while the vertical lines labeled B and A indicate the prevailing best bid and best ask on the CLOB, respectively. The solid curves represent the empirical distributions, while the dashed curves represent the model-implied distributions based on the parameter estimates reported in Table 3. The relative market share between the two venues is already captured as a targeted moment.

implemented by removing the CLOB trading stage from the model, such that the joint distribution satisfies $g^{OTC} = \mu$, holding λ fixed. The results presented in Table 5 show that the current hybrid structure yields higher allocative efficiency than either counterfactual market structure.

Speaking to the policy debate on how to improve the efficiency of opaque over-the-counter markets, the policy implications are twofold. First, proposal to fully centralize the market by eliminating over-the-counter trading is not optimal. This follows from comparing the allocative efficiency of the current hybrid structure with that of the CLOB-only counterfactual (85.54%). Second, introducing a central limit order book alongside a traditional over-the-counter market can be optimal. This follows from comparing the current hybrid structure with the OTC-only counterfactual (65.75%), which increases allocative efficiency by about 21 percentage points. The first policy implication cautions against radical proposals for market de-fragmentation that would consolidate trading into a single centralized venue, while the second supports market designs that have already been introduced in practice. For example, under the Dodd–Frank Act, Swap Execution Facilities (SEFs) allow swaps to be traded through central limit order books as well as through other trading protocols, such as request-for-quote.

Table 5: Allocative Efficiency Across U.S. Treasury Secondary Market Structures

Market Structure	Allocative Efficiency
Current hybrid structure: CLOB and OTC	87.06%
Counterfactual: CLOB only (full centralization)	85.54%
Counterfactual: OTC only (no centralization)	65.75%

4.3 Quasi-Experiment: Tick Size Reduction

I use the tick-size reduction as a quasi-experimental shock to test the model. On November 19, 2018, BrokerTec reduced the tick size for the 2-year on-the-run U.S. Treasury by half. According to CME Group, the purpose of this change was to *"allow more precise hedging and price discovery."*[31] Their argument that a smaller tick size improves price discovery is captured in the model by introducing the tick size as a wedge around the efficient price P^* (Equation 1). Therefore, a smaller tick size brings transaction prices closer to the efficient price and, in this sense, improves price discovery.

I borrow the summary statistics from Table 2 of M. Fleming, Nguyen, and Ruela (2024), which studies this shock using a reduced-form approach. Both their Table 2 and CME Group's report[32] show that the average daily volume on the CLOB increased by approximately 30%, from $20 billion before the tick-size reduction to $26 billion afterward.

I therefore use the model to compute CLOB volume under the original tick size, c, and compare it with the CLOB volume under the reduced tick size, $c' = \frac{1}{2}c$, holding the estimated parameter values fixed at $\hat{\mathbf{\Gamma}}_{\text{Benchmark}}$. The percentage change in CLOB volume is calculated as

$$\text{Change in CLOB volume} = \frac{\text{CLOB Volume}(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}, \frac{1}{2}c) - \text{CLOB Volume}(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}, c)}{\text{CLOB Volume}(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}, c)}$$

The model predicts that reducing the tick size by half increases CLOB volume by 14.16%, which has the same sign as the increase observed in the data, although the model-implied effect is smaller.

5 Normative Policy Analysis

Motivated by the importance of the U.S. Treasury secondary market in global macro-finance, this section uses the model to study two normative questions. First, in Section 5.1, I study the optimal tick size along the yield curve. Recall that, in the model, the parameter α can be interpreted as the volatility (risk) of the asset. Since the volatility of Treasury securities differs across maturities, the optimal tick size should also vary across maturities. In other words, the optimal tick size should itself exhibit a term structure, and I use the estimated model to characterize this optimal term structure. Second, in Section 5.2, I study the optimal size of Federal Reserve asset purchases and their welfare implications. The Treasury secondary market has experienced several episodes of dysfunction over the past decade, during which the Fed intervened by purchasing Treasury securi-

[31] CME Group, "Smaller Ticks, Lower Costs: BrokerTec Reduces 3-Year MPI," available at https://www.cmegroup.com/education/articles-and-reports/smaller-ticks-lower-costs-brokertec-reduces-3-year-mpi.

[32] CME Group, "Smaller Ticks, Lower Costs: BrokerTec Reduces 3-Year MPI," available at https://www.cmegroup.com/education/articles-and-reports/smaller-ticks-lower-costs-brokertec-reduces-3-year-mpi.

ties to support market functioning.[33] I use the model to calculate the amount of Treasury securities that the Fed would need to purchase following a negative aggregate shock to correct the resulting asset misallocation and assess the associated welfare improvement.

These two normative analyses are grounded in the theoretical and quantitative findings of the previous sections and apply to different market environments: normal times and periods of market stress. As Figure 11 and Table 3 show, when aggregate risk is sufficiently small, the current hybrid market structure, in which the two venues coexist, functions well and achieves a high fraction of the first-best allocation. Radical interventions, such as fully centralizing the market, are therefore unnecessary. Instead, the policy focus should be on setting the tick size optimally so that the market itself can achieve high allocative efficiency, in particular because allocative efficiency is mainly achieved by the CLOB and tick size is its design parameter. In contrast, when aggregate risk is sufficiently large, even the coexistence of the two venues cannot generate high allocative efficiency, as shown in Figure 9, Figure 11 and Table 4. In such periods of market stress, improving market design through the choice of the tick size alone may therefore be insufficient, creating a role for direct intervention. Ex post Fed asset purchases can correct post-trade asset misallocation by absorbing assets from inefficient holders.

That said, the calculations in this section are back-of-the-envelope and therefore serve primarily to shed light on the normative policy implications of the model rather than to provide precise quantitative prescriptions. In this sense, these exercises should be interpreted as more qualitative than quantitative. For example, the optimal tick size for the current U.S. Treasury market may differ from that implied by the sample period used in this paper because the level of aggregate risk varies over time. Similarly, a precise quantitative assessment of the appropriate amount of Federal Reserve purchases during a stress episode would ideally use data from that particular period. Therefore, this section primarily provides a welfare rationale for which policy tools are appropriate under different market environments, clarifying several policy debates and the regulatory agenda.

5.1 Term Structure of Tick Sizes

The optimal tick size c^* for each Treasury maturity b is calculated using the benchmark parameter estimates for that maturity, $\hat{\mathbf{\Gamma}}_{b,\text{Benchmark}}$, as follows. First, for each maturity b, I identify common-value change episodes in the data following the procedure outlined in Appendix A and estimate the size of each common-value change. I then compute the empirical frequency of each change size. I denote the resulting distribution for maturity b by $\hat{H}_b(\alpha)$. Next, for each common-value change size α, I conduct a grid search to find the corresponding optimal tick size, $c^*(\alpha; \hat{\mathbf{\Gamma}}_{b,\text{Benchmark}}) =$

[33] For background, see Lorie Logan's speech, available at https://www.newyorkfed.org/newsevents/speeches/2020/log200715. Logan, who served as Manager of the System Open Market Account (SOMA) during the COVID-19 crisis, emphasized the importance of market functioning and the role of Federal Reserve purchases in supporting it.

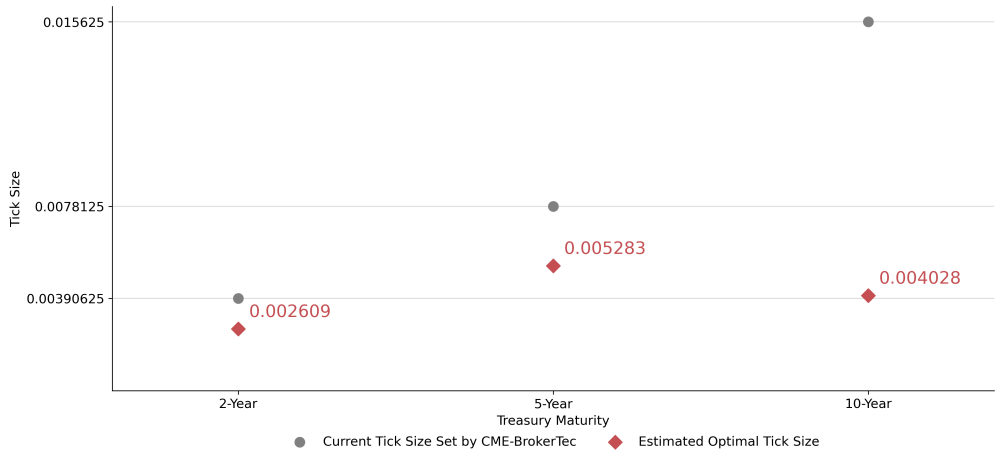


Figure 13: Estimated Optimal Tick Size Across U.S. Treasury Maturities

$\arg\max_c$ Allocative Efficiency$(c, \alpha; \hat{\mathbf{\Gamma}}_{b,\text{Benchmark}})$. Finally, I calculate the average optimal tick size over this distribution as

$$c_b^* = \int_0^{\bar{\alpha}_b} c^*(\alpha; \hat{\mathbf{\Gamma}}_{b,\text{Benchmark}})\, d\hat{H}_b(\alpha),$$

where $\bar{\alpha}_b$ is the largest observed shock size for each maturity b.

Figure 13 presents the estimated optimal term structure of tick sizes for U.S. Treasuries. The optimal tick sizes are smaller than the prevailing tick sizes in the sample period, with the same order of magnitude.[34] However, as emphasized previously, aggregate risk varies over time, whereas tick sizes remain fixed for relatively long periods. This mismatch can lead to misspecified tick sizes and generate inefficiencies, which are partially mitigated by the presence of the over-the-counter.

Implementing the optimal tick size is beyond the scope of this paper, but several approaches are possible in practice. One approach is to mandate the tick size directly through regulation. Alternatively, a market-based approach requires identifying why the observed tick size deviates from the optimum. If a trading platform chooses a tick size that is not optimal because its private objective of maximizing trading volume, and hence commission profits, differs from the social objective of maximizing allocative efficiency, policymakers could use taxes to better align these objectives. By contrast, if reductions in the tick size are constrained by the technological capacity of the trading platform's IT infrastructure, subsidies for technological investment could be helpful.

[34]This figure uses data from Sep 2022. BrokerTec subsequently reduced the tick sizes for the 5-year and 10-year U.S. Treasuries by half on June 28, 2026. The current tick sizes are 0.00390625 for the 5-year Treasury and 0.0078125 for the 10-year Treasury.

5.2 Fed Purchases Program

I measure the optimal Fed purchase quantity by the amount of the asset that remains in the hands of inefficient holders after market trading concludes.[35] The welfare rationale is grounded in the theoretical finding that allocative efficiency is low when aggregate risk is high, thus creating scope for Fed intervention to correct the asset misallocation that market trading fails to resolve, leaving an *ex post misallocation*.[36]

Define the ex post misallocation share in the negative state, $\mathcal{M}_-$, as

$$\mathcal{M}_- \equiv \frac{1}{k}\int_{\underline{\delta}}^{F^{-1}(1-k)} \tilde{\mu}_-(1,\delta)\, d\delta,$$

$$\mathcal{L}_- \equiv \frac{1}{k}\int_{\underline{\delta}}^{F^{-1}(1-k)} \left[F^{-1}(1-k)-\delta\right] \tilde{\mu}_-(1,\delta)\, d\delta,$$

$\mathcal{M}_-$ is the ex post share of the asset held by agents to whom the first-best allocation would not assign it, and $\mathcal{L}_-$ is the corresponding value-weighted welfare loss. Scaling by the asset outstanding amount K, the optimal Fed purchase quantity and the resulting welfare improvement are

$$Q_- = \mathcal{M}_- K, \qquad \Delta W_- = \frac{\mathcal{L}_- K}{100},$$

where the factor 100 accounts for values being quoted per \$100 of par, so that ΔW_- is in dollars.

Figure 14 plots $\mathcal{M}_-$ against the size of the shock using the parameter estimates $\hat{\mathbf{\Gamma}}_{\text{NegativeShock}}$. The ex post misallocation is non-decreasing in aggregate risk and is upper bounded by the ex post misallocation when aggregate risk is sufficiently large that the CLOB becomes inactive, i.e., a market failure in which no trade occurs through the CLOB, and the level of this upper bound is then determined by the search efficiency of the OTC venue.

Scaling by the actual outstanding amount of U.S. Treasuries during the COVID-19 period, Table 6 presents the estimated optimal Fed purchase quantity Q_- and the corresponding welfare improvement ΔW_- during the COVID-19 crisis. To calculate these quantities, I set $K = 36$ billion, the outstanding par amount of the on-the-run 7-year U.S. Treasury during this period. I use this 7-year

[35] This interpretation is conditional on the Fed holding the asset at a lower social cost than the misallocated private agents. In the language of the model, this corresponds to assigning the Fed an ex ante type $\delta^{\text{Fed}} \geq F^{-1}(1-k)$, which can be microfounded by a relatively low holding cost. This seems plausible during the COVID-19 period, when many private agents faced balance-sheet constraints and therefore had sufficiently high holding costs. Under this assumption, Fed purchases transfer the asset from inefficient private holders to the Fed, allowing those private agents to successfully execute the sale of their assets.

[36] This is because fewer limit orders are submitted, limiting the amount of trade that can be executed through the CLOB, and because the quantitative results suggest that search frictions in the OTC venue may also worsen during periods of market stress, further limiting the amount of trade that can be realized through the OTC venue. Together, the market fails and leaves a large ex post misallocation.

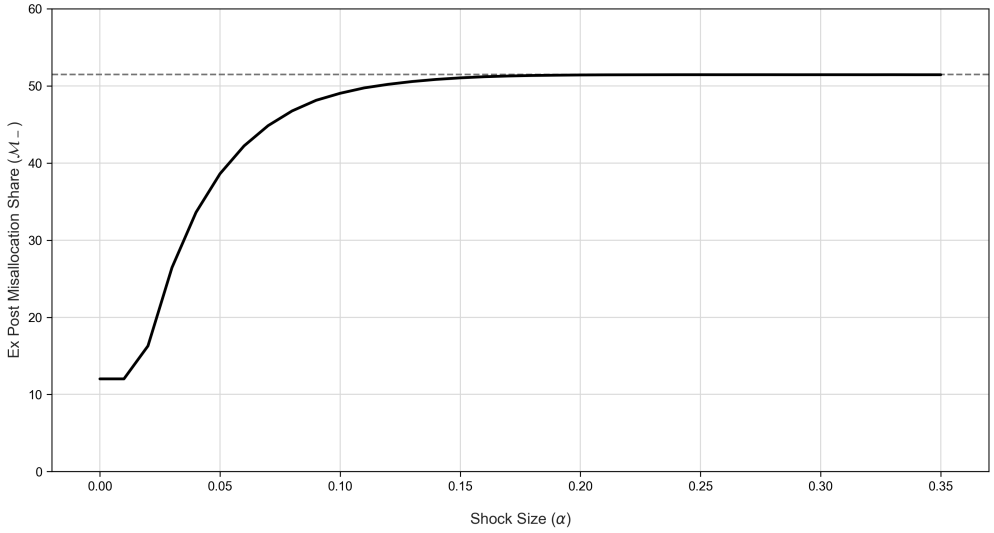


Figure 14: Model-Implied Share of Misallocated Assets, Post-Trade

Treasury because it is the only Treasury security that was on-the-run during the most severe crisis period (March 9–March 23) and was subsequently purchased by the Fed during the main intervention period (March 24–April 30), thereby providing an observed setting in which to evaluate the model's prediction. The choice of these periods follows He, Nagel, and Song (2022) and Duffie, M. J. Fleming, et al. (2023). I estimate the shock size α during the crisis period by calculating the mean of the daily standard deviations of transaction prices.[37]

Table 6: Estimated Optimal Fed Purchase Quantity during the COVID-19 Crisis for One Bond

Estimated Shock $\hat{\alpha}$	Misallocated Share $\hat{\mathcal{M}}_-$	Fed Purchase Quantity $\hat{Q}_-$	Welfare Gain $\Delta\hat{W}_-$
0.33	51.45%	18.52$bn par	9.08 $mn

The results in Table 6 imply an optimal Fed purchase of approximately 18.52 billion of this single Treasury security, generating a welfare improvement of approximately 9.08 million. In the data,[38] the Fed purchased 15.83 billion in par value of this bond during the main intervention period. The model slightly overestimates the actual purchase quantity. This is because, given the estimated shock size of 0.33, the model predicts an inactive CLOB and very small volume of OTC trading at the estimated parameter values $\hat{\mathbf{\Gamma}}_{\text{NegativeShock}}$, whereas the CLOB remains active in the data.[39] Consequently, the model generates a larger amount of ex post asset misallocation (almost as in autarky) and a larger model-implied Fed purchase quantity (which can be viewed as an upper

[37] Recall that, in the model, the variance of the efficient price is α^2, and hence its standard deviation is α.

[38] Data source: Federal Reserve Bank of New York, *Outright Purchases and Sales of Treasury Securities*, https://www.newyorkfed.org/markets/OMO_transaction_data.html. The data suggest that almost all Fed purchases were off-the-run Treasury securities, for which BrokerTec does not provide a CLOB trading platform.

[39] Although CLOB is less active than in the pre-crisis period, as the model predicts.

bound). As discussed above, the primary purpose is to provide a welfare rationale for Fed intervention, i.e., due to the large-scale ex post asset misallocation when the market fails and freezes, rather than a precise quantitative prescription for the size of such an intervention.

6 Conclusion

This paper studies a market-design issue in the U.S. Treasury secondary market, where decentralized over-the-counter trading coexists with the central limit order book and fragments the market. I develop a venue-choice model and use granular Treasury trading data to quantify the allocative efficiency of this fragmented market structure. The main result is that the fragmented structure is optimal and that the two trading mechanisms play complementary roles across different aggregate market conditions. When aggregate risk is low, the centralized venue dominates trading and achieves high allocative efficiency through direct matching, avoiding the inefficiency associated with the search frictions of over-the-counter trading. When aggregate risk is high, trading shifts toward over-the-counter, which sustains market functioning because flexible negotiated pricing is not exposed to the adverse selection generated when rigid quotes become stale and are picked off following changes in the asset's common value. Therefore, the hybrid structure enables gains from trade across different aggregate environments and thus improves market resilience, a concern that has received substantial attention in both the academic literature and policy discussions.

More broadly, the paper shows that fragmentation can be an efficient market structure. This may help rationalize why the coexistence of a central limit order book and decentralized over-the-counter trading is common across countries, including government bond markets in the U.S., Europe, and China. The results should, however, be interpreted in light of several limitations. First, the trading mechanisms considered here may represent only a subset of those used in practice, and the implications for allocative efficiency may change when additional mechanisms coexist. The estimated result that over-the-counter transaction prices exhibit bunching at focal price levels might be empirical evidence that other pricing mechanisms not accounted for here are at work. Second, even within the two venues studied here, several potentially important features are abstracted from, including time priority in limit order book matching and the role of order size, as evident from the regression results. Incorporating these features may require a richer framework, such as a dynamic model with divisible assets. Centralization versus decentralization is a long-standing question in the organization of industries, and the increasing availability of detailed data may allow future research to study this question in greater depth.

References

Adrian, Tobias, Michael Fleming, and Kleopatra Nikolaou (2025). "US Treasury Market Functioning from the Global Financial Crisis to the Pandemic". In: *Annual Review of Financial Economics* 17.

Allen, Jason and Milena Wittwer (2023). "Centralizing over-the-counter markets?" In: *Journal of Political Economy* 131.12, pp. 3310–3351.

Babus, Ana and Cecilia Parlatore (2022). "Strategic fragmented markets". In: *Journal of Financial Economics* 145.3, pp. 876–908.

Biais, Bruno, David Martimort, and Jean-Charles Rochet (2000). "Competing mechanisms in a common value environment". In: *Econometrica* 68.4, pp. 799–837.

Brancaccio, Giulia, Myrto Kalouptsidi, Theodore Papageorgiou, and Nicola Rosaia (2023). "Search frictions and efficiency in decentralized transport markets". In: *The Quarterly Journal of Economics* 138.4, pp. 2451–2503.

Brancaccio, Giulia and Karam Kang (2025). "Search frictions and product design in the municipal bond market". In: *Econometrica* 93.6, pp. 2159–2199.

Brancaccio, Giulia, Dan Li, and Norman Schürhoff (2017). "Learning by trading: The case of the US market for municipal bonds". In: *Unpublished paper. Princeton University*.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan (2014). "High-frequency trading and price discovery". In: *The Review of Financial Studies* 27.8, pp. 2267–2306.

Buchholz, Nicholas (2022). "Spatial equilibrium, search frictions, and dynamic efficiency in the taxi industry". In: *The Review of Economic Studies* 89.2, pp. 556–591.

Budish, Eric, Peter Cramton, and John Shim (2015). "The high-frequency trading arms race: Frequent batch auctions as a market design response". In: *The Quarterly Journal of Economics* 130.4, pp. 1547–1621.

Chen, Daniel and Darrell Duffie (2021). "Market fragmentation". In: *American Economic Review* 111.7, pp. 2247–2274.

Colliard, Jean-Edouard, Thierry Foucault, and Peter Hoffmann (2021). "Inventory management, dealers' connections, and prices in over-the-counter markets". In: *The Journal of Finance* 76.5, pp. 2199–2247.

Correa, Ricardo, Wenxin Du, and Gordon Y Liao (2025). "US banks and global liquidity". In: *Journal of Political Economy Macroeconomics* 3.4, pp. 574–620.

Demsetz, Harold (1968). "The cost of transacting". In: *The quarterly journal of economics* 82.1, pp. 33–53.

Du, Wenxin, Alexander Tepper, and Adrien Verdelhan (2018). "Deviations from covered interest rate parity". In: *The Journal of Finance* 73.3, pp. 915–957.

Duffie, Darrell (2022). *Fragmenting markets: post-crisis bank regulations and financial market liquidity*. Walter de Gruyter GmbH & Co KG.

— (2023). "Resilience redux in the US Treasury market". In: *Jackson Hole Symposium, Federal Reserve Bank of Kansas City*.

— (2025). "How US Treasuries can remain the World' s safe Haven". In: *Journal of Economic Perspectives* 39.2, pp. 195–214.

Duffie, Darrell, Michael J Fleming, Frank M Keane, Claire Nelson, Or Shachar, and Peter Van Tassel (2023). "Dealer capacity and US Treasury market functionality". In: *FRB of New York Staff Report* 1070.

Duffie, Darrell, Nicolae Gârleanu, and Lasse Heje Pedersen (2005). "Over-the-counter markets". In: *Econometrica* 73.6, pp. 1815–1847.

— (2007). "Valuation in over-the-counter markets". In: *The Review of Financial Studies* 20.6, pp. 1865–1900.

Dugast, Jérôme, Thomas Marta, and Fabrice Riva (2026). "Market Depth and Execution Delays". In: *Available at SSRN*.

Dugast, Jérôme, Semih Üslü, and Pierre-Olivier Weill (2022). "A theory of participation in OTC and centralized markets". In: *The Review of Economic Studies* 89.6, pp. 3223–3266.

Fleming, Michael, Giang Nguyen, and Francisco Ruela (2024). "Tick size, competition for liquidity provision, and price discovery: Evidence from the US Treasury market". In: *Management Science* 70.1, pp. 332–354.

Fleming, Michael J and Eli M Remolona (1999). "Price formation and liquidity in the US Treasury market: The response to public information". In: *The journal of Finance* 54.5, pp. 1901–1915.

Foucault, Thierry (1999). "Order flow composition and trading costs in a dynamic limit order market". In: *Journal of Financial markets* 2.2, pp. 99–134.

Foucault, Thierry, Marco Pagano, and Ailsa Röell (2023). *Market liquidity: theory, evidence, and policy*. Oxford University Press.

Gavazza, Alessandro (2016). "An empirical equilibrium model of a decentralized asset market". In: *Econometrica* 84.5, pp. 1755–1798.

Glode, Vincent and Christian C Opp (2020). "Over-the-counter versus limit-order markets: The role of traders' expertise". In: *The Review of Financial Studies* 33.2, pp. 866–915.

Glosten, Lawrence R (1994). "Is the electronic open limit order book inevitable?" In: *The Journal of Finance* 49.4, pp. 1127–1161.

Glosten, Lawrence R and Paul R Milgrom (1985). "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders". In: *Journal of financial economics* 14.1, pp. 71–100.

Hasbrouck, Joel (1991). “Measuring the information content of stock trades”. In: *The Journal of Finance* 46.1, pp. 179–207.

He, Zhiguo, Stefan Nagel, and Zhaogang Song (2022). “Treasury inconvenience yields during the COVID-19 crisis”. In: *Journal of Financial Economics* 143.1, pp. 57–79.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill (2025). *The Economics of Over-the-Counter Markets*. Princeton University Press.

Kwan, Amy, Ronald Masulis, and Thomas H McInish (2015). “Trading rules, competition for order flow and market fragmentation”. In: *Journal of Financial Economics* 115.2, pp. 330–348.

Lee, Tomy and Chaojun Wang (2025). “Regulating Over-the-Counter Markets”. In: *The Journal of Finance* 80.4, pp. 1929–1962.

Liu, Ying, Sebastian Vogel, and Yuan Zhang (2018). “Electronic trading in OTC markets vs. centralized exchange”. In: *Paris December 2018 Finance Meeting EUROFIDAI-AFFI*.

Milgrom, Paul R and Robert J Weber (1982). “A theory of auctions and competitive bidding”. In: *Econometrica: Journal of the Econometric Society*, pp. 1089–1122.

Nakamura, Emi and Jón Steinsson (2018). “High-frequency identification of monetary non-neutrality: the information effect”. In: *The Quarterly Journal of Economics* 133.3, pp. 1283–1330.

Rapp, Andreas C and Martin Waibel (2023). “Managing Regulatory Pressure: Bank Regulation and Corporate Bond Intermediation”. In: *Available at SSRN*.

Weill, Pierre-Olivier (2020). “The search theory of over-the-counter markets”. In: *Annual Review of Economics* 12, pp. 747–773.

Werner, Ingrid M, Barbara Rindi, Sabrina Buti, and Yuanji Wen (2023). “Tick size, trading strategies, and market quality”. In: *Management Science* 69.7, pp. 3818–3837.

Wittwer, Milena and Jason Allen (2026). “Market power and capital constraints”. In: *American Economic Review* 116.4, pp. 1309–1339.

Yoon, Ji Hee (2026). “Endogenous market structure: Over-the-counter versus exchange trading”. In: *Available at SSRN 7483978*.

Internet Appendix

A Appendix: Data Facts

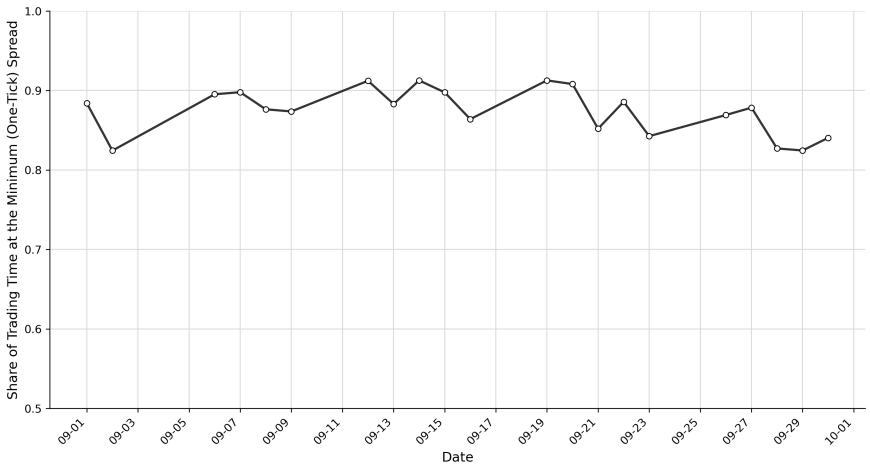


Figure A1: Time-Weighted Share of Trading Time with a One-Tick Quoted Bid–Ask Spread

Notes: This figure uses data for the on-the-run 10-year U.S. Treasury.

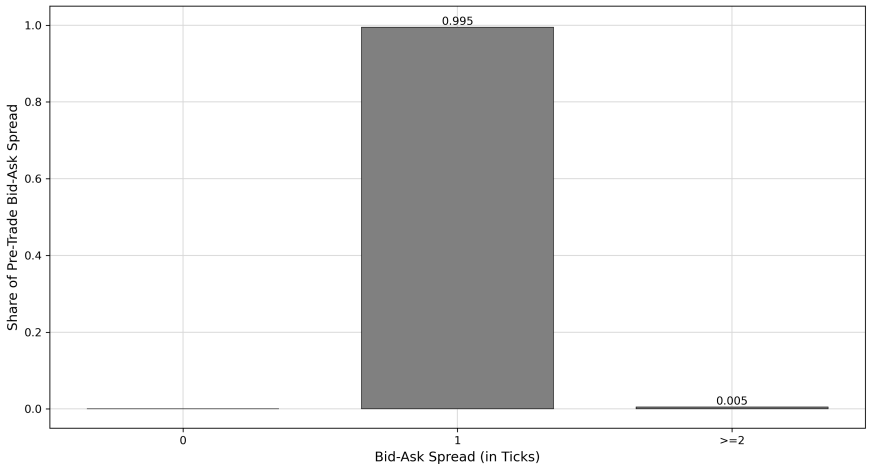


Figure A2: Distribution of the Bid–Ask Spread Immediately Before CLOB Trade Execution

Notes: This figure uses data for the on-the-run 10-year U.S. Treasury.

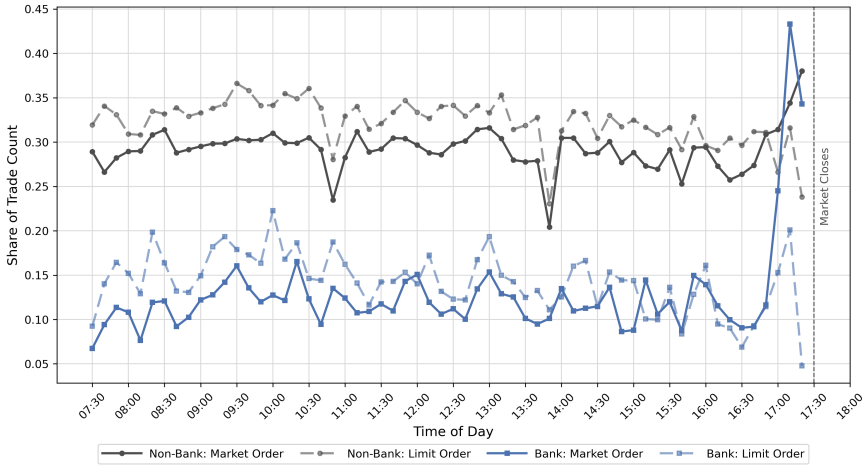


Figure A3: Intraday Venue Shares: Banks versus Non-Banks

Notes: This figure plots the intraday venue shares for bank dealers and non-bank dealers. For each dealer type, the denominator is the total number of trades by that dealer type within each 10-minute interval, and the figure reports the fractions of these trades executed through market orders and limit orders, respectively. As shown in the figure, the market-order share of bank dealers increases sharply during the final 30 minutes before market close (blue solid line), whereas the corresponding increase for non-bank dealers is much smaller. The data construction used in this figure and Table 1 differs from that used in the rest of the empirical analysis because trader-type information is available only in TRACE. I therefore use ATS transactions reported in TRACE as a proxy for CLOB trading and match transaction prices to the best bid and ask observed in BrokerTec at the same second. A transaction through ATS is classified as an aggressive (market) order if it is a sale at or below the best bid or a purchase at or above the best ask, and as a passive (limit) order if it is a sale at or above the best ask or a purchase at or below the best bid. Although ATS trading does not correspond exactly to CLOB trading, the proxy is suitable for the purpose of the analysis here, which is to examine whether traders are willing to pay a price premium (crossing the bid–ask spread) for more certain execution.

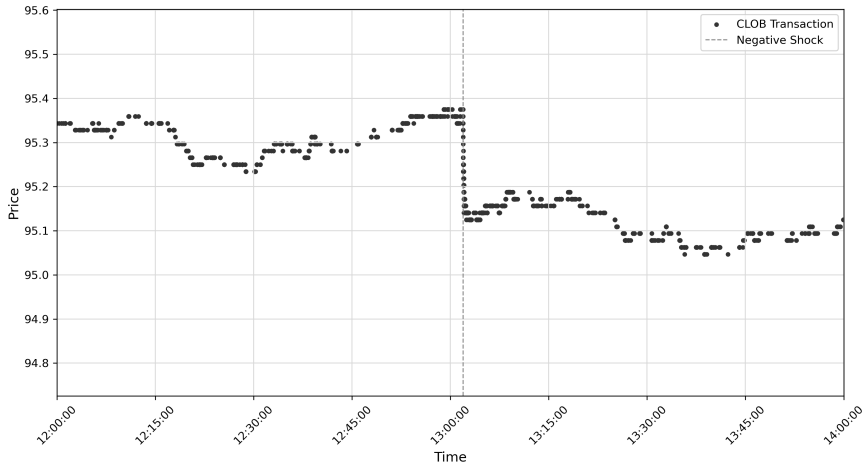


Figure A4: An Example Episode: Time Series of Transaction Prices

Notes: This figure uses CLOB transaction data for the on-the-run 10-year U.S. Treasury on September 12, 2022.

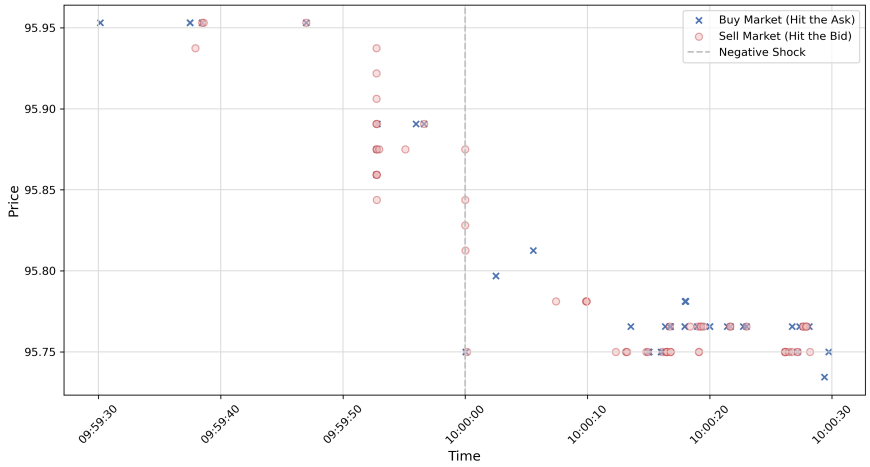


Figure A5: Market Order Flow Following the ISM PMI Release at 10:00, Sep-01-2022

Notes: This figure uses CLOB transaction data for the on-the-run 10-year U.S. Treasury.

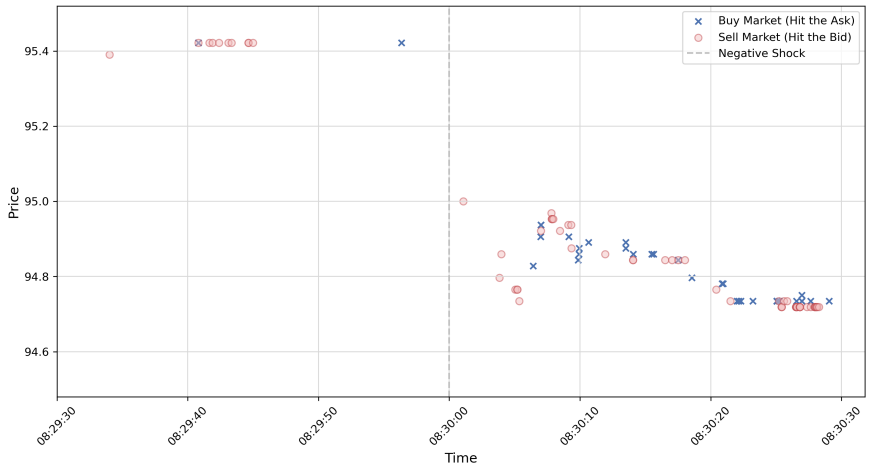


Figure A6: Market Order Flow Following the Official CPI Release at 08:30, Sep-13-2022

Notes: This figure uses CLOB transaction data for the on-the-run 10-year U.S. Treasury.

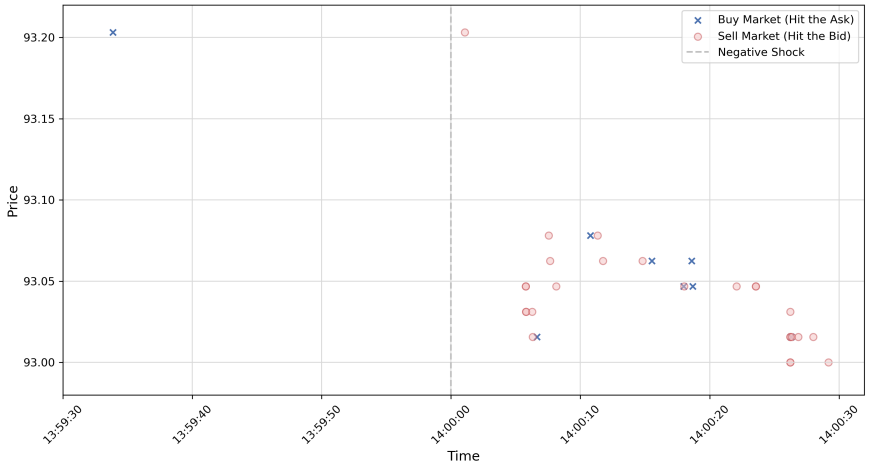


Figure A7: Market Order Flow Following the FOMC Announcement at 14:00, Sep-21-2022

Notes: This figure uses CLOB transaction data for the on-the-run 10-year U.S. Treasury.

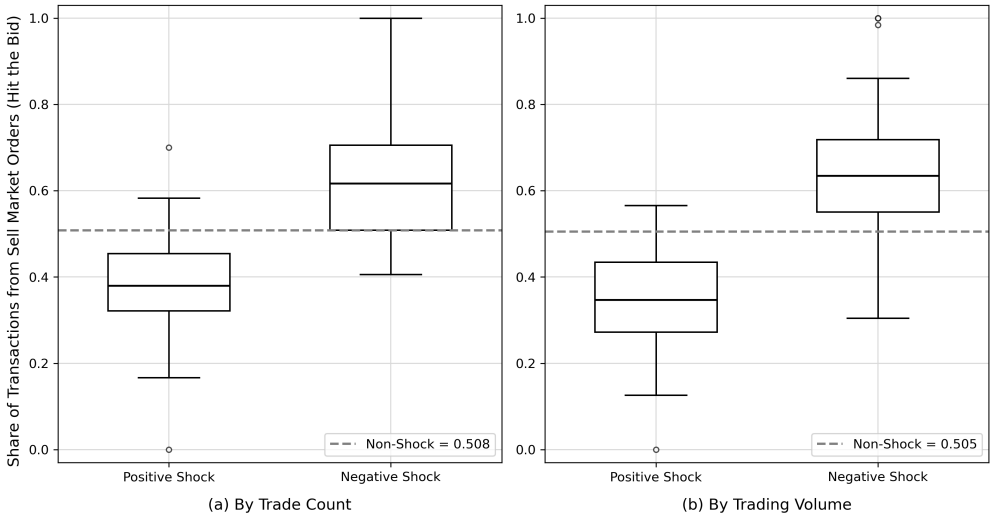


Figure A8: Sell-Market-Initiated Transaction Share on the CLOB Across Event Types

Notes: This figure reports the share of trading initiated by sell market orders across positive-shock, negative-shock, and non-shock periods, using all identified common-value shocks in the sample (with the remaining share initiated by buy market orders). During non-shock periods, market order flow is approximately balanced between buys and sells (dashed horizontal lines, around 0.5). Following negative shocks (right box in both panels), the sell-market-order share rises above one-half, indicating that sell market orders dominate and execute against resting buy limit orders. Following positive shocks (left box in both panels), the pattern reverses: the sell-market-order share falls below one-half, indicating that buy market orders dominate and execute against resting sell limit orders. Thus, common-value shocks systematically generate excess market order flow in the direction of the price change, exposing resting limit orders on the opposite side to adverse selection.

Shock Identification I identify 99 common-value change episodes for the on-the-run 10-year U.S. Treasury over the 21 trading days in September 2022. I first identify monotonic price runs, each beginning at a local price extremum where the price starts moving in one direction and ending at the next local extremum before the direction reverses. I then classify a run as a common-value change episode if it satisfies three criteria: (1) the cumulative price movement exceeds 10 basis points; (2) the movement is not immediately and substantially reversed, defined as an opposite movement beginning within 60 seconds and retracing at least 50% of the original movement; and (3) the episode lasts less than 126 seconds, the average waiting time before execution for limit orders on the on-the-run 10-year U.S. Treasury, ensuring that the identified change occur within the relevant execution horizon and thus capture the payoff-relevant adverse-selection risk faced by limit-order traders. For each common-value change episode, the shock size is measured as the difference between the maximum and minimum prices observed during the episode.

The first criterion excludes ordinary price fluctuations arising from limit order book mechanics, such as bid–ask bouncing, which typically involves movements of only one or two ticks. The second distinguishes common-value changes from transitory price movements driven by immediate liquidity needs, such as large market orders walking the book. The third ensures that the identified episodes occur within the average execution horizon of resting limit orders and are therefore payoff-relevant for limit-order traders. Together, these criteria identify large and persistent price movements over short intervals, which this paper interprets as common-value shocks. The identified episodes include major macroeconomic announcements, such as the PMI release on Sep 1, the CPI release on Sep 13, and the FOMC decision on Sep 21.

I apply the same identification procedure to U.S. Treasuries of other maturities, adjusting only the third criterion to use the corresponding average limit-order waiting time for each maturity.

B Appendix: Derivations and Proofs

Derivation of the Continuation Value in Equation 5. I first explain the OTC matching. The OTC matching proceeds in three steps. First, agents are randomly selected from each side of the OTC venue into a matching pool. Second, the selected owners and non-owners are randomly paired one-to-one. Third, within each pair, trade occurs if and only if there are positive gains from trade.

Let

$$M_s^{OTC} = \lambda \min\left\{N_s^{OTC-O}, N_s^{OTC-NO}\right\}$$

denote the total mass of pairs formed in the OTC venue, where $\lambda \in [0,1]$ is the search efficiency (matching-rate) parameter. That is, the OTC venue forms a fraction λ of the maximum feasible number of pairs, which is determined by the short side of the market.

Given M_s^{OTC}, a mass M_s^{OTC} of agents is randomly selected from each side of the OTC venue into the matching pool. Therefore, the probability that a given owner is selected into the matching pool is

$$\frac{M_s^{OTC}}{N_s^{OTC-O}} = \lambda \min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\},$$

whereas the probability that a given non-owner is selected into the matching pool is

$$\frac{M_s^{OTC}}{N_s^{OTC-NO}} = \lambda \min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\}.$$

The selected owners and non-owners are then randomly paired one-to-one. Conditional on being selected into the matching pool, an agent is therefore randomly paired with an agent from the opposite side of the market, whose valuation is distributed according to the normalized valuation distribution on that side. Importantly, being paired does not necessarily imply trade: after a pair is formed, trade occurs only if there are positive gains from trade.

Consider an owner ($q = 1$) of type δ. Her valuation for holding the asset in state s is

$$v_o = \delta + \epsilon_s.$$

Conditional on being selected into the matching pool, let ν denote the valuation of the non-owner with whom she is randomly paired, where

$$\nu \sim g_s^{OTC-NO}.$$

For a given pair, trade occurs if and only if there are positive gains from trade,

$$\nu > v_o = \delta + \epsilon_s.$$

Under Nash bargaining, conditional on trade occurring, the OTC transaction price is

$$p^{OTC} = \theta v_o + (1-\theta)v.$$

Therefore, the owner's surplus from selling the asset is

$$\begin{aligned} p^{OTC} - v_o &= \theta v_o + (1-\theta)v - v_o \\ &= (1-\theta)(v - v_o) \\ &= (1-\theta)\big(v - (\delta + \epsilon_s)\big). \end{aligned}$$

Since the owner is selected into the matching pool with probability

$$\lambda \min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\},$$

integrating over the possible non-owner counterparties with whom she may be randomly paired, weighted by their density $g_s^{OTC-NO}(v)$, and restricting to pairs for which trade occurs, $v > \delta + \epsilon_s$, gives the owner's expected OTC surplus:

$$\lambda \min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\} \int_{v>\delta+\epsilon_s} (1-\theta)\big(v - (\delta+\epsilon_s)\big) g_s^{OTC-NO}(v)\, dv.$$

Since the OTC term captures only the surplus from selling relative to holding the asset, i.e., the gains from trade, the owner's payoff $\delta + \epsilon_s$ from holding the asset must also be accounted for. Thus, the total continuation value is

$$u(1,\delta,s) = \underbrace{\delta + \epsilon_s}_{\text{flow payoff}} + \underbrace{\lambda(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\} \int_{v>\delta+\epsilon_s} \big(v - (\delta+\epsilon_s)\big) g_s^{OTC-NO}(v)\, dv}_{\text{expected surplus from OTC}}.$$

The continuation value on the non-owner side, $u(0,\delta,s)$, is constructed symmetrically.

□

Proof B.1 (**Proof of Lemma 1**). Fix a state s. Since an individual agent is atomistic, the aggregate equilibrium objects, quantities N_s^{OTC-O}, N_s^{OTC-NO} and the distributions g_s^{OTC-O} and g_s^{OTC-NO}, are taken as given when differentiating with respect to the individual type δ.

First, consider

$$u(1,\delta,s) = \delta + \epsilon_s + \lambda(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\} \int_{\delta+\epsilon_s}^{\infty} (\nu - (\delta+\epsilon_s))\, g_s^{OTC-NO}(\nu)\, d\nu.$$

By the Leibniz integral rule,

$$\begin{aligned}\frac{\partial}{\partial\delta}\int_{\delta+\epsilon_s}^{\infty} (\nu - (\delta+\epsilon_s))\, g_s^{OTC-NO}(\nu)\, d\nu &= -\left[(\nu - (\delta+\epsilon_s))\, g_s^{OTC-NO}(\nu)\right]_{\nu=\delta+\epsilon_s} \\ &\quad + \int_{\delta+\epsilon_s}^{\infty} \frac{\partial}{\partial\delta}\left[(\nu - (\delta+\epsilon_s))\, g_s^{OTC-NO}(\nu)\right] d\nu \\ &= -\int_{\delta+\epsilon_s}^{\infty} g_s^{OTC-NO}(\nu)\, d\nu,\end{aligned}$$

where the boundary term is zero because $\nu - (\delta + \epsilon_s) = 0$ at $\nu = \delta + \epsilon_s$. Therefore,

$$\frac{\partial u(1,\delta,s)}{\partial\delta} = 1 - \lambda(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\} \int_{\delta+\epsilon_s}^{\infty} g_s^{OTC-NO}(\nu)\, d\nu. \tag{A1}$$

Second, consider

$$u(0,\delta,s) = \lambda\theta\min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} \int_{-\infty}^{\delta+\epsilon_s} (\delta + \epsilon_s - \nu)\, g_s^{OTC-O}(\nu)\, d\nu.$$

Applying the Leibniz integral rule again gives

$$\begin{aligned}\frac{\partial}{\partial\delta}\int_{-\infty}^{\delta+\epsilon_s} (\delta+\epsilon_s-\nu)\, g_s^{OTC-O}(\nu)\, d\nu &= \left[(\delta+\epsilon_s-\nu)\, g_s^{OTC-O}(\nu)\right]_{\nu=\delta+\epsilon_s} \\ &\quad + \int_{-\infty}^{\delta+\epsilon_s} \frac{\partial}{\partial\delta}\left[(\delta+\epsilon_s-\nu)\, g_s^{OTC-O}(\nu)\right] d\nu \\ &= \int_{-\infty}^{\delta+\epsilon_s} g_s^{OTC-O}(\nu)\, d\nu,\end{aligned}$$

where, again, the boundary term is zero because $\delta + \epsilon_s - \nu = 0$ at $\nu = \delta + \epsilon_s$. Therefore,

$$\frac{\partial u(0,\delta,s)}{\partial\delta} = \lambda\theta\min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} \int_{-\infty}^{\delta+\epsilon_s} g_s^{OTC-O}(\nu)\, d\nu. \tag{A2}$$

Letting G_s^{OTC-O} and G_s^{OTC-NO} denote the corresponding cumulative distribution functions,

$$\frac{\partial u(1,\delta,s)}{\partial\delta} = 1 - \lambda(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\} \left[1 - G_s^{OTC-NO}(\delta+\epsilon_s)\right],$$

$$\frac{\partial u(0,\delta,s)}{\partial \delta} = \lambda\theta \min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} G_s^{OTC-O}(\delta+\epsilon_s).$$

Thus, for the reservation value $R(\delta,s) \equiv u(1,\delta,s) - u(0,\delta,s)$, I obtain

$$\begin{aligned}
\frac{\partial R(\delta,s)}{\partial \delta} &= \frac{\partial u(1,\delta,s)}{\partial \delta} - \frac{\partial u(0,\delta,s)}{\partial \delta} \\
&= 1 - \lambda(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\}\left[1 - G_s^{OTC-NO}(\delta+\epsilon_s)\right] \\
&\quad - \lambda\theta \min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} G_s^{OTC-O}(\delta+\epsilon_s) \\
&= 1 - \lambda\Bigg[(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\}\left[1 - G_s^{OTC-NO}(\delta+\epsilon_s)\right] \\
&\qquad + \theta \min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} G_s^{OTC-O}(\delta+\epsilon_s)\Bigg]. \qquad \text{(A3)}
\end{aligned}$$

Since

$$\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\}\left[1 - G_s^{OTC-NO}(\delta+\epsilon_s)\right] \in [0,1]$$

and

$$\min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} G_s^{OTC-O}(\delta+\epsilon_s) \in [0,1],$$

the term in brackets in (A3) is a weighted average of two numbers in $[0,1]$, with weights $1-\theta$ and θ, and $\theta \in [0,1]$. Hence,

$$0 \le \left[(1-\theta)\min\left\{1, \frac{N_s^{OTC-NO}}{N_s^{OTC-O}}\right\}\left[1 - G_s^{OTC-NO}(\delta+\epsilon_s)\right] + \theta \min\left\{1, \frac{N_s^{OTC-O}}{N_s^{OTC-NO}}\right\} G_s^{OTC-O}(\delta+\epsilon_s)\right] \le 1.$$

Therefore, for $\lambda \in [0,1]$,

$$\frac{\partial R(\delta,s)}{\partial \delta} \ge 1 - \lambda \ge 0.$$

It follows that $R(\delta,s)$ is weakly increasing in δ.

Proof B.2 (Proof of Lemma 2). By the previous Lemma 1, $R(\delta,s)$ is weakly increasing in δ for every state s. Hence,

$$A - R(\delta,s)$$

is weakly decreasing in δ, while

$$R(\delta,s) - B$$

is weakly increasing in δ.

Since the min $\{\cdot, 1\}$ operators for ϕ_s^{SL} and ϕ_s^{BL} imply that $1-\phi_s^{SL} \geq 0$ and $1-\phi_s^{BL} \geq 0$ for every state s, it follows directly that

$$\Delta(1, \delta; A, c) = \mathbb{E}_s \left[(1-\phi_s^{SL}) \big(A - R(\delta, s) \big) \right] - c$$

is weakly decreasing in δ, whereas

$$\Delta(0, \delta; B, c) = \mathbb{E}_s \left[(1-\phi_s^{BL}) \big(R(\delta, s) - B \big) \right] - c$$

is weakly increasing in δ.

Proof B.3 (**Proof of Proposition 1**)**.** For analytical simplicity, in this proof, I consider a special case where $k = \frac{1}{2}$ and $\delta \sim U[\underline{\delta}, \bar{\delta}]$. The Walrasian price is

$$P^* = F^{-1}(1-k) = \frac{\underline{\delta} + \bar{\delta}}{2}.$$

Consider first a positive tick size $c > 0$. Since $\alpha = 0$ and $\lambda = 0$, an agent of type δ values the asset at δ. Let the bid–ask spread be m ticks of size c, where $m \in \{1, 2, 3, \ldots\}$. The ask and bid are

$$A = P^* + \frac{1}{2}mc, \qquad B = P^* - \frac{1}{2}mc.$$

I first solve the sorting problem conditional on A and B. I conjecture that limit orders are only partially executed, while market orders execute with certainty:

$$\phi^{SL} < 1, \qquad \phi^{BL} < 1, \qquad \phi^{SM} = \phi^{BM} = 1.$$

Consider an initial owner with type δ. The owner has three choices: hold the asset (H), submit a sell market order (SM), or submit a sell limit order (SL). Under the conjectured execution probabilities, the corresponding choice-specific value functions are

$$\begin{aligned} V^H(1, \delta) &= \delta, \\ V^{SM}(1, \delta) &= B, \\ V^{SL}(1, \delta) &= \phi^{SL} A + (1-\phi^{SL})\delta. \end{aligned}$$

The slopes of these three value functions with respect to δ are, respectively,

$$1, \qquad 0, \qquad 1-\phi^{SL}.$$

Under the conjecture that $0 < \phi^{SL} < 1$, these slopes satisfy

$$0 < 1 - \phi^{SL} < 1.$$

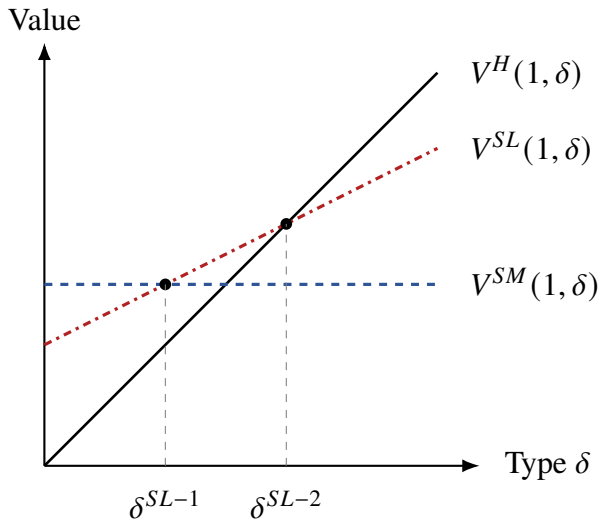


Figure A9: An Illustration of Choice-Specific Value Functions for Owners

Thus, as shown in Figure A9, the sell-market value is the flattest in δ (blue), the value from holding the asset is the steepest (black), and the sell-limit value has an intermediate slope (red). Therefore, the owner's optimal venue choice is characterized by two cutoff types, each determined by a single crossing between two value functions. Employing Definition 4, in this frictionless equilibrium, δ^{SL-1} denotes the type indifferent between submitting a sell limit order and a sell market order, and δ^{SL-2} denotes the type indifferent between submitting a sell limit order and holding the asset. The upper envelope of these value functions therefore implies that owners sort according to

$$\delta < \delta^{SL-1} : SM, \qquad \delta^{SL-1} \leq \delta \leq \delta^{SL-2} : SL, \qquad \delta > \delta^{SL-2} : \text{H}.$$

The two cutoff types satisfy the indifference conditions

$$\phi^{SL} A + (1 - \phi^{SL})\delta^{SL-1} = B, \tag{A4}$$

$$\phi^{SL} A + (1 - \phi^{SL})\delta^{SL-2} = \delta^{SL-2}. \tag{A5}$$

Given the optimal venue choice characterized above and the uniform type distribution, the sell-market and sell-limit order masses expressed in Equations 13 and 14 have the following closed-form expressions

$$N^{SM} = \frac{k}{\bar{\delta} - \underline{\delta}} \left(\delta^{SL-1} - \underline{\delta} \right), \qquad N^{SL} = \frac{k}{\bar{\delta} - \underline{\delta}} \left(\delta^{SL-2} - \delta^{SL-1} \right).$$

By symmetry, which requires $k = 1 - k$ and hence $k = \frac{1}{2}$, I have $N^{BM} = N^{SM}$. Therefore, the consistency condition in Equilibrium Definition 1 requires

$$\begin{aligned}\phi^{SL} &= \frac{N^{BM}}{N^{SL}} = \frac{N^{SM}}{N^{SL}} \\ &= \frac{\frac{k}{\bar{\delta}-\underline{\delta}}\left(\delta^{SL-1} - \underline{\delta}\right)}{\frac{k}{\bar{\delta}-\underline{\delta}}\left(\delta^{SL-2} - \delta^{SL-1}\right)} \\ &= \frac{\delta^{SL-1} - \underline{\delta}}{\delta^{SL-2} - \delta^{SL-1}}. \end{aligned} \tag{A6}$$

Equations (A4)–(A6) jointly determine the two cutoff types and the execution probability through a system of three equations in the three unknowns δ^{SL-1}, δ^{SL-2}, and ϕ^{SL}. From rearranging (A5),

$$\delta^{SL-2} = A.$$

Substitute $\delta^{SL-2} = A$ into Equation (A6)

$$\phi^{SL} = \frac{\delta^{SL-1} - \underline{\delta}}{A - \delta^{SL-1}},$$

and rearrange

$$\phi^{SL}(A - \delta^{SL-1}) = \delta^{SL-1} - \underline{\delta}.$$

Meanwhile, Equation A4 can be rearranged as

$$\phi^{SL}(A - \delta^{SL-1}) = B - \delta^{SL-1}.$$

Therefore,

$$B - \delta^{SL-1} = \delta^{SL-1} - \underline{\delta},$$

and hence

$$\delta^{SL-1} = \frac{B + \underline{\delta}}{2}.$$

Substituting $\delta^{SL-2} = A, \delta^{SL-1} = \frac{B+\underline{\delta}}{2}$ into Equation A6

$$\phi^{SL} = \frac{\frac{B+\underline{\delta}}{2} - \underline{\delta}}{A - \frac{B+\underline{\delta}}{2}} = \frac{B - \underline{\delta}}{2A - B - \underline{\delta}}.$$

Using

$$A = P^* + \frac{mc}{2}, \qquad B = P^* - \frac{mc}{2},$$

this becomes

$$\phi^{SL} = \frac{P^* - \underline{\delta} - \frac{mc}{2}}{P^* - \underline{\delta} + \frac{3mc}{2}}.$$

Since $m \geq 1$ and small $c > 0$, I have $mc > 0$. Therefore, the expressions above imply that $0 < \phi^{SL} < 1$. Recall that the execution probabilities are determined by $\phi^{SL} = \min\left\{\frac{N^{BM}}{N^{SL}}, 1\right\}$, $\phi^{BM} = \min\left\{\frac{N^{SL}}{N^{BM}}, 1\right\}$. Since $\phi^{SL} < 1$, it follows that $\frac{N^{BM}}{N^{SL}} < 1$, and hence $\frac{N^{SL}}{N^{BM}} > 1$, and hence $\phi^{BM} = 1$. By symmetry, I have $\phi^{SM} = \phi^{BM} = 1$. Thus, the conjectured execution pattern is verified.

I next determine the equilibrium bid–ask spread. Employing Corollary 2, the gain from deviation in this frictionless equilibrium is

$$\Delta(1, \delta^{SL-1}; A, c) = (1 - \phi^{SL})\left(A - \delta^{SL-1}\right) - c.$$

Recall by rearranging lower-cutoff indifference condition in Equation A4,

$$B - \delta^{SL-1} = \phi^{SL}(A - \delta^{SL-1}).$$

Therefore,

$$\begin{aligned}(1 - \phi^{SL})(A - \delta^{SL-1}) &= A - \delta^{SL-1} - \phi^{SL}(A - \delta^{SL-1}) \\ &= A - \delta^{SL-1} - (B - \delta^{SL-1}) \\ &= A - B = mc.\end{aligned}$$

Hence,

$$\Delta(1, \delta^{SL-1}; A, c) = mc - c = (m-1)c.$$

For any $c > 0$, if $m > 1$, this deviation is strictly profitable. Therefore, a bid–ask spread wider than one tick cannot be sustained in equilibrium. The equilibrium spread must bind at the minimum feasible positive spread,

$$m^* = 1.$$

Evaluating the equilibrium at $m^* = 1$ gives

$$A^* = P^* + \frac{c}{2}, \qquad B^* = P^* - \frac{c}{2},$$

and

$$\delta^{SL-1,*} = \frac{P^* + \underline{\delta}}{2} - \frac{c}{4}, \qquad \delta^{SL-2,*} = P^* + \frac{c}{2}.$$

By symmetry,

$$\delta^{BL-1,*} = P^* - \frac{c}{2}, \qquad \delta^{BL-2,*} = \frac{P^* + \bar{\delta}}{2} + \frac{c}{4}.$$

The limit-order execution probabilities are

$$\phi^{SL,*} = \frac{P^* - \underline{\delta} - \frac{c}{2}}{P^* - \underline{\delta} + \frac{3c}{2}}, \qquad \phi^{BL,*} = \frac{\bar{\delta} - P^* - \frac{c}{2}}{\bar{\delta} - P^* + \frac{3c}{2}}.$$

Finally, I take the limit as $c \to 0^+$. The equilibrium bid and ask converge to the Walrasian price:

$$\lim_{c\to 0^+} A^* = \lim_{c\to 0^+} B^* = P^*.$$

The cutoff types converge to

$$\lim_{c\to 0^+} \delta^{SL-1,*} = \frac{P^* + \underline{\delta}}{2}, \qquad \lim_{c\to 0^+} \delta^{SL-2,*} = P^*,$$

and

$$\lim_{c\to 0^+} \delta^{BL-1,*} = P^*, \qquad \lim_{c\to 0^+} \delta^{BL-2,*} = \frac{P^* + \bar{\delta}}{2}.$$

The limit-order execution probabilities converge to

$$\lim_{c\to 0^+} \phi^{SL,*} = \lim_{c\to 0^+} \phi^{BL,*} = 1.$$

Moreover,

$$\frac{\partial \phi^{SL,*}}{\partial c} = -\frac{2(P^* - \underline{\delta})}{\left(P^* - \underline{\delta} + \frac{3c}{2}\right)^2} < 0,$$

$$\frac{\partial \phi^{BL,*}}{\partial c} = -\frac{2(\bar{\delta} - P^*)}{\left(\bar{\delta} - P^* + \frac{3c}{2}\right)^2} < 0,$$

Thus, both execution probabilities are strictly decreasing in c. Equivalently, as $c \downarrow 0$, both $\phi^{SL,*}$ and $\phi^{BL,*}$ monotonically increase toward 1 from below.

To verify the market-order execution probabilities in the limit, the order masses satisfy

$$N^{BM,*} = N^{SM,*} = \frac{k}{\bar{\delta} - \underline{\delta}} \left(\frac{P^* - \underline{\delta}}{2} - \frac{c}{4} \right),$$

and

$$N^{SL,*} = \frac{k}{\bar{\delta} - \underline{\delta}} \left(\frac{P^* - \underline{\delta}}{2} + \frac{3c}{4} \right).$$

Therefore,

$$\lim_{c\to 0^+} \phi^{BM,*} = \lim_{c\to 0^+} \min\left\{\frac{N^{SL,*}}{N^{BM,*}}, 1\right\}$$
$$= \min\left\{\lim_{c\to 0^+} \frac{\frac{P^*-\underline{\delta}}{2} + \frac{3c}{4}}{\frac{P^*-\underline{\delta}}{2} - \frac{c}{4}}, 1\right\} = 1.$$

By symmetry,

$$\lim_{c\to 0^+} \phi^{SM,*} = 1.$$

Therefore,

$$\lim_{c\to 0^+} \phi^{SL,*} = \lim_{c\to 0^+} \phi^{BL,*} = \phi^{SM,*} = \phi^{BM,*} = 1.$$

To conclude, in the limit, every owner with $\delta < P^*$ sells the asset: types with $\delta < \frac{P^*+\underline{\delta}}{2}$ sell through market orders, while types with $\frac{P^*+\underline{\delta}}{2} \le \delta < P^*$ sell through limit orders. Similarly, every non-owner with $\delta > P^*$ acquires the asset: types with $P^* < \delta \le \frac{P^*+\bar{\delta}}{2}$ buy through limit orders, while types with $\delta > \frac{P^*+\bar{\delta}}{2}$ buy through market orders. Since all active orders execute with probability one in the limit, the post-trade final allocation is

$$q'(\delta) = \begin{cases} 0, & \delta < P^*, \\ 1, & \delta > P^*. \end{cases}$$

Hence, the asset is ultimately held by agents with valuations above P^*, which is exactly the Walrasian allocation. Therefore, as $c \to 0^+$, the frictionless central limit order book competitive equilibrium converges to the Walrasian equilibrium.

Proof B.4 (**Proof of Proposition 2**)**.** To begin with, notice that since $c > 0$ and $m \ge 1$, the equilibrium bid–ask spread satisfies

$$A - B = cm > 0.$$

Moreover, since $0 < \phi^{SL} < 1$ and $0 < \phi^{BL} < 1$ by assumption, the execution probabilities specified in Equation 3 immediately imply $\phi^{BM} = 1, \phi^{SM} = 1$.

I now prove the proposition. I first prove the venue sorting for the owner side, then for the non-owner side, and finally establish the cutoff ordering.

For an owner ($q = 1$), following Corollary 1, define the Stage-3 market-order cutoff δ^{SM} by

$$R(\delta^{SM}) = B.$$

Since $R(\delta)$ is strictly increasing ($\lambda < 1$ is a sufficient condition), conditional on not submitting a sell limit order in Stage 1, the owner submits a sell market order in Stage 3 if $\delta < \delta^{SM}$ and opts out of the CLOB if $\delta > \delta^{SM}$. Hence, $V(1, \delta \mid \text{Not SL})$ represents the maximal attainable value following the decision not to supply liquidity in Stage 1, taking into account the optimal subsequent choice in Stage 3:

$$V(1, \delta \mid \text{Not SL}) = \max\{B + u(0, \delta),\, u(1, \delta)\} = \begin{cases} B + u(0, \delta), & \delta < \delta^{SM}, \\ u(1, \delta), & \delta > \delta^{SM}, \end{cases}$$

with indifference at $\delta = \delta^{SM}$.

The value from submitting a sell limit order in Stage 1 is

$$V(1, \delta \mid \text{SL}) = \phi^{SL}\big(A + u(0, \delta)\big) + (1 - \phi^{SL})u(1, \delta).$$

Define

$$\pi^{SL}(\delta) \equiv V(1, \delta \mid \text{SL}) - V(1, \delta \mid \text{Not SL}).$$

Economically, $\pi^{SL}(\delta)$ is the profit from supplying liquidity through limit order relative to the maximal attainable value from not supplying liquidity and subsequently choosing optimally in Stage 3. Thus, supplying liquidity is strictly optimal if and only if $\pi^{SL}(\delta) > 0$. Substituting the expressions above gives

$$\begin{aligned} \pi^{SL}(\delta) &= \begin{cases} \phi^{SL}\big(A + u(0, \delta)\big) + (1 - \phi^{SL})u(1, \delta) - \big(B + u(0, \delta)\big), & \delta < \delta^{SM}, \\ \phi^{SL}\big(A + u(0, \delta)\big) + (1 - \phi^{SL})u(1, \delta) - u(1, \delta), & \delta > \delta^{SM}, \end{cases} \\ &= \begin{cases} (1 - \phi^{SL})R(\delta) + \phi^{SL}A - B, & \delta < \delta^{SM}, \\ \phi^{SL}\big(A - R(\delta)\big), & \delta > \delta^{SM}. \end{cases} \end{aligned} \tag{A7}$$

The key is that $\pi^{SL}(\delta)$ is increasing in the first region and decreasing in the second region. To see this property, since $R(\delta)$ is strictly increasing and $0 < \phi^{SL} < 1$, hence,

$$\frac{d\pi^{SL}(\delta)}{d\delta} = \begin{cases} (1 - \phi^{SL})R'(\delta) > 0, & \delta < \delta^{SM}, \\ -\phi^{SL}R'(\delta) < 0, & \delta > \delta^{SM}. \end{cases}$$

That is, the profit from supplying liquidity $\pi^{SL}(\delta)$ is single-peaked in δ, with its peak attained at

δ^{SM}. Moreover, $\pi^{SL}(\delta)$ is continuous at δ^{SM}, since the two branches coincide at δ^{SM}:

$$\pi^{SL}(\delta^{SM}) = \phi^{SL}(A - B) > 0.$$

Therefore, within each of the two regions, $\pi^{SL}(\delta)$ satisfies a strict single-crossing property. The assumed interiority of the cutoff types ensures the existence of a zero in each region, while strict monotonicity of $\pi^{SL}(\delta)$ within each region ensures its uniqueness. Figure A10 illustrates these properties of $\pi^{SL}(\delta)$.

In the first region, the unique cutoff $\delta^{SL-1} < \delta^{SM}$ is determined by

$$\begin{aligned}\pi^{SL}(\delta^{SL-1}) = 0 &\iff (1 - \phi^{SL})R(\delta^{SL-1}) + \phi^{SL}A - B = 0 \\ &\iff R(\delta^{SL-1}) = \frac{B - \phi^{SL}A}{1 - \phi^{SL}}.\end{aligned}$$

Since $\pi^{SL}(\delta)$ is strictly increasing in this region and crosses zero uniquely at δ^{SL-1}, it follows that

$$\pi^{SL}(\delta) < 0 \quad \text{for } \delta < \delta^{SL-1}, \qquad \pi^{SL}(\delta) > 0 \quad \text{for } \delta^{SL-1} < \delta \leq \delta^{SM}.$$

Thus, owners with $\delta < \delta^{SL-1}$ prefer the optimal Stage-3 alternative, a sell market order, whereas owners with $\delta^{SL-1} < \delta \leq \delta^{SM}$ prefer to supply liquidity through a sell limit order in Stage-1.

In the second region, the unique cutoff $\delta^{SL-2} > \delta^{SM}$ is determined by

$$\begin{aligned}\pi^{SL}(\delta^{SL-2}) = 0 &\iff \phi^{SL}\big(A - R(\delta^{SL-2})\big) = 0 \\ &\iff R(\delta^{SL-2}) = A.\end{aligned}$$

Since $\pi^{SL}(\delta)$ is strictly decreasing in this region and crosses zero uniquely at δ^{SL-2}, it follows that

$$\pi^{SL}(\delta) > 0 \quad \text{for } \delta^{SM} \leq \delta < \delta^{SL-2}, \qquad \pi^{SL}(\delta) < 0 \quad \text{for } \delta > \delta^{SL-2}.$$

Thus, owners with $\delta^{SM} \leq \delta < \delta^{SL-2}$ prefer to supply liquidity through a sell limit order in Stage-1, whereas owners with $\delta > \delta^{SL-2}$ prefer the optimal Stage-3 alternative, which is to opt out.

Combining the two regions,

$$\delta < \delta^{SL-1} \iff \text{Sell Market},$$

$$\delta^{SL-1} < \delta < \delta^{SL-2} \iff \text{Sell Limit},$$

$$\delta > \delta^{SL-2} \iff \text{Opt Out CLOB}.$$

At $\delta = \delta^{SL-1}$, the owner is indifferent between Sell Market and Sell Limit, and at $\delta = \delta^{SL-2}$, the owner is indifferent between Sell Limit and Opt Out.

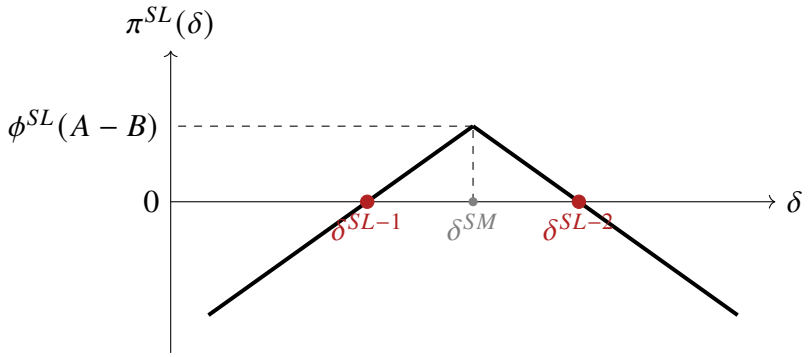


Figure A10: An Illustration of the Properties of the Profit Function $\pi^{SL}(\delta)$

For a non-owner ($q = 0$), following Corollary 1, define the Stage-3 market-order cutoff δ^{BM} by

$$R(\delta^{BM}) = A.$$

Since $R(\delta)$ is strictly increasing, conditional on not submitting a buy limit order in Stage 1, the non-owner opts out of the CLOB if $\delta < \delta^{BM}$ and submits a buy market order in Stage 3 if $\delta > \delta^{BM}$. Hence, $V(0, \delta \mid \text{Not BL})$ represents the maximal attainable value following the decision not to supply liquidity in Stage 1, taking into account the optimal subsequent choice in Stage 3:

$$V(0, \delta \mid \text{Not BL}) = \max\{u(0,\delta),\, u(1,\delta) - A\} = \begin{cases} u(0,\delta), & \delta < \delta^{BM}, \\ u(1,\delta) - A, & \delta > \delta^{BM}, \end{cases}$$

with indifference at $\delta = \delta^{BM}$.

The value from submitting a buy limit order in Stage 1 is

$$V(0, \delta \mid \text{BL}) = \phi^{BL}\big(u(1,\delta) - B\big) + (1 - \phi^{BL})u(0,\delta).$$

Define

$$\pi^{BL}(\delta) \equiv V(0, \delta \mid \text{BL}) - V(0, \delta \mid \text{Not BL}).$$

Economically, $\pi^{BL}(\delta)$ is the profit from supplying liquidity through a limit order relative to the maximal attainable value from not supplying liquidity and subsequently choosing optimally in Stage 3. Thus, supplying liquidity is strictly optimal if and only if $\pi^{BL}(\delta) > 0$. Substituting the expressions

above gives

$$\pi^{BL}(\delta) = \begin{cases} \phi^{BL}\big(u(1,\delta) - B\big) + (1-\phi^{BL})u(0,\delta) - u(0,\delta), & \delta < \delta^{BM}, \\ \phi^{BL}\big(u(1,\delta) - B\big) + (1-\phi^{BL})u(0,\delta) - \big(u(1,\delta) - A\big), & \delta > \delta^{BM}, \end{cases}$$

$$= \begin{cases} \phi^{BL}\big(R(\delta) - B\big), & \delta < \delta^{BM}, \\ A - \phi^{BL}B - (1-\phi^{BL})R(\delta), & \delta > \delta^{BM}. \end{cases} \tag{A8}$$

The key is that $\pi^{BL}(\delta)$ is increasing in the first region and decreasing in the second region. To see this property, since $R(\delta)$ is strictly increasing and $0 < \phi^{BL} < 1$, I have

$$\frac{d\pi^{BL}(\delta)}{d\delta} = \begin{cases} \phi^{BL}R'(\delta) > 0, & \delta < \delta^{BM}, \\ -(1-\phi^{BL})R'(\delta) < 0, & \delta > \delta^{BM}. \end{cases}$$

That is, the profit from supplying liquidity $\pi^{BL}(\delta)$ is single-peaked in δ, with its peak attained at δ^{BM}. Moreover, $\pi^{BL}(\delta)$ is continuous at δ^{BM}, since the two branches coincide at δ^{BM}:

$$\pi^{BL}(\delta^{BM}) = \phi^{BL}(A - B) > 0.$$

Therefore, within each of the two regions, $\pi^{BL}(\delta)$ satisfies a strict single-crossing property. The assumed interiority of the cutoff types ensures the existence of a zero in each region, while strict monotonicity of $\pi^{BL}(\delta)$ within each region ensures its uniqueness.

In the first region, the unique cutoff $\delta^{BL-1} < \delta^{BM}$ is determined by

$$\begin{aligned} \pi^{BL}(\delta^{BL-1}) = 0 &\iff \phi^{BL}\big(R(\delta^{BL-1}) - B\big) = 0 \\ &\iff R(\delta^{BL-1}) = B. \end{aligned}$$

Since $\pi^{BL}(\delta)$ is strictly increasing in this region and crosses zero uniquely at δ^{BL-1}, it follows that

$$\pi^{BL}(\delta) < 0 \quad \text{for } \delta < \delta^{BL-1}, \qquad \pi^{BL}(\delta) > 0 \quad \text{for } \delta^{BL-1} < \delta \le \delta^{BM}.$$

Thus, non-owners with $\delta < \delta^{BL-1}$ prefer the optimal Stage-3 alternative, which is to opt out, whereas non-owners with $\delta^{BL-1} < \delta \le \delta^{BM}$ prefer to supply liquidity through a buy limit order in Stage 1.

In the second region, the unique cutoff $\delta^{BL-2} > \delta^{BM}$ is determined by

$$\pi^{BL}(\delta^{BL-2}) = 0 \iff A - \phi^{BL}B - (1-\phi^{BL})R(\delta^{BL-2}) = 0$$
$$\iff R(\delta^{BL-2}) = \frac{A - \phi^{BL}B}{1-\phi^{BL}}.$$

Since $\pi^{BL}(\delta)$ is strictly decreasing in this region and crosses zero uniquely at δ^{BL-2}, it follows that

$$\pi^{BL}(\delta) > 0 \quad \text{for } \delta^{BM} \le \delta < \delta^{BL-2}, \qquad \pi^{BL}(\delta) < 0 \quad \text{for } \delta > \delta^{BL-2}.$$

Thus, non-owners with $\delta^{BM} \le \delta < \delta^{BL-2}$ prefer to supply liquidity through a buy limit order in Stage 1, whereas non-owners with $\delta > \delta^{BL-2}$ prefer the optimal Stage-3 alternative, a buy market order.

Combining the two regions,

$$\delta < \delta^{BL-1} \iff \text{Opt Out CLOB},$$

$$\delta^{BL-1} < \delta < \delta^{BL-2} \iff \text{Buy Limit},$$

$$\delta > \delta^{BL-2} \iff \text{Buy Market}.$$

At $\delta = \delta^{BL-1}$, the non-owner is indifferent between Opt Out and Buy Limit, and at $\delta = \delta^{BL-2}$, the non-owner is indifferent between Buy Limit and Buy Market.

It remains to establish the ordering of the four cutoff types. From the cutoff indifference conditions derived above,

$$R(\delta^{SL-1}) = \frac{B - \phi^{SL}A}{1-\phi^{SL}}, \qquad R(\delta^{BL-1}) = B,$$

and

$$R(\delta^{SL-2}) = A, \qquad R(\delta^{BL-2}) = \frac{A - \phi^{BL}B}{1-\phi^{BL}}.$$

Since $A > B$, $0 < \phi^{SL} < 1$, and $0 < \phi^{BL} < 1$, I have

$$R(\delta^{SL-1}) = \frac{B - \phi^{SL}A}{1-\phi^{SL}} < \frac{B - \phi^{SL}B}{1-\phi^{SL}} = B = R(\delta^{BL-1}),$$

$$R(\delta^{BL-1}) = B < A = R(\delta^{SL-2}),$$

$$R(\delta^{SL-2}) = A = \frac{A - \phi^{BL}A}{1-\phi^{BL}} < \frac{A - \phi^{BL}B}{1-\phi^{BL}} = R(\delta^{BL-2}).$$

Therefore,

$$R(\delta^{SL-1}) < R(\delta^{BL-1}) < R(\delta^{SL-2}) < R(\delta^{BL-2}).$$

Since $R(\delta)$ is strictly increasing in δ, it follows that

$$\delta^{SL-1} < \delta^{BL-1} < \delta^{SL-2} < \delta^{BL-2}.$$

Proof B.5 (Proof of Proposition 3). Following Corollary 2, the one-shot deviations in the benchmark case are

$$\begin{aligned}\Delta(1, \delta^{SL-1}) &= (1 - \phi^{SL})\big(A - R(\delta^{SL-1})\big) - c,\\ \Delta(0, \delta^{BL-2}) &= (1 - \phi^{BL})\big(R(\delta^{BL-2}) - B\big) - c.\end{aligned}$$

From Proposition 2, the corresponding cutoff conditions are

$$R(\delta^{SL-1}) = \frac{B - \phi^{SL}A}{1 - \phi^{SL}}, \qquad R(\delta^{BL-2}) = \frac{A - \phi^{BL}B}{1 - \phi^{BL}}.$$

Substituting the first cutoff condition gives

$$\begin{aligned}\Delta(1, \delta^{SL-1}) &= (1 - \phi^{SL})\left(A - \frac{B - \phi^{SL}A}{1 - \phi^{SL}}\right) - c\\ &= (1 - \phi^{SL})\left(\frac{A - B}{1 - \phi^{SL}}\right) - c\\ &= A - B - c.\end{aligned}$$

Similarly, substituting the second cutoff condition gives

$$\begin{aligned}\Delta(0, \delta^{BL-2}) &= (1 - \phi^{BL})\left(\frac{A - \phi^{BL}B}{1 - \phi^{BL}} - B\right) - c\\ &= (1 - \phi^{BL})\left(\frac{A - B}{1 - \phi^{BL}}\right) - c\\ &= A - B - c.\end{aligned}$$

Let the bid–ask spread be measured by m ticks of size c, i.e., $A - B = mc$, where $m \in \{1, 2, 3, \ldots\}$. Therefore,

$$\Delta(1, \delta^{SL-1}) = \Delta(0, \delta^{BL-2}) = mc - c = (m - 1)c.$$

Hence, for any $c > 0$, when $m > 1$, both one-shot deviations are strictly profitable, so such a spread cannot be sustained in equilibrium. Only when $m = 1$ is the gain from either deviation equal to

zero. That is, the bid–ask spread binds at the minimum feasible positive spread of one tick.

Proof B.6 (Proof of Proposition 4)**.** When $\lambda = 0$, there is no OTC trading, so the continuation values are

$$u(1, \delta, s) = \delta + \epsilon_s, \qquad u(0, \delta, s) = 0.$$

Hence, the reservation value of the asset is

$$R(\delta, s) = u(1, \delta, s) - u(0, \delta, s) = \delta + \epsilon_s.$$

The market-order decision rules in Corollary 1 therefore become

$$\mathbf{1}_{SM}(\delta, s) = \mathbf{1}\{B > \delta + \epsilon_s\}, \qquad \mathbf{1}_{BM}(\delta, s) = \mathbf{1}\{A < \delta + \epsilon_s\}.$$

Importantly, the ask A and bid B are determined ex ante, before the aggregate state is realized, and are therefore state-invariant. Thus, conditional on the ex-ante equilibrium quotes (A, B), the aggregate state affects the market-order decision only through the state-contingent valuation $\delta + \epsilon_s$. Since $\epsilon_- = -\alpha$ and $\epsilon_+ = +\alpha$, with $\alpha \geq 0$, the negative state shifts every agent's valuation downward relative to the positive state.

Consider first sell market orders. Define the sets of types that optimally submit sell market orders in the two states as

$$\mathcal{D}_-^{SM} \equiv \{\delta : B > \delta - \alpha\}, \qquad \mathcal{D}_+^{SM} \equiv \{\delta : B > \delta + \alpha\}.$$

For any type δ,

$$B > \delta + \alpha \quad \Longrightarrow \quad B > \delta - \alpha.$$

Therefore,

$$\mathcal{D}_+^{SM} \subseteq \mathcal{D}_-^{SM}.$$

That is, every type who finds it optimal to submit a sell market order in the positive state also finds it optimal to submit a sell market order in the negative state, while some additional types may sell by market order only in the negative state (i.e., informed market orders). Equivalently, the corresponding decision rules satisfy, for every δ,

$$\mathbf{1}_{SM}(\delta, -) = \mathbf{1}\{\delta \in \mathcal{D}_-^{SM}\} \geq \mathbf{1}\{\delta \in \mathcal{D}_+^{SM}\} = \mathbf{1}_{SM}(\delta, +).$$

Moreover, the Stage-1 sell-limit-order decision $\mathbf{1}_{SL}(\delta)$ is made before the aggregate state is realized and is therefore state-invariant. Hence, the set of agents available to submit sell market

orders,

$$\{\delta : \mathbf{1}_{SL}(\delta) = 0\},$$

is the same in both states. It follows that

$$\begin{aligned} N_{-}^{SM} &= \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{SL}(\delta))\mathbf{1}_{SM}(\delta, -)kf(\delta)\, d\delta \\ &\geq \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{SL}(\delta))\mathbf{1}_{SM}(\delta, +)kf(\delta)\, d\delta \\ &= N_{+}^{SM}. \end{aligned}$$

Therefore,

$$N_{-}^{SM} \geq N_{+}^{SM}.$$

The following Figure A11 illustrates the key idea of the proof. It also shows where δ_s^{SM} comes from, which appears in Figure 7 and Figure 8. It corresponds to the state-contingent counterpart of δ^{SM} plotted in Figure A10 for the no-risk case.

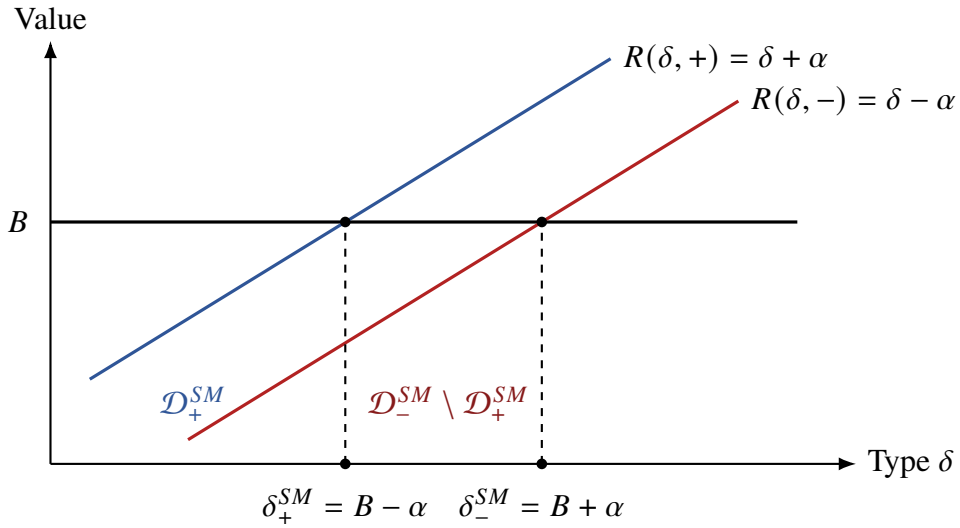


Figure A11: An Illustration of Sell Market Decisions Across Aggregate States

Consider next buy market orders. Define the sets of types that optimally submit buy market orders in the two states as

$$\mathcal{D}_{-}^{BM} \equiv \{\delta : A < \delta - \alpha\}, \qquad \mathcal{D}_{+}^{BM} \equiv \{\delta : A < \delta + \alpha\}.$$

For any type δ,

$$A < \delta - \alpha \quad \Longrightarrow \quad A < \delta + \alpha.$$

Therefore,

$$\mathcal{D}_{-}^{BM} \subseteq \mathcal{D}_{+}^{BM}.$$

That is, every type who finds it optimal to submit a buy market order in the negative state also finds it optimal to submit a buy market order in the positive state, while some additional types may buy by market order only in the positive state. Equivalently, the corresponding decision rules satisfy, for every δ,

$$\mathbf{1}_{BM}(\delta, +) = \mathbf{1}\{\delta \in \mathcal{D}_{+}^{BM}\} \geq \mathbf{1}\{\delta \in \mathcal{D}_{-}^{BM}\} = \mathbf{1}_{BM}(\delta, -).$$

Moreover, the Stage-1 buy-limit-order decision $\mathbf{1}_{BL}(\delta)$ is made before the aggregate state is realized and is therefore state-invariant. Hence, the set of agents available to submit buy market orders,

$$\{\delta : \mathbf{1}_{BL}(\delta) = 0\},$$

is the same in both states. It follows that

$$\begin{aligned} N_{+}^{BM} &= \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{BL}(\delta))\mathbf{1}_{BM}(\delta, +)(1 - k) f(\delta)\, d\delta \\ &\geq \int_{\underline{\delta}}^{\bar{\delta}} (1 - \mathbf{1}_{BL}(\delta))\mathbf{1}_{BM}(\delta, -)(1 - k) f(\delta)\, d\delta \\ &= N_{-}^{BM}. \end{aligned}$$

Therefore,

$$N_{+}^{BM} \geq N_{-}^{BM}.$$

I now consider the corresponding limit-order execution probabilities. For sell limit orders posted at the ask A, the CLOB matching rule gives

$$\phi_{s}^{SL} = \min\left\{\frac{N_{s}^{BM}}{N^{SL}}, 1\right\}.$$

Since N^{SL} is determined at the limit-order stage before the aggregate state is realized, it is state-invariant. Together with $N_{+}^{BM} \geq N_{-}^{BM}$, this implies

$$\begin{aligned} \phi_{+}^{SL} &= \min\left\{\frac{N_{+}^{BM}}{N^{SL}}, 1\right\} \\ &\geq \min\left\{\frac{N_{-}^{BM}}{N^{SL}}, 1\right\} \\ &= \phi_{-}^{SL}. \end{aligned}$$

Hence,

$$\phi_{+}^{SL} \geq \phi_{-}^{SL}.$$

For buy limit orders posted at the bid B, the CLOB matching rule gives

$$\phi_{s}^{BL} = \min\left\{\frac{N_{s}^{SM}}{N^{BL}}, 1\right\}.$$

Since N^{BL} is also determined before the aggregate state is realized and is therefore state-invariant, together with $N_{-}^{SM} \geq N_{+}^{SM}$ this implies

$$\begin{aligned}\phi_{-}^{BL} &= \min\left\{\frac{N_{-}^{SM}}{N^{BL}}, 1\right\} \\ &\geq \min\left\{\frac{N_{+}^{SM}}{N^{BL}}, 1\right\} \\ &= \phi_{+}^{BL}.\end{aligned}$$

Hence,

$$\phi_{-}^{BL} \geq \phi_{+}^{BL}.$$

C Appendix: Numerical Methods

Gain from Deviation Δ

The gain from deviation Δ, defined in Corollary 2, pins down the equilibrium bid–ask spread. Here, in Figures A12 and A13, I plot how the gain from deviation increases with the bid–ask spread, illustrating the economic forces behind undercutting and overbidding the quotes.

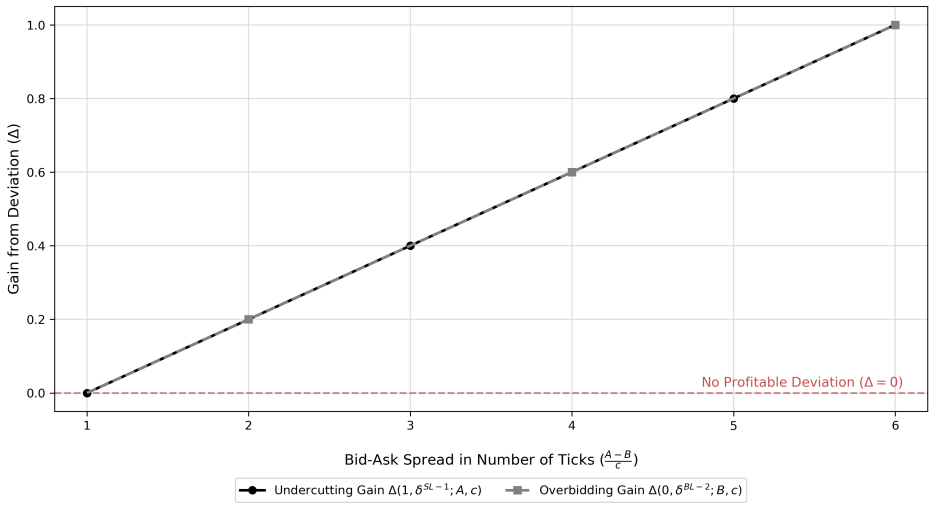


Figure A12: Benchmark Case ($\alpha = 0$): Gain From Deviation, A Numerical Example

Note: This is the partial equilibrium of the numerical example presented in Figure 5. Parameters: tick-size $c = 0.2$, shock size $\alpha = 0$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

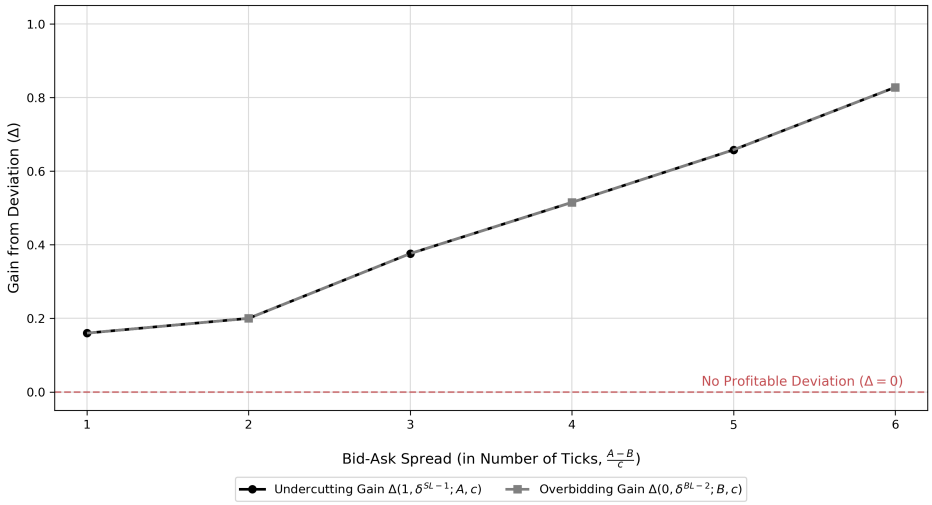


Figure A13: General Case ($\alpha > 0$): Gain From Deviation, A Numerical Example

Note: This is the partial equilibrium of the numerical example presented in Figure 7. Parameters: tick-size $c = 0.2$, shock size $\alpha = 0.3$, bargaining power $\theta = 0.5$, search efficiency $\lambda = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$, asset supply $k = 0.5$.

Grid Search for the Optimal Tick Size

Figure A14 illustrates the numerical method used to solve for the optimal tick size for an arbitrarily chosen shock size, $\alpha = 0.3$. Specifically, I perform a grid search over possible values of c and select the value that maximizes allocative efficiency. Each dot in the figure represents one grid point. The dashed vertical line indicates the optimal tick size for $\alpha = 0.3$, corresponding to the optimal tick size reported in Figure 9.

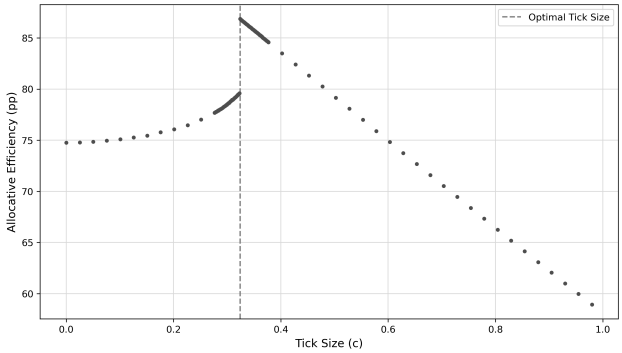


Figure A14: A Numerical Example Illustrating the Grid Search for the Optimal Tick Size ($\alpha = 0.3$)

Parameters: shock size $\alpha = 0.3$, search efficiency $\lambda = 0.5$, bargaining power $\theta = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = kf(\delta)$, asset supply $k = 0.5$.

Figure A15 shows a numerical example of the efficiency loss due to misspecification of the tick size. In this figure, the blue dots represent allocative efficiency when the tick size is set at the optimal level for a given α, which is identical to the bottom panel of Figure 9. The black line represents allocative efficiency calculated using the same risk level α but with an arbitrary fixed tick size, $c = 0.25$. This tick size is optimal only when the risk level happens to be $\alpha \approx 0.24$. Only at this particular value of α does allocative efficiency under the fixed tick size coincide with the optimal one. For all other values of α, allocative efficiency under the fixed tick size is strictly lower than under the optimal tick size, and the gap between the two is the efficiency loss due to tick-size misspecification. The farther the risk level is from the specific value $\alpha \approx 0.24$, the more biased the fixed tick size becomes and the larger the resulting efficiency loss.

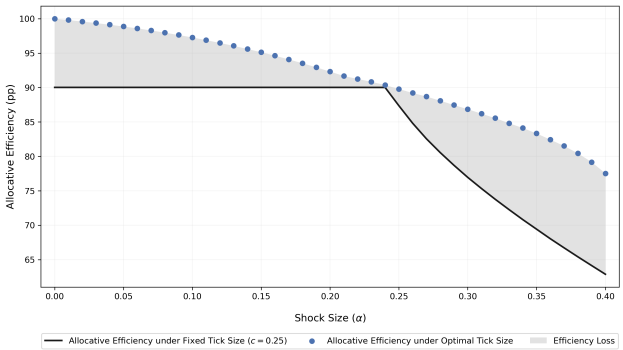


Figure A15: Efficiency Loss from Tick Size Mis-specification

Parameters: search efficiency $\lambda = 0.5$, bargaining power $\theta = 0.5$, type $f(\delta) \sim U[99, 101]$, initial joint distribution $\mu(q = 1, \delta) = kf(\delta)$, asset supply $k = 0.5$.

Parameter Identification The following figure and table discuss the parameter identification.

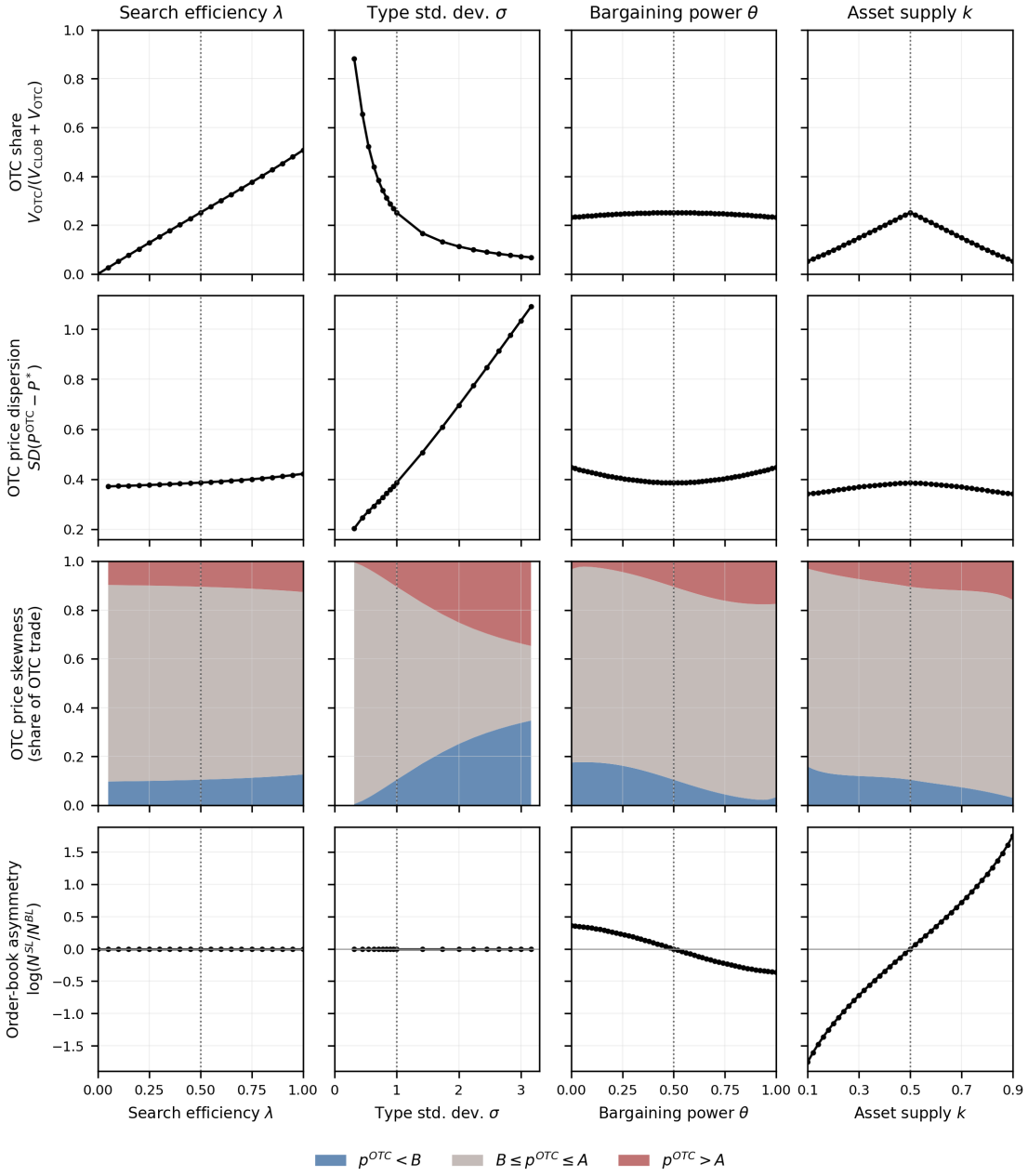


Figure A16: Numerical Comparative Statics for Parameter Identification

Parameters: tick size $c = 1$, type distribution $f(\delta) \sim N(100, \sigma^2)$, and initial joint distribution $\mu(q = 1, \delta) = k f(\delta)$. Each column presents comparative statics of the four moments used for estimation with respect to one specific parameter, holding all other parameter values fixed at their baseline levels. The baseline values in this numerical example are $k = 0.5$, $\theta = 0.5$, $\sigma = 1$, and $\lambda = 0.5$, which are also indicated by the corresponding dashed vertical line.

Because the estimates in Table 3 imply a nearly symmetric equilibrium ($k \approx 0.5$, $\theta \approx 0.5$), Table A1 reports the sign pattern of the Jacobian of the moments with respect to the parameters at

Table A1: Sign Pattern of Moment Responses at the Symmetric Equilibrium

Moment	λ	σ	θ	k
M_1: OTC market share $V_{OTC}/(V_{CLOB} + V_{OTC})$	+	−	≈ 0	≈ 0
M_2: OTC price dispersion $SD(p^{OTC} - p^*)$	+	+	≈ 0	≈ 0
M_3: OTC price skewness $\Pr(p^{OTC} > A) - \Pr(p^{OTC} < B)$	0	0	+	+
M_4: Limit order book asymmetry N^{SL}/N^{BL}	0	0	−	+

Notes: Signs of $\partial M_i/\partial \Gamma_j$ evaluated at $k = \theta = \lambda = 0.5$ and $\sigma = 1$, corresponding to the baseline symmetric equilibrium. Zeros in the lower-left block hold exactly. Entries ≈ 0 in the upper-right block hold locally around the symmetric equilibrium parameter configuration, except for $\partial M_1/\partial k$, which has a kink at $k = 0.5$.

a symmetric baseline. At this baseline the Jacobian is block-diagonal. The first two moments, the OTC share and OTC price dispersion, respond locally only to λ and σ; the last two moments, the skew imbalance and order-book asymmetry, respond only to θ and k. Within each block, the two parameters move the moments in distinct directions: λ and σ both raise dispersion but move the OTC share in opposite directions, while θ and k both raise the skew imbalance but move the order book in opposite directions. Both 2×2 blocks therefore have strictly positive determinants, and the parameters are locally identified.

D Appendix: Estimation Results

Table A2: 5-Year Treasury: Benchmark Model without Risk: Parameter Estimates and Model Fit

Panel A. Targeted Moments	**Data**	**Model**
OTC market share	0.2058	0.2058
OTC price dispersion	0.0049	0.0049
OTC price skewness	-0.0004	-0.0004
Limit order book asymmetry	1.0245	1.0245
Panel B. Estimated Parameters	**Symbol**	**Value**
Search efficiency in OTC	λ	0.8042
Type standard deviation	σ	0.0141
Bargaining power in OTC	θ	0.4992
Asset supply	k	0.5071
Panel C. Non-targeted Moment	**Data**	**Model**
Sell limit order execution probability	12.13%	43.17%
Buy limit order execution probability	12.00%	42.85%
Panel D. Welfare Analysis		
Allocative efficiency		85.00%

Notes: Panel A compares the empirical moments with the model predicted moments. Panel B reports the estimated parameter values. Panel C reports the non-targeted moment as test of model fitness. Panel D reports the model-implied allocative efficiency at the estimated parameter values. The minimized criterion value $Q(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}) = 0$.

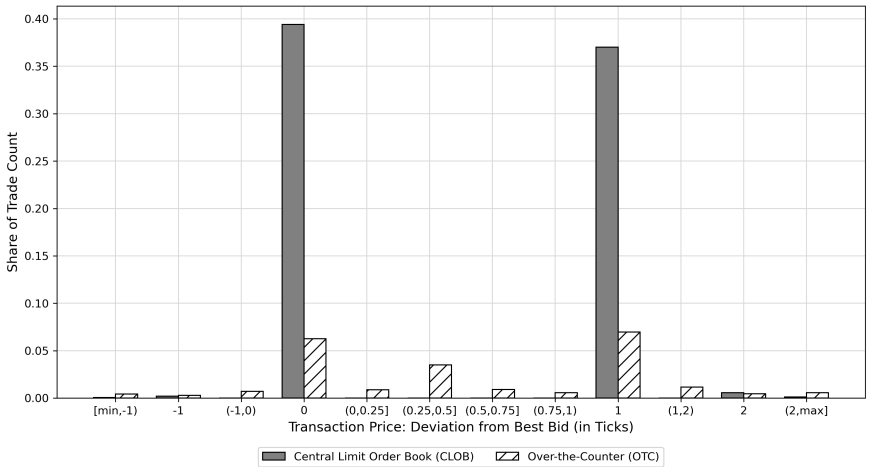


Figure A17: 5-Year Treasury: Distribution of Transaction Prices Relative to the Best Bid: CLOB versus OTC

Table A3: 2-Year Treasury: Benchmark Model without Risk: Parameter Estimates and Model Fit

Panel A. Targeted Moments	**Data**	**Model**
OTC market share	0.1741	0.1170
OTC price dispersion	0.0105	0.0048
OTC price skewness	0.2178	0.2093
Limit order book asymmetry	1.0276	0.8381
Panel B. Estimated Parameters	**Symbol**	**Value**
Search efficiency in OTC	λ	1.0000
Type standard deviation	σ	0.0139
Bargaining power in OTC	θ	0.5636
Asset supply	k	0.5443
Panel C. Non-targeted Moment	**Data**	**Model**
Sell limit order execution probability	11.96%	58.09%
Buy limit order execution probability	11.95%	67.89%
Panel D. Welfare Analysis		
Allocative efficiency		92.21%

Notes: Panel A compares the empirical moments with the model predicted moments. Panel B reports the estimated parameter values. Panel C reports the non-targeted moment as test of model fitness. Panel D reports the model-implied allocative efficiency at the estimated parameter values. The minimized criterion value $Q(\hat{\mathbf{\Gamma}}_{\text{Benchmark}}) = 0.432$. The model's imperfect fit across the four targeted moments is primarily driven by the second moment, OTC price dispersion, which accounts for 67% of the minimized objective function (0.289 out of $Q = 0.432$). Figure A18 illustrates the source of this discrepancy. OTC transaction prices are widely and asymmetrically dispersed: a very large fraction of OTC trades execute at prices more than one tick above the best ask, generating a very long right tail extending well beyond the quoted bid–ask spread. The model cannot fully reproduce this degree of price dispersion and asymmetry, resulting in a relatively poor fit for the second moment.

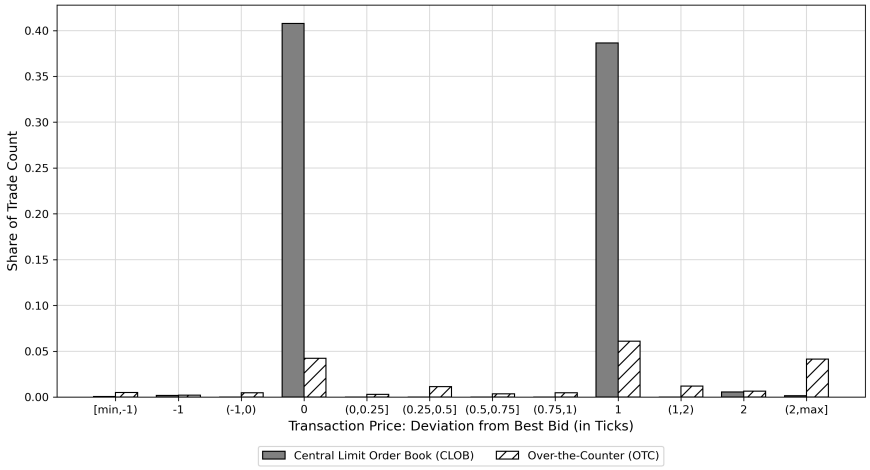


Figure A18: 2-Year Treasury: Distribution of Transaction Prices Relative to the Best Bid: CLOB versus OTC