NBER WORKING PAPER SERIES

SIZE DISCOVERY

Darrell Duffie
Haoxiang Zhu

Working Paper 21696
http://www.nber.org/papers/w21696

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 2015

For helpful discussions and comments, we thank Bruno Biais, Pierre Collin-Dufresne, Songzi Du, Benjamin Junge, Eiichiro Kazumori, Pete Kyle, Emmanuel Moench, Sophie Moinas, Romans Pancs, Mark Ready, Anders Trolle, and Robert Wilson, as well as participants at the American Economics Association annual meeting, University of Bonn, University of Geneva, INSEAD, University of Zurich, Swiss National Bank, HEC Paris, Federal Reserve Board, CFTC, University of Lugano, European University Institute, Toulouse School of Economics, Wilfred Laurier University, EIEF (Banca d'Italia), and the Bundesbank. We are especially grateful for research assistance by Jun Yan of the Stanford Statistics Department and Hyungjune Kang of the MIT Sloan School of Management. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Size Discovery
Darrell Duffie and Haoxiang Zhu
NBER Working Paper No. 21696
November 2015, Revised April 2016
JEL No. D47,D82,G14

ABSTRACT

Size discovery is the use of trade mechanisms by which large quantities of an asset can be exchanged at a price that does not respond to price pressure. Primary examples of size discovery include “workup” in Treasury markets, “matching sessions” in corporate bond and CDS markets, and block-trading “dark pools” in equity markets. By freezing the execution price and giving up on market-clearing, a size-discovery mechanism overcomes large investors' concerns over price impacts. Price-discovery mechanisms clear the market, but cause investors to internalize their price impacts, inducing costly delays in the reduction of position imbalances. We show how augmenting a price-discovery mechanism with a size-discovery mechanism improves allocative efficiency.

Darrell Duffie
Graduate School of Business
Stanford University
Stanford, CA 94305-7298
and NBER
duffie@stanford.edu

Haoxiang Zhu
MIT Sloan School of Management
100 Main Street, E62-623
Cambridge, MA 02142
and NBER
zhuh@mit.edu

1 Introduction

This paper shows that size-discovery mechanisms, by which large transactions can be quickly arranged at *fixed prices*, improve allocative efficiency in markets with imperfect competition and private information over latent supply or demand imbalances.

The issue of market liquidity has received intense attention in the last few years. The Securities and Exchange Commission (2010) and the U.S. Department of the Treasury (2016) raise important questions and concerns about the liquidity and design of markets for U.S. equities and Treasuries. The "Flash Crash" on May 6, 2010 in U.S. equity and futures markets and the "Flash Rally" on October 15, 2014 in U.S. Treasury markets were wake-up calls that even deep liquid markets may experience extreme price movements without obvious fundamental news (Joint CFTC-SEC Advisory Committee, 2011; Joint Staff Report, 2015). There are widespread concerns that dealers are less willing or likely to absorb large trade flows onto their balance sheets (Adrian et al., 2015).

An important aspect of market liquidity is the ability to quickly buy or sell large quantities of an asset with a small price impact. By definition, price impact is primarily a concern of large strategic investors—such as dealers, mutual funds, pension funds, and insurance companies—and not of small or "price-taking" investors. Price impact is a particular concern of major financial intermediaries such as broker-dealers, who often absorb substantial inventory positions in primary issuance markets or from their client investors, and then seek to offload these positions in inter-dealer markets. Duffie (2010) surveys widespread evidence of substantial price impact around large purchases and sales, even in settings with relatively symmetric and transparent information.

To mitigate price impact, investors often split large orders into many smaller pieces and execute them slowly over time. Order splitting is done by computer algorithms in electronic markets and manually in voice markets. As we explain shortly, such piecemeal execution is inefficient from an allocative point of view, given the associated costly delay in reducing undesired positions. Investors could alternatively pass a large to position to a dealer at a price concession, but this strategy has become more costly in recent years, as bank-affiliated dealers are subject to tighter capital and liquidity regulation.

We show that size discovery is an effective way to mitigate the allocative inefficiency caused by strategic avoidance of price impact. Therefore, size discovery acts as a valuable source of block liquidity that substitutes for order-splitting execution strategies and for the supply of market-making services by major dealers. Examples of size-discovery mechanisms used in practice include:

- Workup, a trading protocol by which buyers and sellers successively increase, or "work up," the quantities of an asset that are exchanged at a fixed price. Each participant in a workup has the option to drop out at any time. In the market for U.S. Treasuries, Fleming and Nguyen (2015) find that workup accounts for 43% to 56% of total trading volume on the BrokerTec platform on a typical day.
- "Matching sessions," a variant of workup found in markets for corporate bonds and credit default swaps (CDS). For the most actively traded CDS product, Collin-Dufresne, Junge, and Trolle (2016) find that matching sessions or workup account for 38% of trade volume on GFI, a major swap execution facility.
- Block-crossing "dark pools," such as Liquidnet and POSIT, which are predominantly used in equity markets. In a typical "midpoint" dark pool, buyers and sellers match orders at the midpoint of the best bid price and best offer price shown on transparent exchanges. Dark pools account for about 15% of trading volume in the U.S. equity markets (Zhu, 2014). Certain dark pools offer limited price discovery. Others do not use price discovery at all.

Despite some institutional differences discussed in Section 2, these various forms of size discovery share the key feature of crossing orders at fixed prices, thus without price impact. Although aware of the trade price, market participants conducting size discovery are uncertain of how much of the asset they will be able to trade at that price, which is not sensitive to their demands. One side of the market is eventually rationed, being willing to trade more at the given price. Thus, a size-discovery mechanism cannot clear the market, and is therefore inefficient on its own. Size discovery stands in sharp contrast to "price discovery" trading mechanisms, which find the market-clearing price that matches supply and demand. Nevertheless, precisely by giving up on market clearing, a size-discovery mechanism reduces the adverse effect of investors' strategic incentives to dampen their immediate demands. We show, as a consequence, that a market design combining size discovery and price discovery offers substantial efficiency improvement over a market that relies only on price discovery.

Our modeling approach and the intuition for our results can be roughly summarized as follows. An asset pays a liquidating dividend at a random future time. Before this time, double auctions for the asset are held among n strategic traders at evenly spaced time intervals of some length Δ. Thus, the auctions are held at times 0, Δ, 2Δ, and so on. Before the first of these auctions, the inventory of the asset held by each trader has an undesired component, positive or negative, that is not observable to other traders. Each trader suffers a continuing cost that is increasing in his undesired inventory imbalance. In each of the successive double auctions, traders submit demand schedules. The market operator aggregates these demand schedules and calculates the market-clearing price, at which total demand and supply are matched.

Because there are only finitely many traders, each of them "shades" her demand schedule in order to mitigate her own impact on the market-clearing price. For example, each trader who wishes to sell submits a supply schedule that expresses, at each price, only a fraction of her actual trading interest in order to reduce her own downward pressure on the market-clearing price. The unique efficient allocation is that giving each trader the same magnitude of undesired inventory. At each successive double auction, however, traders' inventories adjust only gradually toward the efficient allocation. As a result, traders with large unwanted positions, whether short of long, may bear significant costs, relative to the efficient allocation. These excess costs are not reduced by holding more frequent auctions. As shown by Vayanos (1999) and Du and Zhu (2015), even if trading is continual, convergence to the efficient allocation is not instantaneous. Strategic incentives to avoid price impact actually reduce the equilibrium rate of reduction of imbalances as the time between trading rounds become smaller.

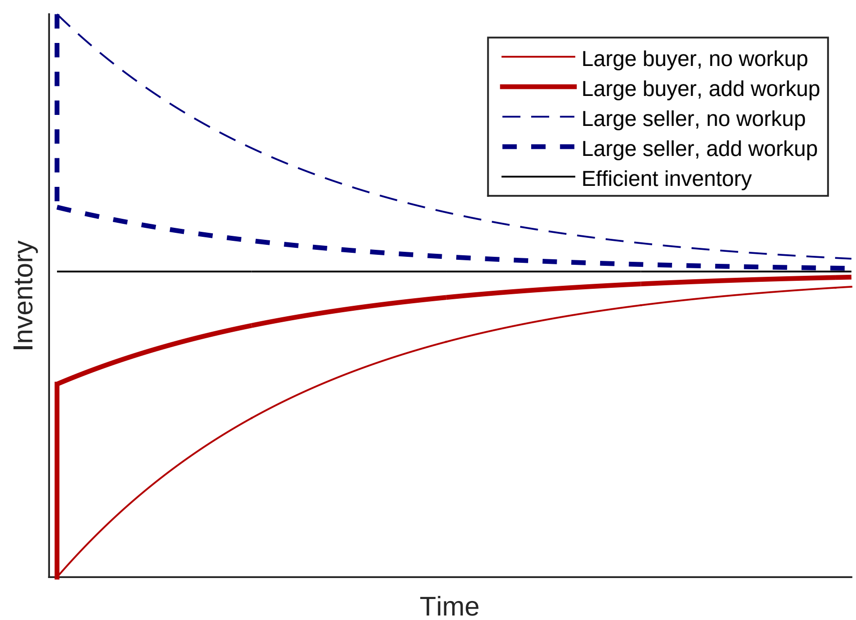

Figure 1: *Inventory paths with and without a workup.* The thin-line plots are the equilibrium inventory paths of a buyer and a seller in sequential-double-auction market. Plotted in bold are the equilibrium inventory paths of the same buyer and seller in a market with a workup followed by the same sequential-double-auction market. This example is plotted for the continuous-time limit of the double-auction market.

In summary, because of strategic bidding behavior and imperfect competition, the sequential-double-auction market is slow in reducing allocative inefficiencies. This point is well recognized in prior work, including the static models of Vives (2011), Rostek and Weretka (2012), and Ausubel, Cramton, Pycia, Rostek, and Weretka (2014), as well as the dynamic models of Vayanos (1999), Rostek and Weretka (2015), and Du and Zhu (2015).

Figure 1 illustrates the time paths of expected inventories of a buyer and a seller for a parametric case of our sequential-double-auction market that we present later in the paper. The two thin-line plotted curves in Figure 1 illustrate the convergence over time of the expected inventories to those of the efficient allocation.

Now, consider an alternative market design in which traders have the opportunity to conduct a size-discovery session, say a workup, before the first double auction. For simplicity of exposition, we first solve the equilibrium for the special case of bilateral workups. Any active bilateral workup session involves a trader with a negative inventory imbalance, the "buyer," and a trader with a positive inventory imbalance, the "seller." Any trader who does not enter a workup participates only in the subsequent double auctions.

The fixed workup price is set at some given level. (We show that our efficiency results are robust to the choice of workup price.) As mentioned, the quantity to be exchanged between the buyer and seller in the workup is raised continually until one of the two traders drops out. That dropout quantity of the asset is then transferred from the seller to the buyer at the fixed workup price. Because the workup price is fixed, neither the buyer nor the seller is concerned about price impact. They are therefore able to exchange a potentially large block of the asset immediately, leading to a significant reduction in the total cost of maintaining undesired inventory over time.

Each workup participant recognizes that the workup process reveals information about his own inventory that is adverse to his interests. In equilibrium, we show that this effect inhibits traders with small inventory imbalances from entering workup. For the same reason, during a workup session a "large trader" continues to work up the size of the trade only until his inventory imbalance falls to some interior endogenous threshold that is determined by two countervailing incentives. On one hand, each trader wishes to minimize the unwanted inventory that is carried into the sequential-double-auction market. These leftover inventories take time to optimally liquidate, involve price-impact costs, and in the meantime are accompanied by holding costs. On the other hand, each trader in a workup faces adverse selection caused by asymmetric information concerning inventory levels. For example, if the buyer's offer to trade an additional unit is accepted, the buyer will have learned that the seller has more to sell than had been expected. The buyer would in that case have missed the chance to buy that unit in a subsequent double auction at a price whose conditional expectation is lowered by the seller's agreement to continue the workup. This risk to the buyer of being adversely selected for trade implies that, at some point in the workup (if the seller has not already dropped out), the buyer should withhold the next additional unit from the workup and reserve it for execution in the double-auction market at a more favorable conditional expected price. In equilibrium, these two effects determine a unique inventory threshold for dropping out of workup, which we calculate explicitly.

The two thick lines in Figure 1 illustrate the welfare-improving effect of augmenting the market design with an initial workup, which causes an instant reduction in inventory imbalances. In a simple parametric setting examined later in the paper, we show that buyers and sellers participating in bilateral workup eliminate between 27.6% and 62.7% of the inefficiency costs they bear from the effect of imperfect competition and the avoidance of price impact.

Comparative statics reveal that workup accounts for a larger fraction of total trade volume if the double auctions are run more frequently, if the arrival of payoff-relevant information is less imminent, or if there are fewer traders in the market. Under any of these conditions, traders are more sensitive to price impact because they will liquidate their inventories more gradually, implying higher inventory holding costs. These conditions therefore increase the welfare improvement allowed by workup.

The same intuition applies in a multilateral workup, in which many buyers and sellers taking turns to trade at the same fixed price. The multilateral workup session begins (if at all) with a workup between the first buyer and seller in their respective queues. Eventually, either the currently active buyer or seller drops out. If, for example, the seller is the first to drop out, it is then revealed whether there is at least one more trader remaining in the seller's queue, and if so whether that seller wishes to continue selling the asset at the same price. Based on this information, the buyer may continue the workup or may choose to drop out and be replaced by another buyer, if there is one, and so on. This process continues until there are no more buyers or no more sellers, whichever happens first. The equilibrium is solved in terms of the dropout threshold for the remaining inventory of an active workup participant, which is updated as each successive counterparty drops out and is replaced with a new counterparty. For example, when a new seller arrives and begins to actively increase the workup quantity, the current buyer's conditional expectation of the total market-wide supply of the asset jumps up, and this causes the buyer's dropout threshold to jump up at the same time by an amount that we compute and that depends on the history of prior workup observations. That is, with the arrival of a new active replacement seller, the buyer infers that the conditional expected double-auction price has become more favorable, and holds back more inventory from the workup, reserving a greater fraction of its trading interest for the double-auction market.

For tractability reasons, our work does not address the endogenous timing of size-discovery trading. Indeed, we are able to solve for equilibria with only an initializing round of size discovery. In practice, size discovery occurs with intermittent timing, presumably whenever position imbalances are sufficiently large on both sides of the market. Our equilibrium solution methods, however, rely on parametric assumptions for size-discovery prices and for the probability distribution of inventory levels entering into size discovery. Replacing these parametric initial conditions with endogenously determined size-discovery conditions is intractable in our framework, and we know of no tractable approaches for a useful equilibrium analysis of intermediate-timed size discovery.

This research is positive rather than normative. Size discovery has existed in Treasury and equity markets for decades. More recently, trade platform operators have introduced size-discovery mechanisms for corporate bonds, CDS, and interest rate swaps. We analyze the allocative effect of this augmentation of conventional price-discovery markets by solving the associated equilibrium and showing how this leads to a Pareto improvement. Traders who execute a positive quantity in size discovery strictly benefit from it; traders who participate only in the price-discovery market are not harmed by the use by others of size-discovery mechanisms. The efficiency improvement is greater to the degree that the initial allocation is worse. That is, size discovery is most helpful when order imbalances are highly concentrated among buyers and sellers. We show, however, that size-discovery mechanisms do not achieve first-best allocations. For example, traders drop out of workups or matching sessions prematurely from a social-welfare viewpoint, based on their equilibrium inference of expected future pricing advantages that are merely transfers.

An alternative research goal would be a normative design of the optimal dynamic mechanism for asset allocation, subject to incentive compatibility and budget balancing. If the inventory-allocation problem were static, a first-best allocation could, under conditions, be achieved by the "AGV" mechanism[1] of Arrow (1979) and d'Aspremont and Gérard-Varet (1979). In a dynamic market with imperfect competition and the stochastic arrival of new inventory shocks, static mechanisms such as AGV are no longer optimal. Solving for an optimal dynamic mechanism is difficult, and well beyond the scope of this paper.[2]

In short, while one could analyze many alternative mechanisms for treating inefficiencies caused by the strategic avoidance of price impact in the presence of large position imbalances, we focus our attention on the equilibrium properties of a class of size-discovery mechanisms that has, in practice, been widely chosen and extremely active.

As far as we are aware, our paper is the first to explicitly model how a size-discovery mechanism reduces allocative inefficiency caused by strategic demand reduction in price-discovery markets. We are also the first to solve for equilibrium behavior in multilateral workup and matching-session markets.

The only prior theoretical treatment of workup, to our knowledge, is by Pancs (2014), who focuses on the entirely different issue of "front-running." In the workup modeled by Pancs (2014), the seller has private information about the size of his desired trade ("block"), whereas the buyer is either a "front-runner" or a dealer. If the seller cannot sell his entire position in the workup, he liquidates the remaining positions to an exogenously given outside demand curve. At any point during the workup, the front-runner may decide to front-run the seller in the same outside demand curve. A dealer does not front-run by assumption. Under parameter conditions, Pancs (2014) characterizes an equilibrium in which each step of the workup transacts the smallest possible incremental quantity. This equilibrium minimizes the front-runner's payoff since it reveals as little information about the seller's block as possible. The key idea of our paper—that by freezing the price, workup mitigates strategic avoidance of price impact in price-discovery markets—is not

[1]In a side communication, Romans Pancs has shown us the explicit AGV mechanism for a simple variant of our model, based on *iid* original inventory positions and the assumption of no subsequent re-trade opportunities. In the Bayes-Nash equilibrium induced by this direct mechanism, each agent truthfully reports his original excess inventory as his type. Agents are assigned balanced-budget payments, based on their reported types.

[2]In a conversation, Bruno Biais suggested the mechanism-design problem of re-allocating inventory at a given point in time, taking as given the subsequent double-auction market.

considered by Pancs (2014).

Block-trading dark pools used in equity markets are also size-discovery mechanisms. The small and parallel literature on dark pools focuses instead on the effect of dark trading on price discovery and liquidity. Relevant papers include Hendershott and Mendelson (2000), Degryse, Van Achter, and Wuyts (2009), Zhu (2014), and Buti, Rindi, and Werner (2015), among others. In these models, each investor's trading interest is one unit, two units, or an infinitesimal amount. By characterizing allocative efficiency in the presence of arbitrarily large trading interests, our model goes substantially beyond existing research on the role of dark pools.

Empirical analyses of workup include those of Boni and Leach (2002, 2004), Dungey and McKenzie (2013), Fleming and Nguyen (2015), and Huang, Cai, and Wang (2002). Empirical studies of dark pools include Buti, Rindi, and Werner (2011), Ready (2014), and Menkveld, Yueshen, and Zhu (2015), among many others.

2 Size Discovery in Practice

In current market practice, size discovery shows up most prominently in three forms of trade mechanisms: workups, matching sessions, and block-crossing dark pools. This section summarizes the institutional settings of these respective mechanisms.

Workup was introduced in the last decades of the 20th century by interdealer voice brokers for U.S. Treasury securities, and is now heavily used on platforms for the electronic trading of Treasuries. The most active of these platforms are BrokerTec and eSpeed. On BrokerTec, for example, workup is fully integrated with central limit order book trading. Once a trade is executed on the limit order book at some price p, a workup session is opened for potential additional trading at the same price. The original buyer and seller and other platform participants may submit additional buy and sell orders that are executed by time priority at this workup price. Trade on the central limit order book is meanwhile suspended. The workup session ends if either *(i)* the workup session has been idle for some specified amount of time, which has been successively reduced in recent years and is now three seconds, or *(ii)* a new aggressive limit order arrives that cannot be matched immediately at the workup price p but can be matched immediately against a standing limit order deeper in the book. In Case *(ii),* the new order establishes a new price, at which point a new workup process may begin. In Case *(i)*, order submission on the limit order book resumes and continues until another limit-order-book trade is executed, kicking off another potential workup trade. This process repeats. A key feature is the integration of workup with the limit order book; when one of these two protocols is in process, the other is interrupted. For more details on BrokerTec's workup protocol, see Fleming and Nguyen (2015), Fleming, Schaumburg, and Yang (2015), and Schaumburg and Yang (2016).

Matching sessions use a trade protocol that is a close variant of workup, and appear most prominently on electronic platforms for trading corporate bonds[3] and credit default swaps (CDS). The markets for corporate bond and CDS are distinguished by much lower trade frequency than those for Treasuries and equities. Matching sessions, correspondingly, are less frequent and of longer duration. For example, matching sessions on Electronfie, a corporate bond trade platform, have a duration of 10 minutes.

A distinctive feature of matching sessions is that the fixed price is typically chosen by the platform operator. Given the incentives of the platform operator to maximize total trading fees, the fixed price seems likely to be designed to maximize expected trading volume. GFI, for example, chooses a matching-session price that is based, according to SIFMA (2016), on "GFI's own data (input from the internal feeds), TRACE data, and input from traders." On the CDS index trade platform operated by GFI, the matching price "shall be determined by the Company [GFI] in its discretion, but shall be between the best bid and best offer for such Swap that resides on the Order Book." Collin-Dufresne, Junge, and Trolle (2016) find that matching sessions or workup account for 38% of trade volume for the most popular CDS index product, known as CDX.NA.IG.5yr, a composite of 5-year CDS referencing 125 investment-grade firms, and 33% of trade volume for the

[3]According to SIFMA (2016), matching sessions are provided on the following corporate bond platforms: Codestreet Dealer Pool (pending release), Electronifie, GFI, Latium (operated by GFI Group), ICAP ISAM (pending release), ITG Posit FI, Liquidity Finance, and Tru Mid.

corresponding high-yield index product.

Trade platforms for interest-rate swaps also commonly incorporate workup or matching-session mechanisms, as described by BGC (2015), GFI (2015), Tradeweb (2014), and Tradition (2015). The importance of workup for the interest-rate swap market is discussed by Wholesale Markets Brokers' Association (2012) and Giancarlo (2015).

Block-trading dark pools operate in equity markets in parallel to stock exchanges, which are also referred to by market participants as "lit" venues. The dominant trade mechanism of stock exchanges is a central limit order book. Lit venues provide the latest bid-ask prices continuously. Dark pools match orders at a price between the most currently obtained bid and ask. Block-trading dark pools such as Liquidnet or POSIT typically use the midpoint of the prevailing bid-ask prices. Most dark pools operate continuously, in that buy and sell orders can be submitted anytime, and matching happens by time priority when both sides are available. When dark pools are executing orders, exchange trading continues. In current U.S. equity markets, only a few dark pools have execution sizes that are substantially larger than those on exchanges. Most dark pools have execution sizes similar to exchanges. For more details on dark-pool trading protocols, see Zhu (2014) and Ready (2014).

3 Dynamic Trading in Double Auctions

This section models dynamic trading in a flexible-price market consisting of a sequence of double auctions. Allocative inefficiency in dynamic double auction markets has already been shown by Vayanos (1999), Rostek and Weretka (2015), and Du and Zhu (2015).[4] This section merely reproduces the key thrust of their contributions in a simpler model. (We use a simple variant of the model of Du and Zhu (2015).) We claim no significant contribution here relative to these three cited papers. Our objective in this section is instead to set up a price-discovery market with imperfect competition as a benchmark. The rest of the paper then analyzes the effect of adding a size-discovery mechanism. Once

[4] Equilibrium models of static demand-schedule-submission games under imperfect competition include those of Wilson (1979), Klemperer and Meyer (1989), Kyle (1989), Vives (2011), and Rostek and Weretka (2012).

we have solved for equilibrium in this price-discovery market, the associated indirect utilities for pre-auction inventory imbalances serve as the terminal utility functions for the prior size-discovery stage, which is modeled in the next section.

We fix a probability space and the time domain $[0, \infty)$. Time 0 may be interpreted as the beginning of a trading day. The market is populated by $n \geq 3$ risk-neutral agents trading a divisible asset. The payoff π of the asset, a random variable with some finite mean v, will be revealed publicly at a random time T that is exponential with parameter r. Thus $E(T) = 1/r$. Everyone has symmetric information about π. At time T, the traders receive their payoffs from the assets and the economy ends.

The n traders' respective asset inventories at the beginning of the double-auction market are given by a vector $z_0 = (z_{10}, z_{20}, \ldots, z_{n0})$ of random variables that have non-zero finite variances. While the individual traders' inventories may be correlated with each other, there is independence among the asset payoff π, the revelation time T, and the vector z_0 of inventories.

At each nonnegative integer trading period $k \in \{0, 1, 2, \ldots\}$ a double auction is used to reallocate the asset. The trading periods are separated by some clock time $\Delta > 0$, so that the k-th auction is held at time $k\Delta$. As the first double auction begins, the information available to trader i includes the initial inventory z_{i0}, but does not include[5] the total inventory $Z_0 = \sum_i z_{i0}$. This allows that some traders may be better informed about Z_0 than others, and may have information about Z_0 going beyond their own respective inventories.

Right before auction $k + 1$, trader i receives an incremental inventory shock $w_{i,k+1}$. The random variables $\{w_{ik}\}$ are *i.i.d.* with mean zero and variance $\sigma_w^2 \Delta$.

At the k-th auction, trader i submits a continuous and strictly decreasing demand schedule. The information available to trader i at period k consists[6] of the trader's

[5]Fixing the underlying probability space $(\Omega, \mathcal{F}, P)$, trader i is endowed with information given by a sub-σ-algebra $\mathcal{F}_{i0}$ of $\mathcal{F}$. The inventory z_{i0} is measurable with respect to $\mathcal{F}_{i0}$, whereas the total inventory Z has a non-zero variance conditional given $\mathcal{F}_{i0}$.

[6]That is, the σ-algebra with respect to which the demand schedule of trader i in the k-th auction must be measurable is the join of the initial σ-algebra $\mathcal{F}_{i0}$ and the σ-algebra generated by $\{p_0, \ldots, p_{k-1}\}$, $\{z_{i1}, \ldots, z_{ik}\}$, and $\{w_{i1}, \ldots, w_{ik}\}$.

initial information, the sequence $p_0, \ldots, p_{k-1}$ of prices observed in prior auctions, as well as the trader's current and lagged inventories, $z_{i0}, \ldots, z_{ik}$. We focus on equilibria in which the demand schedule chosen by trader i optimally depends only on the trader's current pre-auction inventory z_{ik}. That is, trader i submits a demand schedule of the form $x_{ik}(\,\cdot\,; z_{ik}) : \mathbb{R} \to \mathbb{R}$, which is an agreement to buy $x_{ik}(p_k; z_{ik})$ units of the asset at the unique market-clearing price p_k. Whenever it exists, this market clearing price p_k is defined by

$$\sum_i x_{ik}(p_k; z_{ik}) = 0. \tag{1}$$

The inventory of trader i thus satisfies the dynamic equation

$$z_{i,k+1} = z_{ik} + x_{ik}(p_k; z_{ik}) + w_{i,k+1}. \tag{2}$$

The total inventory in the market right before auction k is $Z_k = \sum_i z_{ik}$. The periodic inventory shocks make it impossible to perfectly infer the current total inventory from past prices. Hence, the double-auction game always has incomplete information.

This double-auction mechanism is typical of those used at the open and close of the day on equity exchanges.[7] The double-auction model captures the basic implications of a flexible-price market in which traders are rational and internalize the equilibrium price impacts of their own trades. In practice, participants in a multi-unit auction submit a package of limit orders rather than a demand function. An arbitrary continuous demand function can be well approximated with a large number of limit orders at closely spaced limit prices.

When choosing a demand schedule in period k, each trader maximizes his conditional mean of the sum of two contributions to his final net payoff. The first contribution is trading profit, which is the final payoff of the position held when π is revealed at time T, net of the total purchase cost of the asset in the prior double auctions. The second contribution is a holding cost for inventory. The cost per unit of time of holding q units of inventory is γq^2, for a coefficient $\gamma > 0$ that reflects the costs to the trader of holding

[7]See, for example, http://www.nasdaqtrader.com/content/ProductsServices/Trading/Crosses/fact_sheet.pdf.

risky inventory.[8] For simplicity, we normalize time discounting. This is a reasonable approximation for trader inventory management in practice, at least if market interest rates are not extremely high, because traders lay off excess inventories over relatively short time periods, typically measured in hours or days.

In summary, for given demand schedules $x_{i1}(\cdot), x_{i2}(\cdot), \ldots$, the ultimate net payoff to be achieved by trader i, beginning at period k, is

$$U_{ik} = \pi z_{i,K(T)} - \sum_{j=k}^{K(T)} p_j x_{ij}(p_j; z_{ij}) - \int_{k\Delta}^{T} \gamma z_{i,K(t)}^2 \, dt, \tag{3}$$

where $K(t) = \max\{k : k\Delta \leq t\}$ denotes the number of the last trading period before time t. For given demand schedules, the continuation utility of trader i at the k-th auction, provided it is held before the time T at which the asset payoff is realized, is thus

$$V_{ik} = E\left(U_{ik} \;\middle|\; \mathcal{F}_{ik}\right), \tag{4}$$

where $\mathcal{F}_{ik}$ represents the information of trader i just before the k-th auction.

Based on this calculation, the continuation utility of trader i given by (4) satisfies the recursion

$$V_{ik} = -x_{ik}p_k - \gamma\eta(x_{ik} + z_{ik})^2 + (1 - e^{-r\Delta})(x_{ik} + z_{ik})v + e^{-r\Delta}E(V_{i,k+1} \,|\, \mathcal{F}_{ik}), \tag{5}$$

where we have used the shorthand x_{ik} for $x_{ik}(p_k; z_{ik})$, and where η is the expected duration of time from a given auction (conditional on the event that the auction is before T) until the earlier of the next auction time and the payoff time T:

$$\eta = \int_0^{\Delta} rt\, e^{-rt}\, dt + e^{-r\Delta}\Delta = \frac{1 - e^{-r\Delta}}{r}. \tag{6}$$

[8] Even though they do not have direct aversion to risk, broker-dealers and asset-management firms have extra costs for holding inventory in illiquid risky assets. These costs may be related to regulatory capital requirements, collateral requirements, financing costs, agency costs related to the lack of transparency of the position to higher-level firm managers or clients regarding true asset quality, as well as the expected cost of being forced to raise liquidity by quickly disposing of remaining inventory into an illiquid market. Our quadratic holding-cost assumption is common in models of divisible auctions, including those of Vives (2011), Rostek and Weretka (2012), and Du and Zhu (2015).

The four terms on the right-hand side of (5) represent, respectively, the payment made in the k-th double auction, the expected inventory cost to be incurred in the subsequent period (or until the asset payoff is realized), the expectation of any asset payment to be made in the next period multiplied by the probability that T is before the next auction, and the conditional expected continuation utility in period $k+1$ multiplied by the probability that T is after the next auction.

Each trader takes the total of the demand functions of the other traders as given. By the Bellman principle of dynamic optimality, in each period k, trader i is conjectured to submit a demand schedule $x_{ik}(\,\cdot\,; z_{ik})$ that maximizes the right-hand side of (5) subject to the dynamic equation (2), when taking $V_{i,k+1} = \mathcal{V}(z_{i,k+1})$, for some indirect inventory value function $\mathcal{V} : \mathbb{R} \to \mathbb{R}$. This indirect inventory value function $\mathcal{V}(\,\cdot\,)$ must satisfy (5) for this maximizing demand function, when taking $V_{ik} = \mathcal{V}(z_{ik})$. The equilibrium given by the following proposition is established by a standard Bellman verification argument, which implies the optimality of the corresponding demand functions. For this, we use the fact that, for the candidate optimal demand schedule, $e^{-rk\Delta}\mathcal{V}(z_{ik})$ converges to 0 as k goes to infinity, and the fact that $\limsup_{k\to\infty} e^{-rk\Delta}\mathcal{V}(z_{ik}) \leq 0$ for any feasible demand strategy. Details are in Appendix A.

Proposition 1. *In the game associated with the sequence of double auctions, there exists a stationary and subgame perfect equilibrium, in which the demand schedule of trader i in the k-th auction is given by*

$$x_{ik}(p; z_{ik}) = a_\Delta \left(v - p - \frac{2\gamma}{r} z_{ik} \right), \tag{7}$$

where

$$a_\Delta = \frac{r}{2\gamma} \frac{2(n-2)}{(n-1) + \frac{2e^{-r\Delta}}{1-e^{-r\Delta}} + \sqrt{(n-1)^2 + \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}}}. \tag{8}$$

The equilibrium price in auction k is

$$p_k = v - \frac{2\gamma}{nr} Z_k. \tag{9}$$

The bidding strategies of this equilibrium are ex-post *optimal with respect to all realizations of inventory histories. That is, trader j would not strictly benefit by deviating from the equilibrium strategy even if he were able to observe the history of other traders' inventories,* $\{z_{im} : i \neq j, m \leq k\}$.

The ex-post optimality property of the equilibrium arises from the fact that each trader's marginal indirect value for additional units of the asset depends only on his own current inventory, and not on the inventories of other traders. This property will be useful in solving the workup equilibrium.

The slope a_Δ of the equilibrium supply schedule is increasing in Δ. That is, trading is more aggressive if double auctions are conducted at a lower frequency. We also have

$$\lim_{\Delta \to \infty} a_\Delta = \frac{r(n-2)}{2\gamma(n-1)} < \frac{r}{2\gamma}. \tag{10}$$

Moreover, as Δ goes to 0, a_Δ converges to 0.

The market-clearing price p_k reveals the total inventory Z_k at the moment of the k-th auction. Because the total inventory process $\{Z_0, Z_1, Z_2, \ldots\}$ is a martingale, the price process $\{p_k\}$ is also a martingale.

Although traders have symmetric information about the asset fundamental, uncertainty about the total inventory Z_k generates uncertainty about the market-clearing price. As we will see in the next section, uncertainty over the initial inventory Z_0 is an important determinant of the optimal strategy in the workup stage of the model.

By symmetry and the linearly decreasing nature of marginal values, the efficient allocation immediately assigns each trader the average inventory Z_k/n. The double-auction market, however, merely moves the allocation toward this equal distribution of the asset. Specifically, by substitution, we have

$$x_{ik} = a_\Delta \left(v - p_k - \frac{2\gamma}{r} z_{ik} \right) = -a_\Delta \frac{2\gamma}{r} \left(z_{ik} - \frac{Z_k}{n} \right), \tag{11}$$

$$z_{i,k+1} = z_{ik} + x_{ik} + w_{i,k+1} = z_{ik} - a_\Delta \frac{2\gamma}{r} \left(z_{ik} - \frac{Z_k}{n} \right) + w_{i,k+1}. \tag{12}$$

At auction k, the equilibrium trade x_{ik} eliminates only a fraction $\varphi = a_\Delta 2\gamma/r$ of the "excess inventory" $z_{ik} - Z_{ik}/n$ of trader i. This partial and inefficient liquidation of unwanted inventory is caused by imperfect competition. From (10), we have $\varphi \leq (n-2)/(n-1)$, and this bound is achieved in the limit as $\Delta \to \infty$. As $\Delta \to 0$, we have $a_\Delta \to 0$, and the fractional reduction φ of the mis-allocation of inventory converges to zero.

Since our ultimate objective is to characterize the workup strategy at time 0, we spell out the continuation value of each trader, evaluated at time 0, in the following proposition.

Proposition 2. *Let $V_{i,0+} = E(U_{i0} \mid z_{i0}, p_0)$ denote the initial utility of trader i, evaluated at time 0 after conditioning on the initial market-clearing price p_0, which reveals the initial total inventory $Z \equiv Z_0$. We have:*

$$V_{i,0+} = \left[v\frac{Z}{n} - \frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 \right] + \left(v - 2\frac{\gamma}{r}\frac{Z}{n}\right)\left(z_{i0} - \frac{Z}{n}\right) - \frac{\gamma}{r}\frac{1 - 2a_\Delta\frac{\gamma}{r}}{n-1}\left(z_{i0} - \frac{Z}{n}\right)^2 + \Theta, \tag{13}$$

where $\Theta < 0$ is a constant whose expression is provided in Appendix A.2.

The first term of (13) is the total utility of trader i in the event that trader i already holds the initial efficient allocation Z/n. The second term of (13) is the amount that could be hypothetically received by trader i for immediately selling the entire excess inventory, $z_{i0} - Z/n$, at the market-clearing price, $v - 2\gamma Z/(rn)$. But this immediate beneficial reallocation of the asset does not actually occur because traders strategically shade their bids to account for the price impact of their orders. This price-impact-induced drag on each trader's expected ultimate net payoff, or "utility," is captured by the third term of (13), which is the utility loss caused by the fact that the demand schedule of trader i in each auction is decreasing in a_Δ. The constant Θ captures the additional allocative inefficiency caused by periodic inventory shocks. (If $\sigma_w^2 = 0$, then $\Theta = 0$.) The loss of welfare associated with the initial inventory allocation is proportional to $\sum_i (z_{i0} - Z_0/n)^2$, a natural welfare metric formalized in Appendix C.

Again, because a_Δ is always smaller than $r/(2\gamma)$, full efficiency cannot be achieved.

Moreover, because a_Δ is increasing in Δ, each trader's utility loss gets larger as Δ gets smaller. The basic intuition is as follows. (See Du and Zhu (2015) for a detailed discussion.) Although a smaller Δ gives traders more opportunities to trade, they are also strictly less aggressive in each trading round. A smaller Δ makes allocations less efficient in early rounds but more efficient in late rounds. Traders value early-round efficiency more because of the effective "time discounting" $e^{-r\Delta}$. The net effect is that allocative efficiency is worse if Δ is smaller.

Appendix B provides the continuous-time limit of the discrete-time double auction model, and shows that this limit matches the equilibrium of the continuous-time version of the double-auction model.

4 Introducing Workup for Size Discovery

We saw in the previous section that successive rounds of double auctions move the inventories of the traders toward a common level. This reduction in inventory dispersion is only gradual, however, because at each round, each trader internalizes the price impact caused by his own quantity demands, and thus "shades" his demand schedule so as to trade off inventory holding costs against price impact.

We now examine the effect of introducing at time 0 a size-discovery mechanism, taken for concreteness to be a workup session, that gives traders the opportunity to reduce the magnitudes of their excess inventories without concern over price impact. It would be natural in practice to run a workup session whenever traders' inventories have been significantly disrupted. In the U.S. Treasury market, for example, primary dealers' positions can be suddenly pushed out of balance by unexpectedly large or small awards in a Treasury auction. Individual dealers' inventories could also be disrupted by large surges of demand or supply from their buy-side clients. We show that workup immediately reallocates a potentially large amount of inventory imbalances, which improves allocative efficiency relative to the double-auction market without a workup.

4.1 A model of bilateral workup

For expositional simplicity, we first consider a setting in which each of an arbitrary number of bilateral workup sessions is conducted between an exogenously matched pair of traders, one with negative inventory, "the buyer," and one with positive inventory, "the seller." Any trader not participating in one of the bilateral workup sessions is active only in the subsequent double-auction market. Information held by a pair of workup participants regarding participation in other workup sessions plays no role in our model. That is, the equilibrium for the bilateral workup sessions and the subsequent double auctions is unaffected by information held by the participants in a given workup regarding how many other workup sessions are held and which traders are participating in them. For simplicity, we do not model the endogenous matching of workup partners.

In the next section, we generalize the model to cover a more realistic multilateral workup session.

In a given bilateral workup session, say that between traders 1 and 2, the seller has an initial inventory S^s that is exponentially distributed with parameter μ, thus having a mean of $1/\mu$. The initial inventory S^b of the buyer is negative, with a magnitude $|S^b|$ that independent of, and identically distributed with, the seller's initial inventory. For the purpose of characterizing the equilibrium in a bilateral workup, we only need to assume that the sum of the initial inventories $z_{30}, \ldots, z_{n0}$ of the other $n-2$ traders has mean zero and is independent of S^s and S^b. But for symmetry and a more convenient quantitative assessment of the welfare improvement brought by the bilateral workup, we will additionally assume that $z_{30}, \ldots, z_{n0}$ are *iid* with the density function

$$f(z) = \frac{1}{2}\mu e^{-\mu|z|}, \quad z \in (-\infty, \infty). \tag{14}$$

Thus, everyone has identical second moment of inventory: $E[(S^b)^2] = E[(S^s)^2] = E[z_{i0}^2] = 2/\mu^2$ for all $i \geq 3$.

The workup price $\bar{p}$ is set without the use of information about traders' privately observed inventories, and therefore at some deterministic level $\bar{p}$. We will provide an interval of choices for $\bar{p}$ that is necessary and sufficient for interior equilibrium workup dropout policies. We will also show that the allocative efficiency improvement of workup is invariant to changes in the workup price $\bar{p}$ within this interval. A natural choice for $\bar{p}$ is the unconditional expectation of the asset payoff v, which can be interpreted as the expectation of the clearing price in the subsequent double-auction market, or as the price achieved in a previous round of auction-based trade, before new inventory shocks instigate a desire by traders to lay off their new unwanted inventories.

After each of a given pair of participants in a workup privately observes his own inventory, the workup proceeds in steps as follows:

1. The workup operator announces the workup price $\bar{p}$.

2. The workup operator provides a continual display, observable to buyer and seller, of the quantity $Q(t)$ of the asset that has been exchanged in the workup by time t on the workup "clock." The units of time on the workup clock are arbitrary, and the function $Q(\,\cdot\,)$ is any strictly increasing continuous function satisfying $Q(0) = 0$ and $\lim_{t\to\infty} Q(t) = \infty$. For example, we can take $Q(t) = t$. The workup clock can run arbitrarily quickly, so workup can take essentially no time to complete. This mechanism is essentially the "button mechanism" described in Pancs (2014).

3. At any finite time T_b on the workup clock, or equivalently at any quantity $\mathcal{Q}_b = Q(T_b)$, the buyer can drop out of the workup. Likewise, the seller can drop out at any time T_s or quantity $\mathcal{Q}_s = Q(T_s)$. The workup stops at time $T^* = \min(T_s, T_b)$, at which the quantity $Q^* = Q(T^*) = \min(\mathcal{Q}_b, \mathcal{Q}_s)$ is transferred from seller to buyer at the workup price $\bar{p}$, that is, for the total consideration $\bar{p}\,Q^*$.

After the bilateral workups terminate, all traders enter the sequence of double auctions described in Section 3.

As mentioned in the introduction, the workup procedure modeled here is similar to the matching mechanism used by certain dark pools, such as POSIT and Liquidnet, that specialize in executing large equity orders from institutional investors. In a dark-pool transaction with one buyer and one seller, each side privately submits a desired trade size to the dark pool, understanding that the dark pool would execute a trade for the minimum of the buyer's and seller's desired quantities. In a bilateral setting, workup and dark-pool matching are thus equivalent.

4.2 Characterizing the workup equilibrium

This section characterizes the equilibrium behavior of the two traders in a given bilateral workup session.

Any trader's strategy in the subsequent double-auction market, solved in Proposition 1, depends only on that trader's inventory level. Thus any public reporting, to all n traders, of the workup transaction volume Q^* plays no role in the subsequent double-auction analysis.

We conjecture the following equilibrium workup strategies. The buyer and seller fix deterministic thresholds, M_b and M_s respectively, for the magnitudes of residual inventory at which they drop out of workup. That is, the buyer allows the workup transaction size to increase until the time T_b at which his residual inventory $|S^b + Q(T_b)|$ is equal to M_b. The seller likewise chooses a dropout time T_s at which his residual inventory $S^s - Q(T_s)$ reaches M_s. One trader's dropout is of course pre-empted by the other's. A *threshold equilibrium* is a pair $(M_b, M_s) \in \mathbb{R}^2_+$ with the property that M_b maximizes the conditional expected payoff of buyer given the seller's threshold M_s and conditional on the buyer's inventory S^b, and vice versa. We emphasize that, given M_s, the buyer is not restricted to a deterministic threshold, and vice versa.

The Dropout Strategies. For any $y > 0$, let F_y be the event that the buyer's candidate requested quantity $y > 0$ is filled. That is,

$$F_y = \left\{ 0 \leq -(S^b + y) - M_b, 0 \leq S^s - y - M_s \right\}. \tag{15}$$

The remaining inventory of the buyer, $-(S^b + y)$, is weakly larger than the dropout quantity M_b, for otherwise the buyer would have already dropped out. Similarly, the remaining inventory of the seller, $S^s - y$, is weakly larger than his dropout quantity M_s, for otherwise the seller would have already dropped out.

With the conjectured equilibrium dropout strategies, the memoryless property of the exponential distribution implies that, for the buyer, the seller's inventory in excess of the dropout quantity, which is $W \equiv S^s - y - M_s$, is F_y-conditionally exponential with the same parameter μ. Thus, recalling that Z is the aggregate inventory of the traders, we have

$$E(Z \mid F_y, S^b) = S^b + y + M_s + \frac{1}{\mu}, \tag{16}$$

using the fact that the expected total inventory of all traders not participating in this workup is zero.

By a similar calculation,

$$\begin{aligned} E(Z^2 \mid F_y, S^b) &= E\left[(S^b + y + M_s + W)^2 + \left(\sum_{i=3}^{n} z_{i0} \right)^2 \right] \\ &= (S^b + y + M_s)^2 + E(W^2) + 2(S^b + y + M_s)E(W) + \theta \\ &= (S^b + y + M_s)^2 + \frac{2}{\mu^2} + 2(S^b + y + M_s)\frac{1}{\mu} + \theta, \end{aligned} \tag{17}$$

where

$$\theta = E\left[\left(\sum_{i=3}^{n} z_{i0} \right)^2 \right].$$

On the other hand, given the initial inventory S^b and the candidate quantity $y \geq 0$ to be acquired in the workup, the buyer's conditional expected ultimate value, given $\{Z, S^b\}$, is

$$U^b = -\bar{p}y + \mathcal{V}(S^b + y), \tag{18}$$

where, based on Proposition 2,

$$\mathcal{V}(z) = v\frac{Z}{n} - \frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 + \left(v - 2\frac{\gamma}{r}\frac{Z}{n}\right)\left(z - \frac{Z}{n}\right) - \frac{\gamma}{r}\frac{1 - 2a_{\Delta}\gamma/r}{n-1}\left(z - \frac{Z}{n}\right)^2. \tag{19}$$

Organizing the terms, we get

$$E(U^b \mid F_y, S^b) = -\bar{p}y + v(S^b + y) - \frac{\gamma}{r}C(S^b + y)^2 + 2\frac{\gamma}{r}\left(C - 1\right)\left(S^b + y\right)\frac{E(Z \mid F_y, S^b)}{n}$$
$$- \frac{\gamma}{r}\left(C - 1\right)\frac{E(Z^2 \mid F_y, S^b)}{n^2}, \tag{20}$$

where

$$C = \frac{1 - 2a_\Delta \gamma / r}{n - 1}. \tag{21}$$

Substituting the expressions that we have shown above for $E(Z \mid F_y, S^b)$ and $E(Z^2 \mid F_y, S^b)$ into this expression for $E(U^b \mid F_y)$, we get

$$g(y) \equiv \frac{dE(U^b \mid F_y, S^b)}{dy} = v - \bar{p} - 2\frac{\gamma}{r}C(S^b + y) + 2\frac{\gamma}{r}(C - 1)\frac{1}{n}\left(2(S^b + y) + M_s + \frac{1}{\mu}\right)$$
$$- \frac{\gamma}{r}(C - 1)\frac{1}{n^2}\left(2(S^b + y + M_s) + \frac{2}{\mu}\right). \tag{22}$$

The derivative $g(y)$ is everywhere strictly decreasing in y. Following the conjectured equilibrium, an optimal dropout quantity M_b for the buyer's residual inventory, if the optimum is interior (which we assume for now and then validate), is obtained at a level of y for which this derivative $g(y)$ is equal to zero, and by taking $S^b + y = -M_b$. That is,

$$0 = v - \bar{p} - 2\frac{\gamma}{r}C(-M_b) + 2\frac{\gamma}{r}(C - 1)\frac{1}{n}\left(2(-M_b) + M_s + \frac{1}{\mu}\right)$$
$$- \frac{\gamma}{r}(C - 1)\frac{1}{n^2}\left(2(-M_b + M_s) + \frac{2}{\mu}\right). \tag{23}$$

By completely analogous reasoning, the first-order conditions for the seller's optimal dropout threshold M_s is given by

$$0 = \bar{p} - v + 2\frac{\gamma}{r}CM_s + 2\frac{\gamma}{r}(C - 1)\frac{1}{n}\left(-2M_s + M_b + \frac{1}{\mu}\right)$$
$$- \frac{\gamma}{r}(C - 1)\frac{1}{n^2}\left(-2(M_s - M_b) + \frac{2}{\mu}\right). \tag{24}$$

The Equilibrium Dropout Thresholds. The unique solution to the two first-order necessary and sufficient conditions (23) and (24) is

$$M_b = \frac{n-1}{n + n^2 C/(1-C)} \frac{1}{\mu} + \delta = M + \delta, \tag{25}$$

$$M_s = \frac{n-1}{n + n^2 C/(1-C)} \frac{1}{\mu} - \delta = M - \delta, \tag{26}$$

where

$$M \equiv \frac{n-1}{n + n^2 C/(1-C)} \frac{1}{\mu}. \tag{27}$$

is the dropout quantity for the unbiased price $\bar{p} = v$, and where

$$\delta = \frac{r}{2\gamma} \frac{\bar{p} - v}{C + (1-C)(3n-2)/n^2}. \tag{28}$$

As we see later, the symmetric solution M plays an important role in calculating the comparative statics. To be consistent with the initial conjecture that the equilibrium is interior, that is, $M_b > 0$ and $M_s > 0$, the workup price $\bar{p}$ must have the property that $|\delta| \leq M$, or equivalently,

$$|\bar{p} - v| \leq \frac{2\gamma M[C + (1-C)(3n-2)/n^2]}{r}. \tag{29}$$

We will only treat prices satisfying this interior-dropout condition.

It is intuitive that a biased workup price causes asymmetric dropout behavior. If $\bar{p} > v$, the workup price is less favorable than the double auction price for the buyer, but more favorable for the seller. Thus, the buyer is more cautious than the seller in the workup, in that the buyer's dropout level is higher than the seller's. The opposite is true if $\bar{p} < v$.

This result is summarized by the following proposition.

Proposition 3. *Suppose that the workup price $\bar{p}$ satisfies (29). The workup session has a unique equilibrium in deterministic dropout-inventory strategies. The buyer's and seller's dropout levels, M_b and M_s, for residual inventory are given by (25) and (26), respectively. That is, in equilibrium, the buyer and seller allow the workup quantity to increase until* *the magnitude of their residual inventories reach M_b or M_s, respectively, or until the other trader has dropped out, whichever comes first.*

Figure 2 illustrates the impact of the workup on the undesired inventory levels of two traders. In this simple example, there are $n = 5$ traders and one bilateral workup session. The workup price is $\bar{p} = v$. The two workup participants have mean inventory size $1/\mu = 1$. We calculate the equilibrium outcome of the workup when the outcome of the workup buyer's pre-workup inventory S^b is -2, the outcome of the workup seller's pre-workup inventory S^s is 1.5, and the outcomes of all of the other traders' initial inventories are zero. The outcome for the efficient allocation of all traders is $Z/n = -0.1$. We focus on the continuous-time sequential-double-auction market. The equilibrium workup dropout quantity is in this case $M = 0.3$. Because we have the outcome that $|S^b| > |S^s|$, the seller exits the workup first, after executing the quantity $1.5 - 0.3 = 1.2$. The seller's inventory after the workup is 0.3, whereas the buyer's inventory after the workup is $-(2 - 1.2) = -0.8$.

Proposition 3 shows that as long as the workup price is not too biased, the two workup participants do not generally attempt to liquidate all of their inventories during the workup (in that $M_b > 0$ and $M_s > 0$). Their optimal target inventories are determined by two countervailing incentives. On one hand, because of the slow convergence of a trader's inventory to efficient levels during the subsequent double-auction market, each trader has an incentive to execute large block trades in the workup. On the other hand, a trader faces adverse selection regarding the total inventory Z and the double-auction prices. For example, if the buyer's expectation of the future auction price is lower than the workup price $\bar{p}$, the buyer would be better off buying some of the asset in the subsequent auction market, despite the associated price impact. This incentive encourages inefficient "self-rationing" in the workup. A symmetric argument holds for the seller. Depending on a trader's conditional expectation of the total market excess inventory Z, which changes as the workup progresses, the trader sets an endogenous dropout inventory threshold such that the two incentives are optimally balanced. In setting his optimal target inventory, a trader does not attempt to strategically manipulate the other trader's inference of the

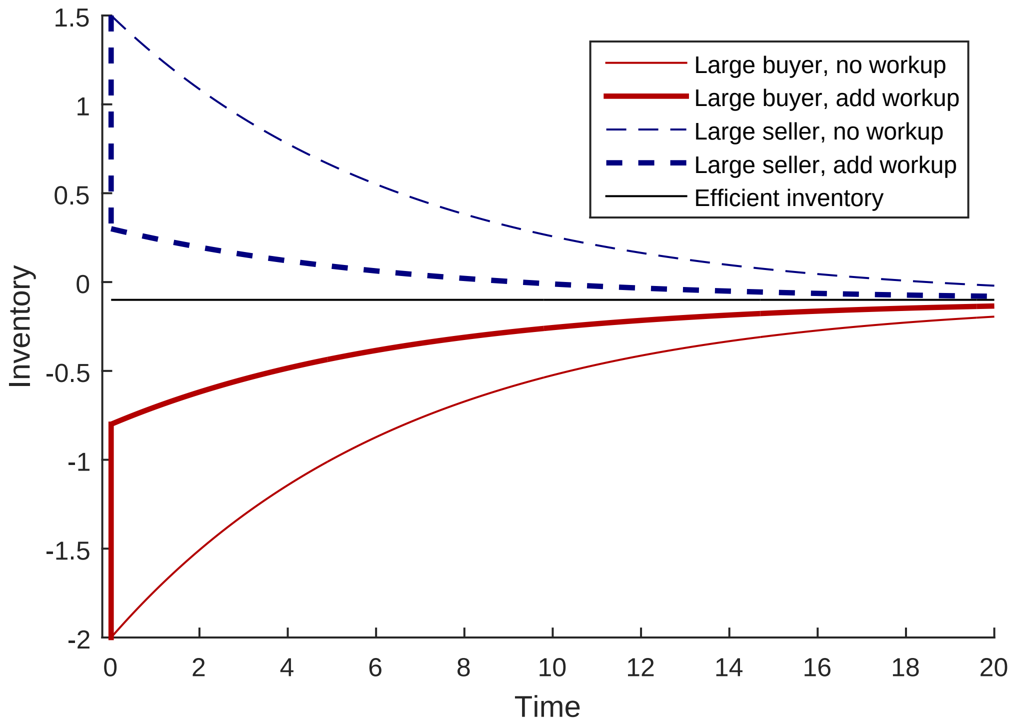

Figure 2: *Immediate inventory imbalance reduction by workup.* Parameters: $n = 5$, $\mu = 1$, $r = 0.1$, $\gamma = 0.05$, $\Delta = 0$, $S^b = -1.5$, $S^s = 2$. The outcomes of the inventories of traders not entering workup are zero.

total inventory Z, because optimal auction strategies do not depend on conditional beliefs about Z.

4.3 Properties of the workup equilibrium

We now describe additional properties of the workup equilibrium of Proposition 3.

Pareto improvement in efficiency. Although adding workups does not lead to the efficient allocation, it causes a sudden and beneficial reduction in inventory imbalances. Because workup participation is voluntary, traders can only increase their utilities (at least weakly) by participating in workups. Traders suffer no loss in expected net benefit (relative to a market without workup) from not participating in workups, as can be checked from (13). Thus, adding workup sessions is a Pareto improvement among all traders, and offers a strict improvement to those who, in expectation, execute a strictly positive quantity in a workup. There is generally a non-trivial welfare improvement associated with adding workups, measured by the increase in the sum of the total of trades' ex-ante expected net payoff.

Proposition 4. *Adding workup sessions before the sequential-double-auction market is a Pareto improvement. That is, every trader's ex-post utility is increased, and the ex-post utility of participating traders who, in expectation, execute strictly positive workup quantities is strictly increased, over that associated with a sequential-double-auction market that is not preceded by workup.*

Equilibrium outcomes. Now we discuss the outcomes of the bilateral workup equilibrium, including the probability of active workup participation, the expected trading volume, and welfare improvement between the buyer and the seller. We further show that these quantities are invariant to variation in the workup price $\bar{p}$ provided that $\bar{p}$ is in the interval satisfying condition (29).

The probability of triggering an active workup is

$$P(S^s > M_s, |S^b| > M_b) = e^{-\mu(M+\delta)}e^{-\mu(M-\delta)} = e^{-2\mu M}, \tag{30}$$

which is decreasing in M and does not depend on $\bar{p}$, within the range of interior solutions.

Moreover, by substituting in (27), we obtain:

$$P(S^s > M_s, |S^b| > M_b) = \exp\left(-\frac{2(n-1)}{n + n^2C/(1-C)}\right). \tag{31}$$

That is, the probability of having an active workup does not depend on the average inventory sizes in the market. This probability of active workup depends instead on the competitiveness of the double-auction market (which is captured by the number n of traders), the mean arrival rate r of price-relevant information, and the auction-market frequency $(1/\Delta)$. We will discuss these comparative statics in detail shortly.

Similarly, the expected trading volume in the workup is

$$\begin{aligned}
&E\left[\max\left(\min\left(|S^b| - M_b, S^s - M_s\right), 0\right)\right] \\
&\quad = \int_{x=M+\delta}^{\infty}\int_{y=M-\delta}^{\infty} \mu e^{-\mu x}\mu e^{-\mu y}\min(x-(M+\delta), y-(M-\delta))\, dx\, dy. \qquad (32)
\end{aligned}$$

By the change in variables $u = x - M - \delta$ and $w = y - M + \delta$, the integral can be re-expressed as

$$\begin{aligned}
&\int_{u=0}^{\infty}\int_{w=0}^{\infty} \mu e^{-\mu(u+M+\delta)}\mu e^{-\mu(w+M-\delta)}\min(u,w)\, du\, dw \\
&\quad = e^{-2\mu M}\int_{u=0}^{\infty}\int_{w=0}^{\infty} \mu e^{-\mu u}\mu e^{-\mu w}\min(u,w)\, du\, dw, \qquad (33)
\end{aligned}$$

which is decreasing in M and is invariant to δ in the interval $[0, M]$. Thus, within the range of workup prices at which dropout thresholds are interior, the expected workup trade volume does not depend on the workup price $\bar{p}$.

By a similar but slightly longer calculation, Appendix C shows that for $\delta \in [0, M]$, the fraction of the total inefficiency costs of the buyer and the seller that is eliminated by their participation in the bilateral workup is

$$R = \frac{n}{2(n-1)} e^{-2M\mu}(1 + M\mu). \qquad (34)$$

Because $e^{-2M\mu}(1 + M\mu)$ is decreasing in M, which in turn is increasing in n, this proportional cost reduction R decreases with the number n of market participants. That is, in terms of its relative effectiveness in eliminating inventory-cost inefficiencies caused by imperfect competition in price-discovery markets, workup is more valuable for markets with fewer participants.

For the continuous-time version of the double-auction market (or in the limit in Δ goes to zero), we have simply

$$R = \frac{3n-2}{4(n-1)} e^{-(n-2)/n}. \qquad (35)$$

For $n = 3$, this cost-reduction ratio is $R = 0.627$. As n gets large, $R \to 0.75e^{-1} = 0.276$. So, buyers and sellers participating in bilateral workup eliminate between 27.6% and 62.7% of the inefficiency costs caused by imperfect competition and avoidance of price impact. Figure 3 shows how R declines with the number n of market participants.

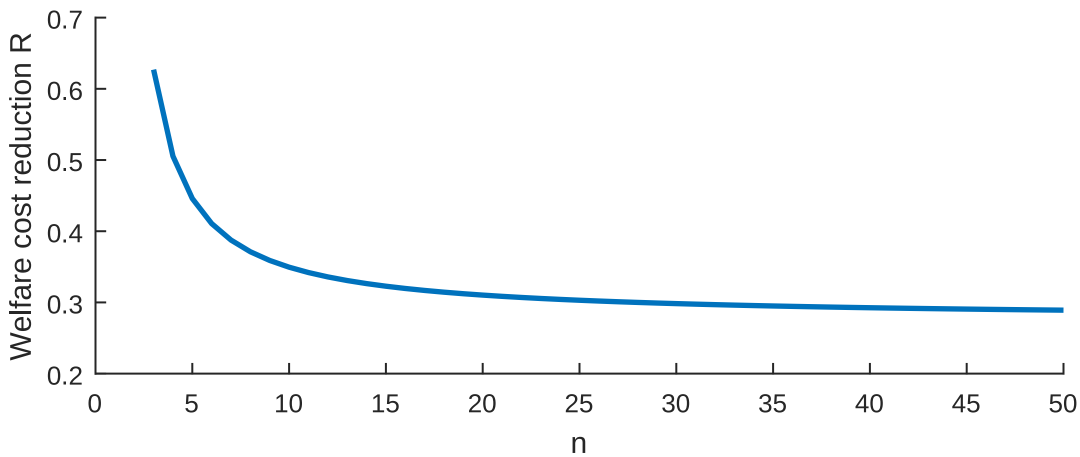

Figure 3: *The proportional welfare improvement of traders participating in workup.* The plot shows the fraction R of the total inefficiency cost of the buyer and the seller that is eliminated by their participation in bilateral workup.

Comparative statics. We have just shown that the probability of triggering an active workup and the expected workup trading volume are decreasing in the workup inventory dropout threshold M. Now, we discuss how M varies with respect to Δ, r, and n. These comparative statics reveal how the attractiveness of the size-discovery workup mechanism varies with market conditions.

We have

$$\zeta(\Delta) \equiv 1 - 2a_\Delta \frac{\gamma}{r} = \frac{\sqrt{(n-1)^2(1-e^{-r\Delta})^2 + 4e^{-r\Delta}} - (n-1)(1-e^{-r\Delta})}{2e^{-r\Delta}}. \tag{36}$$

By Proposition 2, $\zeta(\Delta)$ is the fraction of excess inventory that remains after each successive double auction. The smaller is this quantity, the more aggressive are traders' submitted demand schedules. The constant γ that scales the quadratic inventory cost does not in itself affect $\zeta(\Delta)$ or M. This is perhaps surprising, and applies in light of the fact that the aggressiveness of demand schedules fully offsets the effect of γ, causing $a_\Delta \gamma$ to be invariant to γ.

By calculation,

$$\zeta'(\Delta) = \frac{re^{r\Delta}(n-1)}{2} \left(\frac{\sqrt{(n-1)^2(e^{r\Delta}-1)^2 + 4e^{r\Delta} - 4\left(1 - \frac{1}{(n-1)^2}\right)}}{\sqrt{(n-1)^2(e^{r\Delta}-1)^2 + 4e^{r\Delta}}} - 1 \right) < 0. \quad (37)$$

Because M is decreasing in $C = \zeta(\Delta)/(n-1)$, M is therefore increasing in Δ. That is, the smaller is Δ (the more frequent the double auctions), the smaller is M, and the more active is workup. Intuitively, a small Δ discourages aggressive trading in the double auctions because traders have frequent opportunities to trade, which implies that the convergence to efficient inventory levels is slow. (This welfare cost of fast trading is also discussed by Vayanos (1999) and Du and Zhu (2015).) Workup is therefore more attractive to "large traders," that is, those tending to have large inventory imbalances, because workup can lead to more dramatic reductions in excess inventories, despite the adverse-selection effect of workup.

As the double auctions become more frequent, that is as Δ goes to zero, we know that $a_\Delta \to 0$ and thus $C \to 1/(n-1)$. In this case, M converges downward to the continual-auction limit

$$\frac{n-2}{2n}\frac{1}{\mu}, \quad (38)$$

and the probability of triggering a workup becomes maximal, at $\exp(-(n-2)/n)$. At this continuous-time limit, which is the same as the behavior of the corresponding continuous-auction model shown in Appendix B, the probability of triggering a workup decreases in n. Intuitively, the double auction market becomes more efficient as the number n of participants grows, getting closer and closer to price-taking competitive behavior. Hence, as n grows, there is less allocative benefit from size discovery. In fact, we can show that M increases with n regardless of the model parameters. (For details, see Appendix A.3.)

For example, in a market with $n = 20$ traders, if workup is preceded by a continuous-time auction market, the probability of active workup is $e^{-18/20} \approx 0.41$. With only $n = 5$

traders, this active-workup probability increases to $e^{-3/5} \approx 0.55$.

We also have

$$\frac{d}{dr}\left(1 - 2a_\Delta \frac{\gamma}{r}\right) < 0. \tag{39}$$

That is, the lower is the mean arrival rate of the asset payoff information, the smaller is M, and the more likely it is that an active workup is triggered. Intuitively, the more delayed is the final determination of asset payoffs, the less aggressive are traders in their double-auction demand schedules, which in turn increases the attractiveness of using workup to quickly reduce inventory imbalances.

These comparative statics are summarized by the following proposition.

Proposition 5. *All else equal, the probability of triggering a workup and the expected workup trade volume are higher (that is, M is lower) if:*

1. *The frequency of subsequent double auctions is higher (Δ is smaller).*
2. *The number n of traders is lower.*
3. *The mean arrival rate of asset payoff news r is lower.*

4.4 Empirical implications

The most direct empirical implication of our results is that size-discovery mechanisms are more likely to be active in markets with highly concentrated position imbalances, and especially in markets with relatively few active market participants. Workup and matching-session mechanisms exist in practice under just these circumstances. In particular, interdealer markets for Treasuries and swaps are populated by a small number of major dealers whose role as market makers is to take large positions from their customers and then to lay off net position imbalances with other major dealers, some of whom are likely to have large position imbalances of the opposite sign. Likewise, corporate bond and CDS markets tend to have very few active market participants at a given time, and position imbalances tend to be large relative to daily trade volume.

Within a given market, our results predict that the fraction of total trade volume that is executed by size discovery is increasing in the degree of large position imbalances. In Treasury markets, for example, our results suggest that workups would be particularly active immediately following a Treasury auction, in which some dealers are awarded much larger amounts of a newly issued security than the net trade requests of their clients, while other dealers are awarded much smaller positions than the net demand of their own clients.

For illustrative simplicity, consider the continuous-time version of our model in Appendix B without periodic inventory shocks. The total absolute quantity of position imbalances at a given time t is $D(t) = \sum_i |z_{it} - Z/n|$, where z_{it} is the position of dealer i at time t. Using (150), this aggregate position imbalance $D(t)$ can be calculated from the equilibrium (one-sided) trade volume

$$q(t) = \frac{1}{2} \sum_i \left| \frac{d}{dt} z_{it} \right| = \frac{(n-2)r}{4} D(t). \tag{40}$$

A slightly more complicated but similar linear expression applies in discrete time, or with random inventory shocks.

Because the incentive to conduct size discovery is (weakly) increasing in the dispersion measure $D(t)$, it follows from (40) that size discovery is more likely to be active when central-limit-order-book (CLOB) volume is high. Indeed, this is consistent with the evidence shown in Table 9 of Fleming and Nguyen (2015), that workup is more likely on BrokerTec if trade volume is higher. Likewise, in our model, expected workup volume is increasing in immediately subsequent CLOB trade volume. Fleming and Nguyen (2015) also find that workup is more likely to happen if the pre-workup CLOB has a higher depth and a lower spread, which is consistent with the rationale that wider dispersion of undesirable inventories creates two-way trading interests and better liquidity on CLOB. (Although their evidence is consistent with the prediction of our model for the case of a single round initial of workup, our model does not incorporate dynamic choice of workup.)

This positive correlation between workup activity and CLOB trade volume (and other measures suggesting better liquidity) might not be apparent from a partial-equilibrium perspective, by which traders would purportedly see less advantage in conducting size discovery whenever they note that they can take advantage of high CLOB volume to reduce their position imbalances quickly. Our equilibrium model reflects the correct causality, namely that high position imbalances induce higher volumes in *both* size-discovery and price-discovery mechanisms.

5 Multilateral Workups

In Section 4 we solved the equilibrium for bilateral workup sessions, and showed that workup provides size-discovery welfare benefits. This section extends our results to dynamic multilateral workups, which are more commonly used in practice, for example on electronic trading platforms. The intuition for the allocative efficiency benefits of size-discovery is similar to that for the simpler case of bilateral workup. Moreover, additional insights are gained from the equilibrium dynamic dropout policies in multilateral workups.

We take the numbers N_b of buyers and N_s of sellers to be initially unobservable, independent, and having the same geometric distribution. Specifically, for any non-negative integer k,

$$P(N_b = k) = P(N_s = k) = f(k) \equiv q^k(1-q), \tag{41}$$

for some $q \in (0,1)$. We have $E(N_b) = E(N_s) = q/(1-q)$. The interpretation is that after each buyer exits the workup, there is a new buyer with probability q, and likewise for sellers. (The multilateral workup model is difficult to solve with a deterministic number of traders.[9])

Although it is natural that the number of institutional investors and financial intermediaries seeking to trade large positions is unobservable and stochastic, as we have assumed here, we are forced for reasons of tractability to assume that once trading in the double-auction market begins, the total number of market participants is revealed to all. (Otherwise, the analysis of the double auction market would be overly complicated.)

[9]The bilateral workup model can be solved if the number of buyers and the number of sellers are geometrically distributed. The explicit calculations are more involved but available upon request.

Pre-workup inventories are positive for sellers and negative for buyers. For both buyers and sellers, the absolute magnitudes of pre-workup inventory sizes are *iid* exponentially distributed, with parameter μ, thus with mean $1/\mu$. The numbers of buyers and sellers and the pre-workup inventory sizes are independent. Before participating in workup, each trader observes only his own inventory.

It follows from the independence assumptions and the memoryless property of the geometric distribution that, conditioning on all information available to a trader during his turn at workup, the conditional distribution of the numbers of buyers and sellers that have not yet entered workup retain their original independent geometric distributions.

As in Section 4, the workup session takes place before the start of the double-auction market. The workup begins by pairing the first buyer and first seller. During the workup, the exit from workup of the i-th buyer causes the $(i+1)$-st buyer to begin workup, provided $N_b > i$. The $(i+1)$-st buyer can then choose whether to begin actively buying or to immediately drop out without trading. Similarly, when seller j exits, he is replaced with another seller if $N_s > j$. The exit of a trader, whether a buyer or a seller, and the replacement of the trader is observable to everyone when it occurs. (The identities of the exiting traders are irrelevant, and not reported, beyond whether they are buyers or sellers.) The quantities executed by each departing trader are also observable. In particular, the event that a trader drops out of workup without executing any quantity is also observable. The workup ends when buyer number N_b exits or when seller number N_s exits, whichever is first.

Throughout this section, we assume for simplicity that the workup price $\bar{p}$ is set at the expectation of the subsequent auction price p_0, which is v.

We will show that at any given point during the workup, the state vector on which the equilibrium strategies depend is of the form (m, X, y), where:

- m is the total number of buyers and sellers that have already entered workup, including the current buyer and seller.
- X is the total conditional expected inventory held by previously exited participants,

given all currently available information. Given our information structure, this conditional expectation is common to all workup participants.

y is the quantity that the current workup pair has already executed. We emphasize that $y = 0$ corresponds to a state in which the current workup pair have yet to execute any trade, allowing for the positive-probability event that at least one of them may drop out of workup without executing any quantity.

We let $\mathcal{M}_b(m, X) > 0$ and $\mathcal{M}_s(m, X) > 0$ denote the conjectured dropout thresholds of the current buyer and seller, respectively, in a workup state (m, X, y) that is *active*, meaning $y > 0$. That is, we conjecture that, when the workup state is active, the current buyer drops out once the absolute magnitude of his remaining inventory has been reduced to $\mathcal{M}_b(m, X)$. We conjecture and later verify an equilibrium in which these thresholds depend only on (m, X), and not on a trader's current inventory or on other aspects of the observable history of the game. We call any equilibrium of this form an "equilibrium in Markovian threshold dropout strategies."

The distinction between an active workup pair ($y > 0$) and a matched but currently inactive pair ($y = 0$) is important to the equilibrium policies. Suppose, for example, that we are in an active state for the first buyer and first seller. That is, the first buyer and the first seller have executed a positive quantity y in the workup, and nothing else has yet happened. As we will show later, because $X = 0$, the buyer and the seller use a common dropout threshold, say $\mathcal{M}_0$. If, for example, the buyer exits, then every workup participant infers that the buyer's residual inventory level is $-\mathcal{M}_0$. By contrast, at an inactive state, if the buyer immediately exits, then everyone else learns that the buyer's inventory size is *at most* $\mathcal{M}_0$, and in particular is distributed with a truncated exponential distribution, with the conditional expectation $\nu(\mathcal{M}_0)$, where, for any positive number y,

$$\nu(y) \equiv \frac{\int_{x=0}^{y} x\mu e^{-\mu x}\, dx}{1 - e^{-\mu y}} < y. \tag{42}$$

Thus, whether a trader exits without trading a strictly positive amount affects the inference of all traders.

The key step to solving the equilibrium is to calculate the dropout thresholds $\mathcal{M}_b(m, X)$ and $\mathcal{M}_s(m, X)$. For this, we first consider the problem of the current active buyer, whose initial inventory is S^b, on the event that the buyer has to this point executed some quantity $y > 0$ and the active seller has not yet exited. Let n_b and n_s denote the number of buyers and the number of sellers yet to enter the workup, respectively, excluding the current pair. We will calculate the first-order optimality conditions by artificially including the residual queue sizes n_b and n_s in the buyer's conditioning information, and then later averaging with respect to the conditional distribution of (n_b, n_s). By the same logic used in Section 4, the buyer has conditional mean and variance[10] of aggregate market-wide inventory given by

$$E(Z \mid S^b, m, X, y, n_b, n_s) = S^b + y + \mathcal{M}_s(m, X) + \frac{1}{\mu} + X + (n_s - n_b)\frac{1}{\mu}, \tag{43}$$

$$E(Z^2 \mid S^b, m, X, y, n_b, n_s) = (S^b + y + \mathcal{M}_s(m, X))^2 + 2(S^b + y + \mathcal{M}_s(m, X))\frac{1}{\mu} + \Gamma_b(n_b, n_s), \tag{44}$$

where $\Gamma_b(n_b, n_s)$ is a quantity that does not depend on y. Relative to the calculation (16) for the case of bilateral workup, the conditional mean $E(Z \mid m, X, y, n_b, n_s)$ includes the extra terms X and $(n_s - n_b)/\mu$. The exact level of the second moment $E(Z^2 \mid m, X, y, n_b, n_s)$ does not affect the equilibrium threshold, because it plays no role in the first-order optimality condition for the choice of y at which the buyer drops out.

Omitting the arguments of $\mathcal{M}_s$ and $\mathcal{M}_b$, we have

$$\begin{aligned} \frac{dE(U^b \mid S^b, m, X, y, n_b, n_s)}{dy} &= v - \bar{p} - 2\frac{\gamma}{r}C(n)(S^b + y) \\ &+ 2\frac{\gamma}{r}(C(n) - 1)\frac{1}{n}\left(2(S^b + y) + \mathcal{M}_s + \frac{1}{\mu} + X + (n_s - n_b)\frac{1}{\mu}\right) \\ &- \frac{\gamma}{r}(C(n) - 1)\frac{1}{n^2}\left(2(S^b + y + \mathcal{M}_s) + \frac{2}{\mu}\right), \end{aligned} \tag{45}$$

[10] The event of executing y units has probability zero, but the stated conditional moments make sense when applying a regular version of the conditional distribution of Z given the executed quantity and given X.

where $n = m + n_b + n_s$ and where

$$C(n) = \frac{1 - 2a_\Delta \gamma / r}{n - 1}. \tag{46}$$

By the law of iterated expectations, we can average with respect to the product distribution of (n_b, n_s), to obtain

$$\begin{aligned} \frac{dE(U^b \mid S^b, m, X, y)}{dy} &= v - \bar{p} - \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) 2 \frac{\gamma}{r} C(n)(S^b + y) \\ &+ \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) 2 \frac{\gamma}{r} (C(n) - 1) \frac{1}{n} \left(2(S^b + y) + \mathcal{M}_s + \frac{1}{\mu} + X + (\ell - k) \frac{1}{\mu} \right) \\ &- \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) \frac{\gamma}{r} (C(n) - 1) \frac{1}{n^2} \left(2(S^b + y + \mathcal{M}_s) + \frac{2}{\mu} \right), \end{aligned} \tag{47}$$

where $n = m + k + \ell$.

The first-order condition for optimal y should hold with equality if $S^b + y = -\mathcal{M}_b$, that is,

$$\begin{aligned} 0 &= \frac{dE(U^b \mid S^b, m, X, y)}{dy} \Big|_{S^b + y = -\mathcal{M}_b} \\ &= v - \bar{p} - \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) 2 \frac{\gamma}{r} C(n)(-\mathcal{M}_b) \\ &\quad + \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) 2 \frac{\gamma}{r} (C(n) - 1) \frac{1}{n} \left(2(-\mathcal{M}_b) + \mathcal{M}_s + \frac{1}{\mu} + X + (\ell - k) \frac{1}{\mu} \right) \\ &\quad - \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} f(k) f(\ell) \frac{\gamma}{r} (C(n) - 1) \frac{1}{n^2} \left(2(-\mathcal{M}_b + \mathcal{M}_s) + \frac{2}{\mu} \right), \end{aligned} \tag{48}$$

where $n = m + k + \ell$.

By a completely analogous calculation, the seller, whose initial inventory is S^s, stays in workup until the buyer has exited or the workup quantity has reached a level y satisfying the seller's first-order condition, whichever comes first. This occurs when the seller's remaining inventory reaches the threshold $S^s - y = \mathcal{M}_s$. Thus, the first-order condition for y takes the form

$$
\begin{aligned}
0 &= \frac{dE(U^s \mid S^s, m, X, y)}{dy}\Big|_{S^s - y = \mathcal{M}_s} \qquad (49)\\
&= \bar{p} - v + \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} f(k)f(\ell)2\frac{\gamma}{r}C(n)\mathcal{M}_s \\
&\quad + \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} f(k)f(\ell)2\frac{\gamma}{r}(C(n)-1)\frac{1}{n}\left(-2\mathcal{M}_s + \mathcal{M}_b + \frac{1}{\mu} - X - (\ell - k)\frac{1}{\mu}\right) \\
&\quad - \sum_{k=0}^{\infty}\sum_{\ell=0}^{\infty} f(k)f(\ell)\frac{\gamma}{r}(C(n)-1)\frac{1}{n^2}\left(-2(\mathcal{M}_s - \mathcal{M}_b) + \frac{2}{\mu}\right),
\end{aligned}
$$

where $n = m + k + \ell$. As the sum of two *iid* geometric random variables, $n_b + n_s$ has the negative binomial conditional distribution[11] with mass function

$$g(k) = (k+1)q^k(1-q)^2. \qquad (50)$$

Substituting in $\bar{p} = v$, the pair of linear first-order necessary and sufficient conditions for optimality, (48) and (49), have the unique solutions

$$\mathcal{M}_b(m, X) = M^*(m) + L(m)X, \qquad (51)$$

$$\mathcal{M}_s(m, X) = M^*(m) - L(m)X, \qquad (52)$$

where, taking $n = m + k$,

$$M^*(m) = \frac{1}{\mu}\frac{\sum_{k=0}^{\infty} g(k)(1 - C(n))\frac{n-1}{n^2}}{\sum_{k=0}^{\infty} g(k)\left(C(n) + \frac{1-C(n)}{n}\right)} \qquad (53)$$

and

$$L(m) = \frac{\sum_{k=0}^{\infty} g(k)\frac{1-C(n)}{n}}{\sum_{k=0}^{\infty} g(k)\left(C(n) + \frac{(1-C(n))(3n-2)}{n^2}\right)}. \qquad (54)$$

The symmetric and opposite roles of X for the buyer and the seller thresholds is intuitive. In a multilateral workup, the role of the conditional expected total inventory X of those traders who have already exited workup is similar to the role of the workup price

[11] This can be shown from the fact $g(\,\cdot\,)$ is the convolution $f * f$.

"bias" $\bar{p} - v$ in the bilateral workup equilibrium described by Proposition 3. For example, an increase in X makes the current buyer more cautious, setting a higher dropout quantity, and makes the current seller more aggressive, setting a lower dropout quantity. This is so because as X rises, the conditional expected market-clearing price of the subsequent double auctions falls. This encourages the current buyer to wait for the double-auction market to cover his inventory shortfall, and encourages the seller to reserve less inventory for sale in the double-auction market. The opposite is true for a decrease in X.

In order for the above conjectured strategies to be consistent, we need to prove that the thresholds of incumbents are weakly increasing with each dropout, and that the thresholds are always nonnegative. That is, we want to show that

$$\mathcal{M}_b(m+1, X') \geq \mathcal{M}_b(m, X), \tag{55}$$

$$\mathcal{M}_s(m+1, X') \geq \mathcal{M}_s(m, X), \tag{56}$$

$$\mathcal{M}_b(m, X) \geq 0, \tag{57}$$

$$\mathcal{M}_s(m, X) \geq 0, \tag{58}$$

for any possible successive outcomes X and X' of the conditional expected inventory of departed workup participants (before and after a dropout). If these conditions fail, a trader's optimal dropout threshold may depend on his current inventory or the past threshold of his counterparty, among possibly other variables. These complications would render the problem intractable.

The monotonicity and positivity properties of (55)–(58) are satisfied if $e^{-r\Delta} > 1/2$, a relatively unrestrictive condition. For example, taking a day as the unit of time, if payoff-relevant information arrives once per day ($r = 1$) and the double auctions are held at least twice per day ($\Delta \leq 0.5$), we would have $e^{-r\Delta} \geq e^{-0.5} \approx 0.61 > 0.5$, and (55)–(58) are satisfied.[12]

Proposition 6. *The coefficients $M^*(m)$ and $M^*(m)/L(m)$ are always weakly increasing*

[12]We have also checked that if Δ is large enough, then $L(m)$ is not monotone increasing in m. Although the non-monotonicity of $L(m)$ for large Δ blocks our particular proof method when Δ is sufficiently large, it does not necessarily rule out other approaches to demonstrating equilibria in threshold strategies for large Δ.

in m *for* $m \geq 2$*. If* $e^{-r\Delta} > 1/2$*, then* $L(m)$ *is also weakly increasing in* m *for* $m \geq 2$*. If* $e^{-r\Delta} > 1/2$*, the monotonicity and nonnegativity conditions of* (55)–(58) *are satisfied.*

The property that $M^*(m)$ and $L(m)$ are increasing in m is intuitive. As more traders drop out (that is, as m increases), the expected total number of traders in the subsequent double-auction market goes up, by the memoryless property of the geometric distribution. Since the double-auction market becomes more competitive as more traders participate, and the associated inefficiency related to price impact thus becomes smaller, there is less advantage to using workup, so $M^*(m)$ goes up.

In addition, traders who have already exited the workup will enter the double-auction market with their residual inventories, causing a predictable shift in the double-auction price relative to the workup price. For example, if the conditional expected inventory X of past workup participants is positive, then the double-auction price is expected to be lower than v, a favorable condition for the workup buyer. Again, because a larger number m of past and current workup participants makes the double-auction market more competitive in expectation, those who have already exited the workup will be more aggressive in liquidating their residual inventories, thus front-loading their sales in the relatively early rounds. The workup buyer, therefore, expects to purchase the asset in the double-auction market sooner and at more favorable prices. Consequently, the buyer will set an even higher dropout threshold. By a symmetric argument, conditional on $X > 0$, a higher m means that the seller sets an even lower threshold. That is, a higher m means a higher sensitivity of the thresholds $\mathcal{M}_b$ and $\mathcal{M}_s$ to the total expected inventory X of those who have already exited workup.

The monotonicity of the thresholds, (55)–(56), means that after the exit of a trader, his counterparty's dropout threshold (weakly) increases. For example, if the current seller j exits before the current buyer i, then X goes up, so the new threshold $M^*(m) + L(m)X$ of buyer i goes up. Likewise, after each exit of a buyer, X goes down, and the dropout threshold of the seller who remains in the workup increases. Thus, after the exit of a counterparty, the incumbent either drops out immediately because of his increased threshold, or he stays in despite his new higher threshold. Conditional on the latter event, for other traders, the incumbent's remaining inventory in excess of his new, increased threshold is again an exponentially-distributed variable with mean $1/\mu$. The non-negativity of the thresholds, (57)–(58), implies that no trader wishes to "overshoot" across the zero inventory boundary. These properties ensure stationarity and are in fact needed for tractability of this general approach to solving for equilibria.

Summarizing, the following proposition provides a complete description of Markovian workup equilibrium.[13] The equilibrium workup strategy of each player depends on that player's privately observed pre-workup asset inventory and on the publicly observable Markov process[14] (m, X, y).

Proposition 7. *Suppose that $e^{-r\Delta} > 1/2$. The multilateral dynamic workup game associated with workup price $\bar{p} = v$ has a unique equilibrium in Markovian threshold dropout strategies. This equilibrium is characterized by the following recursive determination of the workup state and of traders' equilibrium dropout strategies. Here, z_i^b and z_j^s denote the pre-workup inventories of the i-th buyer and the j-th seller, respectively. The initial workup state is $(m, X, y) = (2, 0, 0)$.*

1. *At any inactive workup state $(m, X, 0)$:*

   (a) *If $|z_i^b| \leq \mathcal{M}_b(m, X)$ and $z_j^s > \mathcal{M}_s(m, X)$, where $\mathcal{M}_b(m, X)$ and $\mathcal{M}_s(m, X)$ are given by (51) and (52) respectively, then the buyer, and only the buyer, exits*

[13] Because of the continuum of agent types and actions, we cannot formally apply the standard notion of perfect Bayesian equilibrium for dynamic games with incomplete information, because that would call for conditioning on events that have zero probability, such as a counterparty dropping out of workup after executing a trade of a specific size. In our setting, actions are commonly observable and there is no issue concerning off-equilibrium-path conjectures, so almost any natural extension of simple perfect Bayesian equilibrium to our continuum action and type spaces leads to our equilibrium. For example, we could apply the notion of open sequential equilibrium of Myerson and Reny (2015). For the present draft of the paper, we refer simply to an "equilibrium" in the sense that every agent applies Bayes' Rule based on a regular version of the conditional distribution of Z given the observed variables, in order to compute its optimal threshold strategy, given the threshold strategies of other agents. As stated, there is a unique such equilibrium in threshold strategies because: *(i)* given the other traders' threshold strategies, a given trader's threshold strategy is uniquely determined by its first order necessary condition for optimality (which is sufficient because of concavity), and *(ii)* there is a unique solution for the pair of first-order conditions for the equations for the threshold strategies $\mathcal{M}_b$ and $\mathcal{M}_s$.

[14] Specifically, $(m, X, y) = (m_t, X_t, y_t)_{(t \geq 0)}$ is a continuous-time Markov process with state space $\mathbb{N} \times \mathbb{R} \times \mathbb{R}$, where $\mathbb{N}$ is the space of natural numbers. To be precise, one can add an artificial independent exponential "wait time" after each transition to an inactive state. This ensures that the state cannot jump twice at the same time on the workup clock when making a transition from an inactive state to an inactive state after an immediate dropout.

immediately (that is, without trading any quantity). Unless $N_b = i$*, the workup state then evolves to* $(m+1, X - \nu(\mathcal{M}_b(m,X)), 0)$.

(b) *If* $|z_i^b| > \mathcal{M}_b(m,X)$ *and* $z_j^s \leq \mathcal{M}_s(m,X)$*, then the seller, and only the seller, exits immediately. Unless* $N_s = j$*, the workup state evolves to* $(m+1, X + \nu(\mathcal{M}_s(m,X)), 0)$.

(c) *If* $|z_i^b| \leq \mathcal{M}_b(m,X)$ *and* $z_j^s \leq \mathcal{M}_s(m,X)$*, then both sides exit immediately, without trading any quantity. Unless* $N_b = i$ *or* $N_s = j$*, the workup state evolves to* $(m+2, X - \nu(\mathcal{M}_b(m,X)) + \nu(\mathcal{M}_s(m,X)), 0)$.

(d) *If* $|z_i^b| > \mathcal{M}_b(m,X)$ *and* $z_j^s > \mathcal{M}_s(m,X)$*, then the current buyer* i *and seller* j *enter an active workup. That is, the workup state evolves to* $(m, X, 0)$.

(e) *If, at any of the transitions above,* $N_b = i$ *or* $N_s = j$*, then the workup ends.*

2. *At any active workup state* (m, X, y)*, the current buyer* i *and seller* j *remain in the workup as their traded quantity rises until the earlier of the two following events (a) and (b):*

   (a) *The remaining inventory of the buyer (which is negative) rises to the threshold* $-\mathcal{M}_b(m,X) = -(M^*(m) + L(m)X)$*. At this point, the buyer exits. Unless* $N_b = i$*, the workup state evolves to* $(m+1, X - (M^*(m) + L(m)X), 0)$.

   (b) *The remaining inventory of the seller falls to the threshold* $\mathcal{M}_s(m,X) = M^*(m) - L(m)X$*. At this point, the seller exits. Unless* $N_s = j$*, the state evolves to* $(m+1, X + M^*(m) - L(m)X, 0)$.

   (c) *On the zero-probability event that (a) and (b) occur simultaneously, the state evolves to* $(m+2, X - 2L(m)X, 0)$ *unless* $N_b = i$ *or* $N_s = j$.

   (d) *If, at either or both of (a) or (b), we have* $N_b = i$ *or* $N_s = j$*, then the workup ends.*

6 Concluding Remarks

This paper demonstrates the welfare benefit of size discovery, and adds to the general understanding of how market designs have responded in practice to frictions associated with imperfect competition.

Price-discovery markets are efficient in an idealized price-taking competitive market, for example one in which traders are infinitesimally small, as in Aumann (1964). The First Welfare Theorem of Arrow (1951), by which market clearing allocations are efficient, is based on the price-taking assumption. In many functioning markets, however, price taking is a poor approximation of trading behavior because of traders' awareness of their own price impact, and efficiency is lost. For instance, in inter-dealer financial markets, there are often heavy concentrations of inventory imbalances among a relatively small set of market participants. These are large dealers, hedge funds, and other asset management firms that are extremely conscious of their potential to harm themselves by price impact. In the case of U.S. Treasury markets, for instance, government auctions often leave a small number of primary dealers with significant position imbalances. Some dealers are surprised by being awarded substantially more bonds in the auction than needed to meet their customer commitments and desired market-making inventories. Some receive significantly less than desired. Fleming and Nguyen (2015) explain how dealers exploit workups to lay off their imbalances.

We have shown that under imperfect competition, adding a size-discovery mechanism such as a workup improves allocative efficiency over a stand-alone price-discovery mechanism, such as sequential double auctions. Precisely because a workup freezes the transaction price, it avoids the efficiency losses caused by the strategic avoidance of price impact in price-discovery mechanisms. Workup participants are therefore willing to trade large blocks of an asset almost instantly, leading to a quick reduction of inventory imbalances and improvement in allocative efficiency.

We have also shown that equilibrium optimal workup strategies trade off the benefit of quickly eliminating large undesired positions against the cost of adverse selection associated with information about the total market inventory. As a result, only traders with large inventory imbalances actively participate in workups, and workup participants set an endogenous threshold for the level of remaining inventory at which they drop out.

We emphasize that the welfare benefit of size discovery is higher if it is used in combination with a price-discovery mechanism. In fact, if size discovery were the available only trading mechanism, it would be even less efficient than a price-discovery-only market. Appendix D shows that, in terms of (ex ante expected) welfare, the three possible market structures can be ranked as follows:

$$\text{workup + double auctions} \ \succeq\ \text{double auctions only} \ \succeq\ \text{workup only}, \tag{59}$$

where "$\succeq$" means "more efficient than," in the sense of total social surplus.

It would be natural to extend our model so as to incorporate more general workup timing and an endogenous workup price. In our current model, a single multilateral workup (or multiple bilateral workups) is added before the opening of the sequential-double-auction market. A natural interpretation is that the workup occurs at the beginning of each trading day, say at the closing price of the previous day. We have shown that this form of "low-frequency" size discovery improves welfare. A useful extension would allow "higher-frequency size discovery," for example a multilateral workup before each double auction, at the previous double-auction price. (The first workup could be done at the closing price of the previous day.) An extension of this sort is likely to be extremely complex to analyze. For instance, this timing introduces an incentive for each large trader to "manipulate" double-auction prices in order to profit from subsequent workups. Indeed, the same challenge may apply to any extension in which the workup price is endogenously "discovered" by strategic traders. Although it would be interesting and useful to characterize the resulting equilibrium behavior, we have not yet found a tractable way to do so.

Appendix

A Proofs

This appendix contains proofs of results stated in the main text.

A.1 Proof of Proposition 1

As in the text, we simplify the notation by writing "x_{ik}" in place of "$x_{ik}(p_k; z_{ik})$," and conjecture an equilibrium strategy of the form

$$x_{ik} = av - bp_k + dz_{ik}. \tag{60}$$

Under this conjecture, and because the inventory shocks have mean zero, the equilibrium price p_k is a martingale because the total inventory Z_k is a martingale.

Trader i in round k effectively selects the optimal execution price p_k. Adapting the method of Du and Zhu (2015), we write the first-order optimality condition of trader i as

$$(n-1)b\Bigg[(1-e^{-r\Delta})\left(v - \frac{2\gamma}{r}(x_{ik}+z_{ik}) + \sum_{j=1}^{\infty} e^{-rj\Delta}(1+d)^j(v - \frac{2\gamma}{r}E_k(z_{i,k+j}+x_{i,k+j}))\right)$$
$$- p_k - \sum_{j=1}^{\infty} e^{-rj\Delta}(1+d)^{j-1} d\, E(p_{k+j})\Bigg] - x_{ik} = 0, \tag{61}$$

where $E_k(\,\cdot\,)$ denotes conditional expectation given z_{ik} and p_k.

By the evolution equation for the inventory $\{z_{ik}\}$, we have, for all $j \geq 1$,

$$z_{i,k+j} + x_{i,k+j} = (1+d)^j(z_{ik}+x_{ik}) + \sum_{l=1}^{j-1}(av - bp_{k+l} + w_{i,k+l+1})(1+d)^{j-l} \tag{62}$$

$$+ (av - bp_{k+j}) + w_{i,k+1}(1+d)^j. \tag{63}$$

Since all shocks have mean zero and the prices are martingales, we have

$$E_k(z_{i,k+j}+x_{i,k+j}) = (1+d)^j(z_{ik}+x_{ik}) + (av - bp_k)\left(\frac{(1+d)^j}{d} - \frac{1}{d}\right). \tag{64}$$

The above equation is linear in x_{ik}, v, p_k, and z_{ik}. Matching the coefficients with those of the conjectured strategy $x_{ik} = av - bp_k + dz_{ik}$ and solving the three equations, we have

$$b = a, \tag{65}$$

$$d = -\frac{2\gamma}{r}a, \tag{66}$$

$$a = a_\Delta \equiv \frac{r}{2\gamma}\left(1 + \frac{(n-1)(1-e^{-r\Delta}) - \sqrt{(n-1)^2(1-e^{-r\Delta})^2 + 4e^{-r\Delta}}}{2e^{-r\Delta}}\right). \tag{67}$$

A.2 Proof of Proposition 2

Our proof strategy consists of two steps. First, we calculate $V_{i,0+}$ under the assumption that $\sigma_w = 0$ (that is, no periodic inventory shocks after time 0). This gives the first three terms in the expression of $V_{i,0+}$. Then, we calculate the last term Θ, the contribution of periodic inventory shocks to the indirect utility.

Step 1: No periodic inventory shocks. With $w_{ik} = 0$ for all i and $k \geq 1$, and given the equilibrium price p^*, we can write the law of motion of the inventory of trader i as

$$\begin{aligned} z_{i,k+1} &= z_{ik} + a_\Delta \left(v - p^* - \frac{2\gamma}{r} z_{ik} \right) \\ &= z_{ik} - a_\Delta \frac{2\gamma}{r} \left(z_{ik} - \frac{Z}{n} \right), \end{aligned} \tag{68}$$

which implies that

$$z_{i,k+1} - \frac{Z}{n} = \left(1 - a_\Delta \frac{2\gamma}{r} \right) \left(z_{ik} - \frac{Z}{n} \right). \tag{69}$$

We let

$$V_{i,0+} = \sum_{k=0}^{\infty} e^{-r\Delta k} E \left[-x_{ik} p^* + (1 - e^{-r\Delta}) \left(v(x_{ik} + z_{ik}) - \frac{\gamma}{r} (x_{ik} + z_{ik})^2 \right) \, \Bigg| \, z_{i0}, Z \right]. \tag{70}$$

The inventories evolve according to

$$z_{i,k+1} - \frac{Z}{n} = \left(1 - a_\Delta \frac{2\gamma}{r} \right) \left(z_{ik} - \frac{Z}{n} \right) = \left(1 - a_\Delta \frac{2\gamma}{r} \right)^{k+1} \left(z_{i0} - \frac{Z}{n} \right). \tag{71}$$

It follows that, in equilibrium,

$$x_{ik} = a_\Delta \frac{2\gamma}{r} \left(\frac{Z}{n} - z_{ik} \right) = a_\Delta \frac{2\gamma}{r} \left(1 - a_\Delta \frac{2\gamma}{r} \right)^k \left(\frac{Z}{n} - z_{i0} \right). \tag{72}$$

The price-related term in (70) is

$$\begin{aligned} \sum_{k=0}^{\infty} e^{-r\Delta k} p^* x_{ik} &= \sum_{k=0}^{\infty} e^{-r\Delta k} \left(v - \frac{2\gamma}{nr} Z \right) a_\Delta \frac{2\gamma}{r} \left(1 - a_\Delta \frac{2\gamma}{r} \right)^k \left(\frac{Z}{n} - z_{i0} \right) \\ &= \left(v - \frac{2\gamma}{nr} Z \right) \left(\frac{Z}{n} - z_{i0} \right) \frac{a_\Delta \frac{2\gamma}{r}}{1 - e^{-r\Delta}(1 - a_\Delta \frac{2\gamma}{r})}. \end{aligned} \tag{73}$$

In (70), the term that involves v is

$$\begin{aligned} & v \sum_{k=0}^{\infty} \left(1 - e^{-r\Delta} \right) e^{-r\Delta k} \left[\frac{Z}{n} + \left(1 - a_\Delta \frac{2\gamma}{r} \right)^{k+1} \left(z_{i0} - \frac{Z}{n} \right) \right] \\ &= v \frac{Z}{n} + \frac{(1 - a_\Delta \frac{2\gamma}{r}) \left(1 - e^{-r\Delta} \right)}{1 - e^{-r\Delta}(1 - a_\Delta \frac{2\gamma}{r})} v \left(z_{i0} - \frac{Z}{n} \right). \end{aligned} \tag{74}$$

In (70), the term that involves γ is

$$
\begin{aligned}
&-\frac{\gamma}{r}\sum_{k=0}^{\infty}\left(1-e^{-r\Delta}\right)e^{-r\Delta k}\left[\frac{Z}{n}+\left(1-a_\Delta\frac{2\gamma}{r}\right)^{k+1}\left(z_{i0}-\frac{Z}{n}\right)\right]^2 \\
&\quad=-\frac{\gamma}{r}\left(\frac{Z}{n}\right)^2-\frac{\left(1-a_\Delta\frac{2\gamma}{r}\right)\left(1-e^{-r\Delta}\right)}{1-e^{-r\Delta}(1-a_\Delta\frac{2\gamma}{r})}\frac{2\gamma Z}{nr}\left(z_{i0}-\frac{Z}{n}\right)-\frac{\gamma}{r}\frac{1-a_\Delta\frac{2\gamma}{r}}{n-1}\left(z_{i0}-\frac{Z}{n}\right)^2.
\end{aligned}
\tag{75}
$$

Adding up the three terms, we get the first, second, and third term in the expression for $V_{i,0+}$.

Step 2: Add the effect of periodic inventory shocks. We now calculate the terms in the indirect utility caused by the extra terms $\{w_{ik}\}$, where $k \geq 1$.

For any integer $t \geq 0$, we let s_t be the coefficient of w_{il} in the expression of $z_{i,l+t}$ and let u_t be the coefficient of w_{jl} in the expression of $z_{i,l+t}$, where $j \neq i$. Clearly, $s_0 = 1$ and $u_0 = 0$.

For simplicity of expressions, write $c_\Delta = a_\Delta \frac{2\gamma}{r}$.

We can write (12) more explicitly as

$$
z_{i,k+1} = (1-c_\Delta)\,z_{ik} + c_\Delta\frac{z_{ik}+\sum_{j\neq i} z_{jk}}{n} + w_{i,k+1}. \tag{76}
$$

Thus, we get recursive equations of $\{u_t\}$ and $\{s_t\}$:

$$
u_{t+1} \;=\; (1-c_\Delta)u_t + c_\Delta\left(\frac{n-1}{n}u_t+\frac{1}{n}s_t\right) = \left(1-\frac{c_\Delta}{n}\right)u_t + \frac{c_\Delta}{n}s_t, \tag{77}
$$

and

$$
s_{t+1} \;=\; (1-c_\Delta)s_t + c_\Delta\left(\frac{n-1}{n}u_t+\frac{1}{n}s_t\right) = \left(1-\frac{(n-1)c_\Delta}{n}\right)s_t + \frac{(n-1)c_\Delta}{n}u_t. \tag{78}
$$

These recursive equations have the solution (using $s_0 = 1$ and $u_0 = 0$):

$$
s_t = \frac{1+(n-1)(1-c_\Delta)^t}{n}, \quad u_t = \frac{1-(1-c_\Delta)^t}{n}. \tag{79}
$$

Fixing i: Let's first calculate the difference caused by the w terms in the expression of $E\left[-x_{ik}p_k \mid z_{i0}, Z\right]$. From (79) and the recursive equations for u_t and s_t, we see that the coefficient of w_{il} $(l \leq k)$ in the expression of x_{ik} is

$$
-c_\Delta\left(s_{k-l}-\frac{s_{k-l}+(n-1)u_{k-l}}{n}\right) = c_\Delta\frac{n-1}{n}(u_{k-l}-s_{k-l}), \tag{80}
$$

and the coefficient of w_{jl} $(l \leq k)$ in the expression of x_{ik} is

$$
-c_\Delta\left(u_{k-l}-\frac{s_{k-l}+(n-1)u_{k-l}}{n}\right) = c_\Delta\frac{1}{n}(s_{k-l}-u_{k-l}). \tag{81}
$$

Similarly, the coefficient of w_{il} and w_{jl} $(l \leq k,\ j \neq i)$ in the expression of p_k is

$$-\frac{2\gamma}{r}\frac{s_{k-l}+(n-1)u_{k-l}}{n}. \tag{82}$$

Since each w term has (conditional and unconditional) mean of zero, all expectation terms linear in w_{il} or w_{jl} are zero. Moreover, because the inventory shocks are independent of each other, all quadratic terms—expect those of the form w_{ml}^2, where $m \in \{1, 2, ..., n\}$ and $l \leq k$—are also zero. These imply that the contribution of the periodic inventory shocks to $E\left[-x_{ik}p_k \mid z_{i0}, Z\right]$ is:

$$\begin{aligned}
& \sum_{l=1}^{k} -\left(c_\Delta \frac{n-1}{n}(u_{k-l}-s_{k-l})\right)\left(-\frac{2\gamma}{r}\frac{s_{k-l}+(n-1)u_{k-l}}{n}\right) E\left[w_{il}^2 \mid z_{i0}, Z\right] \\
& + \sum_{l=1}^{k}\sum_{j\neq i} -\left(c_\Delta \frac{1}{n}(s_{k-l}-u_{k-l})\right)\left(-\frac{2\gamma}{r}\frac{s_{k-l}+(n-1)u_{k-l}}{n}\right) E\left[w_{jl}^2 \mid z_{i0}, Z\right] \\
= \;& \sigma_w^2\Delta\left(-\frac{2\gamma}{r}\frac{s_{k-l}+(n-1)u_{k-l}}{n}\right)\sum_{l=1}^{k}\left[-c_\Delta\frac{n-1}{n}(u_{k-l}-s_{k-l})-(n-1)\cdot c_\Delta\frac{1}{n}(s_{k-l}-u_{k-l})\right] \\
= \;& 0.
\end{aligned} \tag{83}$$

Obviously, the w terms make no difference to the term $E\left[(x_{ik}+z_{ik}) \mid z_{i0}, Z\right]$ because the inventory shocks have mean zero.

Now let's turn to the difference caused by the w terms in the expression of $E\left[(x_{ik}+z_{ik})^2 \mid z_{i0}, Z\right]$. From (80), the coefficient of w_{il} $(l \leq k)$ in the expression of $x_{ik}+z_{ik}$ is

$$c_\Delta\frac{n-1}{n}(u_{k-l}-s_{k-l})+s_{k-l},$$

and from (81), the coefficient of w_{jl} $(l \leq k)$ in the expression of $x_{ik}+z_{ik}$ is

$$c_\Delta\frac{1}{n}(s_{k-l}-u_{k-l})+u_{k-l}.$$

Again, because the w terms have mean zero and are mutually independent, the difference caused by the w terms in the expression of $E\left[(x_{ik}+z_{ik})^2 \mid z_{i0}, Z\right]$ is:

$$\begin{aligned}
& \sum_{l=1}^{k}\left[c_\Delta\frac{n-1}{n}(u_{k-l}-s_{k-l})+s_{k-l}\right]^2 E\left[w_{il}^2 \mid z_{i0}, Z\right] \\
& + \sum_{l=1}^{k}\sum_{j\neq i}\left[c_\Delta\frac{1}{n}(s_{k-l}-u_{k-l})+u_{k-l}\right]^2 E\left[w_{jl}^2 \mid z_{i0}, Z\right] \\
= \;& \sigma_w^2\Delta\left(\sum_{t=0}^{k-1}\left[c_\Delta\frac{n-1}{n}(u_t-s_t)+s_t\right]^2+(n-1)\sum_{t=0}^{k-1}\left[c_\Delta\frac{1}{n}(s_t-u_t)+u_t\right]^2\right).
\end{aligned}$$

Thus, the difference caused by the w terms in the expression of $V_{i,0+}$ is

$$\Theta \equiv -\sigma_w^2 \Delta \frac{\gamma}{r}(1-e^{-r\Delta}) \tag{84}$$
$$\cdot \sum_{k=1}^{\infty} e^{-r\Delta k} \left(\sum_{t=0}^{k-1} \left[c_\Delta \frac{n-1}{n}(u_t - s_t) + s_t \right]^2 + (n-1) \sum_{t=0}^{k-1} \left[c_\Delta \frac{1}{n}(s_t - u_t) + u_t \right]^2 \right),$$

which is a constant that does not depend on $\{z_{i0}\}$ or Z.

A.3 Proof of Proposition 5

The comparative statics of M with respect to r and Δ are provided in the text. The only item left is to show that M increases in n.

Define

$$A \equiv \frac{2e^{-r\Delta}}{1-e^{-r\Delta}}, \tag{85}$$
$$B \equiv \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}, \tag{86}$$
$$\alpha_n = (n-1) + A + \sqrt{(n-1)^2 + B}. \tag{87}$$

Then, we can write

$$C(n) = \frac{1}{n-1}\left(1 - \frac{2(n-2)}{\alpha_n}\right). \tag{88}$$

To show that M increases in n, it is equivalent to show that

$$\frac{n\left(1 + \frac{nC(n)}{1-C(n)}\right)}{n-1} > \frac{(n+1)\left(1 + \frac{(n+1)C(n+1)}{1-C(n+1)}\right)}{n}, \tag{89}$$

which, after simplification, is equivalent to

$$\frac{(n+1)^2}{n-1}\frac{1}{1+\frac{2}{\alpha_{n+1}}} - \frac{n^2}{n-2}\frac{1}{1+\frac{2}{\alpha_n}} < 1. \tag{90}$$

Note that α_n is increasing n, fixing other parameters. So,

$$\alpha_{n+1} - \alpha_n = 1 + \sqrt{n^2 + B} - \sqrt{(n-1)^2 + B} < 1 + \frac{(2n-1)}{2\sqrt{(n-1)^2 + B}} < \frac{4n-3}{2(n-1)}. \tag{91}$$

Using the above inequality, we can show that

$$\frac{1}{1+\frac{2}{\alpha_{n+1}}} - \frac{1}{1+\frac{2}{\alpha_n}} < \frac{1}{(\alpha_n+2)^2}\frac{4n-3}{n-1}. \tag{92}$$

Applying the above inequality, we can show that the left-hand side of (90) satisfies:

$$\frac{(n+1)^2}{n-1}\frac{1}{1+\frac{2}{\alpha_{n+1}}}-\frac{n^2}{n-2}\frac{1}{1+\frac{2}{\alpha_n}}$$
$$<\frac{n^2-3n-2}{(n-1)(n-2)}\frac{\alpha_n}{\alpha_n+2}+\frac{(n+1)^2(4n-3)}{(n-1)^2}\frac{1}{(\alpha_n+2)^2}\equiv\psi(\alpha_n). \tag{93}$$

We have

$$\psi'(\alpha_n)=\frac{2}{(\alpha_n+2)^2}\underbrace{\left[\frac{n^2-3n-2}{(n-1)(n-2)}-\frac{1}{\alpha_n+2}\frac{(n+1)^2(4n-3)}{(n-1)^2}\right]}_{\lambda(\Delta)}. \tag{94}$$

We now fix n and r, and consider changes in α_n through the changes in $\Delta > 0$. For any fixed n and r, α_n and $\lambda(\Delta)$ decrease in Δ. In particular, as $\Delta \to 0$, we have $\alpha_n \to \infty$, and $\lambda(0) > 0$. But as $\Delta \to \infty$, we have $\alpha_n \downarrow 2(n-1)$, and $\lambda(\Delta)$ converges to

$$\frac{n^2-3n-2}{(n-1)(n-2)}-\frac{1}{2n}\frac{(n+1)^2(4n-3)}{(n-1)^2}<0. \tag{95}$$

Therefore, as a function of α_n and in the domain $[2(n-1),\infty)$, $\psi(\alpha_n)$ first decreases in α_n and then increases in α_n. To show that $\psi(\alpha_n) < 1$, it suffices to verify that $\lim_{\alpha_n\to\infty}\psi(\alpha_n) < 1$ and $\psi(2(n-1)) < 1$. As $\alpha_n \to \infty$, the second term of $\psi(\alpha_n)$ vanishes and the first term converges to $\frac{n^2-3n-2}{(n-1)(n-2)} < 1$. At $\alpha_n = 2(n-1)$,

$$\begin{aligned}\psi(2(n-1)) &= \frac{n^2-3n-2}{(n-1)(n-2)}\frac{n-1}{n}+\frac{(n+1)^2(4n-3)}{(n-1)^2}\frac{1}{4n^2}\\ &= 1+\frac{1}{n}\left(\frac{(n+1)^2(4n-3)}{4n(n-1)^2}-\frac{n+2}{n-2}\right)\\ &< 1+\frac{1}{n}\left(\frac{(n+1)^2}{(n-1)^2}-\frac{n+2}{n-2}\right)=1+\frac{1}{n}\left(\frac{4n}{(n-1)^2}-\frac{4}{n-2}\right)<1.\end{aligned}$$

This completes the proof.

A.4 Proof of Proposition 6

Monotonicity of $M^*(m)$, $L(m)$, and $M^*(m)/L(m)$

We first prove the following lemma.

Lemma 1. *If for positive real numbers $\{\lambda_i, \alpha_i, \beta_i : i \geq 0\}$, we have $\frac{\lambda_i}{\lambda_{i+1}} < \frac{\lambda_{i+1}}{\lambda_{i+2}}$ and $\frac{\alpha_i}{\beta_i} < \frac{\alpha_{i+1}}{\beta_{i+1}}$, then for any positive integer k,*

$$\frac{\sum_{i=0}^k \lambda_i\alpha_i}{\sum_{i=0}^k \lambda_i\beta_i}<\frac{\sum_{i=0}^k \lambda_i\alpha_{i+1}}{\sum_{i=0}^k \lambda_i\beta_{i+1}}.$$

Proof. Because $\frac{\alpha_i}{\beta_i} < \frac{\alpha_{i+1}}{\beta_{i+1}}$ for $i \geq 0$, it is easy to see that

$$\frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}} = \frac{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1} \frac{\alpha_{i+1}}{\beta_{i+1}}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}} \leq \frac{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1} \frac{\alpha_k}{\beta_k}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}} = \frac{\alpha_k}{\beta_k},$$

which implies

$$\frac{\alpha_{k+1}}{\beta_{k+1}} > \frac{\alpha_k}{\beta_k} \geq \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}}. \tag{96}$$

From (96) we have

$$\begin{aligned} \frac{\sum_{i=0}^{k} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k} \lambda_i \beta_{i+1}} &= \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1} + \lambda_k \beta_{k+1} \frac{\alpha_{k+1}}{\beta_{k+1}}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1} + \lambda_k \beta_{k+1}} \\ &\geq \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1} + \lambda_k \beta_{k+1} \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1} + \lambda_k \beta_{k+1}} \\ &= \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}}. \end{aligned} \tag{97}$$

Similarly, we can prove that

$$\frac{\sum_{i=0}^{k} \lambda_i \alpha_i}{\sum_{i=0}^{k} \lambda_i \beta_i} \leq \frac{\sum_{i=1}^{k} \lambda_i \alpha_i}{\sum_{i=1}^{k} \lambda_i \beta_i}. \tag{98}$$

Equations (97) and (98) imply that in order to prove Lemma 1 we only need to show that

$$\frac{\sum_{i=1}^{k} \lambda_i \alpha_i}{\sum_{i=1}^{k} \lambda_i \beta_i} < \frac{\sum_{i=0}^{k-1} \lambda_i \alpha_{i+1}}{\sum_{i=0}^{k-1} \lambda_i \beta_{i+1}},$$

which is equivalent to

$$\frac{\sum_{i=1}^{k} \lambda_i \alpha_i}{\sum_{i=1}^{k} \lambda_i \beta_i} < \frac{\sum_{i=1}^{k} \lambda_{i-1} \alpha_i}{\sum_{i=1}^{k} \lambda_{i-1} \beta_i}. \tag{99}$$

Notice that

$$\frac{\sum_{i=1}^{k} \lambda_{i-1} \alpha_i}{\sum_{i=1}^{k} \lambda_{i-1} \beta_i} - \frac{\sum_{i=1}^{k} \lambda_i \alpha_i}{\sum_{i=1}^{k} \lambda_i \beta_i} = \frac{\left(\sum_{i=1}^{k} \lambda_{i-1} \alpha_i\right)\left(\sum_{i=1}^{k} \lambda_i \beta_i\right) - \left(\sum_{i=1}^{k} \lambda_i \alpha_i\right)\left(\sum_{i=1}^{k} \lambda_{i-1} \beta_i\right)}{\left(\sum_{i=1}^{k} \lambda_{i-1} \beta_i\right)\left(\sum_{i=1}^{k} \lambda_i \beta_i\right)}. \tag{100}$$

So it suffices to prove the numerator of (100) is positive. By expansion we have

$$\begin{aligned}
& \left(\sum_{i=1}^{k} \lambda_{i-1}\alpha_i\right)\left(\sum_{i=1}^{k} \lambda_i\beta_i\right) - \left(\sum_{i=1}^{k} \lambda_i\alpha_i\right)\left(\sum_{i=1}^{k} \lambda_{i-1}\beta_i\right) \\
= & \sum_{1\le s<t\le k} (\lambda_{s-1}\alpha_s\lambda_t\beta_t + \lambda_{t-1}\alpha_t\lambda_s\beta_s) + \sum_{i=1}^{k} \lambda_{i-1}\alpha_i\lambda_i\beta_i \\
& - \sum_{1\le s<t\le k} (\lambda_s\alpha_s\lambda_{t-1}\beta_t + \lambda_t\alpha_t\lambda_{s-1}\beta_s) - \sum_{i=1}^{k} \lambda_{i-1}\beta_i\lambda_i\alpha_i \\
= & \sum_{1\le s<t\le k} (\lambda_{s-1}\alpha_s\lambda_t\beta_t + \lambda_{t-1}\alpha_t\lambda_s\beta_s - \lambda_s\alpha_s\lambda_{t-1}\beta_t - \lambda_t\alpha_t\lambda_{s-1}\beta_s)\,. \qquad (101)
\end{aligned}$$

Because, for all $s < t$,

$$\begin{aligned}
& \lambda_{s-1}\alpha_s\lambda_t\beta_t + \lambda_{t-1}\alpha_t\lambda_s\beta_s - \lambda_s\alpha_s\lambda_{t-1}\beta_t - \lambda_t\alpha_t\lambda_{s-1}\beta_s \\
= & \lambda_s\lambda_t\beta_s\beta_t\left(\frac{\lambda_{s-1}}{\lambda_s}\frac{\alpha_s}{\beta_s} + \frac{\lambda_{t-1}}{\lambda_t}\frac{\alpha_t}{\beta_t} - \frac{\lambda_{t-1}}{\lambda_t}\frac{\alpha_s}{\beta_s} - \frac{\lambda_{s-1}}{\lambda_s}\frac{\alpha_t}{\beta_t}\right) \\
= & \lambda_s\lambda_t\beta_s\beta_t\left(\frac{\lambda_{s-1}}{\lambda_s} - \frac{\lambda_{t-1}}{\lambda_t}\right)\left(\frac{\alpha_s}{\beta_s} - \frac{\alpha_t}{\beta_t}\right) \\
> & 0, \qquad (102)
\end{aligned}$$

the right side of (101) is positive, and the proof of the Lemma is complete. □

Letting $k \to \infty$ in Lemma 1, we get

$$\frac{\sum_{i=0}^{\infty} \lambda_i\alpha_i}{\sum_{i=0}^{\infty} \lambda_i\beta_i} \le \frac{\sum_{i=0}^{\infty} \lambda_i\alpha_{i+1}}{\sum_{i=0}^{\infty} \lambda_i\beta_{i+1}}. \qquad (103)$$

Letting $\lambda_i = g(i) = (i+1)q^i(1-q)^2$, we have

$$\frac{\lambda_i}{\lambda_{i+1}} = \frac{(i+1)}{(i+2)q} < \frac{(i+2)}{(i+3)q} = \frac{\lambda_{i+1}}{\lambda_{i+2}}. \qquad (104)$$

Monotonicity of $M^*(m)$. Given Lemma 1, to show that $M^*(m) \le M^*(m+1)$, it suffices to show that

$$\frac{(1-C(n))\frac{n-1}{n^2}}{C(n) + \frac{1-C(n)}{n}} \qquad (105)$$

is increasing in n for $n \ge 2$.

In the continuous-time double-auction market of Appendix B, we have $\Delta = 0$ and $C(n) = 1/(n-1)$, so the ratio (105) simplifies to

$$\frac{n-2}{2n},$$

which is increasing in n.

For $\Delta > 0$, we denote

$$D(n) = 2 - \frac{2(n-2)}{(n-1) + \frac{2e^{-r\Delta}}{1-e^{-r\Delta}} + \sqrt{(n-1)^2 + \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}}}. \tag{106}$$

It is easy to see that $D(n)$ is decreasing in n. Using $C(n) = (1 - 2a_\Delta\gamma/r)/(n-1)$, we have

$$\frac{(1-C(n))\frac{n-1}{n^2}}{C(n) + \frac{1-C(n)}{n}} = \frac{1 - D(n)/n}{D(n)} = \frac{1}{D(n)} - \frac{1}{n}. \tag{107}$$

Since $D(n)$ is decreasing in n, the right-hand side of the above expression is increasing in n, and the proof for the monotonicity of $M^*(m)$ is complete.

Monotonicity of $L(m)$ if $e^{-r\Delta} > 1/2$. Given Lemma 1, to show that $L(m) \leq L(m+1)$, it suffices to show that

$$\frac{\frac{1-C(n)}{n}}{C(n) + \frac{(1-C(n))(3n-2)}{n^2}} \tag{108}$$

is increasing in n for $n \geq 2$.

In the continuous-time double-auction market, with $\Delta = 0$ and $C(n) = 1/(n-1)$, the ratio (108) simplifies to

$$\frac{(n-2)n}{4(n-1)^2},$$

which is indeed increasing in n.

For $\Delta > 0$, we define

$$t = \frac{2e^{-r\Delta}}{1 - e^{-r\Delta}}, \tag{109}$$

and

$$R(n) = \sqrt{(n-1)^2 + \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}} - (n-1). \tag{110}$$

Now we can write

$$C(n) = \frac{1}{n-1}\left(1 - \frac{2(n-2)}{n-1+t+n-1+R(n)}\right). \tag{111}$$

We claim that

$$0 \leq R(n) \leq t, \text{ and } R(n) \text{ decreases in } n. \tag{112}$$

It is obvious that $R(n)$ is non-negative and decreases in n. To see that $R(n) \leq t$, we can directly calculate

$$\frac{\frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}}{\sqrt{(n-1)^2 + \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}} + n - 1} \leq \frac{\frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}}{\sqrt{1 + \frac{4e^{-r\Delta}}{(1-e^{-r\Delta})^2}} + 1} = t.$$

Using (109) and (110), we can write

$$\frac{\frac{(1-C(n))}{(n)}}{\left(C(n)+\frac{(1-C(n))(3n-2)}{(n)^2}\right)} = \frac{(n-2)n(2n+t+R(n))}{2(n-1)(3n^2+2n(t-2)-2t+2(n-1)R(n))}. \tag{113}$$

We denote the numerator and denominator of the right-hand side of (113) by $Y_1(n)$ and $Y_0(n)$, respectively. To show monotonicity, it is enough to prove that

$$Y_1(n+1)Y_0(n) - Y_1(n)Y_0(n+1) > 0.$$

After expansion, we get

$$\begin{aligned} &Y_1(n+1)Y_0(n) - Y_1(n)Y_0(n+1) \\ &\quad = 2(R(n)-R(n+1))n^5 + 2(t-2+R(n+1))n^4 \\ &\qquad + (24-4t-6R(n)+2R(n+1))\, n^3 + 2(6+13t+6R(n)+7R(n+1))n^2 \\ &\qquad +8(-2+t^2+(t-1)R(n+1)+R(n)(t+1+R(n+1)))n \\ &\qquad -4(t+R(n))(t+2+R(n+1)). \end{aligned} \tag{114}$$

Using (112), lower bounds on the coefficients of each term in the polynomial on the right-hand side are as follows:

$$n^5: \quad 2(R(n)-R(n+1)) \geq 0. \tag{115}$$

$$n^4: \quad 2(t-2+R(n+1)) \geq 2(t-2). \tag{116}$$

$$n^3: \quad 24-4t-6R(n)+2R(n+1) \geq 24-10t. \tag{117}$$

$$n^2: \quad 2(6+13t+6R(n)+7R(n+1)) \geq 2(6+13t). \tag{118}$$

$$n: \quad 8(-2+t^2+(t-1)R(n+1)+R(n)(t+1+R(n+1))) \geq 8(t^2-2). \tag{119}$$

$$\text{Constant}: \quad -4(t+R(n))(t+2+R(n+1)) \geq -8t(2t+2). \tag{120}$$

With the above inequalities, we get

$$\begin{aligned} &Y_1(n+1)Y_0(n) - Y_1(n)Y_0(n+1) \qquad (121) \\ &\geq 2(t-2)n^4 + (24-10t)\, n^3 + 2(6+13t)n^2 + 8(t^2-2)n - 8t(2t+2) \\ &= 2(t-2)n^2(n^2-5n+6.5) + 4n^3 + (38+13t)n^2 + 8(t^2-2)n - 8t(2t+2) \\ &= 2(t-2)n^2\left((n-2.5)^2+0.25\right) + 4n^3 + 38n^2 - 16n + t^2(8n-16) + t(13n^2-16). \end{aligned}$$

Under the condition $e^{-r\Delta} > \frac{1}{2}$, we have $t > 2$. It is easy to see the right-hand side of (121) is positive for $n \geq 2$.

Monotonicity of $M^*(m)/L(m)$. We can write

$$\frac{M^*(m)}{L(m)} = \frac{1}{\mu}\frac{\sum_{k=0}^{\infty} g(k)(1-C(n))\frac{n-1}{n^2}}{\sum_{k=0}^{\infty} g(k)\frac{1-C(n)}{n}} \cdot \frac{\sum_{k=0}^{\infty}\left(C(n)+\frac{(1-C(n))(3n-2)}{n^2}\right)}{\sum_{k=0}^{\infty} g(k)\left(C(n)+\frac{1-C(n)}{n}\right)}. \tag{122}$$

Given Lemma 1, to show that $M^*(m)/L(m)$ increases in m, it suffices to show that

$$\frac{(1-C(n))\frac{n-1}{n^2}}{\frac{1-C(n)}{n}} \quad \text{and} \quad \frac{C(n)+\frac{(1-C(n))(3n-2)}{n^2}}{C(n)+\frac{1-C(n)}{n}}$$

are both increasing in n.

Monotonicity in n of the first expression is obvious, for

$$\frac{(1-C(n))\frac{n-1}{n^2}}{\frac{1-C(n)}{n}} = 1-\frac{1}{n}.$$

The second expression can be expressed as

$$\frac{C(n)+\frac{(1-C(n))(3n-2)}{n^2}}{C(n)+\frac{1-C(n)}{n}} = 1+2\frac{(1-C(n))\frac{n-1}{n^2}}{C(n)+\frac{1-C(n)}{n}}. \tag{123}$$

The last term in the above expression is increasing in n, as shown in the proof of monotonicity of $M^*(m)$.

Proof of (55)–(58)

Suppose that $e^{-r\Delta} > 1/2$. We have shown that in this case $M^*(m)$, $L(m)$ and $M^*(m)/L(m)$ are all increasing in m for $m \geq 2$. We now prove (55)–(58) by induction. We let $X_{i,j}$ denote the X element of the state vector (m, X, y) that applies for buyer i and seller j.

At $i = j = 1$ and $X_{1,1} = 0$. Clearly, both thresholds are equal to $M^*(2)$ and are positive at this initial state. Moreover, if the buyer exits, then $X_{2,1} < 0$, and

$$\mathcal{M}_s(3, X_{2,1}) = M^*(3) - L(3)X_{2,1} > M^*(2) = \mathcal{M}_s(2, X_{1,1}). \tag{124}$$

If the seller exits, then $X_{1,2} > 0$, and

$$\mathcal{M}_b(3, X_{1,2}) = M^*(3) + L(3)X_{1,2} > M^*(2) = \mathcal{M}_b(2, X_{1,1}). \tag{125}$$

At generic (i, j) and $X_{i,j}$. By symmetry, it suffices to prove these inequalities for the exit of seller j. By the conjectured update rule,

$$X_{i,j+1} - X_{i,j} = \begin{cases} M^*(i+j) - L(i+j)X_{i,j}, & \text{if seller } j \text{ traded positive quantity} \\ \nu(M^*(i+j) - L(i+j)X_{i,j}), & \text{if seller } j \text{ traded zero quantity} \end{cases},$$

where the last line follows from the induction step

$$\mathcal{M}_s(i+j, X_{i,j}) = M^*(i+j) - L(i+j)X_{i,j} \geq 0.$$

In this case, we want to show that

$$M^*(i+j+1) + L(i+j+1)X_{i,j+1} \geq M^*(i+j) + L(i+j)X_{i,j}, \tag{126}$$
$$M^*(i+j+1) - L(i+j+1)X_{i,j+1} \geq 0. \tag{127}$$

If established, the first inequality (126) would imply that the incumbent buyer's new threshold remains positive if the old threshold is positive. Since it is the seller who exited, the inequality for the "incumbent seller" is irrelevant.

To show (126), we calculate

$$
\begin{aligned}
&M^*(i+j+1) + L(i+j+1)X_{i,j+1} - M^*(i+j) - L(i+j)X_{i,j} \\
&\quad\geq M^*(i+j+1) - M^*(i+j) + (L(i+j+1) - L(i+j))X_{i,j} \\
&\quad\geq M^*(i+j)\frac{L(i+j+1)}{L(i+j)} - M^*(i+j) + (L(i+j+1) - L(i+j))X_{i,j} \\
&\quad= \frac{L(i+j+1) - L(i+j)}{L(i+j)}(M^*(i+j) + L(i+j)X_{i,j}) \geq 0, \qquad (128)
\end{aligned}
$$

where the last inequality follows from the induction step that $\mathcal{M}_b(i+j, X_{i,j}) \geq 0$ and the monotonicity of $L(m)$, and the penultimate inequality follows from the monotonicity of $M^*(m)/L(m)$.

To show (127), we calculate

$$
\begin{aligned}
&M^*(i+j+1) - L(i+j+1)X_{i,j+1} \\
&\quad\geq M^*(i+j+1) - L(i+j+1)\left(X_{i,j} + M^*(i+j) - L(i+j)X_{i,j}\right) \\
&\quad\geq M^*(i+j)\frac{L(i+j+1)}{L(i+j)} - L(i+j+1)M^*(i+j) - L(i+j+1)(1 - L(i+j))X_{i,j} \\
&\quad= \frac{L(i+j+1)(1 - L(i+j))}{L(i+j)}\left(M^*(i+j) - L(i+j)X_{i,j}\right) \geq 0, \qquad (129)
\end{aligned}
$$

where the last inequality follows from the induction step that $\mathcal{M}_s(i+j, X_{i,j}) \geq 0$ and the fact[15] that $L(\,\cdot\,) < 1/2$, and the penultimate inequality follows from the monotonicity of $M^*(m)/L(m)$.

B Continuous-Time Double-Auction Market

This appendix states the continuous-time limit of the discrete-time double auction market, and independently solves for the equilibrium in the corresponding continuous-time double auction market. We thereby show that these two settings have identical equilibrium behavior. Since the periodic inventory shocks after time 0 merely add a constant to a trader's indirect utility at time 0 (see Proposition 2), these shocks do not affect the equilibrium strategies in the workup or double auctions. Consequently, in the calculations below we will avoid introducing inventory shocks after time zero.

[15] The ratio of a pair of terms in the numerator and denominator in the expression of $L(m)$ is

$$
\frac{\frac{1}{n}}{\frac{C(n)}{1-C(n)} + \frac{3n-2}{n^2}} < \frac{\frac{1}{n}}{\frac{3n-2}{n^2}} = \left(3 - \frac{2}{n}\right)^{-1} \leq \frac{1}{2}
$$

for any $n \geq 2$. Thus, $L(m) < 1/2$.

Continuous-time limit of the discrete-time double auction market

Corollary 1. *Suppose that the inventory shocks are zero after time 0. As $\Delta \to 0$, the equilibrium of Proposition 1 converges to the following continuous-time limit.*

1. *The limit demand schedule*[16] *of trader i at time t is*

$$x_{it}^{\infty}(p; z_{it}^{\infty}) = a^{\infty}\left(v - p - \frac{2\gamma}{r} z_{it}^{\infty}\right), \tag{130}$$

where

$$a^{\infty} = \frac{(n-2)r^2}{4\gamma} \tag{131}$$

and where the limiting inventory position of trader i at time t is

$$z_{it}^{\infty} = \frac{Z_t}{n} + e^{-(n-2)rt/2}\left(z_{i0} - \frac{Z}{n}\right). \tag{132}$$

The equilibrium price at time t is

$$p^* = v - \frac{2\gamma}{nr} Z. \tag{133}$$

2. *The limiting expected net payoff of trader i at time 0, conditional on z_{i0} and the initial auction price p^*, is*

$$V_{i,0+}^{\infty} = v\frac{Z}{n} - \frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 + \left(v - \frac{2\gamma}{r}\frac{Z}{n}\right)\left(z_{i0} - \frac{Z}{n}\right) - \frac{\gamma}{r(n-1)}\left(z_{i0} - \frac{Z}{n}\right)^2. \tag{134}$$

Proof. The only nontrivial part of the proof is the limit of the convergence rate. Because $1 - a_{\Delta} 2\gamma/r$ is the convergence factor per auction period, the associated convergence factor per unit of time is

$$\left(1 - a_{\Delta}\frac{2\gamma}{r}\right)^{1/\Delta}.$$

Here, we ignore the effect of partial integer periods per unit of time, which is irrelevant in the limit as Δ goes to zero. Finally, we have the limiting convergence rate

$$\lim_{\Delta \to 0} \frac{\log(1 - a_{\Delta}\frac{2\gamma}{r})}{\Delta} = -\lim_{\Delta \to 0} \frac{a_{\Delta}\frac{2\gamma}{r}}{\Delta} = -\frac{(n-2)r}{2}. \tag{135}$$

□

[16] In a continuous-time setting, a demand schedule at time t can be expressed by a demand "rate function" $D_t(\,\cdot\,)$, which means that if the time path of prices is given by some function $\phi : [0, \infty) \to \mathbb{R}$, then the associated cumulative total quantity purchased by time t is $\int_0^t D_s(\phi(s))\, ds$, whenever the integral is well defined. In our case, the discrete-period demand schedule $x_{ik}(\,\cdot\,; z_{ik})$ has the indicated limit demand schedule, as a demand rate function, because $z_{i,K(t)} \to z_{it}^{\infty}$ and because, for any fixed price p and fixed inventory level z,

$$\lim_{\Delta \downarrow 0} \frac{a_{\Delta}}{\Delta}\left(v - p - \frac{2\gamma}{r} z\right) = a^{\infty}\left(v - p - \frac{2\gamma}{r} z\right).$$

Continuous-time double auction market

We fix a probability space and the time domain $[0, \infty)$. The setup for the joint distribution of the exponential payoff time T, the payoff π of the asset, and the initial inventories $z_0 = (z_{10}, z_{20}, \ldots, z_{n0})$ of the $n \geq 3$ of traders is precisely the same as that for the discrete-time auction model of Section 3. The initial information structure is also as in Section 3. In our application of this model in Section 5, the number n of traders is an outcome of the random workup population size $N_b + N_s$. The outcome of $N_b + N_s$ is publicly known when workup is complete. So, it is enough to solve the continuous-time auction model for any fixed trader population size n.

In our new continuous-time setting, a demand schedule at time can be expressed by a demand "rate function" $D : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, representing the rate of demand $D_t(p)$ of asset per unit of time at time t and at price p. This means that if the time path of prices is given in some state of the world by some function $\phi : [0, \infty) \to \mathbb{R}$, then by time t the cumulative quantity purchased is $\int_0^t D_s(\phi(s))\, ds$ and the total price paid is $\int_0^t \phi(s) D_s(\phi(s))\, ds$, whenever these integrals are well defined.

We will consider an equilibrium in which demand $D_{it}(p)$ of trader i at time t and price p is continuous in both t and p and strictly decreasing in p, and such that the market clearing price $\phi(t)$ at time t, when well defined, is the solution in p of the market-clearing equation:

$$\sum_i D_{it}(p) = 0. \tag{136}$$

This market clearing price, when well defined, is denoted $\Phi(\sum_i D_{it})$.

An equilibrium is a collection $(D_1, \ldots, D_n)$ of demand functions such that, for each time t the market-clearing price $\Phi(\sum_i D_{it})$ is well defined and such that, for agent i, the demand function D_i solves the problem, taking $D_{-i} = \sum_{j \neq i} D_j$ as given,

$$\sup_D E\left[z_i^D(T)\pi - \int_0^T \left[\gamma z_i^D(t)^2 + D_t\left[\Phi\left(D_t + D_{-i,t}\right)\right] \Phi\left(D_t + D_{-it}\right)\right]\, dt \right], \tag{137}$$

where $\gamma > 0$ is a holding-cost parameter and

$$z_i^D(t) = z_{i0} + \int_0^t D_s\left[\Phi\left(D_s + D_{-i,s}\right)\right]\, ds$$

is the inventory of agent i at time t.

We will look for an equilibrium in which the initial price $p(0)$ instantly reveals the total market supply Z and in which the demand function of trader i depends only his current inventory z_{it}.

We will conjecture and verify the equilibrium given by

$$D_{it}(p) = a\left(v - p - \frac{2\gamma}{r} z_{it}\right), \tag{138}$$

where

$$a = \frac{(n-2)r^2}{4\gamma}. \tag{139}$$

The unique associated market-clearing price at any time t is

$$p^* = v - \frac{2\gamma}{nr} Z. \tag{140}$$

From this, the inventory position of trader i at time t can be calculated as

$$z_{it} = \frac{Z}{n} + e^{-(n-2)rt/2} \left(z_{i0} - \frac{Z}{n} \right). \tag{141}$$

Given this conjectured equilibrium, for any agent i, the sum $D_{-i,t}$ of the demand functions of the other agents at time t is

$$D_{-i,t}(p) = \mathcal{D}_{-i}(p; z_{it}, Z) \equiv \sum_{j \neq i} a \left(v - p - \frac{2\gamma}{r} z_{jt} \right) = (n-1)a(v-p) - \frac{2a\gamma}{r}(Z - z_{it}).$$

Based on this calculation, the continuation utility $V(z)$ for the inventory level of any trader at any time $t < T$ who has the inventory level z satisfies the Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{aligned} 0 = \sup_D \big[&-D(\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z))\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z)) + V'(z) D(\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z)) \big] \\ &- \gamma z^2 + r(vz - V(z)). \end{aligned} \tag{142}$$

The first term on the right-hand side of (142) is the rate of cost of acquiring inventory in auctions, that is, the quantity rate $D(\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z))$ multiplied by the price $\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z))$. The second term is the marginal value $V'(z)$ of inventory multiplied by the rate $D(\Phi(D + \mathcal{D}_{-i}(\,\cdot\,; z, Z))$ of inventory accumulation. The sum of these first two terms is optimized by choosing some demand function D. The next term accounts for the rate of inventory holding cost, γz^2. The final term is the product of the mean rate r of arrival of the time of the asset payoff and the expected change $vz - V(z)$ in the trader's indirect utility if that payoff were to occur immediately.

Because Z is constant and observable after time 0, the HJB equation does not pin down a unique optimizing demand function $D(\,\cdot\,)$. Instead, the HJB equation makes the demand problem for agent i equivalent to picking the quantity x the agent wishes to buy, and then submitting any demand function $D(\,\cdot\,)$ with the property that $D(p) = x$, where p solves $x + \mathcal{D}_{-i}(p; z, Z) = 0$. In order to avoid degenerate behavior of this type, we require that the submitted demand function $D_i(\,\cdot\,)$ must depend only on the inventory z_{it} trader i and of course the price p. That is, we require that $D_{it}(p) = f_t(p, z_{it})$ for some function $f_t : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Nevertheless, in equilibrium, the resulting demand will turn out to be optimal even if the class of demand functions is expanded to allow dependence on Z.

We will conjecture and verify that, in equilibrium,

$$V(z) = v\frac{Z}{n} - \frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 + \left(v - 2\frac{\gamma}{r}\frac{Z}{n}\right)\left(z - \frac{Z}{n}\right) - \frac{\gamma}{r}\frac{1}{n-1}\left(z - \frac{Z}{n}\right)^2. \tag{143}$$

We use the fact that V is quadratic and concave, thus bounded above.

Proposition 8. *Suppose, for a given trader i, that the demand function D_j for any trader $j \neq i$ is given by (138). The function V given by (143) satisfies the HJB equation (142). Given this choice for V, the optimization problem posed within the HJB equation* *is satisfied by the demand function D_{it} of (138). The optimal demand problem (137) for agent i is also solved by (138).*

This result is shown as follows. With (143), the HJB equation, applied to agent i at time t, is equivalent to solving, for each outcome of Z, the optimal demand

$$\sup_x \left[-x\mathcal{D}_{-i}^{-1}(-x; z_{it}, Z) + V'(z_{it})x\right], \tag{144}$$

where $\mathcal{D}_{-i}^{-1}(q; z_{it}, Z)$ is the inverse total demand of the other agents at any quantity q, meaning that price p for which

$$q = (n-1)a(v-p) - \frac{2a\gamma}{r}(Z - z_{it}).$$

Solving,

$$\mathcal{D}_{-i}^{-1}(q; z_{it}, Z) = v - \frac{1}{a(n-1)}\left[q + \frac{2a\gamma}{r}(Z - z_{it})\right].$$

Thus, the demand problem of agent i is

$$\sup_x \left[-x\left(v - \frac{1}{a(n-1)}\left[-x + \frac{2a\gamma}{r}(Z - z_{it})\right]\right) + V'(z_{it})x\right]. \tag{145}$$

The first-order necessary condition for optimality of x^* is

$$-v + \frac{2\gamma}{r(n-1)}(Z - z_{it}) + V'(z_{it}) - 2x^*\frac{1}{a(n-1)} = 0,$$

where

$$V'(z_{it}) = v - 2\frac{\gamma}{r}\frac{Z}{n} - 2\frac{\gamma}{r}\frac{1}{n-1}\left(z_{it} - \frac{Z}{n}\right).$$

The unique solution x^* of this first-order condition also satisfies the second-order sufficiency condition, and is given by

$$x^* = \frac{2a\gamma}{r}\left(\frac{Z}{n} - z_{it}\right).$$

The associated market clearing price is

$$p^* = \mathcal{D}_{-i}^{-1}(-x^*; z_{it}, Z) = v - \frac{1}{a(n-1)}\left[-x^* + \frac{2a\gamma}{r}(Z - z_{it})\right] = v - 2\frac{\gamma}{r}\frac{Z}{n}. \tag{146}$$

We now verify that the postulated demand function D_{it} for agent i achieves the above demand x^*, regardless of the outcome of Z. We have

$$D_{it}(p^*) = a\left(v - p^* - \frac{2\gamma}{r}z_{it}\right) = a\left(v - \left(v - 2\frac{\gamma}{r}\frac{Z}{n}\right) - \frac{2\gamma}{r}z_{it}\right) = \frac{2a\gamma}{r}\left(\frac{Z}{n} - z_{it}\right),$$

which is indeed equal to the optimal demand x^*.

In order to prove that the proposed indirect utility function V satisfies the HJB equation, we substitute our expressions for $V(z)$, p^*, and $D_{it}(p^*)$ into the right-hand-side of the HJB equation (142). To confirm that (142) is satisfied, we must show that for all real

z and Z,

$$0 = -\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)\left(v-2\frac{\gamma}{r}\frac{Z}{n}\right)+V'(z)\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)+r(vz-V(z))-\gamma z^2. \quad (147)$$

To see that (147) holds, note that

$$V'(z)\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right) = v\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)-2\frac{\gamma}{r}\frac{Z}{n}\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)+2\frac{\gamma}{r}\frac{1}{n-1}\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)^2$$

and that

$$\begin{aligned} r(vz-V(z)) &= rvz - rv\frac{Z}{n}+r\frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 - r\left(v-\frac{2\gamma}{r}\frac{Z}{n}\right)\left(z-\frac{Z}{n}\right)+r\frac{\gamma}{r(n-1)}\left(z-\frac{Z}{n}\right)^2 \\ &= r\frac{\gamma}{r}\left(\frac{Z}{n}\right)^2+r\frac{2\gamma}{r}\frac{Z}{n}\left(z-\frac{Z}{n}\right)+r\frac{\gamma}{r(n-1)}\left(z-\frac{Z}{n}\right)^2. \end{aligned} \quad (148)$$

The right-hand side of (147) is thus computed as

$$\begin{aligned} &-\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)\left(v-2\frac{\gamma}{r}\frac{Z}{n}\right)+V'(z)\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)+r(vz-V(z))-\gamma z^2 \\ &= \frac{4a\gamma}{r}\left(\frac{Z}{n}-z\right)\frac{\gamma}{r}\frac{Z}{n}+-2\frac{\gamma}{r}\frac{Z}{n}\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)+2\frac{\gamma}{r}\frac{1}{n-1}\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)^2 \\ &\quad +\gamma\left(\frac{Z}{n}\right)^2+2\gamma\frac{Z}{n}\left(z-\frac{Z}{n}\right)+\frac{\gamma}{n-1}\left(z-\frac{Z}{n}\right)^2-\gamma z^2. \end{aligned}$$

Substituting $a=(n-2)r^2/4\gamma$, we have

$$\begin{aligned} &-\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)\left(v-2\frac{\gamma}{r}\frac{Z}{n}\right)+V'(z)\frac{2a\gamma}{r}\left(\frac{Z}{n}-z\right)+r(vz-V(z))-\gamma z^2 \\ &= (n-2)\gamma\left(\frac{Z}{n}-z\right)\frac{Z}{n}+-(n-2)\gamma\frac{Z}{n}\left(\frac{Z}{n}-z\right)+\frac{(n-2)\gamma}{n-1}\left(\frac{Z}{n}-z\right)^2 \\ &\quad +\gamma\left(\frac{Z}{n}\right)^2+2\gamma\frac{Z}{n}\left(z-\frac{Z}{n}\right)+\frac{\gamma}{n-1}\left(z-\frac{Z}{n}\right)^2-\gamma z^2. \end{aligned}$$

So, V satisfies the HJB equation because

$$\frac{(n-2)}{n-1}\left(\frac{Z}{n}-z\right)^2+\left(\frac{Z}{n}\right)^2+2\frac{Z}{n}\left(z-\frac{Z}{n}\right)+\frac{1}{n-1}\left(z-\frac{Z}{n}\right)^2-z^2=0.$$

Thus, using the fact that the demand function D_{it} solves the maximization problem of the HJB equation, and using the fact that V solves the HJB equation, an application of Ito's formula to the process J defined by $J(t)=V(z_{it})$ for $t<T$, and by $J(t)=\pi z_{iT}$ for $t \geq T$ implies that

$$V(z_{i0}) = E\left[z_{iT}(T)\pi - \int_0^T \left[\gamma z_{it}^2 + D_{it}\left[\Phi\left(D_{it}+D_{-i,t}\right)\right]\Phi\left(D_{it}+D_{-it}\right)\right]\,dt\right].$$

For any other demand function D for agent i, the HJB equation and Ito's formula implies that implies that

$$V(z_{i0}) \geq E\left[z_i^D(T)\pi - \int_0^T \left[\gamma z_i^D(t)^2 + D_t \left[\Phi\left(D_t + D_{-i,t}\right)\right] \Phi\left(D_t + D_{-it}\right)\right] \, dt\right].$$

Thus D_i is indeed optimal for trader i given D_{-i}, and $V(z)$ is indeed the indirect utility of any agent with inventory z. This proves Proposition 8.

We can now recapitulate the main results of this section:

Proposition 9. *An equilibrium of the continuous-time double-auction market is as follows.*

1. *The demand function D_{it} of trader i at time t is given by:*

$$D_{it} = \frac{(n-2)r^2}{4\gamma}\left(v - p - \frac{2\gamma}{r} z_{it}\right), \tag{149}$$

where the equilibrium inventory of trader i at time t is

$$z_{it} = \frac{Z}{n} + e^{-(n-2)rt/2}\left(z_{i0} - \frac{Z}{n}\right). \tag{150}$$

The equilibrium price at time t is constant at

$$p^* = v - \frac{2\gamma}{nr} Z. \tag{151}$$

2. *The indirect utility $V(z)$ of any agent i for inventory z at any time $t > 0$ that is before the asset payoff time T is given by*

$$V(z) = v\frac{Z}{n} - \frac{\gamma}{r}\left(\frac{Z}{n}\right)^2 + \left(v - \frac{2\gamma}{r}\frac{Z}{n}\right)\left(z - \frac{Z}{n}\right) - \frac{\gamma}{r(n-1)}\left(z - \frac{Z}{n}\right)^2. \tag{152}$$

C Welfare and Squared Asset Dispersion

A reallocation of the inventory vector $(z_{10}, \ldots, z_{in})$ is an allocation $z' = (z'_1, \ldots, z'_n)$ with the same total Z. A reallocation z' is a Pareto improvement if, when replacing z_{i0} with z'_i, the equilibrium utility $E(V_{i,0+})$ before entering the sequential-double-auction market is weakly increased for every i and strictly increased for some i. We have the following corollary of Proposition 2.

Corollary 2. *The total expected ex-ante utility $W(z_0) = \sum_{i=1}^n E(V_{i,0+})$ is one-to-one and strictly monotone decreasing (in fact linear) in the sum of mean squared excess asset positions,*

$$D(z_0) = E\left(\sum_{i=1}^n \left(z_{i0} - \frac{Z}{n}\right)^2\right).$$

Thus, if a reallocation $z' = (z'_1, \ldots, z'_n)$ is a Pareto improvement, then $D(z') < D(z_0)$.

This result follows from the fact that $W(z_0)$ is a constant plus the product of $D(z_0)$ and a negative constant.

Because traders' preferences are linear with respect to total net pecuniary benefits, $W(\cdot)$ is a reasonable social welfare function. This follows from the fact that for any allocations z' and z with $W(z') > W(z)$, the allocation z' is Pareto preferred to z after allowing for transfer payments.

The magnitude of welfare improvement offered by the bilateral workup, conditional on the double auctions, can be calculated explicitly. We focus on the welfare of the buyer and the seller in the bilateral workup under consideration, and assume zero inventory shocks after time 0. Start from any pre-workup inventory levels $-x < 0$ for the buyer and $y > 0$ for the seller, where x and y are exponentially distributed with mean $1/\mu$. The workup volume is $V \equiv \max(0, \min(x - (M + \delta), y - (M - \delta)))$, and the post-workup inventories are $-x + V$ and $y - V$. By Corollary 2, the ex-ante welfare improvement induced by a single bilateral workup is proportional to (with multiplier $(\gamma/r)C$, by (13)) the reduction in total mean-squared inventory dispersion for the buyer-seller workup pair, which is

$$\begin{aligned} &E\left[(-x - Z/n)^2 + (y - Z/n)^2 - (-x + V - Z/n)^2 - (y - V - Z/n)^2\right] \\ &\quad = E\left[-2V^2 + 2(x + y)V\right] \\ &\quad = \int_{x=M+\delta}^{\infty} \int_{y=M-\delta}^{\infty} \mu e^{-\mu x} \mu e^{-\mu y} (-2V^2 + 2(x + y)V)\, dx\, dy. \end{aligned} \tag{153}$$

A change of variables, taking $u = x - M - \delta$ and $w = y - M + \delta$, allows one to re-express the integral as

$$\int_{u=0}^{\infty} \int_{w=0}^{\infty} \mu^2 e^{-2\mu M} e^{-\mu(u+w)} \left[-2\min(u, w)^2 + 2(u + w + 2M)\min(u, w)\right] du\, dw. \tag{154}$$

Further simplification reduces this integral to

$$\frac{2e^{-2M\mu}(1 + M\mu)}{\mu^2}, \tag{155}$$

which is decreasing in M and invariant to δ in the interval $[0, M]$.

On the other hand, without the workup, the expected welfare cost of the buyer and the seller that arises from strategic avoidance of price impact is proportional to (also with multiplier $(\gamma/r)C$, by (13))

$$\begin{aligned} &E[(-x - Z/n)^2 + (y - Z/n)^2] \\ &= E\left[\left(-\frac{n-1}{n}x - \frac{1}{n}y - \frac{1}{n}\sum_{i=3}^{n} z_{i0}\right)^2 + \left(\frac{1}{n}x + \frac{n-1}{n}y - \frac{1}{n}\sum_{i=3}^{n} z_{i0}\right)^2\right] \\ &= \frac{n-1}{n}\frac{4}{\mu^2}, \end{aligned} \tag{156}$$

where we use the facts that $E[x^2] = E[y^2] = E[z_{i0}^2] = 2/\mu^2$ and that $(x, y, \{z_{i0}\})$ are mutually independent.

Therefore, the fraction of welfare cost between the buyer and the seller that is eliminated by the bilateral workup is

$$R = \frac{n}{2(n-1)} e^{-2M\mu}(1 + M\mu). \tag{157}$$

D Comparing Various Market Structures

In the main body of the paper we have shown that adding a single workup before the sequential double auction market improves allocative efficiency. In this appendix, we solve an alternative market structure with only a size discovery mechanism—bilateral workups—at time 0. This size-discovery-only market presents an interesting tradeoff: the lack of future trading opportunities rules out after-workup inventory rebalancing, but it also encourages traders to work up more quantities during the workup. For simplicity, we focus on the case with unbiased workup price, i.e., $\bar{p} = v$. We will also restrict attention to zero periodic inventory shocks, since the presence of inventory shocks do not change traders' strategies (see Section 3).

No trading at all. It is easy to see that without any trading, the welfare of trader i is given by an expression similar to (13), except that $\Theta = 0$ and the penultimate term is

$$-\frac{\gamma}{r}\left(z_{i0} - \frac{Z}{n}\right)^2, \tag{158}$$

rather than

$$-\frac{\gamma}{r}\frac{1 - 2a_\Delta \gamma / r}{n-1}\left(z_{i0} - \frac{Z}{n}\right)^2. \tag{159}$$

This is our benchmark.

Only workup. If there is only a single workup and no double auctions, there would be no price discovery. This means that the total inventory Z is never disclosed or anyone. In this case, a trader's expected utility after the workup, from holding inventory z, is

$$\mathcal{V}(z) = vz - \frac{\gamma}{r}z^2. \tag{160}$$

In a bilateral workup, the buyer's utility is simply

$$U^b = -\bar{p}y + \mathcal{V}(S^b + y) = vS^b - \frac{\gamma}{r}(S^b + y)^2, \tag{161}$$

where we have used $\bar{p} = v$. Note that the total inventory Z is irrelevant here because the buyer has no further opportunities to trade. Taking the first-order condition with respect to y and equating it to zero at $S^b + y = -M_b$, we get

$$M_b = 0. \tag{162}$$

By a symmetric calculation, the seller's dropout threshold is

$$M_s = 0. \tag{163}$$

Zero dropout thresholds imply that self-rationing we saw in Section 4 does not happen here, because there are no further trading opportunities. Thus, in this alternative structure, everyone who receives any inventory shock wishes to participate in bilateral workups.

Next, we calculate the welfare improvement of having a single workup, relative to the no-trade benchmark. If the buyer's inventory size is $x > 0$ and the seller's inventory size is $y > 0$, then the workup volume is $V = \min(x, y)$. The improvement in allocative efficiency (relative to the no-trading benchmark) for this pair is

$$E[x^2 + y^2 - (-x + V)^2 - (y - V)^2]. \tag{164}$$

Inspecting (153) of Appendix C, we see that the above expectation can be simplified by taking the special case of $Z = 0$ and $M = \delta = 0$ in (153), and the expectation simplifies to $2/\mu^2$. Since the coefficient in front of the squared inventory is γ/r, the improvement in allocative efficiency for each buyer-seller pair is

$$\frac{\gamma}{r}\frac{2}{\mu^2}. \tag{165}$$

Since there are n traders, we have at most $\lfloor n/2 \rfloor$ buyer-seller pairs. Thus, if bilateral workups are the only opportunities to trade, the expected efficiency improvement, relative to the no-trade benchmark, is at most

$$\frac{\gamma}{r}\frac{2}{\mu^2}\lfloor \frac{n}{2} \rfloor. \tag{166}$$

Only double auctions. If we only add the double auctions (and no workup), the efficiency improvement (relative to the no-trading benchmark) is

$$\frac{\gamma}{r}\left(1 - \frac{1 - 2a_\Delta \gamma/r}{n-1}\right) E\left[\sum_i \left(z_{i0} - \frac{Z}{n}\right)^2\right] = \frac{\gamma}{r}\left(1 - \frac{1 - 2a_\Delta \gamma/r}{n-1}\right)(n-1)\frac{2}{\mu^2}, \tag{167}$$

where $\frac{2}{\mu^2}$ is the variance of each z_{i0} term. Since $a_\Delta \geq 0$, the efficiency improvement brought by the double auctions has a lower bound of

$$\frac{\gamma}{r}\frac{2}{\mu^2}(n-2). \tag{168}$$

Comparison among market structures. Clearly, for all $n \geq 3$, $n - 2 \geq \lfloor \frac{n}{2} \rfloor$. Thus, a market structure with only double auctions weakly dominates the market structure with only a single workup. This inequality is actually strict because in expectation, n traders generate fewer than $\lfloor n/2 \rfloor$ bilateral workup pairs. Thus, we have the following ranking of expected welfare:

$$\text{workup + double auctions } \succeq \text{double auctions only } \succeq \text{ workup only}. \tag{169}$$

References

Adrian, T., M. Fleming, O. Shachar, and E. Vogt (2015): "Has U.S. Corporate Bond Market Liquidity Deteriorated?" *Liberty Street Economics, Federal Reserve Bank of New York*.

Arrow, K. (1951): "An Extension of the Basic Theorems of Classical Welfare Economics," in *Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, Berkeley: University of California Press, 507–532.

——— (1979): "The Property Rights Doctrine and Demand Revelation Under Incomplete Information," in *Economics and Human Welfare*, ed. by M. Boskin, Academic Press.

Aumann, R. (1964): "Markets with a Continuum of Traders," *Econometrica*, 32, 39–50.

Ausubel, L. M., P. Cramton, M. Pycia, M. Rostek, and M. Weretka (2014): "Demand Reduction, Inefficiency and Revenues in Multi-Unit Auctions," *Review of Economic Studies*, 81, 1366–1400.

BGC (2015): "BGC Derivative Markets, L.P. Rules," Tech. rep.

Boni, L. and C. Leach (2004): "Expandable Limit Order Markets," *Journal of Financial Markets*, 7, 145–185.

Boni, L. and J. C. Leach (2002): "Supply Contraction and Trading Protocol: An Examination of Recent Changes in the U.S. Treasury Market," *Journal of Money, Credit, and Banking*, 34, 740–762.

Buti, S., B. Rindi, and I. M. Werner (2011): "Diving into Dark Pools," Working paper, Fisher College of Business, Ohio State University.

——— (2015): "Dark Pool Trading Strategies, Market Quality and Welfare," Working paper, University of Toronto.

Collin-Dufresne, P., B. Junge, and A. B. Trolle (2016): "Market Structure and Transaction Costs of Index CDSs," Tech. rep., EPFL (to appear).

d'Aspremont, C. and L. Gérard-Varet (1979): "Incentives and Incomplete Information," *Journal of Public Economics*, 11, 25–45.

Degryse, H., M. Van Achter, and G. Wuyts (2009): "Dynamic Order Submission Strategies with Competition between a Dealer Market and a Crossing Network," *Journal of Financial Economics*, 91, 319–338.

Du, S. and H. Zhu (2015): "Welfare and Optimal Trading Frequency in Dynamic Double Auctions," NBER working paper 20588.

Duffie, D. (2010): "Presidential Address: Asset Price Dynamics with Slow-Moving Capital," *Journal of Finance*, 65, 1237–1267.

Dungey, Mardi, Ólan. H. and M. McKenzie (2013): "Modelling Trade Duration in U.S. Treasury Markets," *Quantitative Finance*, 13, 1431–1442.

Fleming, M. and G. Nguyen (2015): "Order Flow Segmentation and the Role of Dark Trading in the Price Discovery of U.S. Treasury Securities," Working paper, Federal Reserve Bank of New York.

Fleming, M., E. Schaumburg, and R. Yang (2015): "The Evolution of Workups in the U.S. Treasury Securities Market," Liberty Street Economics, Federal Reserve Bank of New York.

GFI (2015): "GFI Swaps Exchange LLC Rulebook," Tech. rep.

Giancarlo, J. C. (2015): "Pro-Reform Reconsideration of the CFTC Swaps Trading Rules: Return to Dodd-Frank," White paper.

Hendershott, T. and H. Mendelson (2000): "Crossing Networks and Dealer Markets: Competition and Performance," *Journal of Finance*, 55, 2071–2115.

Huang, R. D., J. Cai, and X. Wang (2002): "Information-Based Trading in the Treasury Note Interdealer Broker Market," *Journal of Financial Intermediation*, 11, 269–296.

Joint CFTC-SEC Advisory Committee (2011): "Recommendations Regarding Regulatory Responses to the Market Event of May 6, 2010," Summary Report, Securities Exchange Commission and Commodity Futures Trading Commission, Washington, D.C.

Joint Staff Report (2015): "The U.S. Treasury Market on October 15, 2014," Tech. rep., U.S. Department of the Treasury, Board of Governors of the Federal Reserve System, Federal Reserve Bank of New York, U.S. Securities and Exchange Commission, and U.S. Commodity Futures Trading Commission, Washington, D.C.

Klemperer, P. D. and M. A. Meyer (1989): "Supply Function Equilibria in Oligopoly under Uncertainty," *Econometrica*, 57, 1243–1277.

Kyle, A. S. (1989): "Informed Speculation with Imperfect Competition," *Review of Economic Studies*, 56, 317–355.

Menkveld, A. J., B. Z. Yueshen, and H. Zhu (2015): "Shades of Darkness: A Pecking Order of Trading Venues," Working paper, MIT.

Myerson, R. and P. Reny (2015): "Sequential Equilibria of Multi-Stage Games with Infinite Sets of Types and Actions," Working paper, University of Chicago.

Pancs, R. (2014): "Workup," *Review of Economic Design*, 18, 37–71.

Ready, M. J. (2014): "Determinants of Volume in Dark Pool Crossing Networks," Working paper, University of Wisconsin-Madison.

Rostek, M. and M. Weretka (2012): "Price Inference in Small Markets," *Econometrica*, 80, 687–711.

——— (2015): "Dynamic Thin Markets," *Review of Financial Studies*, 28, 2946–2992.

Schaumburg, E. and R. Yang (2016): "The Workup, Technology, and Price Discovery in the Interdealer Market for U.S. Treasury Securities," Liberty Street Economics, Federal Reserve Bank of New York.

Securities and Exchange Commission (2010): "Concept Release on Equity Market Structure; Proposed Rule," *Federal Register*, 75, 3593–3614.

SIFMA (2016): "SIFMA Electronic Bond Trading Report: US Corporate & Municipal Securities," Securities Industry and Financial Markets Association, Technical Report.

Tradeweb (2014): "Market Regulation Advisory Notice – Work-Up Protocol," Tech. rep.

Tradition (2015): "Tradition SEF Platform Supplement," Tech. rep.

U.S. Department of the Treasury (2016): "Notice Seeking Public Comment on the Evolution of the Treasury Market Structure," Docket No. TREAS-DO-2015-0013, Department of the Treasury, Washington, D.C.

Vayanos, D. (1999): "Strategic Trading and Welfare in a Dynamic Market," *The Review of Economic Studies*, 66, 219–254.

Vives, X. (2011): "Strategic Supply Function Competition with Private Information," *Econometrica*, 79, 1919–1966.

Wholesale Markets Brokers' Association (2012): "Comment for Proposed Rule 77 FR 38229," Tech. rep.

Wilson, R. (1979): "Auctions of Shares," *Quarterly Journal of Economics*, 93, 675–689.

Zhu, H. (2014): "Do Dark Pools Harm Price Discovery?" *Review of Financial Studies*, 27, 747–789.