Do high-frequency market makers share risks?

Corey Garriott, Vincent van Kervel, and Marius Zoican*

Abstract

We document low cross-sectional correlations between high-frequency market maker (MM) inventory positions, suggesting poor risk sharing. Using a unique data set on Canadian futures markets, a simple inventory cost estimate is 300% above the optimal benchmark. Our model explains how heterogeneity in MM inventories impacts their liquidity provision in time-priority markets. Depth is maximized if large inventory MMs arrive later in the queue. Random queue positions allow us to cleanly distinguish between inventory frictions and adverse selection using quote sizes. In our sample, both frictions are equally large and imperfect risk sharing explain up to 8% of depth variability.

Keywords: limit order markets, time priority, risk sharing, adverse selection, inventory risk
JEL Codes: G11, G12, G14

*Corey Garriott (corey.garriott@tmx.com) is affiliated with the Montréal Exchange (part of the TMX Group). Vincent van Kervel (vivankervel@uc.cl) is affiliated with Pontifica Universidad Católica de Chile. Marius Zoican (marius.zoican@rotman.utoronto.ca) is affiliated with University of Toronto Mississauga and Rotman School of Management. We have greatly benefited from discussion of this research with Sean Foley, Mina Lee, Maureen O'Hara, Andriy Shkilko, Patrik Såndas (discussant), Elvira Sojli, Tālis Putniņš, and Semih Üslü. We are grateful to conference participants at the 2022 Western Finance Association, 2022 Northern Finance Association, 2022 Mid-Atlantic Research Conference, as well as to seminar participants at the TMX Group, Microstructure Exchange (Asia-Pacific) and Tianjin University. We thank the Montréal Exchange and the TMX Group for data support. Marius Zoican gratefully acknowledges funding support from the Canadian Securities Institute Research Foundation and the Canadian Social Sciences and Humanities Research Council (SSHRC) through an Insight Development research grant (430-2020-00014).

Do high-frequency market makers share risks?

Abstract

We document low cross-sectional correlations between high-frequency market maker (MM) inventory positions, suggesting poor risk sharing. Using a unique data set on Canadian futures markets, a simple inventory cost estimate is 300% above the optimal benchmark. Our model explains how heterogeneity in MM inventories impacts their liquidity provision in time-priority markets. Depth is maximized if large inventory MMs arrive later in the queue. Random queue positions allow us to cleanly distinguish between inventory frictions and adverse selection using quote sizes. In our sample, both frictions are equally large and imperfect risk sharing explain up to 8% of depth variability.

Keywords: limit order markets, time priority, risk sharing, adverse selection, inventory risk
JEL Codes: G11, G12, G14

1 Introduction

Modern exchanges are typically organized as limit order books with price and time priority rules (Parlour and Seppi, 2008; O'Hara, 2015). Time priority implies that limit orders on a given price level execute sequentially, with older orders executing before newer ones. A consequence of sequential limit order execution is that it prevents perfect risk sharing among market makers. The market maker at the front of the queue is the first to trade against any incoming market orders, while market markers further back in the queue only trade after the queue has cleared. The classical literature starts from the premise that the market making sector uniformly absorbs the incoming order flow, implying strongly correlated inventories across market makers (e.g., Grossman and Miller, 1988; Reiss and Werner, 1998; Hansch, Naik, and Viswanathan, 1998). In contrast, modern exchanges with time priority mechanically generate inventory divergence across market makers, at least at high-frequencies.

How well do high-frequency market-making algorithms share risks on fast-paced modern markets with time priority? Unlike the human-driven markets of past decades, modern exchanges are anonymous and dominated by fast-paced algorithms. This limits the potential to directly share risks in real time. Indeed, for Canadian futures markets, we find that trading between high-frequency market makers accounts for only 12.3% of volume, despite HFT market-makers being involved in 76% of the trades. For comparison, inter-dealer trading activity ranges from 24% of volume in the London Stock Exchange in the 1990s (Reiss and Werner, 1998) to 28-42% in electronic corporate bond markets between 2010-2017 (O'Hara and Zhou, 2021).

Our first contribution is to empirically document stylized facts on cross-sectional and time-series patterns in market maker inventories. We use proprietary trade and quote data from January to August 2021 on three futures products traded on the Montréal Exchange: 5- and 10-year Canadian Government Bonds Futures and the S&P/TSX 60 equity index, respectively, and identify market maker accounts using similar criteria as Brogaard, Hendershott, and Riordan (2014) and Kirilenko, Kyle, Samadi, and Tuzun (2017). We document three stylized facts. First, inventory frictions are important even at high frequencies. Market makers unwind their position often—the average inventory half-life is below 10 minutes. At the same time, they end the day flat (i.e., with zero inventory change) 72% of the time. Second, there is significant heterogeneity in inventories at high frequencies. The average pairwise correlation between market maker inventories, sampled every 30 seconds, is only 11.3%. Third, market maker positions in the queue of limit orders is weakly correlated with their inventory – the correlation coefficient is just 10.1% across snapshots. That is, market makers with larger inventories to unwind do not seem to obtain better queue positions.

These three observations paint a striking picture: market makers seem to strongly care about their risk positions as witnessed by the quick reversals and the desire to end the day flat. Yet, their positions have a low cross-sectional correlation, suggesting imperfect risk sharing. Further, in a

market with time priority, the market makers who would benefit the most of being early in the queue to unwind positions do not seem to be able to do so. This raises several questions. How do the market maker limit order strategies depend on queue position and on cross-sectional variation in inventories? Is there an effect on aggregate liquidity provision? And, can we quantify the magnitude of risk sharing inefficiency?

To answer these questions, our second contribution is a simple limit order book model, where we extend the adverse selection framework of Sandås (2001) by introducing market maker inventory costs. Rather than a competitive market making sector, we study imperfect competition between a limited number of market makers who arrive sequentially to submit limit orders. To zoom in on the role of time priority, we follow Parlour (1998) and assume a single-tick market with two possible prices (a bid and an ask). The assumption is realistic for the many stocks that are tick size constrained, which according to Yao and Ye (2018) happens 41% of the time for Nasdaq 100 stocks in 2010.

Market makers face two exogenous frictions: adverse selection and inventory costs. Analyzing them jointly is important, since adverse selection affects the cost of unwinding inventory with limit orders. Incoming market orders are drawn from an exponential distribution, with larger orders generating higher (linear) price impact due to adverse selection. Since larger market orders are more informed on average (Hasbrouck, 1991),[1] limit orders earlier in the queue face lower adverse selection risk (Glosten, 1994). The second friction is inventory risk (Ho and Stoll, 1981): As intermediaries, market makers aim to connect natural buyers and sellers without taking an extreme position in the asset. Each market maker faces an inventory penalty over her squared (post-trade) position, which implies a marginal inventory cost that increases linearly in the holdings.

The equilibrium condition is that each consecutive market maker places a limit order such that the *marginal* share earns zero expected profits. In competitive limit order book models, this condition only holds for the last share in the queue (Glosten, 1994; Back and Baruch, 2013; Biais, Martimort, and Rochet, 2000). The equilibrium order size depends on the queue position of the market marker (i.e., the depth already offered when he arrives to the exchange) and on his pre-trade inventory holdings. Crucially, the solution can be directly mapped into an empirical linear regression model that estimates the relative magnitude of inventory and adverse selection frictions from the limit orders by market makers.

The solution shows that additional market makers submit smaller quantities. The quantities reduce because adverse selection worsens deeper in the book, yet remain positive to the extent that each additional market maker has a low initial marginal inventory cost. Models with perfect competition between risk neutral market makers (e.g., Sandås, 2001) obtain that the first market maker completely fills the book at each level. Introducing inventory concerns allows for multiple

[1] The increased adverse selection also holds for multiple small market orders trading in the same direction as the price impact accumulates.

market makers being present at the best price. An important model result is that heterogeneity in market maker inventories, combined with a random arrival sequence, affects liquidity as measured by market depth. The reason is that the *inventory* position of market makers early in the queue impacts the *adverse selection* cost for market makers further back in the queue. Individual order sizes, and by consequence aggregate depth, depend on the sequence with which market makers, and their associated inventories, appear in the queue.

Depth is maximized when market makers with large inventories, and consequently pressing trading needs, arrive at the back of the queue. The rationale is that such market makers post large orders to mean-revert their position. If they would arrive early in the queue, their large orders would increase adverse selection for subsequent market makers and crowd out their liquidity provision. Instead, when the large orders appear last in the queue, the crowding out effect is minimal – yielding larger quoted depth.

To maximize risk-sharing across market makers, the opposite arrival sequence is optimal. Efficient risk sharing implies that the market maker most eager to unwind inventory appears first in the queue of limit orders. That is, priority should be given to the market maker with the most extreme inventory to minimize her penalty in expected utility. Precisely because this investor is eager to unwind, she submits a large limit order. But the large limit order first in the queue (optimal for risk-sharing) crowds-out liquidity provision by all subsequent market makers because of adverse selection. To the extent that market makers arrive quasi-randomly, markets oscillate between more effective risk sharing across market makers and better liquidity, as measured by quoted depth.[2]

As our third contribution, we take the model to the data. We observe (anonymized) account identifiers for all trades in the sample. In addition to trades, our data contains 30-second snapshots of the top level in the limit order book for all futures contracts and maturities. For each snapshot, we observe all orders with their respective time priority levels and account identifier. We identify market maker accounts following Kirilenko, Kyle, Samadi, and Tuzun (2017) and further require that accounts have a resting quote at the top of the book in at least 20% of the snapshots. We obtain 12 market-maker accounts, all of whom are high-frequency traders. To identify adverse selection and inventory frictions, we regress the sizes of market makers' limit orders on their queue position and inventory holdings. The regression specification corresponds directly to the theoretical model, which tells us that the coefficient on queue position represents the adverse selection component of the marginal cost of liquidity supply, while the coefficient on the inventory holdings represents the inventory component. Both variables have a statistically significant impact on limit order size at the best price level. A one standard deviation increase in the queue position reduces quote

[2]In the model, we only allow market-makers to post limit orders. In practice, they can also use market orders to reduce inventory risk, but this is costly as it requires paying the bid-ask spread. The bid-ask spread might serve as an upper bound for the cost of inefficient risk sharing.

size by 0.146 contracts (7.93% relative to the average quote size of 1.84 contracts). A one-standard deviation increase in market maker inventory leads to orders being larger by 0.152 contracts (8.2% in relative terms). These results give direct insight in the components of the marginal cost of liquidity provision, and we see that both frictions appear about equally large in the liquid Canadian futures markets.

Our theoretical framework allows us to map cross-sectional inventory divergence into a measure of imperfect risk sharing. A back-of-the-envelope calculation showcases that aggregate inventory costs (using the quadratic functional form in the model) are four times larger than under a benchmark where market makers optimally share their inventory.[3]

Lastly, we test two predictions that are specific to our model and focus on the *interaction* between the inventory and adverse selection frictions at high frequencies. In line with the model, we find that aggregate depth increases if the market makers with the largest inventories to unwind arrive at the back of the queue. Randomness in the market maker arrival sequence generates (predictable) variation in market depth, up to a level of 8.4%. We further confirm the crowding-out effect of limit orders early in the queue on the sizes of subsequent orders. The marginal impact of an inventory shock on aggregate market depth is 13% higher for a market maker at the fourth position in the queue than for a market maker at the top of queue. All results are highly symmetric across both sides of the order book.

We relate to a literature aiming to estimate inventory and adverse selection costs. We believe we are the first to estimate inventory frictions based on limit order sizes of individual market makers. This contributes to the literature that estimates inventory effects based on transitory price pressures (see, e.g., Kraus and Stoll, 1972; Brogaard, Hendershott, and Riordan, 2014; Hendershott and Menkveld, 2014). Our analysis complements this work, as it offers an alternative identification approach for the same economic mechanism. We also contribute to the literature that uses the slope of the limit order book to estimate adverse selection (Sandås, 2001; Hollifield, Miller, and Sandås, 2004) by extending the theoretical framework to account for inventory frictions. Madhavan and Smidt (1991, 1993) use trade prices for NYSE specialists to distinguish between the two frictions and finds that adverse selection concerns are much stronger than inventory costs; we find that the two frictions have similar magnitudes in Canadian futures quote data.

We also contribute to a growing literature exploring the role of priority rules in limit order books. Parlour and Seppi (2008) provide an excellent survey of the early studies. More recently, Yao and Ye (2018) and Li, Wang, and Ye (2021) argue that binding tick sizes constrain liquidity provision and allocates rents to high-frequency traders who are well positioned to profit from time

[3]For this calculation, we assume the market makers continuously face an (unobservable) penalty on their squared inventory holdings. Dividing the actual squared positions by the square of the theoretical position under perfect risk sharing shows that the inventory cost could be four times lower. This calculation is adjusted for variation in market maker risk aversion, as proxied by the standard deviation of their inventory holdings. The details are in Section 4.3.

priority rules. Degryse and Karagiannis (2019) shows that if the tick size is tight, internalization under price-broker-time priority improves welfare relative to price-time priority. Finally, Budish, Cramton, and Shim (2015) argue that time priority rules in markets dominated by high-frequency traders stimulates toxic arbitrage. We add to this literature by exploring the interaction between (high-frequency) inventory frictions and adverse selection on modern markets with time priority.

Finally, our work contributes to the literature studying risk-sharing across intermediaries and the impact of inventory constraints on liquidity. Reiss and Werner (1998) use data from London Stock Exchange in 1991 (when the exchange operated as a dealer market) and show that inter-dealer trading represents 24% of the aggregate volume. Further, more than 60 percent of all inter-dealer trades are motivated by risk sharing, in that they simultaneously reduce inventory imbalances. On the Canadian futures market, we find less risk sharing in the cross-section of high-frequency market makers. Comerton-Forde, Hendershott, Jones, Moulton, and Seasholes (2010) show that large inventory positions for NYSE specialists translate to wider spreads and lower liquidity. Hendershott and Seasholes (2007) show that specialist inventories are negatively correlated with returns across multiple days. We complement this literature by emphasising the impact of inventory heterogeneity on aggregate depth.

2 Data and stylized facts

In this section, we first describe our trade-and-quote data, emphasizing how it allows us to sharply identify inventories and queue position at high frequencies for each trader account. Next, we identify market making accounts using a methodology derived from Kirilenko, Kyle, Samadi, and Tuzun (2017). Consistent with existing literature (Menkveld, 2013), we show that market makers tightly manage their inventory: they end most days completely flat and unwind their position frequently throughout the day.

We document two novel stylized facts. First, there is substantial cross-sectional heterogeneity in inventories for market makers with quotes at the top of the order book. Second, the arrival sequence of market maker limit orders in the book is largely uncorrelated with their inventory. That is, market makers with large positions to unwind are not more likely to be at the front of the queue. Since the adverse selection cost is a function of the queue position of an order (Sandås, 2001), our finding implies that adverse selection and inventory frictions are to a large extent orthogonal to each other.

In Section 3, we start from these facts and build a simple model that leverages inventory heterogeneity and random arrival to cleanly identify adverse selection and inventory frictions in order book data. We test the empirical predictions of the model in Section 4.

2.1 Data

We use trade and quote data from the Montréal Exchange between January 1st and August 18th, 2021 on three derivatives products: Ten-Year Government of Canada Bond Futures (CGB), Five-Year Government of Canada Bond Futures (CGF), as well as S&P/TSX 60 Index Standard Futures (SXF). All products are traded on a continuous-time limit order book with strict price-time priority.[4]

First, we observe masked trading account identifications (IDs) for each transaction, as well as the trade price and quantity, a trade initiator flag (i.e., either `maker` or `taker`), and an order type flag (i.e., whether the buyer and seller in each trade submitted a `market` order, a `market-on-open` order, or a `limit` order). Importantly, the anonymized identities are the same across contracts and over time. Trades have millisecond-level timestamps. Second, the quote data consists of 30-second snapshots of the top level in the limit order book. Each snapshot contains all outstanding limit orders at the best bid and ask prices. For each order, we observe its limit price, the submitter's trading account ID, as well as the order's execution priority. Since all orders at the best bid (ask) are identically priced, the execution priority corresponds exactly to the order's time priority. The data includes all iceberg orders (i.e., where only part of the order is visible to the market); there are no fully dark orders on the Montréal Exchange.

The government bond futures are the most liquid bond instruments in Canada, and the five-year and ten-year contracts in practise determine the yield curve. Each futures contract offers an exposure to bonds of about CAD100,000 in our sample. The equity futures contract offers an exposure of about CAD250,000 to the Canadian equity index.

2.2 Who are the market makers?

A natural first step in our analysis is to identify which ones out of the 3,511 unique accounts in our data trade in a manner consistent with market-making strategies.

2.2.1 Market maker definition in our sample

Grossman and Miller (1988) define market makers (or market intermediaries) as "traders who can fill gaps arising from imperfect synchronization between the arrival of buyers and sellers." In line with this definition, we follow the data-driven classification of Kirilenko, Kyle, Samadi, and Tuzun (2017) (KKST) and Brogaard, Hendershott, and Riordan (2019) to identify firms operating as intra-day intermediaries. We then identify the subset of market makers among the intermediaries by adding a fourth criteria based on how frequently they post limit orders at the top of the book:

[4]While both Montréal Exchange (derivatives) and Toronto Stock Exchange (equities) both belong to the TMX Group, they use different trading engines – that is, SOLA and Quantum, respectively. One key difference is that Toronto Stock Exchanges matches orders based on price-*broker*-time priority, whereas Montréal Exchange uses price-time priority.

(i) First, on average the account must participate in 50 or more daily trades in a particular instrument. This criterion allows us to filter out small traders that do not have a sufficient volume of activity to be labeled as market makers.

(ii) Second, the end-of-day net position for a market maker account should not exceed, on average, 5% of her daily traded dollar volume. For a given account j, the relative net position in instrument i at the end of day d is

$$\text{Net position}_{ijd} = \frac{\| \sum q_{t,i,d} p_{t,i,d} \|}{\text{DollarVolume}_{ijd}}, \tag{1}$$

where t runs over transactions by j in day d, $p_{t,i,d}$ is the transaction price of trade t and $q_{t,i,d}$ is the signed quantity of trade t (i.e., $q > 0$ for buys and $q < 0$ for sells).

(iii) Third, KKST require market maker accounts to mean revert inventories at a relatively high frequency (see, e.g., Ho and Stoll, 1983). That is, the average across days and instruments of account j's minute-by-minute inventory deviations from the end of day position should be at most 2.5%. Formally, we require that market maker accounts satisfy

$$\frac{1}{D \times I} \sum_{d}^{D} \sum_{i}^{I} \left[\sqrt{\frac{1}{\text{Mins}} \sum_{t}^{\text{Mins}} \left(\frac{NP_{tjid} - NP_{\text{end-of-day}} ijd}{\text{DollarVolume}_{ijd}} \right)^2} \right] \leq 2.5\%, \tag{2}$$

where D and I is the number of days and traded instruments in the sample, *Mins* is the number of minutes in the trading day, t runs over minutes, and *NP* is the account's net dollar position at a particular point in time.

(iv) The average daily top-of-the-book presence, defined as the share of snapshots where an account has a resting limit order at either the best bid, the best ask, or both, is at least 20%. That is,

$$\frac{1}{D} \sum_{d}^{D} \frac{\sum_{s}^{S_d} \mathbb{1}_{\text{account at best bid or ask},s,d}}{S_d} \geq 20\%, \tag{3}$$

where d runs over the D days in the sample, S_d is the number of snapshots with at least one resting quote in day d across all instruments, and $\mathbb{1}_{\text{account at best bid or ask}}$ takes the value one if the account has at least one resting top-of-the-book limit order in a given snapshot and zero else.

Based on the first three criteria from Kirilenko, Kyle, Samadi, and Tuzun (2017) we identify 22 candidate market-maker accounts. However, the three criteria focus on inventory patterns and trading volume, while ignoring whether traders use mostly passive (i.e., resting limit) or aggressive (i.e., market or marketable limit) orders. The fourth criteria captures this distinction, which is

important in the context of our model as it focuses specifically on the quoting strategies of market markers.

We find that 12 accounts out of 22 also fulfill the fourth top-of-book filter, which we label $\mathbf{MM}_1$ through $\mathbf{MM}_{12}$. While the last step leads to a seemingly large drop from 22 to 12 accounts, the latter set amounts to 96.26% of the limit orders across the snapshots (i.e., for 8m out of 8.3m orders over the full sample). In fact, four intermediaries accounts post fewer than 100 top-of-the-book limit orders each (to compare, the most active three accounts post more than 1 million quotes each). We cross-check the final sample with the Montréal Exchange and confirm that all **MM** accounts are also high-frequency traders.

One limitation of our approach is that a single firm may use multiple trading accounts. For example, each individual trader at the firm may have their own account to simplify performance metrics. The concern is that some individual accounts may not meet the criteria of inventory reversals, while the aggregation of accounts across the firm would. In this case, our analysis ignores some accounts that in fact are market makers.

We are purposefully taking a conservative approach to defining market-makers accounts. That is, we are requiring **MM** accounts to simultaneously fulfill the Kirilenko, Kyle, Samadi, and Tuzun (2017) criteria and the top-of-the-book filter, on an instrument-by-instrument level. Ten accounts fulfill the intraday intermediation criteria but have a low presence in the book; another 13 accounts are present in at least 20% of the top-of-the-book snapshots, but take large intraday or end-of-day positions. There is only one additional account that meets criteria 1,2, and 4; but not the intraday mean-reverting criteria 3. Applying the criteria on an instrument-by-instrument level has the drawback that it might leave out market makers that net their position across products: e.g., take a long position in 10-year bonds and hedge it with a short position in 5-year bonds. However, we argue this is not a first-order concern in our sample: the Montréal Exchange charges margins on a product-by-product basis and does not provide relief for offsetting fixed income positions. That is, end-of-day inventory is costly even if market makers offset their inventory across products, as they need to post collateral on each individual position.

2.2.2 Market maker trading patterns

Table 1 shows summary statistics at the instrument-day-trading account level. These statistics are used to classify accounts into **MM** and non-**MM** according to the criteria of Section 2.2.1. A salient feature is that market maker accounts are significantly different from non-**MM** accounts across a number of dimensions. On a median day, market-makers end completely flat (zero inventory) and mean-revert often throughout the day (on average, the inventory deviates only 0.37% from the end-of-day target). Further, market-makers have a resting quote at the top of the order book more than half the time. In contrast, non-**MM** accounts trade much less than **MM**s (on average

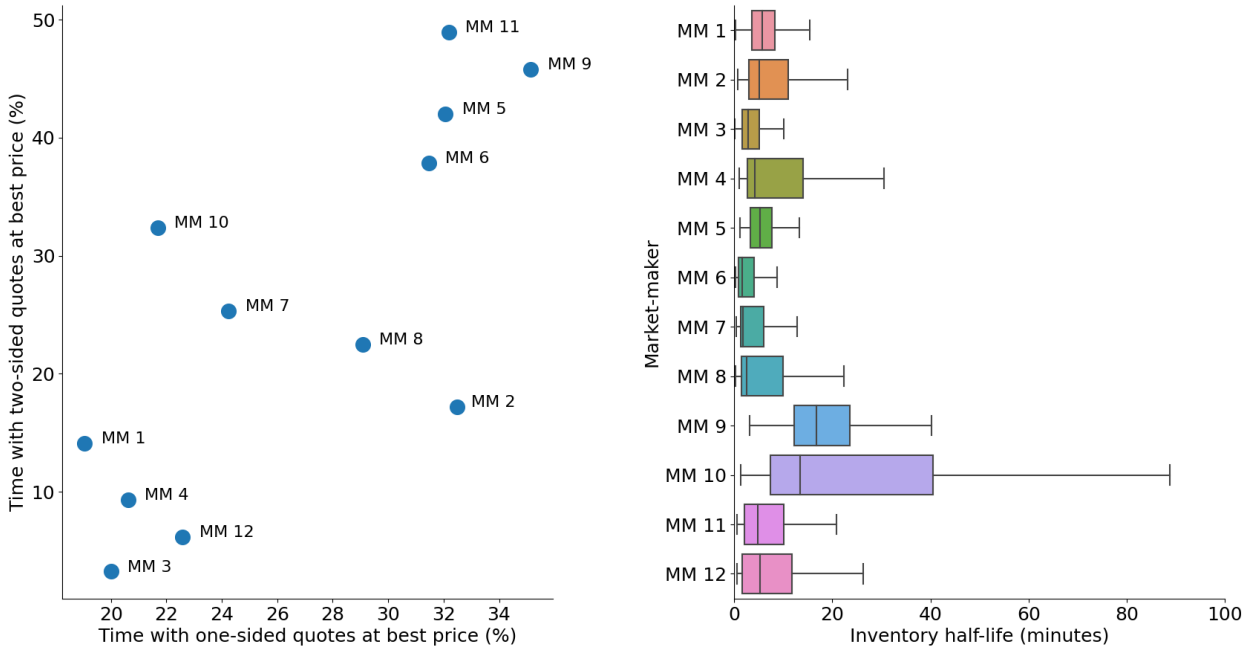
6% of market-maker volume, i.e., 229.78/3870.19) as well as more directional. On a median day, non-market makers either only buy or only sell, leading to a net position that is equal to the trading volume. Additionally, non-**MM**s do not continuously provide liquidity, as they have outstanding quotes at the top of the book only throughout 2% of a typical trading day.

[Insert Table 1 here]

Figure 1 further illustrates the **MM** trading and quoting patterns. The left panel plots the one-sided and two-sided presence in the top-of-the book for each market-maker. The most active market-makers have resting quotes at the top of the book more than 80% of the time (50% of the time on both sides of the book, and 30% on either the bid or the ask side of the book). Note that one-sided market making is consistent with inventory frictions (i.e., a long **MM** might only offer liquidity on the ask side).

Figure 1: Market-maker quote and trading patterns

This figure illustrates salient quote and trading patterns for market-maker accounts. The left panel plots, for each **MM** account, the percentage of snapshots where the **MM** had a single-sided quote at the top of the book against the percentage of snapshots where **MM** had a two-sided quote at the top of the book. The right panel plots, for each market-maker, the instrument-day distribution of inventory half-lives (measured in minutes). For each box, we display the mean (vertical line inside the box) and inter-quartile range (edges of the box). The whiskers extend to cover the rest of the distribution.



The right panel of Figure 1 shows that the average half-life of inventory shocks is about 15 minutes, meaning that market-makers mean-revert their positions very quickly throughout the

day. To estimate the half-life, we first sample inventories for each account with a 30 second frequency. Then, for each account m, instrument i and day d in the sample, we estimate a one-lag auto-regressive process for inventories (see, e.g., page 410 in Hendershott and Menkveld, 2014):

$$\text{Inventory}_{m,i,d,t+30} = \phi_{0,m,i,d} + \phi_{1,m,i,d}\text{Inventory}_{m,i,d,t} + \text{error}. \tag{4}$$

The estimated inventory half-life is a function of the estimated AR coefficient $\hat{\phi}_1$, that is,[5]

$$\text{Half-Life}_{m,i,d} = -\frac{\log(2)}{\log\left(\|\hat{\phi}_{1,m,i,d}\|\right)}. \tag{5}$$

A lower half-life corresponds to faster mean reversion of inventory. Figure 1 plots the **MM** inventory half-life distribution across days and instruments. In our sample, the average inventory half-life across accounts is only 15 minutes, which is quite similar to the 5-minute half-life for the HFT studied in Menkveld (2013). Further, inventory mean-reversion speed is consistently high across different **MM** accounts: six out of twelve accounts have an average inventory half-life of less than 10 minutes. In other words, market-makers in our sample are relatively homogeneous and turn over inventory at very similar frequencies. They all seem very concerned about managing inventories.

2.2.3 Market-maker inventory divergence

Sequential trading and queue priority imply that market makers cannot perfectly share risks, i.e., absorb an equal share of every incoming market order. In this section, we show that intraday market-maker inventories have very low correlations, which is suggestive of poor risk sharing. Market maker inventories are set to zero at the start of the sample and subsequently tracked through time.

To systematically study inventory divergence, we compute for each day and traded product in the data the average pairwise correlation between market-maker inventories for three sampling frequencies: 30 seconds, 30 minutes, and two hours. We show the results in the top-left panel of Figure 2. The average inventory correlation is consistently low: between 10% with two-hour sampling and 11.3% with 30-second sampling. This surprisingly low correlation is consistent with the short half-lives of inventory shocks, in the sense that each market maker seems to quickly build positions up and down—seemingly independent of other market makers. The top-right and bottom-left panels in Figure 2 show that the pattern is consistent across products (the average

[5] To see this, consider a general demeaned AR(1) process $y_t = \phi y_{t-1}$. We are looking for the number of lags h where the process halves the distance to its stationary mean: $\mathbb{E}_t y_{t+h} = \frac{1}{2} y_t$. However, since $\mathbb{E}_t y_{t+h} = \phi^h y_t$, h is the solution to $\phi^h y_t = \frac{1}{2} y_t$.

inventory correlation is 5.9% for equity index futures and 15.4% for 5-year bond futures) as well as throughout our sample period between January and August 2021. We interpret this result as cross-sectional inventory divergence. In a market where all **MM**s could trade with each incoming order, such as a batch or double auction, market-maker inventories could move in perfect lockstep, implying cross-sectional inventory correlations of 1.

Figure 2: Inventory divergence across market makers

This figure illustrates short-term inventory divergence across market-maker (**MM**) accounts. Panel (a) shows the distribution of average pairwise correlations of market maker inventories, sampled over different frequencies: 30 seconds, 30 minutes, and 2 hours, respectively. The pairwise correlations are calculated daily for each contract and all pairs of the 12 **MM**s. Panel (b) shows the 30-minute average pairwise correlation across market maker inventories across the three futures contracts in the sample: equity index, 5-year, and 10-year Government of Canada bonds. Finally, Panel (c) tracks the 30-minute average pairwise correlation across all days in our sample, i.e., between January 1st and August 18, 2021.

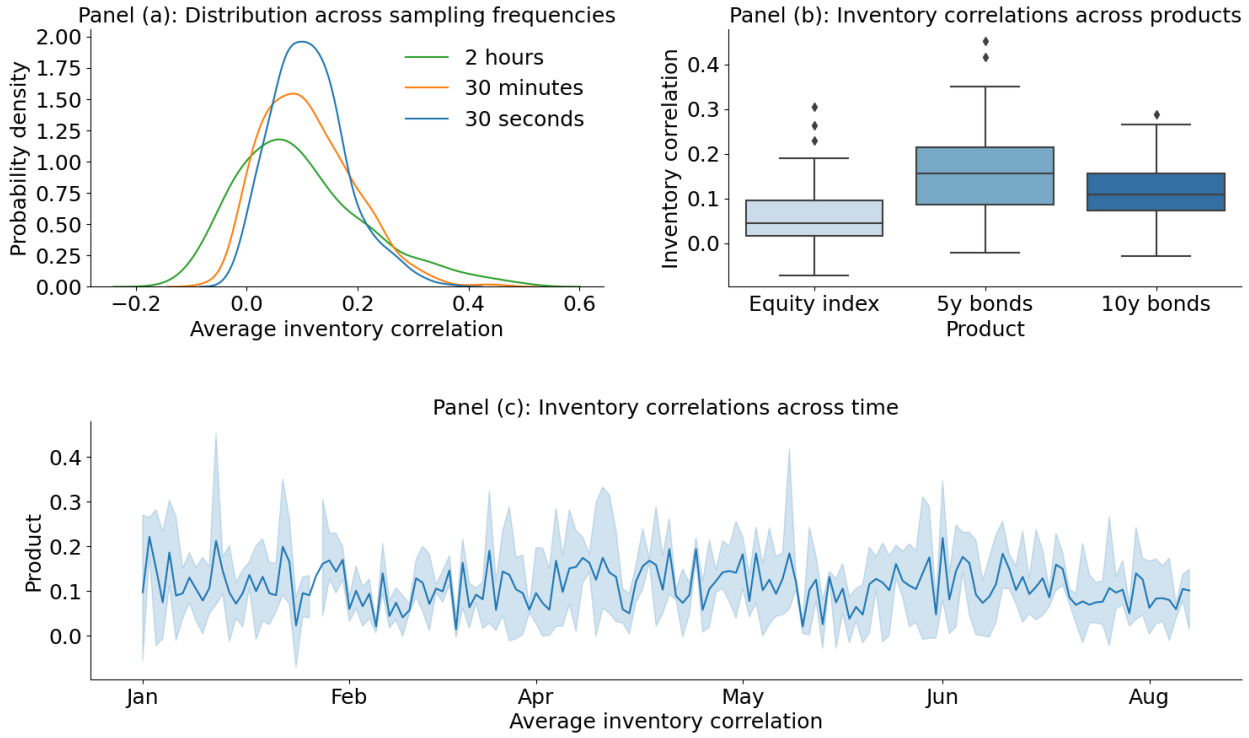


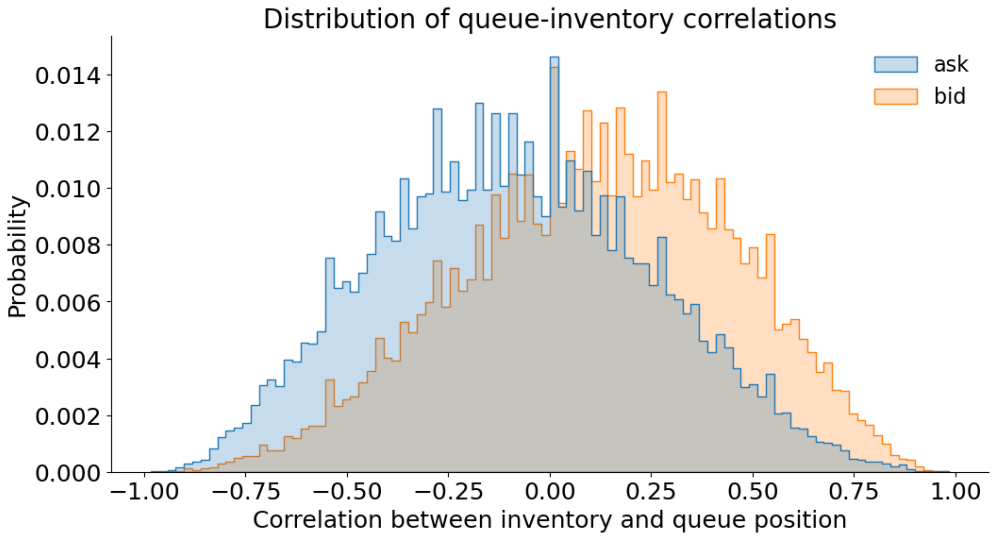
Figure 3 illustrates that cross-sectional variation in market maker inventory is orthogonal to the position of their limit orders in the execution queue. This result suggests that, while market makers have heterogeneous inventories, their limit orders arrive seemingly randomly to the exchange queue – it is not the case, for example, that market makers with large inventories to unwind manage to systematically enjoy better time priority. In the figure, we compute the rank correlation between the queue position of each **MM**'s marginal order and their inventory conditional upon all orders ahead of it being executed. The correlation coefficient between queue position and inventory is

only -8.8% for ask orders and 11.5% for bid orders. To account for the fact that a market maker may have multiple orders in the queue, we focus on the marginal limit order for each **MM**, that is the order with the lowest time priority.

In Section 3 we build a model that leverages this queue arrival randomness to pin down adverse selection and inventory frictions from order size data. The model also makes sharp predictions about how aggregate depth depends in equilibrium upon the sequence with which market makers with heterogeneous inventories arrive in the queue.

Figure 3: Correlation between queue priority and market-maker inventory

We plot the probability distribution of the Spearman correlation coefficient between the queue priority of the marginal order of a market maker and the market maker's inventory conditional upon all orders ahead of it being executed. This correlation is calculated for 128,180 30-second order book snapshots with 9 or more active market-makers (i.e., with presence of at least 75% of all market-maker accounts). We plot the distribution separately for ask-side and bid-side snapshots.



3 A model of liquidity supply with heterogeneous inventories

3.1 Model primitives

Asset. Consider a three-period economy, where time is indexed by $t \in \{0, 1, 2\}$. There is a single risky asset, which pays off a stochastic dividend $\tilde{v}$ at $t = 2$. At the start of the game $t = 0$, the expected value of the stochastic dividend is $\mathbb{E}_0\tilde{v} = v$.

Trading environment. The asset is traded on a limit order market with time priority and a two-price grid (i.e., a one-tick market as in Parlour, 1998). A limit order is a price quote to either buy

(a bid) or to sell (an ask) a particular number of contracts. A market order trades immediately against the outstanding limit orders. Limit orders are queued under a time priority rule: that is, if multiple limit orders are at the same price, the order with an earlier time trades first. The market is transparent: Traders observe the state of the order book as well as the entire history of limit orders and trades.

Traders can post buy and sell limit orders at one of two potential price levels: either p_{-1} or p_1. We assume that $p_{-1} < v < p_1$, such that a trader with no private value or additional information about the asset only submits buy orders at p_{-1} (earning $v - p_{-1} > 0$ per unit) and only submits sell orders at p_1 (earning $p_1 - v$ per unit following execution). It is straightforward to see that submitting sell orders at p_{-1} or buy orders at p_1 yields an expected loss, and is therefore not optimal in the absence of private values or signals about $\tilde{v}$.

Agents. There are two types of traders in the economy: $J \geq 2$ market makers (**MM**s) and a mass of impatient traders. At $t = 0$, each **MM** j starts with a position I_j in the risky asset, i.e., her inventory. A positive (negative) I_j corresponds to a long (respectively, short) position in the asset. Market makers start with zero wealth, do not have either private values or additional information about the asset, and submit limit orders at $t = 0$ to maximize:

$$\mathbb{E}_0 \text{Utility}_{\mathbf{MM}} = \mathbb{E}_0 \left[\text{Wealth}_{t=2} - \frac{\gamma}{2} I_{t=2}^2 \right], \tag{6}$$

where $\gamma > 0$ is a measure of risk-aversion. Equivalently, market makers have a target inventory, and $I_{t=2}$ represents the deviation over which they face a convex penalty at $t = 2$. The expected wealth at $t = 2$ corresponds to the net cash flow from trading.

We model impatient traders as in Sandås (2001). An impatient trader arrives at the market at $t = 1$ and is equally likely to submit a marketable buy order at the ask price p_1 or a marketable sell order at the bid price p_{-1}. The order size x is drawn from a two-sided exponential distribution,

$$f(x) = \begin{cases} \frac{1}{2\phi} \exp(-\frac{x_t}{\phi}) & \text{if } x_t > 0 \text{ (market buy);} \\ \frac{1}{2\phi} \exp(\frac{x_t}{\phi}) & \text{if } x_t < 0 \text{ (market sell).} \end{cases} \tag{7}$$

The expected impatient trader's order size, $\mathbb{E}[\|x\|]$, is $\phi > 0$, which is symmetric across buy and sell orders. Further, the impatient trader's order size is informative about the asset's true value:

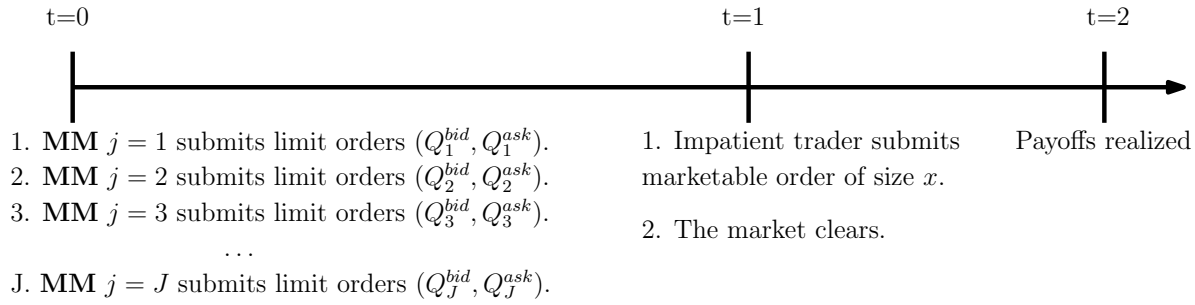
$$\mathbb{E}[\tilde{v} \mid x] = v + \lambda x. \tag{8}$$

The linear price impact parameter $\lambda > 0$ maps to the informational content of impatient traders' orders: Buys typically contain positive information with respect to the fundamental value; similarly,

sales contain negative information. Larger orders contain more information than small orders (at rate λ), generating greater price impact.

Timeline. Figure 4 illustrates the timeline of the model. At $t = 0$, the J market makers arrive sequentially to the market and and place limit orders in the book. Without loss of generality, we index the market makers in order of their arrival, and therefore of their place in the queue. For example, $\mathbf{MM}_1$ is the first market maker to arrive and place a limit order in the book.

Figure 4: Model timeline



We note that, since market makers only place buy orders at the bid price p_{-1} and sell orders at the ask price p_1, no trade occurs at $t = 0$. Let Q_j^{bid} and Q_j^{ask} denote the size of the limit order placed by market maker j at the best bid (i.e., at price p_{-1}) and best ask (price p_1), respectively. At $t = 1$, an impatient trader arrives and submits a marketable limit order of size x_t, equally likely to be a buy or or a sell order. At $t = 2$, the stochastic dividend and market maker utilities are realized. All limit orders still in the book expire.

Equilibrium. The only strategic decisions in the game occur at $t = 0$. An equilibrium in this game is a sequence of quantity pairs $\left(Q_j^{\star,\text{bid}}, Q_j^{\star,\text{ask}}\right)$ for each market maker $j \in \{1, 2, \ldots, J\}$ that maximize the **MM** utility in (6) conditional on the initial inventory I_j and of the state of the order book upon j's arrival, that is all orders $\left(Q_k^{\star,\text{bid}}, Q_k^{\star,\text{ask}}\right)$ for $k < j$.

The game is symmetric across buy and sell orders. For clarity of exposition, in the remainder of the section we are going to focus on the ask side of the book (i.e., **MM** sell orders) and drop superscripts referring to order direction.

Discussion. First, to derive sharp predictions on the impact of time priority, we shut down price priority in our model by assuming a simple, two-price, grid. That is, market-makers in our model cannot "jump" ahead of the queue by quoting a more aggressive price (Parlour, 1998).

Consequently, our model best describes markets in highly-liquid assets, for which the tick size is binding and market makers compete on queue position rather than prices.

Second, in the model market makers face inventory penalties after each trade. If the game were to be played repeatedly, then our inventory penalty maps to a holding cost per unit of time as in Du and Zhu (2017) or Duffie and Zhu (2017). Such inventory penalties can be motivated by risk aversion and capital constraints: market makers need to unwind inventories by the end of the trading day to avoid posting collateral on their position. A larger current inventory leads to a higher cost of successfully unwinding it on time: we capture this mechanism with a single parameter γ. Further, this assumption is consistent with the evidence of section 2.1, that high-frequency market-makers have inventory half-lives of less than 15 minutes throughout the day.

Third, time priority introduces a first-mover advantage for **MM**s who happen to arrive early in the queue. Each market maker submits an order size such that the profit on the last (marginal) share is zero. Adding additional shares would yield an expected marginal loss, while adding fewer foregoes expected profits. Note that the expected profit of the whole limit order is strictly positive,[6] because time priority prevents Bertrand competition between market makers. Therefore, it is valuable to secure an early position in the queue. One implication is that our model is robust to extending the strategy space: for example, a "staggered" strategy of submitting a limit order, waiting for other market makers to arrive, and then submitting another limit order is sub-optimal. In equilibrium, market makers will only submit limit orders once upon their first arrival, to maximize the value of the queue position.

3.2 Equilibrium analysis

Consider the first market maker to arrive at the market at $t = 0$, that is $\mathbf{MM}_1$. The expected profit for the last unit offered at the ask price, Q_1, is

$$\mathbb{E}U_{\mathbf{MM}1}\left(Q_1\right) = \int_{Q_1}^{\infty} \left[p_1 - v - \lambda x + \gamma(I_1 - Q_1)\right] \frac{1}{2\phi} \exp(-\frac{x}{\phi}) \mathrm{d}x. \tag{9}$$

First off, the order execution probability is

$$\int_{Q_1}^{\infty} \frac{1}{2\phi} \exp(-\frac{x}{\phi}) \mathrm{d}x = \frac{1}{2} \exp\left(-\frac{Q_1}{\phi}\right), \tag{10}$$

[6] To see this, suppose the optimal order is 100 shares. The last marginal share of this limit order (at location 100) earns zero expected profits. This means that shares 1 to 99 earn positive expected profits because they are earlier in the queue and i) face lower adverse selection risk and ii) face a lower marginal inventory cost.

that is, the probability that the impatient trader submits a buy order for more than Q_1 units. Conditional on execution, the market maker cashes in the ask price p_1 and gives away a security with expected payoff $v + \lambda x$ (the buy trade generates market impact). After selling Q_1 units, the market maker's inventory becomes $I_1 - Q_1$, and she incurs a quadratic penalty for any non-zero position at $t = 2$. Note that as the market maker is selling, a positive pre-trade holding will be reduced towards zero which increases utility. Conversely, for negative holdings the additional selling reduces utility. The marginal impact of a trade of size Q_1 on the inventory penalty is

$$\frac{\partial}{\partial Q_1}\left(-\frac{\gamma}{2}\left(I_1 - Q_1\right)^2\right) = \gamma\left(I_1 - Q_1\right). \tag{11}$$

Finally, the expected order size x conditional on the marginal unit Q_1 being executed is $\phi + Q_1$. We can write the expected utility on the marginal unit for the first market maker as

$$u^{\mathbf{MM1}}\left(Q_1\right) = \frac{1}{2}\exp\left(-\frac{Q_1}{\phi}\right)\left[p_1 - v - \lambda(Q_1 + \phi) + \gamma\left(I_1 - Q_1\right)\right]. \tag{12}$$

The first market maker is maximizing her profit, and therefore sets Q_1 such that the marginal unit breaks even. The equilibrium quantity $Q_1^\star$ is pinned down by setting the expression in (12) equal to zero:

$$Q_1^\star = \frac{p_1 - v + \gamma I_1 - \lambda\phi}{\gamma + \lambda}. \tag{13}$$

We turn next to the optimal quote for subsequent market-makers. For market maker j, the marginal limit order is at location $\sum_{k=1}^{j-1} Q_k^\star + Q_j$, behind the orders submitted by the first $j-1$ market makers. Therefore, the expected profit on the marginal unit is

$$u^{\mathbf{MM}j}\left(Q_j\right) = \frac{1}{2}\exp\left(-\frac{\sum_{k=1}^{j-1} Q_k^\star + Q_j}{\phi}\right)\left[p_1 - v - \lambda(\sum_{k=1}^{j-1} Q_k^\star + Q_j + \phi) + \gamma\left(I_j - Q_j\right)\right]. \tag{14}$$

There are two key differences between equations (12) and (14). First, the execution probability for market-maker j is lower since it requires a larger order from the impatient trader. Second, conditional on execution, market maker j faces a larger adverse selection cost (i.e., $\lambda(\sum_{k=1}^{j-1} Q_k^\star + Q_j + \phi)$). We equate the marginal utility to zero and solve for the optimal quote in Proposition 1.

Proposition 1. *The market makers $j = \{1, \ldots, J\}$ arrive consecutively to the exchange, observe the state of the limit order book upon arrival, and submit to the best ask price level an optimal limit order of size*

$$Q_j^\star = \frac{p_1 - v + \gamma I_j - \lambda\left(\sum_{k=1}^{j-1} Q_k^\star + \phi\right)}{\gamma + \lambda}. \tag{15}$$

This recursive solution can also be solved explicitly to obtain

$$Q_j^\star = \frac{p_1 - v - \lambda\phi}{\gamma}\left(\frac{\gamma}{\gamma+\lambda}\right)^j + \underbrace{\frac{\gamma}{\gamma+\lambda} I_j}_{\substack{\textit{direct}\\ \textit{effect}}} - \underbrace{\sum_{k=2}^{j}\left(I_{j-k+1}\frac{\lambda}{\gamma}\left(\frac{\gamma}{\gamma+\lambda}\right)^{k-1}\right)}_{\textit{crowding-out effect if } j \geq 2}. \tag{16}$$

Proposition 1 states the optimal ask quote size for any market maker (i) decreases in the magnitude of adverse selection λ and inventory penalty γ, (ii) increases in inventory position I_j making the market maker more eager to sell, and (iii) depends on the *entire* cross-sectional distribution of inventories by market makers ahead in the queue in Equation (16). We note that equation (15) is recursive, in that the optimal quantity for market maker j depends on the quantities submitted by her predecessors, whereas Equation (16) shows the same solution as a function of the individual inventory positions of the predecessors.

The solution reveals that adverse selection limits a market makers' ability to perfectly unwind inventories: for a one unit increase in initial inventory, she increases her sell quote by only $\frac{\gamma}{\gamma+\lambda} < 1$ units. For a higher price impact λ, she faces a larger cost to mean-revert inventory. Further, all else fixed, market makers in the back of the queue post smaller orders. The rationale is that these face higher adverse selection costs as they only execute against very large incoming market orders which are more informed on average. If we switch off the inventory penalty (i.e., set $\gamma = 0$), the result in Proposition 1 converges to the liquidity provision equilibrium of Sandås (2001), that is

$$Q_1 = \frac{p_1 - v - \lambda\phi}{\lambda}. \tag{17}$$

In Sandås (2001), the market making sector is risk neutral and perfectly competitive: the first market maker fills the book with the equilibrium quantity and therefore time priority becomes irrelevant. Only by adding inventory frictions the sequence of arrivals matters.

Equation (16) re-states the equilibrium limit order size solved explicitly rather than recursively. The first term, $\frac{p_1 - v - \lambda\phi}{\gamma}$ is essentially the size of the (risk-adjusted) pie: the realized spread defined as the ask price minus the expected asset value conditional on execution, scaled by the inventory penalty. Each market maker gets a slice of the pie, $\left(\frac{\gamma}{\gamma+\lambda}\right)^j$, which is decreasing in the queuing order. The first market maker cannot capture the entire profit, as it requires submitting a large limit order which gets too costly because of inventory risk. She scales back her order, leaving a slice for her successor, who in turn does the same.

We decompose the impact of inventory on optimal quote size into a *direct effect* and a *crowding-out effect*. The direct effect is the impact of a market maker's *own* inventory on optimal quote size. Intuitively, each market maker would like to fully offload their inventory position I_j, but needs to make an adjustment for the adverse selection cost, which is reflected in the second term in

equation (16). Indeed, fully unwinding inventory requires posting a large limit order which raises the adverse selection cost, and therefore the market maker scales back.[7]

The third term illustrates a novel *crowding-out effect*: that is, the impact of inventories of market makers ahead in the queue reduce market maker j's optimal quote size. Note that last term sums over the $j-1$ predecessors. Each of them adjust their order size to manage *inventory risk* proportional to $\frac{\gamma}{\gamma+\lambda}$, which is the inventory component of their marginal liquidity cost. In turn, these changes in order sizes affect the queue location and *adverse selection risk* of subsequent market makers, who adjust their quotes according to the last term. For example, the inventory shock of market maker 1 affects Q_1 directly, but also Q_2 linearly through adverse selection. The changes in Q_1 and Q_2 together affect Q_3 quadratically in I_1 (hence the power term $(k-1)$).

In Corollary 1, we sum up individual order sizes over all j, to generate additional predictions on the aggregate quoted depth.

Corollary 1. *The aggregate quoted depth $\overline{Q}_J = \sum_{j=1}^{J} Q_j^\star$ at the best ask price is*

$$\overline{Q}_J = \sum_{k=1}^{J} \left[\left(\frac{p_1 - V - \lambda v \phi}{\gamma} + I_{J-k+1} \right) \left(\frac{\gamma}{\gamma+\lambda} \right)^k \right]. \tag{18}$$

Corollary 1 generates a sharp prediction on how aggregate quoted depth depends on the queuing sequence of market makers and their initial inventory holdings. The key is that each market makers' inventory affects her own order and those of her successors, but not the orders by market makers ahead in the queue. It is for this reason that the queuing sequence matters, which we further illustrate with a numerical example in Figure 5. We consider three market makers, indexed by their arrival order, across four different inventory distribution scenarios. For all scenarios, the aggregate maker maker inventory is kept constant at six units.

In the first scenario (first bar in Figure 5), $\mathbf{MM}_1$ has a long position of six units. She posts a large sell order (7 units) as she has a dual motive for trade: capture the bid ask spread *and* unwind the large inventory. The large order increases the adverse selection risk for $\mathbf{MM}_2$ and $\mathbf{MM}_3$, who have lower incentives to trade. Since $\mathbf{MM}_2$ and $\mathbf{MM}_3$ also have no inventory, they post very small quotes to earn a small fraction of the bid-ask spread (0.5 and 0.25 units, respectively). The large inventory position of $\mathbf{MM}_1$ crowds out the liquidity provision by $\mathbf{MM}_2$ and $\mathbf{MM}_3$, and the aggregate depth in this scenario is 7.75 units.

Consider now the last scenario (fourth bar in Figure 5). The first market-maker is short two units ($I_1 = -2$). A large sell order would push her inventory further away from zero. However, being first in the queue does generate an opportunity to capture the spread while bearing low

[7] Conversely, a short market maker wants to shrink his limit order. However, shrinking to the full extent of the short position reduces the adverse selection cost, which in turn creates a profitable market making opportunity, thereby increasing the limit order partially.

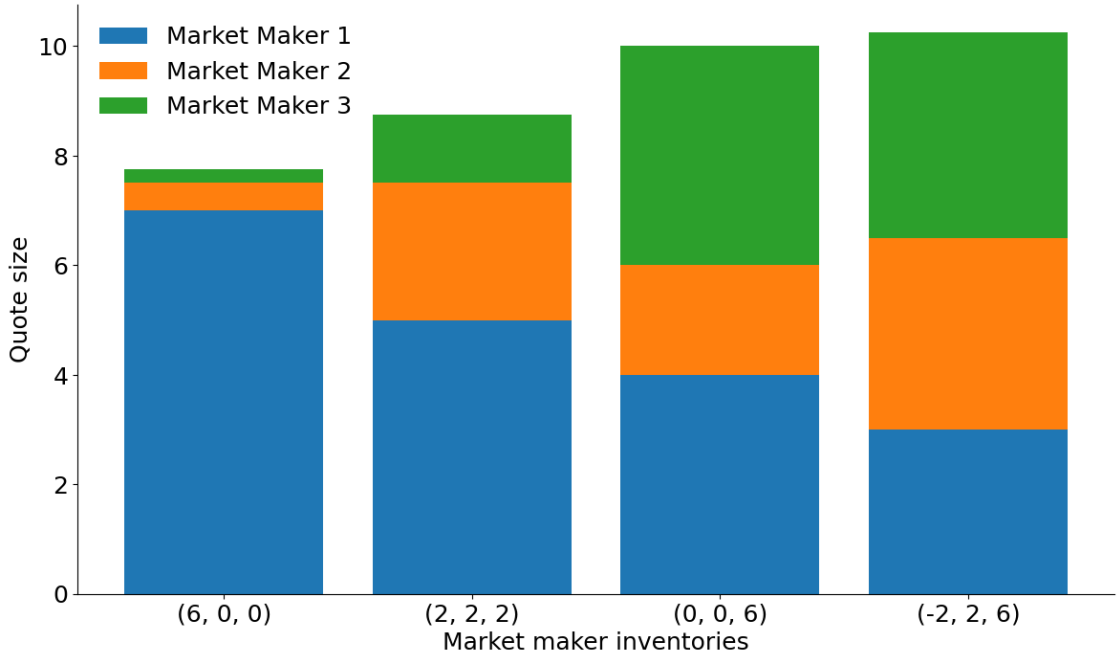
adverse selection risk: $\mathbf{MM}_1$ posts an order to sell 3 units. The smaller-sized order reduces the crowding-out effect for $\mathbf{MM}_2$ and $\mathbf{MM}_3$, incentivizing them to post larger orders. Moreover, $\mathbf{MM}_2$ and $\mathbf{MM}_3$ have long inventories that they wish to unwind, which further increases the sizes of their sell orders. The aggregate depth in this scenario is 10.25 units, even though the aggregate inventory for the market-making sector is unchanged.

An additional result of Corollary 1 and Figure 5 is that the highest liquidity level (i.e., largest quoted depth) is achieved precisely when risk sharing across market makers is lowest. To maximize depth, market makers with the largest inventories are placed at the back of the queue to minimize the crowding-out effect on successors. However, as they are the least likely to trade, this leads to further inventory divergence across market makers and less efficient risk sharing.

The last implication of Corollary 1 is that the marginal impact of an inventory shock for $\mathbf{MM}_j$ on total quoted depth is equal to $\left(\frac{\gamma}{\gamma+\lambda}\right)^{J-j}$; i.e., the shock gets attenuated by the $J-j$ successors. Thus, the aggregate depth is more sensitive to the inventory of back-of-the-queue market makers. The reason is that the crowding-out effect is stronger for investors earlier in the queue. A positive inventory shock to a market maker early in the queue will be partially offset by a reduction in the order sizes by many subsequent market makers who face higher adverse selection.

Figure 5: Equilibrium ask limit order sizes as a function of inventory and queue position.

We show the equilibrium sizes of ask limit orders based on equation (16) for an economy with three market makers who arrive sequentially, with **MM**1 being the first and **MM**3 being the last. In aggregate, they hold an inventory of 6 contracts. For each bar we vary the inventory positions across market makers. In the first bar, **MM**1 is long 6 contracts, while **MM**2 and **MM**3 have zero. In the second bar they all have a medium inventory position of 2 contracts. In the third bar **MM**1 and **MM**2 have zero, while **MM**3 has 6 contracts. In the fourth bar, **MM**1 is short two shares, while **MM**2 and **MM**3 are long 2 and 6 shares, respectively. The best ask price is \$10.01 and the ex ante expectation of the fundamental value $v = \$10.00$ and the average trade size ϕ is normalized to 1. The price impact parameter is $\lambda = 1/900$, and inventory cost parameter is $\gamma = 1/900$. We tabulate the equilibrium quote sizes under the graph.



Market maker	1	2	3	Total
Case 1: Initial inventory holding	6	0	0	6
Equilibrium liquidity supply	7.00	0.5	0.25	7.75
Case 2: Initial inventory holding	2	2	2	6
Equilibrium liquidity supply	5.00	2.50	1.25	8.75
Case 3: Initial inventory holding	0	0	6	6
Equilibrium liquidity supply	4.00	2.00	4.00	10.00
Case 4: Initial inventory holding	-2	2	6	6
Equilibrium liquidity supply	3.00	3.50	3.75	10.25

Figure 6: Marginal impact of inventory at different queue positions

We show the marginal impact of an inventory shock of $\frac{\gamma+\lambda}{\gamma}$ units (which corresponds to a unitary increase in limit order size) for the market maker at queue position j on aggregate depth. The baseline price impact parameter is $\lambda = 1/900$ and inventory cost parameter is $\gamma = 2/900$; the dashed line shows a setting with a higher level of adverse selection with $\lambda = 2/900$. The figure reveals a crowding-out effect: the increase in limit order size of a market maker early in the queue due to an inventory shock gets offset by a reduction of liquidity provision by market makers later in the queue due to higher adverse selection. This crowding-out effect reduces the impact of the inventory shock on aggregate depth; and the effect disappears for market maker #6, the last market maker in the queue.

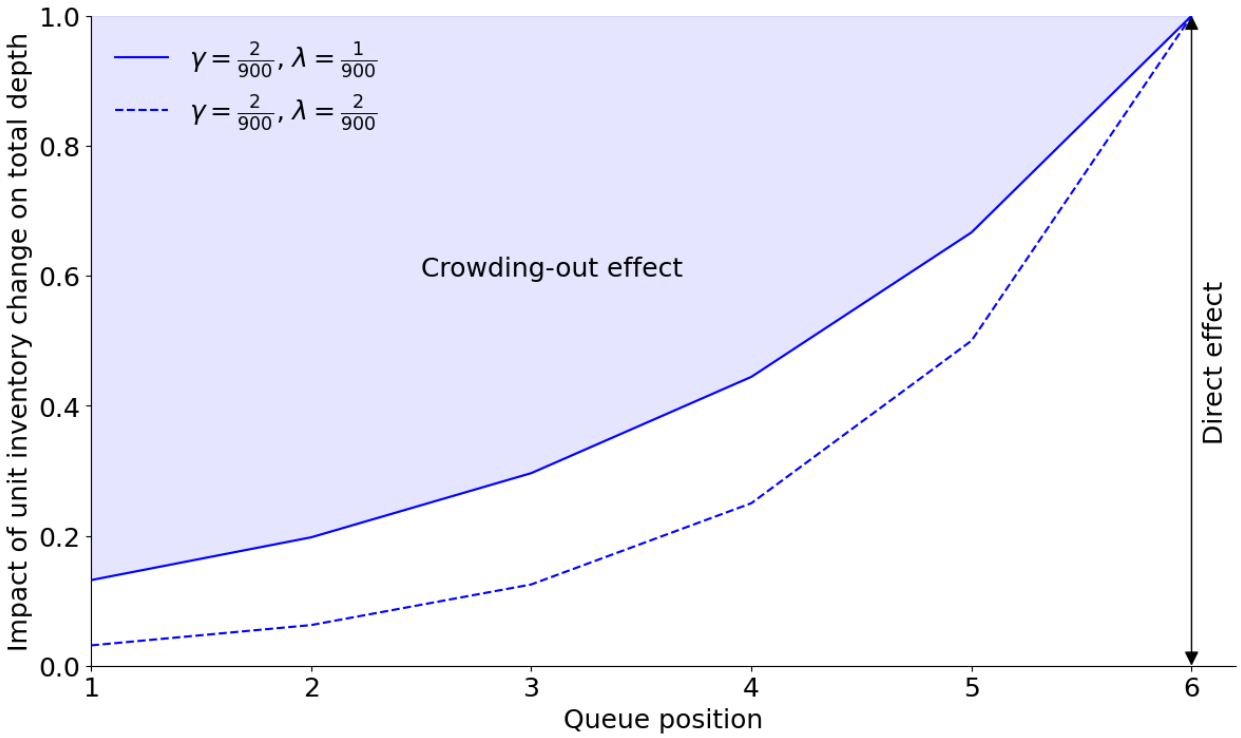


Figure 6 illustrates this result for a six-**MM** queue. We shock, in turn, each market-makers' inventory position by $\frac{\gamma+\lambda}{\gamma}$ units, which results in a unit increase in limit order size. We notice that the crowding-out effect is stronger for market makers earlier in the queue (i.e., those with many successors), but much weaker for market makers later in the queue (who have only few successors). The last market maker in the queue does not crowd out liquidity by any subsequent market makers, and therefore impacts aggregate depth only through the direct (unitary) effect. The figure further confirms that the crowding-out effect goes through adverse selection, and thus is larger when when adverse selection is high (the dashed-line).

3.3 Testable predictions

Proposition 1 and Corollary 1 yield four testable predictions, which we enumerate for clarity in this section. We proceed to test these prediction in the next Section 4.

Prediction 1: The size of a market makers' limit order decreases in the order's queue position due to adverse selection.

Prediction 2: The size of a market makers' sell (buy) limit order increases (decreases) in her inventory level.

Predictions 1 and 2 follow directly from Proposition 1. Back-of-the-queue market makers face larger adverse selection costs, and therefore decrease their liquidity provision. Further, all market makers aim to unwind their inventory: they want to sell more when they hold long positions and buy more when they hold short positions. The two predictions claim that inventory effects and adverse selection are individually important for liquidity provision, as measured by posted quote sizes.

We generate two additional idiosyncratic predictions starting from Corollary 1. The following two predictions focus on the *interaction* between adverse selection and inventory constraints on markets with time priority. They correspond to the crowding-out results illustrated in Figures 5 and 6.

Prediction 3: Quoted depth is larger if the market-makers with the highest inventories to unwind are located at the back of the queue.

Prediction 4: Quoted depth is more sensitive to the inventory of back-of-the-queue market makers.

4 Empirical evidence

4.1 Adverse selection and inventory effects

In this section we test two empirical implications of Proposition 1. From equation (15), the size of market-maker limit orders: (i) decreases in the queue length ahead of the order, due to heightened adverse selection risk and (ii) increases in the absolute magnitude of the **MM**'s long (short) inventory for quotes on the ask (bid) side. Equivalently, market-makers weigh both adverse selection and inventory concerns in setting the optimal order size.

The main panel. To test the two effects in the Montréal Exchange data, we build a panel across snapshots and market makers. Each limit order book snapshot corresponds to a particular traded

instrument, side of the book (i.e., bid or ask), and 30-second timestamp. Next, we transform the data in four ways:

(i) Empirically, as trading progresses, a market maker may have multiple limit orders in the queue of the best price level. Proposition 1 and the zero-profit condition (15) are defined in terms of the *marginal* order size. To match the model, the unit of observation is the last limit order in the queue (i.e., the marginal quote) for each snapshot and market maker account.

(ii) The challenge of multiple limit orders in the queue also requires a modification of the inventory to correctly match the model. Let I_j denote the inventory of market maker with queue priority j at the time of the snapshot. The modified inventory then is the inventory conditional upon all orders by j earlier in the queue being executed:

$$\tilde{I}_j = I_j - d_{\text{ask}} \times \sum_{k=1}^{j-1} Q_k \times d_{j,k}, \tag{19}$$

where Q_k is the quote size at priority k, $d_{j,k}$ is a dummy variable taking value one if order k belongs to market maker j and zero else, and d_{ask} takes value one for ask-side snapshots and negative one for bid-side snapshots (to reflect that the market maker is either selling or buying).

(iii) Finally, we note that equation (15) can in principle yield negative optimal quote sizes. Since in practice market-makers cannot submit negative quantities, we would empirically observe that they do not submit limit orders. To deal with this selection bias, if a market-maker is only active on one side of the book in a given snapshot, we add an order with a zero quantity at the end of the queue on the inactive side. This replaces the selection bias with a censoring bias, where we observe zero values even though the theoretical zero-profit order size is negative.

(iv) The panel only includes marginal quotes of market makers; that is, we do not study the impact of adverse selection and inventory concerns on limit orders submitted by non-market maker account. However, when computing the queue position and the quantity ahead of a **MM** order, we take into account both **MM** and non-**MM** orders.

The resulting panel has about eight million snapshot-market maker quotes, over 158 trading days, with 3,478 snapshots per day and 14.5 quotes per snapshot on average. We describe its main summary statistics in Table 2. We see that there is very little variation in the sizes of submitted limit orders, as indeed 82.4% of limit orders are either for 0, 1 or 2 contracts.

[Insert Table 2 here]

Market-makers hold small inventories on average (0.64 contracts). However, the inventory distribution is rather fat-tailed: with small probabilities, market-makers hold large inventories (up to 1,000 contracts – either long or short). As a result, the inventory measure has a sizeable standard deviation (48.94 contracts).

The three traded instruments in our sample are extremely liquid, as the bid-ask spread is 0.99 basis points on average with a book depth of 49.59 contracts (relative to an average trade size of 4.68 shares[8]). The tick size binds in 51.3% of order book snapshots.[9] Finally, on average a market maker resting quote sits behind 9 other limit orders with higher time priority.

In the remainder of the section, we use the snapshot-market maker panel to investigate the impact of inventory concerns and adverse selection (as a function of queue priority) on quote sizes.

Inventory concerns. We turn first to a preliminary analysis of inventory concerns. Figure 7 showcases that market makers' limit order size is strongly correlated with their inventory. We bin market makers' signed inventory into five quintiles. Market makers in the bottom quantile (i.e., holding large short positions) post significantly larger orders on the bid side of the book than on the ask side. That is, they are primarily interested in buying the asset and converging towards a zero net position. Conversely, market makers in the top quintile (i.e., holding sizeable long positions) post larger orders on the ask side in an attempt to reduce their position. The difference between top and bottom quintile is about 0.9 contracts, which is sizeable relative to the average limit order of 1.84 contracts. The left panel of Figure 7 illustrates the relationship using raw (modified) inventory values, whereas the right panel uses the residuals from regressing inventory on the queue size ahead of the **MM**'s quote (in contracts), the length of the queue for the given snapshot, market depth, as well as instrument and date fixed effects.

[8] We consider all trades initiated by the same account, in the same direction, and with the same second timestamp as coming from a single order.

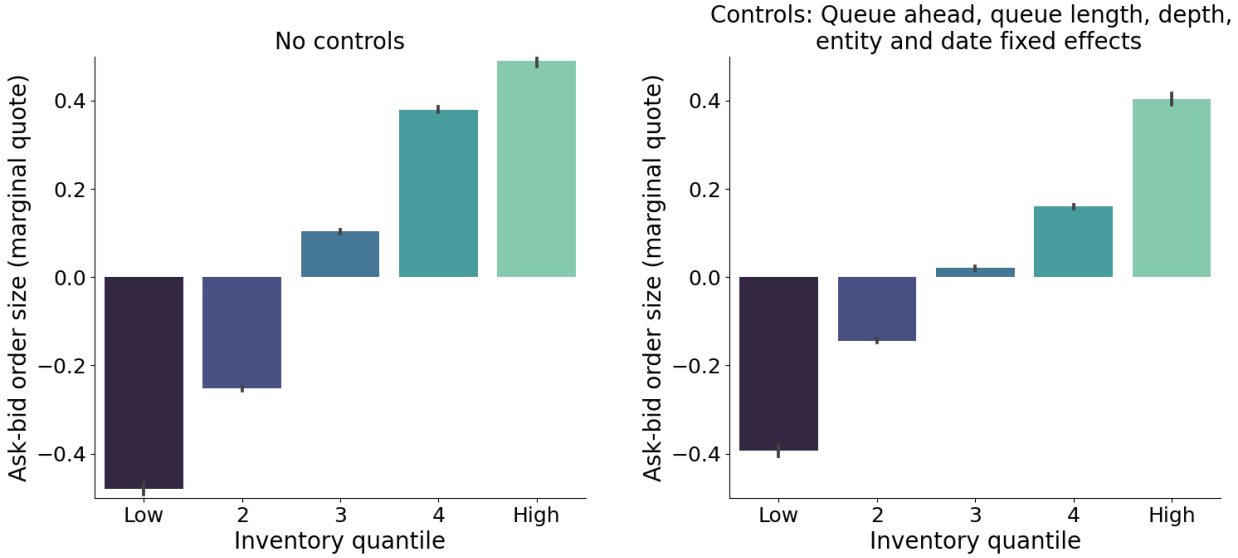
[9] The tick size is CA$0.01 per CA$100 nominal value for bonds futures and 0.10 index points for equity futures.

Figure 7: Market maker inventory and limit order sizes

This figure plots the difference between market maker ask and bid order sizes, as a function of the market maker's modified inventory. If, for a given snapshot, a market maker has a quote only on the bid (ask) side of the book, the ask (bid) order size is set to zero. The market maker modified inventory is defined as in equation (19): it is negative if the market maker has a short position and positive otherwise. In the left panel, we generate inventory quantiles based on the raw modified inventory (in contracts). In the right panel, we use the residuals from the following regression model:

$$\text{Mod. Inventory}_{it} = \phi_0 + \phi_1 \text{Queue ahead}_{it} + \phi_2 \text{Queue length}_i + \phi_3 \text{Market depth}_i + \text{Entity F.E.} + \text{Date F.E.} + \text{error},$$

where i runs over market-makers, t runs over snapshot, the queue ahead a given quote is measured in contracts, queue length for the snapshot is measured in number of orders, and the market depth is measured in contracts across both sides of the order book.



Queue position. Next, we investigate how the market maker's position in the queue affects quote sizes. A key prediction of the model is that limit orders deeper in the book face higher adverse selection risk and should therefore be smaller. Figure 8 provides preliminary evidence of adverse selection by plotting the average limit order size as a function of queue priority. The left-side panel shows the unconditional order size against the position in the time-priority queue. We notice an U-shaped pattern: at first, orders further back in the queue are smaller, consistent with an adverse selection mechanism. However, further back in the queue, the average order size increases in the queue position. We find that the U-shaped pattern is explained away by controlling for the length of the queue. We propose the following intuition: most snapshots have only a few orders in the queue – indeed, 41% of snapshots have fewer than 5 orders. Snapshots with many orders correspond to periods with attractive market making conditions (e.g., periods of low volatility and adverse selection), meaning market-makers quote larger sizes. After controlling for this market condition (total depth), we find that the quote size decreases monotonically in the queue position (right-side panel of Figure 8). Moving from the first to the second position in the queue has the

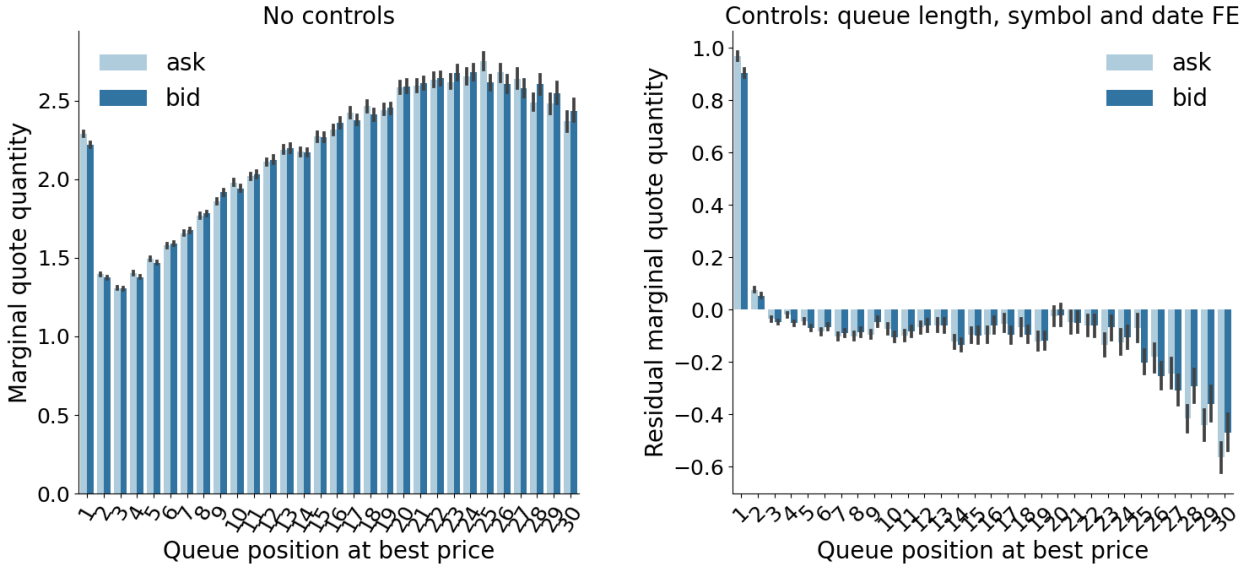
largest impact on quote size – approximately 0.87 contracts. On average, market-makers at front of the queue submit orders 1.01 contracts larger than everyone else. The effect is comparable to the quote size impact of moving from the top to the bottom quintile of inventory (i.e., 0.80 contracts, as illustrated in the right panel of Figure 7).

Figure 8: Queue position and limit order size

This figure plots the average marginal quote size for market makers against the queue position. A larger queue position corresponds to a lower execution priority. If, for a given snapshot, a market maker has a quote only on the bid (ask) side of the book, the ask (bid) order size is set to zero. In the left panel, we plot raw quote sizes (in contracts). In the right panel, we use the residuals from the following regression model:

$$\text{Quote size}_{it} = \phi_0 + \phi_1 \text{Queue length}_i + \text{Entity F.E.} + \text{Date F.E.} + \text{error},$$

where i runs over market-makers, t runs over snapshots, and the queue length for the snapshot is measured in number of orders.



Regression analysis. We formally test the model predictions by mapping the expression for equilibrium order size in Proposition 1 to a linear regression framework. First, let Q_j^{ASK} and Q_j^{BID} denote the last limit order (i.e., with the lowest priority) of market maker j on the best ask or bid price level. From Equation (15), they satisfy:

$$Q_j^{ASK} = \frac{p_1 - V + \gamma \tilde{I}_j - \lambda \left(\sum_{k=1}^{j-1} Q_k + \phi \right)}{\gamma + \lambda} \text{ and } Q_j^{BID} = \frac{V - p_{-1} - \gamma \tilde{I}_j - \lambda \left(\sum_{k=1}^{j-1} Q_k + \phi \right)}{\gamma + \lambda}. \tag{20}$$

We note that the index j identifies a market maker according to the priority rank of his last limit order in the queue in a particular snapshot and side of the book. In other snapshots, and on the other side of the book, j might identify another market maker, depending on the queuing sequence

of that particular snapshot and side.

To estimate a single regression model we sign the order size and introduce

$$\tilde{Q}_j = \begin{cases} Q_j^{ASK} & \text{if quote on the ask side, and} \\ -Q_j^{BID} & \text{if quote on the bid side.} \end{cases} \tag{21}$$

After careful manipulation, we rewrite Equation (21) as a linear function of several variables,[10] which can then be estimated directly with the following linear regression:

$$\tilde{Q}_{j,t} = a + bD_t^{\text{ask}} + c \times D_t^{\text{ask}} \times \mathit{QuotedSpread}_t + d \times D_t^{\text{side}} \times \sum_{k=1}^{j-1} Q_{k,t} + e \times \tilde{I}_{j,t} + \varepsilon_{j,t}, \tag{23}$$

where t runs over snapshots and j over market-makers active in the snapshot (in order of time priority). The dummy D_t^{side} takes value one for ask limit orders and minus one for bid limit orders, respectively. The dummy D_t^{ask} takes value one for ask limit orders and zero otherwise. Each regression coefficient maps to an exact quantity in the model. To bridge the gap between theory and data, we add an error term such that the model prediction holds approximately. Finally, fixed income and equity futures contracts have different average dollar settlement values.[11] To be able to meaningfully pool observations across asset classes, we standardize the queue ahead ($\sum_{k=1}^{j-1} Q_{k,t}$) and inventory ($\tilde{I}_{j,t}$) variables at traded instrument level. Standardizing the key variable for each asset allows us to control for the fact that different instruments might have different volatility levels, which would impact the magnitude of risk-aversion (γ) and adverse selection (λ) coefficients.

Mapping the equilibrium order size in (20) to the regression model (23), we identify coefficients $d = -\frac{\lambda}{\lambda+\gamma}$ and $e = \frac{\gamma}{\lambda+\gamma}$.[12] These correspond to two key predictions of the model: the limit order size decreases in the length of the queue ahead due to adverse selection concerns (i.e., $d < 0$) and increases in the magnitude of inventory to mean-revert ($e > 0$). We formally test the two hypotheses and present the results in Table 3.

In Table 3, we document that both adverse selection and inventory concerns matter for the size of marginal quotes posted by market-makers, with the expected signs. Queue sizes and

[10] Specifically, the model states that

$$\tilde{Q}_j = \frac{p_1 - V + \lambda\phi}{\gamma+\lambda} - \frac{2\lambda\phi}{\gamma+\lambda} D_t^{\text{ask}} + \frac{1}{\gamma+\lambda} \times D_t^{\text{ask}} \times \mathit{QuotedSpread}_t - \frac{\lambda}{\gamma+\lambda} \times D_t^{\text{side}} \times \sum_{k=1}^{j-1} Q_{k,t} + \frac{\gamma}{\gamma+\lambda} \times \tilde{I}_{j,t} \tag{22}$$

[11] For example, an S&P/TSX 60 futures contract (SXF) has a multiplier of 200 over the index value, corresponding to a settlement value of approximately CA$250,000. In contrast, one contract in either five or ten-year Government of Canada bond futures corresponds to for a nominal bond value of CA$100,000.

[12] The OLS regression only allows us to estimate the relative magnitudes of the inventory friction versus adverse selection friction. Identifying γ and λ individually would require additional structure on the coefficients (for example, with GMM). We prefer the OLS specification which has both a simple and a structural interpretation.

inventories are standardized, for ease of interpretation. The coefficient on *queue ahead* $\times d_{side}$ equals -0.146: that is, a one-standard deviation increase in the queue size ahead of an order reduces the order size by 0.146 contracts – a 7.9% drop relative to the sample average of 1.84 contracts. The structural interpretation of this coefficient is $-\frac{\lambda}{\lambda+\gamma}$ and reflects adverse selection (λ) scaled by the total marginal cost of liquidity supply $(\lambda + \gamma)$.

The coefficient on *Inventory* is 0.152: if a market maker's inventory increases by one standard deviation, the limit order size on the ask (bid) side increases (decreases) by 0.152 contracts. The structural interpretation of this coefficient is $\frac{\gamma}{\lambda+\gamma}$, reflecting the inventory component of the marginal cost of liquidity supply. The two coefficient have similar magnitudes, suggesting that adverse selection and inventory concerns are, on average, equally important factors in the Canadian Government Bond and Equity Index futures market.[13]

The economic magnitudes we find are economically relevant. The signs are consistent with the intuition of the model, although technically they violate it. According to the model, $d = -\frac{\lambda}{\lambda+\gamma}$ and $e = \frac{\gamma}{\lambda+\gamma}$, such that $e - d = 1$, while our estimates are $0.152 - (-0.146) = 0.298$; or about 30% of the theoretical prediction. We still think the 30% is sizeable, given that the theory makes strong assumptions on the quadratic inventory penalty and ignores, for example, dynamic considerations or that any trader may submit limit orders—not just the market makers.

[Insert Table 3 here]

4.2 Time priority and quoted depth

In Section 4.1, we document that adverse selection and inventory concerns individually impact quote sizes. We argue that the interaction between the two channels is particularly relevant in markets with time priority: adverse selection represents a cost to unwind inventory, and adverse selection worsens for orders in the back of the queue. Our theoretical framework makes sharp and novel predictions about the impact of time priority rules on liquidity. In particular, if market makers have heterogeneous inventory levels then the aggregate best-price quoted depth depends on their arrival sequence (i.e., on the time priority for execution). From Corollary 1, quoted depth on the ask (bid) side of the book is maximized when the market-makers with largest long (short) inventories are at the back of the queue. The intuition is that a limit order increases adverse selection of all subsequent orders in the queue. When a market maker with a long inventory arrives early in the queue, his large order will crowd out the orders by subsequent market makers.

To quantify the effect, we compute the Spearman rank correlation coefficient (ρ) between market-maker modified inventory and the queue position of their marginal quote for each order book

[13] There is heterogeneity across asset classes. Adverse selection is more salient for stock futures ($d = -0.176$) than for bonds ($d = -0.095$), whereas the opposite is true for inventory frictions—$e = 0.195$ for bonds and $e = 0.07$ for stock futures.

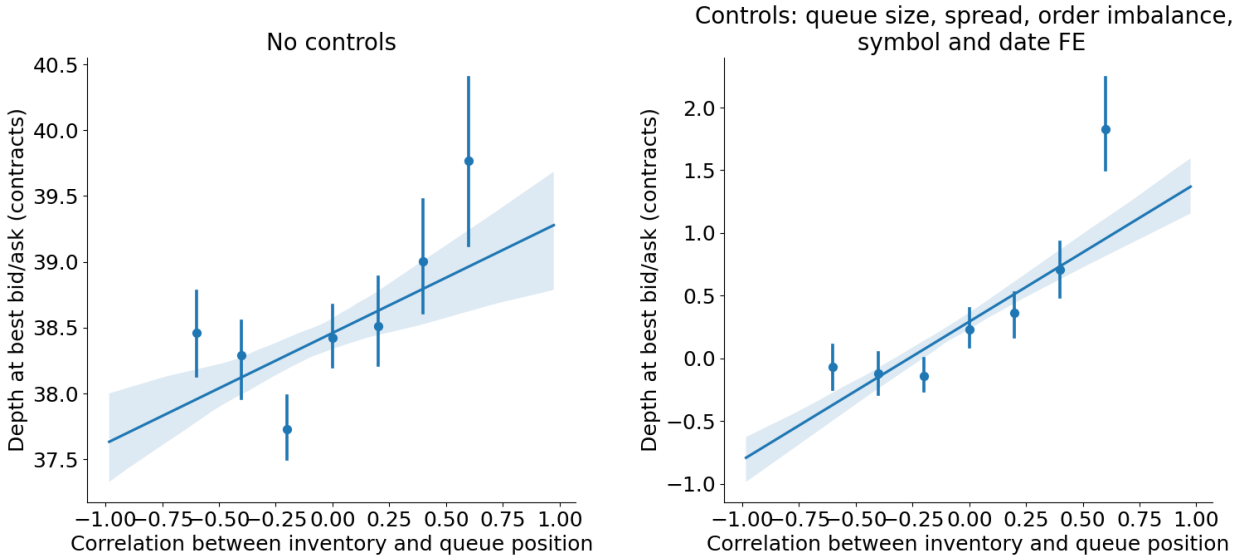
snapshot in our data – the same measure we used in Figure 3. We then test whether this correlation is positively related to the quoted depth, as predicted by theory. To make sure the correlation is estimated on sufficient observations in a given snapshot, we only select snapshots where at least three-quarters of market makers are present at the market: that is, we require nine market makers with resting quotes at the top level in the book. The final sample consists of 125,242 order book snapshots across both the bid and the ask side (out of 1,697,746 snapshots).

Figure 9: Market-maker arrival sequence and aggregate depth

We plot the market depth posted by market-makers against $\rho \times d_{\text{side}}$, where d_{side} is a dummy variable taking the value one (minus one) for ask-side (bid-side) quotes. Market depth is defined as the sum of all quote sizes posted by market-makers j in the respective order book side snapshot t (i.e.,$\sum_{tj}$ Quote size$_j$) In the left panel, we use the raw depth in contracts. In the right panel, we use the residuals from the following regression model:

$$\sum_{tj} \text{Quote size}_j = \phi_0 + \phi_1 \text{Queue length}_i + \phi_2 \text{Quoted spread} + \phi_3 \text{Order imbalance}_i + \text{Entity F.E.} + \text{Date F.E.} + \text{error},$$

where j runs over market-makers, t runs over snapshots, and the queue length for the snapshot is measured in number of orders. The quoted spread (in basis points) is defined as the difference between prevailing ask and bid prices, scaled by the midpoint. Order imbalance is defined as the the difference between ask and bid market depths, scaled by their sum.



In Figure 9 we plot the market depth posted by **MM**s against the modified correlation coefficient, and find a positive relation. Recall that in Section 2.2.3 and Figure 3, we documented that arrival sequence is very weakly correlated with inventory, implying significant queuing randomness. Does randomness in arrival sequence translate to predictable variation in market depth? Figure 9 suggests that it does. To facilitate interpretation, we pool observations on the bid and ask side by multiplying the quote-inventory correlation by a dummy (d_{side}) taking value one for ask-side snapshots and minus one for bid-side snapshots. A higher value of the modified correlation now

implies that **MM**s with largest inventory to mean-revert are at the back of the queue. In line with the model, market depth increases in the modified correlation coefficient, both when looking at raw values (left panel) and when we control for factors such as the queue size, quoted spread, and order imbalance (right panel).

To formally test the model prediction, we estimate the following regression model:

$$\sum_j Q_{j,s,t} = \psi \times \rho_{s,t} \times d_{\text{side}} + \text{Controls} + \text{Symbol and Date FE} + \text{error}, \tag{24}$$

where j runs over market-makers, s is the side of the book (ask or bid), and t indexes snapshots. We compute market depth by summing all quote sizes across market-makers j present at the market. The d_{side} dummy is one (or minus one) for ask-side (bid-side) snapshots.

We show the results in Table 4. In line with the model, the coefficient on $\rho \times d_{\text{side}}$ is positive and statistically significant. If we transition from a setting where market-makers arrive at random ($\rho = 0$) to one where market-makers with highest incentive to trade are placed at the back of the queue ($\rho = 1$), market depth would increase by 1.1 contracts. The magnitude corresponds to a 4.2% increase relative to the average **MM**-posted depth of 26 contracts. Further, model (5) in Table 4 suggests the effect is highly symmetric across the bid and ask sides of the book.

[Insert Table 4 here]

From a normative lens, there is a 8.4% market depth gap between a priority sequencing that maximizes market-maker utility ($\rho = -1$, i.e., **MM**s with the most pressing trading needs execute first) and the sequencing that maximizes market depth ($\rho = 1$).

Corollary 1 yields a related testable prediction: the marginal impact of inventory on aggregate depth is higher for limit orders at the back of the time priority queue. For each snapshot with at least four **MM** resting limit orders, we regress the market-maker quoted depth on the modified inventory for the first, second, third, and fourth market-makers in the queue:

$$\sum_j Q_{j,s,t} = \delta_1 \tilde{I}_1 + \delta_2 \tilde{I}_2 + \delta_3 \tilde{I}_3 + \delta_4 \tilde{I}_4 + \text{Controls} + \text{Symbol and Date FE} + \text{error}, \tag{25}$$

where t runs over snapshots, j runs over market-makers active in the snapshot (in order of time priority). We estimate the regression separately for the ask and bid sides of the book and show the results in Table 5. As expected, all δ coefficients are positive for the ask-side regressions (high inventory increases **MM** willingness to sell) and negative for the bid-side regressions (high inventory reduces **MM** willingness to buy). In line with Corollary 1, the δ coefficient estimates increase in absolute value, almost monotonically, for orders further back in the queue. For example, a one-standard deviation increase in inventory for the first **MM** in the queue leads to a increase in

ask-side depth of 0.287 contracts. However, the marginal effect of inventory is 13% higher for the fourth market-maker in the queue: a one-standard deviation inventory jump translates to a 0.326 increase in depth.

[Insert Table 5 here]

4.3 A measure of risk sharing inefficiency

Inventory divergence likely translates to inefficient risk-sharing across market makers.[14] We follow the model of Section 3, and assume market makers constantly incur a linear penalty γ_j over their quadratic holdings. A back-of-the-envelope measure of risk-sharing inefficiency is the ratio between actual quadratic holdings (summed over all market makers) divided by the holdings if the market makers would share risks perfectly:

$$\textit{RiskSharingInefficiency}_{i,d} = \frac{\sum_{s\in d}\sum_{j=1}^{J}\gamma_j I_j^2}{\sum_{s\in d}\sum_{j=1}^{J}\gamma_j\left(\sum_{j=1}^{J}\omega_j I_j\right)^2}, \tag{26}$$

where i, d, and s run over traded instruments, days, and 30-second snapshots respectively. There are J market-makers indexed by j, with inventory penalty γ_j and inventory level I_j (measured in contracts). Finally, ω_j is the optimal share of the aggregate inventory held by **MM** j such that inventory costs are minimized across the market maker sector.

The numerator in equation (26) stands for the sum of quadratic inventory penalties across all market-makers j for all snapshots on day d. The denominator is the sum of counterfactual penalties, assuming that the J market makers each take an equal share of the aggregate inventory. A value larger than one indicates imperfect risk sharing, as the aggregate inventory costs are larger than under the benchmark of equal inventories. By Jensen's inequality, the measure is equal to one only under perfect risk sharing.

One challenge is to appropriately choose values for the inventory penalty parameters γ_j. A simple approach, following the model, is to assume all market makers are equally risk averse and share the same γ. The advantage of this assumption is that the unknown parameter γ in (26) drops out (to minimize aggregate costs, each **MM** holds a fraction $\omega_j = \frac{1}{J}$ of the aggregate inventory). We compute the risk-sharing inefficiency for each product and day in the sample and obtain that realized inventory costs are almost nine times as large as under perfect risk sharing.

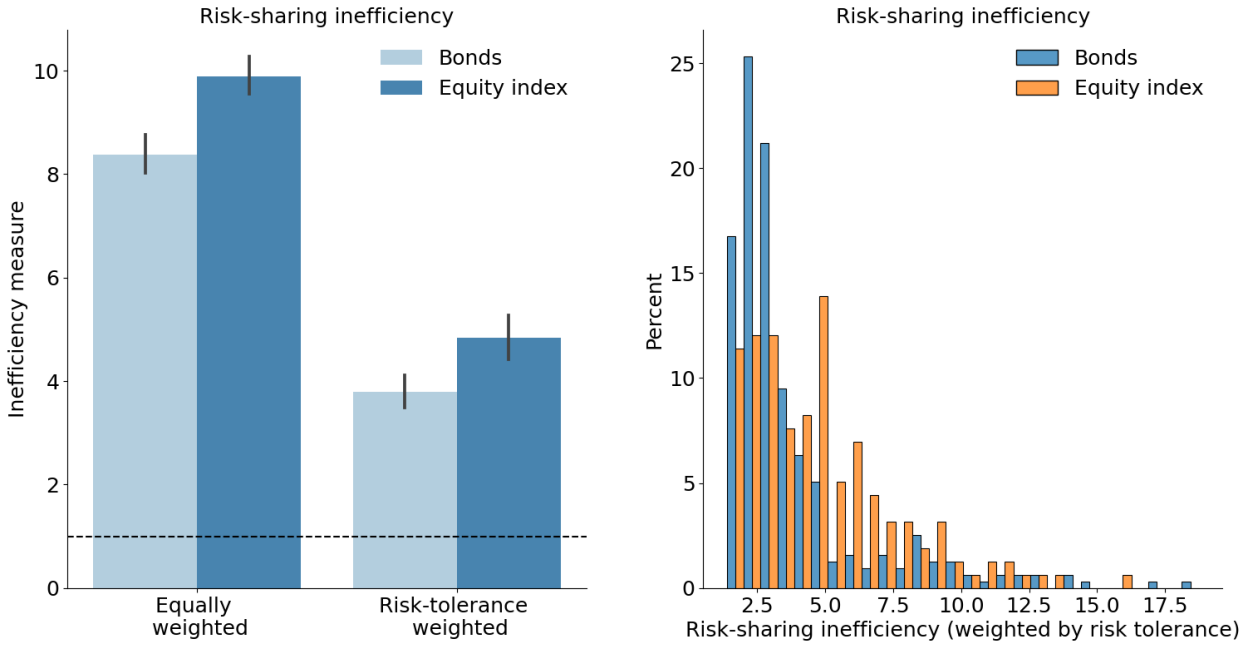
[14] We admit that we do not observe the entire portfolio holdings of the market makers, so there may be offsetting positions in correlated assets. We are not overly worried, since the market makers end the day with a flat inventory more than 72% of the time. In case of offsetting positions in correlated assets there would be no need to end the day exactly flat.

However, assuming homogeneous inventory penalties is likely to over-estimate risk sharing inefficiency. If some market makers are less risk averse, they naturally accommodate larger inventory swings. To account for this effects, we assume that each γ_j is inversely proportional to the standard deviation of inventory for account j, or $\gamma_j = \frac{\gamma^*}{SD(I_j)}$, for a constant cost γ^* that applies to all. That is, market-makers with larger inventory swings are considered to be more risk tolerant and therefore have a lower value of γ_j. To minimize aggregate costs, market maker j holds a fraction $\omega_j = \frac{SD(I_j)}{\sum_j SD(I_j)}$ of the aggregate inventory – that is, proportional to her risk tolerance. With this approach, the unobservable γ^* drops out. We find that the average "risk-tolerance weighted" inefficiency is 4.13: that is, realized inventory costs are more than 300% larger than the case of optimal risk sharing. In Figure 10, we show that average inefficiency is of similar magnitude for future contracts on both bonds and the equity index. The right panel of Figure 10 plots the distribution of our inefficiency measure across product-days and illustrates that large deviations from optimal risk sharing (i.e., a 10-times larger cost) are not uncommon, particularly for equity indices.

The market makers seem to share risks poorly, which is puzzling for two reasons. First, open inventory positions are costly to market makers, and they put significant effort in closing the positions as witnessed by the short half-lives of inventory shocks and market makers frequently ending the day flat. The low level of risk sharing across market makers suggests an inefficiency. At least partly, this inefficiency is caused by time priority, which prevents the market makers most eager to trade to appear first in the queue. A second puzzle is that the literature often considers the market making sector as a group of intermediaries who absorb the incoming order flow. Thus implies that in periods of high buying (selling) demand, all market makers should be selling (buying) together, generating positive time-series correlations in inventory positions. Our results on high-frequency market makers cast doubts on this interpretation.

Figure 10: Inefficient risk-sharing across market makers

This figure illustrates the average risk-sharing inefficiency, defined as in equation (26), across asset classes. In the left panel, we compute an equally-weighted inefficiency (each market maker has the same inventory penalty γ) as well as risk-tolerance-weighted inefficiency (for a market maker, γ_j is the inverse of her inventory standard deviation on a given day). The right panel plots the distribution of the risk-tolerance-weighted inefficiency across products and days, separately for each underlying asset class (bonds and equities).



5 Conclusion

This paper sheds light on the cross-section of market maker inventories in modern markets. We document that, while market makers quickly revert their positions within minutes, their inventories are uncorrelated in the cross-section. To better understand the implications of cross-sectional heterogeneity in inventories, we build a model to study the interaction between informational and inventory frictions on high-frequency markets. The model includes time priority, which, unlike other trading protocols, prevents market makers to perfectly share the risk of the incoming order flow. The market makers arrive sequentially, and limit orders in the front of the time-priority queue impose a negative adverse selection externality on back-of-the-queue quotes, as those trade with more informed market orders on average. The magnitude of this externality, however, depends on the inventory of market makers at the front of the queue. Those with large positions to unwind post larger orders, generating higher adverse selection for subsequent arrivals.

We use a proprietary data set from Montréal Exchange to test our model. It offers a novel identification approach to separate adverse selection from inventory frictions.

For the Canadian futures market, we find that both frictions have a similar quantitative impact on order size. In support of our model, we document that depth is 4.1% larger when high-inventory market-markers are at the back of the queue compared to a random arrival sequence. Further, a simple and crude estimate of the risk-sharing inefficiency suggests inventory costs may be four times larger than under an optimal risk-sharing benchmark.

Our model has potential implications for market design. Time priority prevents perfect competition between market makers and permit positive expected profits to fund their business model. Yet, time priority also opens a wedge between the objectives of higher liquidity (maximum depth) and optimal risk-sharing in the cross-section of market makers: the two objectives map to opposite arrival sequences in the queue. We also note that time priority can, by itself, generate heterogeneous inventories. The first market maker in the queue executes her entire order before any market makers trade, such that post-trade inventories tend to diverge. In contrast, optimal risk sharing would entail that all market makers absorb a fraction of each incoming trade.[15] An interesting topic for future research is to determine whether alternative clearing rules (i.e., batch auctions or price-quantity-time priority) can successfully eliminate the liquidity risk and align the two objectives.

[15]Optimal risk sharing could be implemented in batch and double auctions, and in a *pro-rata* limit order book setting common in some futures markets. There, each market maker executes a fraction of the incoming trade proportional to the size of his limit order (see Field and Large, 2008, for a discussion).

References

Back, Kerry, and Shmuel Baruch, 2013, Strategic liquidity provision in limit order markets, *Econometrica* 81, 363–392.

Biais, Bruno, David Martimort, and Jean-Charles Rochet, 2000, Competing mechanisms in a common value environment, *Econometrica* 68, 799–837.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High-Frequency Trading and Price Discovery, *The Review of Financial Studies* 27, 2267–2306.

——— , 2019, Price discovery without trading: Evidence from limit orders, *The Journal of Finance* 74, 1621–1658.

Budish, Eric, Peter Cramton, and John Shim, 2015, The high-frequency trading arms race: Frequent batch auctions as a market design response, *Quarterly Journal of Economics* 130, 1547–1621.

Comerton-Forde, Carole, Terrence Hendershott, Charles M. Jones, Pamela C. Moulton, and Mark S. Seasholes, 2010, Time variation in liquidity: The role of market-maker inventories and revenues, *The Journal of Finance* 65, 295–331.

Degryse, Hans, and Nikolaos Karagiannis, 2019, Priority Rules, CEPR Discussion Papers 14127 C.E.P.R. Discussion Papers.

Du, Songzi, and Haoxiang Zhu, 2017, What is the Optimal Trading Frequency in Financial Markets?, *The Review of Economic Studies* 84, 1606–1651.

Duffie, Darrell, and Haoxiang Zhu, 2017, Size Discovery, *The Review of Financial Studies* 30, 1095–1150.

Field, Jonathan, and Jeremy Large, 2008, Pro-rata matching and one-tick futures markets, CFS Working Paper Series 2008/40 Center for Financial Studies (CFS).

Glosten, Lawrence R., 1994, Is the electronic open limit order book inevitable?, *The Journal of Finance* 49, 1127–1161.

Grossman, Sanford J., and Merton H. Miller, 1988, Liquidity and market structure, *The Journal of Finance* 43, 617–633.

Hansch, Oliver, Narayan Y. Naik, and S. Viswanathan, 1998, Do inventories matter in dealership markets? evidence from the london stock exchange, *The Journal of Finance* 53, 1623–1656.

Hasbrouck, Joel, 1991, Measuring the information content of stock trades, *The Journal of Finance* 46, 179–207.

Hendershott, Terrence, and Albert J. Menkveld, 2014, Price pressures, *Journal of Financial Economics* 114, 405–423.

Hendershott, Terrence, and Mark S. Seasholes, 2007, Market maker inventories and stock prices, *American Economic Review* 97, 210–214.

Ho, Thomas, and Hans R. Stoll, 1981, Optimal dealer pricing under transactions and return uncertainty, *Journal of Financial Economics* 9, 47–73.

Ho, Thomas S. Y., and Hans R. Stoll, 1983, The dynamics of dealer markets under competition, *The Journal of Finance* 38, 1053–1074.

Hollifield, Burton, Robert A. Miller, and Patrik Sandås, 2004, Empirical analysis of limit order markets, *Review of Economic Studies* 71, 1027–1063.

Kirilenko, Andrei, Albert S. Kyle, Mehrdad Samadi, and Tugkan Tuzun, 2017, The flash crash: High-frequency trading in an electronic market, *The Journal of Finance* 72, 967–998.

Kraus, Alan, and Hans Stoll, 1972, Price impacts of block trading on the new york stock exchange, *Journal of Finance* 27, 569–88.

Li, Sida, Xin Wang, and Mao Ye, 2021, Who provides liquidity, and when?, *Journal of Financial Economics* 141, 968–980.

Madhavan, Ananth, and Seymour Smidt, 1991, A bayesian model of intraday specialist pricing, *Journal of Financial Economics* 30, 99–134.

——— , 1993, An analysis of changes in specialist inventories and quotations, *The Journal of Finance* 48, 1595–1628.

Menkveld, Albert J., 2013, High frequency trading and the new market makers, *Journal of Financial Markets* 16, 712–740 High-Frequency Trading.

O'Hara, Maureen, 2015, High frequency market microstructure, *Journal of Financial Economics* 116, 257–270.

——— , and Xing Zhou, 2021, The electronic evolution of corporate bond dealers, *Journal of Financial Economics* 140, 368–390.

Parlour, Christine, and Duane Seppi, 2008, Limit order markets: A survey, *Handbook of financial intermediation and banking* 5, 63–95.

Parlour, Christine A., 1998, Price dynamics in limit order markets, *The Review of Financial Studies* 11, 789–816.

Reiss, Peter C., and Ingrid M. Werner, 1998, Does risk sharing motivate interdealer trading?, *The Journal of Finance* 53, 1657–1703.

Sandås, Patrik, 2001, Adverse Selection and Competitive Market Making: Empirical Evidence from a Limit Order Market, *The Review of Financial Studies* 14, 705–734.

Yao, Chen, and Mao Ye, 2018, Why Trading Speed Matters: A Tale of Queue Rationing under Price Controls, *The Review of Financial Studies* 31, 2157–2183.

Table 1: Trading statistics for market maker (**MM**) and non-market maker (non-**MM**) accounts

This table displays trading summary statistics at the instrument-day-trading account level, separately for market makers (top panel) and non-market makers (bottom panel). Volume is measured in contracts traded. Net position and inventory variation are defined as in equations (1) and (2), respectively. Time at the best bid or offer (BBO) is computed as in equation (3). The statistics in this table are used to classify accounts into **MM** and non-**MM**.

Panel (a): Market maker accounts

Statistic	Trade count	Volume	Net pos. (%)	Inventory variation (%)	Time at BBO (%)
Mean	2,025.91	3,870.19	0.42	0.91	52.92
St. Dev.	2,321.61	4,833.90	2.66	2.32	19.53
Pctl(25)	323	621.5	0	0.16	33.17
Median	1,190	1,792	0	0.37	51.57
Pctl(75)	3,006.8	5,641.8	0.05	0.82	69.31
N	4,784	4,784	4,784	4,784	4,784

Panel (b): Non-market maker accounts

Statistic	Trade count	Volume	Net pos. (%)	Inventory variation (%)	Time at BBO (%)
Mean	82.70	229.78	73.25	55.50	1.92
St. Dev.	258.77	802.75	38.61	31.00	4.32
Pctl(25)	3	5	39.3	27.0	0.14
Median	14	25	100	66.1	0.53
Pctl(75)	62	138	100	78.7	1.92
N	137,606	137,606	137,606	137,606	130,119

Table 2: Summary statistics for the market-makers' marginal quote panel

Statistic	Quote size	Inventory	Book depth	Quoted spread (bps)	Queue position
Mean	1.84	0.64	49.59	0.99	8.95
St. Dev.	3.62	48.94	57.95	0.48	7.51
Pctl(25)	0	−5	8	0.69	3.00
Median	1	0	34	0.79	7.00
Pctl(75)	2	6	71	0.97	13.00
N	8,002,806	8,002,806	8,002,806	8,002,790	8,002,798

Data from 30-second order book snapshots with active market-makers

Table 3: Impact of queue length and inventory on limit order size

This table reports the coefficients of the following regression (equation 23):

$$\tilde{Q}_{j,t} = a + bD_t^{\text{ask}} + c \times D_t^{\text{ask}} \times QuotedSpread_t + d \times D_t^{\text{side}} \times \sum_{k=1}^{j-1} Q_{k,t} + e \times \tilde{I}_{j,t} + \varepsilon_{j,t},$$

where the dependent variable $\tilde{Q}_{j,t}$ is the marginal quote of the market-maker with priority j in snapshot t, signed as in equation (21). Quoted spread is measured in basis points, as the difference between the ask and the bid prices scaled by the midpoint prevailing in the snapshot. D_t^{ask} and D_t^{side} are two dummy variables taking value one for ask quotes and zero (respectively, minus one) for bid quotes. The queue ahead of a marginal quote is measured in contracts. $\tilde{I}_{j,t}$ is the modified inventory for market-maker j, using the methodology in equation (19). We standardize both the queue ahead and inventory variables at instrument-day level. The sample includes all marginal quotes from market-maker accounts, sampled in 30 second snapshots, from January 1st to August 18th, 2021. The data covers all maturity contracts in three futures instruments traded on the Montréal Exchange: on Five- and Ten-Year Government of Canada Bond futures, as well as the S&P/TSX 60 blue-chip equity index.

	Dependent variable:				
	Marginal quote size				
	(1)	(2)	(3)	(4)	(5)
d_{ask}	1.808**	1.808**	1.811**	1.817**	2.031***
	(3.147)	(3.147)	(3.168)	(3.197)	(3.743)
quoted spread $\times d_{\text{ask}}$	0.00005	0.00005	0.0002		0.0001**
	(1.506)	(1.482)	(0.155)		(2.355)
queue ahead $\times d_{\text{side}}$	−0.146***	−0.146***	−0.145***	−0.144***	
	(−3.560)	(−3.561)	(−3.610)	(−3.630)	
order priority $\times d_{\text{side}}$					−0.034*
					(−2.212)
Inventory	0.152***	0.152***	0.153***	0.153***	0.144***
	(3.733)	(3.733)	(3.688)	(3.681)	(3.864)
queue length	0.071***	0.071***	0.071***	0.071***	0.086***
	(3.961)	(3.964)	(3.852)	(3.676)	(3.915)
book depth $\times d_{\text{side}}$	0.009	0.009			−0.016
	(0.108)	(0.108)			(−0.205)
order imbalance $\times d_{\text{side}}$	0.018**				
	(2.604)				
Symbol FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Trader FE	Yes	Yes	Yes	Yes	Yes
Observations	8,002,790	8,002,790	8,002,790	8,002,798	8,002,790
Adjusted R^2	0.234	0.234	0.234	0.234	0.235

Note: *p<0.1; **p<0.05; ***p<0.01

Standard errors clustered at symbol, date, and trader level.

Table 4: Time priority sequence and market depth

This table reports the coefficients of the following regression (equation 24):

$$\sum_j Q_{j,s,t} = \psi \times \rho_{s,t} \times d_{\text{side}} + \text{Controls} + \text{Symbol and Date FE} + \text{error},$$

where the dependent variable $\sum_j Q_{j,s,t}$ is the sum of marginal quotes across all market-makers j in snapshot t on book side s. $\rho_{s,t}$ is the Spearman correlation coefficient between queue priority and modified market-maker inventory in snapshot t on book side s. d_{ask} and d_{side} are two dummy variables taking value one for ask quotes and zero (respectively, minus one) for bid quotes. d_{bid} is defined as $1 - d_{\text{ask}}$. The queue size is measured in contracts. Order imbalance is defined as the the difference between ask and bid market depths, scaled by their sum. The aggregate inventory is defined as the sum of modified inventories across market-makers ($\sum_j \tilde{I}_{j,s,t}$). The sample includes all order book snapshots where at least 75% of market-making accounts are active, from January 1st to August 18th, 2021. The data covers all maturity contracts in three futures instruments traded on the Montréal Exchange: on Five- and Ten-Year Government of Canada Bond futures, as well as the S&P/TSX 60 blue-chip equity index.

	Dependent variable:				
	Market-maker quoted depth (contracts) on book side				
	(1)	(2)	(3)	(4)	(5)
$\hat{\rho}$(queue, inventory) $\times$ d_{side}	0.520	0.981**	0.972**	1.100***	
	(0.701)	(3.617)	(3.543)	(4.356)	
$\hat{\rho}$(queue, inventory) $\times$ d_{ask}					1.118***
					(4.207)
$\hat{\rho}$(queue, inventory) $\times$ d_{bid}					−1.082**
					(−3.960)
queue size		1.772***	1.772***	1.693***	1.693***
		(16.927)	(16.906)	(18.348)	(17.955)
order imbalance			0.004	0.004	0.004
			(1.913)	(1.882)	(1.859)
quoted spread (bps)				11.534***	11.534***
				(8.738)	(8.738)
aggregate inventory (x 100)				−0.122	−0.123
				(−0.987)	(−0.955)
Symbol FE	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes
Observations	125,242	125,242	125,242	125,242	125,242
Adjusted R^2	0.188	0.677	0.678	0.690	0.690

Note: *p<0.1; **p<0.05; ***p<0.01
Standard errors clustered at symbol, date, and timestamp level

Table 5: Time priority and the marginal impact of inventory on depth

This table reports the coefficients of the following regression (equation 25):

$$\sum_j Q_{j,s,t} = \delta_1 \tilde{I}_1 + \delta_2 \tilde{I}_2 + \delta_3 \tilde{I}_3 + \delta_4 \tilde{I}_4 + \text{Controls} + \text{Symbol and Date FE} + \text{error},$$

where the dependent variable $\sum_j Q_{j,s,t}$ is the sum of marginal quotes across all market-makers j in snapshot t on book side s. $\tilde{I}_j$ is the modified inventory of market-maker j, defined as in equation (19). We standardize inventory at symbol and trader level to account for the fact that some market-makers tolerate larger inventory swings than others. Priority levels are defined across market-makers: that is, priority #4 implies this is the fourth market-maker in the queue and not necessarily the fourth trader overall. Quoted spread is measured in basis points, as the difference between the ask and the bid prices scaled by the midpoint prevailing in the snapshot. Order imbalance is defined as the the difference between ask and bid market depths, scaled by their sum. The sample includes all order book snapshots where at least four market-maker accounts are active, from January 1st to August 18th, 2021. The data covers all maturity contracts in three futures instruments traded on the Montréal Exchange: on Five- and Ten-Year Government of Canada Bond futures, as well as the S&P/TSX 60 blue-chip equity index.

	Dependent variable:					
	Market-maker quoted depth					
	Ask side			Bid side		
	(1)	(2)	(3)	(4)	(5)	(6)
Inventory: Priority #1	0.187***	0.189***	0.287***	−0.185***	−0.186***	−0.262***
	(3.411)	(3.374)	(5.301)	(−3.519)	(−3.666)	(−5.797)
Inventory: Priority #2	0.212**	0.215**	0.303***	−0.229***	−0.232***	−0.301***
	(3.166)	(3.140)	(4.471)	(−5.239)	(−5.481)	(−6.924)
Inventory: Priority #3	0.240***	0.244***	0.316***	−0.277***	−0.280***	−0.343***
	(4.098)	(4.021)	(5.155)	(−5.012)	(−5.267)	(−5.627)
Inventory: Priority #4	0.251**	0.254**	0.326***	−0.244**	−0.247**	−0.298***
	(3.118)	(3.212)	(3.740)	(−3.137)	(−3.238)	(−4.013)
quoted spread (bps)		0.617	1.089		0.641	0.982
		(0.724)	(1.244)		(0.727)	(1.068)
order imbalance			0.023**			−0.018*
			(2.876)			(−2.200)
Symbol FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	372,447	372,447	372,447	373,430	373,430	373,430
Adjusted R^2	0.319	0.319	0.334	0.324	0.324	0.334
Note:						*p<0.1; **p<0.05; ***p<0.01
						Standard errors clustered at symbol, date, and timestamp level