**USC FBE FINANCE SEMINAR**
*presented by Brian Weller*
**FRIDAY**, Feb. 15, 2013
**10:30 am – 12:00 pm**, **Room: JKP-104**

# Liquidity and High Frequency Trading

Brian Weller[⇤]

Job Market Paper

November 10, 2012

Abstract

I explore how liquidity provision varies among intermediaries in asset markets. Intermediating high frequency traders (HFTs) in particular appear fundamentally different from other market makers. Using trader-identified transaction data from the Commodity Futures Trading Commission for gold, silver, and copper futures markets, I shed light on variation in liquidity supply and high frequency trading. I show market maker "speed" can explain the cross-section of intermediary spreads, volumes, and profits observed in the data. I also find a symbiotic risk-sharing relationship between HFTs and slower market makers. The resulting retrading among market makers generates long sequences of intermediaries between terminal sellers and buyers of assets and delinks trading volume from liquidity. I rationalize these results in the context of a model of a speed hierarchy. The model additionally suggests speed heterogeneity may increase short-run intermediation capacity but diminish long-run capacity via a high frequency arms race.

JEL classification: D53, D82, G10, G12, G20.

Keywords: High Frequency Trading, Intermediation Chains, Liquidity, Speed Arms Race, Intermediation Capacity.

[⇤] Booth School of Business and Department of Economics, University of Chicago. bweller@chicagobooth.edu. Tel: +1 216 469 2610. Website: http://home.uchicago.edu/ bweller/. I would like to thank my dissertation committee for their frequent advice and guidance: John Cochrane, Bryan Kelly, Toby Moskowitz, and Lubos Pastor. I also would like to thank Eric Budish, Stefano Giglio, Richard Haynes, Andrei Kirilenko, Jennifer La'O, Juhani Linnainmaa, Amit Seru, Pietro Veronesi, Robert Vishny, and finance seminar participants at Chicago Booth. I gratefully

acknowledge financial support from the John and Serena Liew Fellowship Fund at the Fama-Miller Center for Research in Finance and the University of Chicago Booth School of Business as well as data support from the Commodity Futures Trading Commission.

# 1 Introduction

Intermediaries in asset markets, or market makers, improve the ability of traders to complete trades quickly with small transaction costs. However, market makers do not provide liquidity equally. Market makers often charge different fees and intermediate orders selectively. Some intermediate large volumes while maintaining small positions, whereas others primarily reintermediate trades to absorb the risk of other market makers.

Speed—or the ability of market makers to access and generate orders quickly—is among the most salient dimensions of market maker heterogeneity. The fastest market makers are high frequency traders (HFTs), a category defined by its use of algorithms, low-latency technology, and high message rates.[1] Despite their small numbers, HFTs comprise roughly half of all trading volume in U.S. stock and futures markets.[2] Few groups have provoked more heated debate among regulators and market participants on their effect on market quality. Proposals to restrict HFT activity range from minimum quote lifetimes in the United States to financial transaction taxes in Europe. Nevertheless, despite a concerted effort to regulate HFTs,[3] few studies of high frequency trading exist in the academic literature. This work adds to the HFT debate by establishing empirical facts on their characteristics and shedding light on their function in the intermediation ecosystem.

In this paper, I show heterogeneity is an important feature of the market making sector. In particular, I study dispersion among market maker speeds to explain several features of aggregate and cross-sectional intermediation activity. I find long chains of intermediaries of varying speeds separating terminal sellers and buyers, and strong relations between market maker speeds and spreads, volumes, and profits. I unify these results in a model of a speed hierarchy in which fast market makers have privileged access to oncoming order flow, and slow market makers exist to intermediate marginal trades and absorb risk. The model also suggests speed heterogeneity can improve market liquidity in the short run, but can worsen long-run liquidity by incentivizing a potentially destructive high frequency arms race.

I use audit-trail data from the Commodity Futures Trading Commission (CFTC) to achieve millisecond resolution on intraday intermediation activity. The data capture all trades

[1] CFTC Technology Advisory Committee, Sub-Committee on Automated and High Frequency Trading Working Group 1, October 30, 2012, http://www.cftc.gov/.

[2] Remarks by Commodity Futures Trading Commission Commissioner Scott D. O'Malia, June 19, 2012, http://www.cftc.gov/.

[3] The U.S. Securities and Exchange Commission, the U.S. Commodity Futures Trading Commission, the European Securities and Market Authority, and Hong Kong's Securities and Futures Commission are among the many regulatory agencies considering or implementing limits on high frequency trading activity. Several legislatures, including the U.S. Senate, and large exchanges also

are contemplating additional controls or restrictions on these traders.

Page 3

in gold, silver, and copper commodity futures contracts traded on the Chicago Mercantile Exchange (CME) during the week of December 12-16, 2011. The CFTC also provides account identifying information for both parties involved in each trade. This unique feature enables comprehensive analysis at the level of individual traders and market makers.

The data reveal considerable speed heterogeneity among market makers. By one measure, the fastest 5% of active market makers, which I label high frequency traders, are 40-50 times faster than the slowest 50%. HFTs play a central role in market making and comprise 45%, 44%, and 34% of intermediation dollar volume in gold, silver, and copper futures, respectively. Their outsized volumes relative to their numbers suggest speed plays a central role in modern market making. In fact, in a regression of market maker outcomes on speeds, I find a one unit increase in log speed is associated with a 6%-12% improvement in spreads, a 59%-80% increase in volume, and a 60%-84% growth of profits. Given these pronounced differences in volumes and profits, the coexistence of such a wide range of electronic market makers poses a puzzle.

Even in highly liquid centralized markets, transitory liquidity demands can occupy intermediation capacity for a large fraction of the trading day. I document for the first time the degree to which processing orders can be a long and liquidity-intensive process. Ten percent of intermediated futures contracts require at least five hours and five to six intermediaries to transit from fundamental sellers to fundamental buyers. High trading volume may signal not the presence of liquidity, but its absence as market makers shuffle open contracts amongst themselves on intermediation chains. The existence and prevalence of these chains refute standard models of liquidity provision as a zero or one market maker process. Moreover, roughly 40% of intermediation chains have at least one fast and one slow market maker, suggesting a synergistic relationship among intermediaries of differing speeds.

I formulate a speed hierarchy model to interpret speed's contribution to cross-sectional variation in market maker outcomes as well as patterns of inter-market maker trading. In the model, fast market makers intermediate the least informed orders, thereby sharpening adverse selection faced by slower market makers. As a result, high frequency traders can make markets for more fundamental trading volume at lower spreads than can slower market makers. The residual order flow available to slower market makers is significantly less profitable. The speed hierarchy model thus makes precise an indirect channel by which high frequency traders make slower market makers worse off.

In the model, slower market makers coexist with high frequency traders primarily to absorb their inventory risk and to intermediate highly informed, or "toxic," order flow that HFTs would otherwise consider costly. Empirically, I observe prevalent risk sharing between fast and slow intermediaries: faster market makers are 1.5%-2.4% more likely to close their positions when trading against slower market makers than the reverse. To appreciate the economic magnitude of this volume imbalance, the differential in gold futures is sufficient to close out the sum of maximal HFT positions almost threefold. In line with a risk-sharing interpretation, I find no evidence of HFTs profiting directly at slower market makers' expense.

All else equal, the model also predicts speed heterogeneity increases liquidity provision and facilitates price discovery if retrading activity is not too high. However, endogenous market maker speeds may result in an arms race associated with overinvestment in speed. The discontinuity of profits in absolute speeds—only relative speed or speed rank matters in the hierarchy model—generates leapfrogging incentives and explains the apparent puzzle of high compensation for millisecond improvements in order execution times. Moreover, the arms race can sideline slower, more risk tolerant intermediaries, resulting in reduced aggregate liquidity provision even if total volume remains unchanged.

The paper proceeds as follows. Section 2 discusses related literature. Section 3 introduces the CFTC data and provides summary statistics on market making activity. Section 4 develops the model of speed heterogeneity. Section 5 documents intermediation chains and tests cross-sectional predictions of the model. Section 6 evaluates theoretical effects of speed heterogeneity on liquidity and describes the high frequency arms race. Section 7 offers additional empirical support for underlying assumptions of the speed hierarchy model. Section 8 concludes. In addition to tables and figures, the appendix provides a guide to notation, a worked example of the model, and additional notes on data sources and construction.

## 2 Related Literature

This paper relates primarily to three strands of literature. First, it sheds light on the activity of high frequency traders and rationalizes several existing results on the effect of HFTs on market quality. Second, it joins a handful of other papers in explicitly considering heterogeneity within the market making sector, particularly on the dimension of speed. Third, it extends the inter-dealer risk sharing literature and empirically grounds its natural consequence of intermediation chains.

A small but rapidly growing literature addresses algorithmic and high frequency trading's effects on market quality. Chaboud, Chiquoine, Hjalmarsson and Vega (2011), Hasbrouck and Saar (2010), Hendershott and Riordan (2011) and Hendershott, Jones and Menkveld (2011) are among the first to systematically study high frequency trading. High frequency trading activity is associated with smaller bid-ask spreads, more efficient quotes, and enhanced liquidity and price discovery on the Deutsche Börse (Hendershott and Riordan (2011)), NYSE (Hendershott et al. (2011)) and NASDAQ (Hasbrouck and Saar (2010)), and in foreign exchange markets (Chaboud et al. (2011)). Brogaard, Hendershott and Riordan (2012) complement these studies by finding high frequency traders contribute directly to price discovery as HFTs impound information aggressively and permanently into prices. I extend this literature by relating the source of HFTs' low spreads and high volumes provision to their speed. I also show cross-sectional variation on these measures is high even among fast market makers, despite appreciable correlation in their trading activity (Chaboud et al. (2011)).

Cartea and Penalva (2012) introduce speed heterogeneity in market making by supplementing high frequency traders with slow market makers ("professional traders") in a Grossman and Miller (1988) type model of market making. Cartea and Penalva (2012) show theoretically that HFTs reduce welfare by increasing spreads for liquidity traders and extracting profits from other market makers. I refute both results theoretically and empirically. I find that HFTs offer the lowest spreads, in keeping with the studies of Chaboud et al. (2011) and others, and do not directly make profits from slower intermediaries. However, I offer theoretical support for Cartea and Penalva (2012)'s broader argument that high frequency trading can harm slower liquidity providers, albeit indirectly by sharpening the information asymmetry to which they are exposed.

Biais, Foucault and Moinas (2011) consider speed heterogeneity in the context of equilibrium entry of fundamental traders into high frequency trading. The authors employ a model in which the strengthening of adverse selection on non-HFT traders generates a strategic complementarity in speed. Crowding out of slow traders and overinvestment in speed occurs in one of the five equilibria considered. The speed hierarchy I develop provides a natural means of explaining and testing the differences in counterparties and profits that pervasively generate the adverse selection and arms race results in Biais, Foucault and Moinas (2011). When speed is endogenous, I reach a similar conclusion that large traders are more likely to invest in speed in an arms race, all else being equal, but also that ordering of speed by size can be overturned if heterogeneity in costs of speed is high enough.

Locke and Sarajoti (2001) do not explicitly address speed heterogeneity, but they do undertake a similar approach to this paper in testing a reduced-form model of a skill-based hierarchy of traders in commodity futures markets. I provide an alternative interpretation of "skill"—speed—and the associated theoretical underpinnings. Where the empirical implications overlap, I find direct transactions between fast and slow market makers are not profitable for the intermediaries higher in the hierarchy, by contrast with Locke and Sarajoti (2001)'s positive profitability of skilled traders at the expense of unskilled ones. Nevertheless, Locke and Sarajoti (2001)'s findings suggest a speed-based intermediation hierarchy may be pervasive across market structures, because their study addresses a period during which traditional dealers, rather than HFTs, are the central market makers.

In the model, speed heterogeneity with retrading among market makers generates excess volume and long intermediation chains. Retrading, or reintermediation of new volume, enables fast market makers to share risk with slower market makers, who in turn are compensated by the fast intermediary for taking on inventory and informational risk. Inventory risk sharing among market makers in the academic literature dates to Ho and Stoll (1983)'s model of inter-dealer trading. Several subsequent works confirm a high degree of inter-dealer, or "hot potato," trading and risk sharing (e.g., Lyons (1997), Vogler (1997), Hansch, Naik and Viswanathan (1998), and Reiss and Werner (1998)). In evaluating retrading in commodities futures markets, I build on the methodology of Hansch, Naik and Viswanathan (1998), who find traders with large inventories tend to close positions against traders with smaller or offsetting inventories. However, heterogeneous risk tolerance can generate permanent differences in position sizes and contract flows toward high inventory intermediaries. To account for variable risk tolerance and assess risk sharing only along the speed dimension, I test whether fast intermediaries are more likely to close positions when trading against slower intermediaries than the reverse, regardless of current holdings.

# 3 Data Description

## 3.1 Data on Account-Level Trading Activity

I use audit-trail data provided by the Commodity Futures Trading Commission (CFTC) for three Chicago Mercantile Exchange (CME) futures contracts during the week of December 12-16, 2011.[4] The CME-CFTC data set captures all trades made in these contracts. Observations are at the transaction level with millisecond timestamps. Where trades are made in the same millisecond, I use the CME Globex matching engine's unique order numbers to reconstruct precedence. Each record provides an extensive list of properties of the generating order, trade, trade venue, and trading parties. Anonymized customer identifiers, including a flag for the initiator ("aggressor") of a trade, enable comprehensive tracking of trading activity by account. Trades can also be grouped by an order identifier to accommodate partial fills of limit orders and assess additional characteristics of account holders. Table 1 provides summary information on the contracts selected and their relation to other available futures contracts on these commodities. Appendix D provides additional information on data sources and cleaning procedures.

[4] Products trade Sunday to Friday, 5:00 p.m. - 4:15 p.m. (CT). By convention, trading days are defined as starting at 5:00 p.m. on the previous day. The Monday, December 12, trading day runs from 5:00 p.m. on Sunday, December 11, through 4:15 p.m. on Monday, December 12.

Most futures trading in each individual metal on the CME COMEX and NYMEX exchanges is concentrated in a single contract. The CME contracts are the dominant futures contracts in gold and silver, but not in copper. Available copper contract data offer an incomplete picture of that market because close substitutes are readily available on the London Metal Exchange (66.6% of futures volume) and the Shanghai Futures Exchange (18.9% of futures volume). The data set is also incomplete to the extent that traders participate in the spot and ETF markets. The sample thus excludes liquidity transfers and hedging across trading venues, but these omissions likely have little impact on the cross-sectional relations and intermediation chains analyzed here.

I choose gold, silver, and copper futures for several reasons. As the three most active metals contracts, they have high volume and trade on similar underlying commodities. Their associated options market activity is nevertheless relatively light, unlike the equity index, foreign exchange, interest rates, and energy products. As a result, futures data alone capture a greater share of trading in metals than they do in many other liquid products. Metals also are subject to comparable exchange rules and settlement procedures, which facilitates augmenting a short time series by comparing results in the cross-section.

Table 2 provides basic summaries of the cross-section of activity in these markets. Gold is far more active than silver and copper both in volume and in the number of accounts trading regularly. Average trade and order sizes are similar across markets, being roughly 1 and 2, respectively. Orders shredded to secure better terms of trade dominate exchange activity.

## 3.2 Classifying Market Participants

I now turn to the issue of assessing intermediary participation in these commodity markets. CME-CFTC data extracts scramble customer account identifiers to preserve trader anonymity. Limited identifying information is provided in addition to the customer account, making classification into types such as intermediaries and fundamental traders a non-trivial task. Nonetheless, distinguishing between typical liquidity demanders (fundamental traders) and suppliers (market makers) is important for understanding the roles of different actors in the intermediation process.

I classify accounts into four behavioral categories: fundamental traders, market makers, small traders, and opportunistic traders. Fundamental traders take large directional positions relative to their intraday volume. I adopt a threshold for the ratio of end-of-day position to volume, above which the trader is classified as a fundamental buyer or seller. I select a 20% cutoff because a small breakpoint is present at that level in the ratio's empirical distribution. This definition makes the natural assumption that fundamental sellers primarily sell contracts and fundamental buyers primarily purchase them.[5]

Second, market makers actively manage their inventory risk. I define market makers as those accounts whose cumulative intraday positions cross zero at least twice. Zero crossings simultaneously capture the strong mean-reversion in positions and the modest directional bets characteristic of market making. I further segment the market maker category by sorting on the number of times positions cross zero and constructing quantiles to roughly capture the corresponding empirical distribution. This dimension of variation corresponds roughly with speed. Fast traders intuitively are those best equipped to minimize their inventory risk by capturing advantageous opposing orders. I label accounts in the top quantile of zero crossings as high frequency traders (HFTs). In addition to representing the right tail of speed by the zero crossing metric, this group appears qualitatively different from other market maker categories on the basis of volume, spreads, and other speed measures analyzed throughout this work.[6]

Third, small traders have low volumes. This category captures retail trader participation and one-off position adjustments. The role of small traders in liquidity provision or liquidity taking is unclear, so they are generally set aside for empirical work. Finally, opportunistic traders are the residual active accounts that do not fit into other categories. I summarize the classification methodology in Table 3a.

Classifications are recomputed daily by commodity and are independent of activity in other markets. In this sense, I classify accounts by observed trading activity rather than by CME identifier. An account that acts like an HFT in gold on December 13 may be classified as a opportunistic trader in silver on December 13 or in gold on December 14. Where accounts satisfy the requirements for multiple classifications, I use the higher classification in Table 3a. For example, a trader with low volume in a given direction is labeled small rather than fundamental because the small quantity of her trading is plausibly the more salient feature for understanding how her orders are intermediated.

Table 3b reports the robustness of classifications across days in the sample. The first result that emerges is that classifications are not highly stable. Market makers tend to remain market makers but transition with non-trivial probability into other categories. The

[5] Kirilenko et al. (2011) also use this criterion (with a 15% threshold) for classification in E-Mini S&P 500 futures. Results are similar or slightly stronger using their ad hoc classification methodology.

[6] Fast traders that take on non-market making roles may rightly be called HFTs for purposes other than assessing their role as intermediaries. I focus primarily on intermediation activity rather than on the activity of high frequency traders as a group. Fast traders that are not primarily market makers are classified instead as fundamental or opportunistic traders, depending on the nature of their activity.

breakdown by market making category is more revealing, however. Market makers that cross zero frequently tend to continue doing so, whereas those who cross only occasionally comprise the bulk of the exits from market making activity. This persistence is unsurprising given the high fixed costs associated with high frequency trading. Traders that take large directional positions often continue to take large positions but also transition into the small trading category or inactivity. This finding accords with the intuition that large directional position changes are often followed by fine-tuning in subsequent days once the bulk of the intended position has been accumulated. Small traders tend to remain small or exit the market. These traders are likely either retail investors or larger participants that enter as fundamental traders and later make infrequent position adjustments. Overall, the patterns of transition support the initial classifications.

## 3.3 Summary Statistics by Participant Type

Table 4 summarizes the cross-section of aggregate intermediation activity across commodity futures contracts along three dimensions. Panel A presents the number of participants in each market by type. The averages understate the number of high frequency participants because several fast market makers oscillate around the zero crossing cutoff despite the category's relative stability. Panel B reveals a handful of HFTs have become dominant in terms of market share. HFT volume alone is comparable to or larger than fundamental trader volume and an order of magnitude larger than small trader volume. Volume drops off sharply in zero crossings, but total market maker volume is nonetheless considerably larger than fundamental trading volume. On a per-trader basis, high frequency accounts are far more active than any other type of market participant. Traditional market makers also trade more than the typical fundamental trader in the sample. Panel C shows fundamental traders are net providers of liquidity in the sense of primarily executing trades using resting, or passive, limit orders. HFTs, by contrast, appear to be liquidity consumers by this measure. This finding challenges the notion of liquidity provision as leaving resting orders in the limit order book to facilitate other parties' aggressive trading.

Table 5 presents averages for three new dimensions of market participant characteristics. Panel A describes average maximal net positions within each type. Although potentially hedged elsewhere, HFT positions in the observed commodities are large relative to those accumulated by other non-fundamental traders. These positions typically are drawn down to nearly zero at market close, but appreciable inventory risk may nonetheless be taken in the interim. Other market makers are roughly equal on this dimension.

Panels B and C report two measures of speed. Panel B introduces a new speed measure.

I take the 10th percentile of durations between orders that switch the direction (parity) of the previous order. I invert this duration value, adding a second to ensure boundedness, to obtain a switching frequency. The switch frequency measure proxies for a lower bound on the time required to generate intentional trading decisions. Frequent, rapid oscillation between buy and sell order executions suggests the use of high frequency technology. High frequency traders distinguish themselves from slower intermediaries by switching more frequently than other market maker categories. The marked drop at the 80th percentile of zero crossings, especially in copper, may indicate qualitative speed or technological differences among types of market makers. For this reason, I perform separate regressions for both the fastest 20% of market makers as well as for all market makers to discern whether speed plays a measurably different role in fast market making strategies.

Panel C reveals the distribution of zero crossings has a long right tail. Building and decumulating positions more than a hundred times a day is a uniquely modern phenomenon among market makers and further supports the interpretation of the top quantile as HFTs. Because all intermediary types typically cross the zero position threshold several times each day and end each trading day with near-zero inventories, I lose little in assuming all intermediaries enter the sample with zero positions, provided large accumulations do not occur prior to the sample period. Fundamental traders that typically build directional positions mechanically only rarely cross zero. Panel D confirms higher numbers of zero crossings accompany the higher switching frequencies of fast traders. I use Spearman rank correlations to assess their relation because the mapping between speed measures is not linear.

Figure 1 presents empirical distributions of log switching frequency and log zero crossing speed measures. Histogram counts are averaged across dates in the sample. The bottom histogram is truncated on the right to preserve market maker anonymity in accordance with CFTC guidelines. The rightmost bar aggregates the long tail, which would otherwise extend past the 4.5 tick mark suggested by Table 5. Both figures reveal high speeds are concentrated among relatively few market makers, and these intermediaries coexist with a large number of very slow market makers. Developing aggregate and cross-sectional implications of this speed heterogeneity is the primary contribution of the model I present in the following section.

# 4 A Model of Speed Heterogeneity in Intermediation

In this section, I introduce a model relating intermediation chains and the cross-section of intermediary activity to speed heterogeneity in the market making sector. Several of this work's empirical results can be interpreted in isolation. However, the speed hierarchy developed here provides a candidate mechanism for explaining observed empirical results, offering additional theoretical predictions that cannot be tested using the data available, and generating further implications for risk sharing activity and order flow predictability.

## 4.1 The Economy

### 4.1.1 Traders and Market Makers

The economy consists of a positive mass continuum □ of sellers and N market makers (intermediaries). Sellers trade because of liquidity and informational motives. The liquidity motive is the source of surplus in the economy, whereas information is zero-sum between buyers and sellers taking opposing sides of a transaction. Sellers compensate market makers for facilitating trade by paying them a fee, or spread. I do not model direct seller-buyer transactions or the eventual transfer of contracts to buyers.

Throughout, I often refer to sellers as fundamental traders to distinguish them from market makers that sell to other market makers. Sellers in the model correspond to the fundamental trader category of Section 3.2. By construction, these market participants trade dollar volumes exceeding $1 million during the trading day, and hence primarily represent institutional rather than retail investors. In keeping with this interpretation, I assume sellers are risk neutral. By contrast, market makers are risk averse with idiosyncratic risk aversion parameter $\kappa_j$ 2 $(0, \kappa)$ and mean-variance utility over terminal wealth. This assumption captures the inventory risk management motive of asset market intermediaries.[7]

Assets are non-fungible contracts in elastic supply. For tractability, I impose a standard assumption that fundamental traders can sell either zero or one contract. All sellers possess information about their specific contracts. To avoid modeling asset price dynamics, information takes the simple, zero-sum form of generating positive value $x_i$ for sellers and $x_i$ for buyers, where $x_i$ denotes the informedness of seller i. In this sense, informed contracts are said to be “toxic” to the purchasing market makers. Buying contracts from informed traders generates losses to market makers, but these losses can be passed on to other market makers by retrading the asset before the information is impounded into the price. Contracts are non-fungible only to ensure that the information from one seller has no bearing on the actions of other traders or on market makers’ willingness to participate in other transactions.

Fundamental trader informedness $x_i$ is drawn i.i.d. from some known distribution with (improper) density $f(x)$ defined over an interval $[A, B]$ □ $R_+$. Because traders exist on a continuum, the empirical density of traders’ x’s converges to $f(x)$. I assume all market

[7] Bounding $\kappa_j$ by $\kappa < 1$ imposes little restriction on κs but is a necessary technical condition in Theorem 1.

makers can perfectly forecast the informedness of their potential counterparties.[8] I offer empirical evidence supporting this assumption when discussing order flow predictability in Section 7.2. Known $x_i$ enables the market maker to condition on $x_i$ when setting the spread $c(x_i)$.

Contracts have a common fundamental value component v that is shared by all agents in the economy. v is stochastic with mean $v = 0$ and standard deviation $\sigma = 1$.[9] Sellers also have a common liquidity demand $l > 0$. I shut down variation across liquidity demands to isolate the treatment of sellers of differing informedness, because variation across informedness is essential to understanding the source of differences in spreads and volumes as a function of market maker speed. Fundamental traders' ex post values of selling a contract are the sum of the fundamental value v, the common liquidity demand l, and the information profits $x_i$, minus the market making spread $c(x_i)$

$$v_i = v + l + x_i \qquad c(x_i)$$

Because sellers are risk neutral, this expression reduces to an ex ante value of trading of

$$E\,[v_i] = v + l + x_i \qquad c(x_i) = l + x_i \qquad c(x_i)$$

I normalize the value of not trading to $v_i = 0$ because neither the liquidity nor the informational motives are satisfied and no payment is made to the market maker.

In addition to taking on variably toxic order flows, market makers are subject to inventory risk associated with unpredictable variation in prices. Inventory risk is quadratic in position size because innovations in fundamental values v are perfectly correlated across contracts. Convexity of inventory risk prevents individual intermediaries from dominating market making activity. Consequently, variation in market maker risk aversion offers another margin on which intermediates can earn profits. Market makers with high speeds profit directly from spreads charged to fundamental traders and, as I describe in the next section, market makers with

[8] I do not anticipate imperfect forecasting of trader types to alter qualitatively any of the results presented here, provided the forecasting ability does not vary across market makers. Adding noise to inference will amplify adverse selection faced by market makers, because traders with higher valuations than the posted prices will not transact. I conjecture, but do not prove, that adding noise to forecasting is to a first approximation equivalent to shifting the informedness distribution to the right: low informedness sellers will no longer transact, and the transactions that do occur are more harmful to market makers.

[9] This distributional assumption is without loss of generality. v 6= 0 increases the price by exactly v and therefore can be netted out in sellers' trading decisions. Non-unit standard deviations are captured by redefining $\kappa_0$ as $\kappa\sigma$. The model can be extended by imposing asset price dynamics, but they are inessential here.

high risk capacities collect risk sharing fees from faster intermediaries. Figure 2 illustrates both strategies and the flow of contracts and payments in the economy.

Market makers take the opposite side of informed trades to net $c(x_i)$ $x_i$ for each contract intermediated, excluding incremental inventory risk costs. In determining the spread $c(x_i)$, marker makers bargain over information rents but can extract the entirety of liquidity demands l. For simplicity, I assume all market makers have fixed bargaining power $\gamma$ 2 (0, 1) over information rents $x_i$, so that market makers cannot recoup the entire informational loss associated with trade. $l + \gamma x_i$ is then the maximum incentive-compatible spread market makers can charge to sellers. I assume further that fundamental sellers are non-strategic and therefore accept any spread $c(x_i)$ □ $l + \gamma x_i$. Alternatively, one could view $\gamma < 1$ as the outcome of sellers' strategic decisions to have orders executed by the earliest available market makers. The (1 $\gamma$) $x_i$ wedge between the spread captured by intermediaries and the cost to facilitating informed trading gives that highly informed order flow can be too toxic for any market maker to handle profitably despite the cross-subsidy generated by liquidity demands.

### 4.1.2 Timing

Market makers are sequential monopolists in that the jth market maker can only intermediate the residual order flow of the preceding 1,...,j 1 market makers. Speed is the position of a market maker in the intermediation ordering. Fast market makers are first in the ordering and consequently have an earlier opportunity to intermediate order flow. For now I assume market makers know their speeds and the positions are exogenous. I relax the latter assumption in Section 6.2. The intermediation hierarchy described by the ordering acts like a set of filters for order flow. Order flow encounters each filter in series. The most attractive orders are intermediated, leaving only the less desirable potential trades to filter through to the next market maker. Hence each market maker worsens the order flow available to slower market makers. I summarize these relationships graphically in Figure 3.

For exposition, let the set of sellers be countable with identifiers 1, 2,... for each trader i. Traders arrive sequentially in the market and decide whether to trade at posted prices. For trade to occur, the market maker's spread cannot exceed the liquidity demand l plus the implicitly negotiated share of information rents $\gamma x_i$. I abstract from optimization over the initiator of limit orders in this setting and assume that non-intermediary orders are aggressive. Timing proceeds as follows.

The first market maker places passive limit orders at a spread c(x) for all x in her optimal intermediation interval î$q_1$ ,$q_1$ó □ [A, B], where $q_1$ and $q_1$ denote the lower and upper bounds on informedness levels intermediated, respectively.[10] The first market maker forecasts the

[10] I show that the optimal intermediated set is indeed an interval in Lemma 1. I slightly abuse

imminent order of trader 1 and cancels all unfavorable limit orders. Trader 1 then arrives at the trading venue with informedness $x_1$. In equilibrium, the market maker will withdraw liquidity provision if $x_1 > q_1$ and leave a sole passive order of $1+\gamma x_1$, the maximum extractable quantity, otherwise. If an incentive-compatible resting order remains, trader 1 executes against it. No optimization over timing occurs in this model: the trader acts as a price taker and finds only the first intermediary as an available counterparty. In this sense, speed provides monopoly power to the market maker.

If the first market maker chooses not to intermediate, the second market maker places passive limit orders as a function of the filtered order flow. As before, the fundamental trader will sell at price $1+\gamma x_1$ or be excluded by the second market maker if $x_1 /2$ îq $_2$,$q_2$ó. This process iterates until either the trade is intermediated or the slowest intermediary has declined to trade, whereby the first seller leaves the market without trading. The order book then replenishes for the second trader and the process repeats until the set of traders has been exhausted. The economy ends once all fundamental traders have arrived to market and all retrading activity has ceased. At this time, v is revealed and information is impounded into prices at the expense of market makers still holding contracts.

I analyze two closely related models in this paper: a baseline in which market makers act only as passive liquidity suppliers and an extension in which market makers trade again with slower intermediaries to share inventory risk. This risk sharing generates the long intermediation chains described in Section 5.1. In the latter model, the jth intermediary can share inventory risk by retrading with slower market makers. If the jth intermediary chooses to retrade to offload contracts to the j + 1st intermediary, the j + 1st intermediary faces an identical decision vis-`a-vis the j + 2nd intermediary. The process continues until risk sharing through retrading is no longer desired or the set of market makers is exhausted. I denote the volume reintermediated from market maker j by market maker j + 1 as $Q_{j+1}$.

To keep the reintermediation decision tractable, I assume the jth intermediary can only retrade with the j + 1st intermediary.[11] As a conservative assumption, I also endow the slower intermediary with all bargaining power in inter-market maker transactions. Under this assumption, the particular contracts that are retraded among the $Q_{j+1}$ are immaterial because faster market makers pay a premium of $x_i$ to slower ones as a component of the marginal cost of holding a contract. Giving the slower intermediary complete bargaining power places a lower bound on reintermediation activity; incomplete bargaining power over reintermediation rents would create an incentive for faster traders to accumulate and retrade excess volumes. All predictions of the model should remain qualitatively similar in either

notation in that seller i's contract is intermediated rather than seller i's informedness $x_i$.

[11] This assumption holds if faster market makers trade aggressively against slower ones, and the j + 1st intermediary places a limit order before the jth intermediary decides to retrade and other market makers arrive.

case.

## 4.2 Equilibrium and Optimal Intermediation Activity

Before characterizing equilibrium, I show the market maker's optimization problem can be simplified considerably because the optimal set of fundamental orders for each market maker to intermediate is a single interval.

Lemma 1 (Interval Intermediation). If $f(x)$ is defined over a set $[A, B] \square R_+$ with $f(x)$ 2 (0, 1) 8x 2 $[A, B]$, then the optimal intermediation policy consists of choosing an interval to fully intermediate, $hq_j, q_j i$, up to measure zero deviations.

Proof. I provide intuition that is formalized in the Mathematical Appendix. Assume there is a positive quantity of intermediation activity for which the intermediated set $D_j$ is not an interval. Taking this quantity as fixed, left-shifting intermediated contracts to close a gap between subintervals is feasible because support on the interval is convex. Inventory risk costs are identical because the quantity intermediated remains fixed, but revenues are higher because less informed contracts have a smaller wedge between the spread collected and the informational cost to market makers. Hence the resulting set dominates $D_j$ and a single interval must be optimal.

Lemma 1 holds because market makers are unable to capture with spreads the full cost of holding informed contracts. Profits are decreasing in informedness as the wedge between the payment extracted from fundamental traders and the contract's toxicity to intermediaries increases. Market makers will therefore leave no gaps in the optimal intermediated set by taking on contracts more toxic than is necessary, holding total volume fixed. With Lemma 1 in hand, I can now define equilibrium in the speed hierarchy model.

Definition 1 (Equilibrium). A speed hierarchy equilibrium with reintermediation is a set of spreads $1c_j(x_i)l_{i2I}$, intermediation intervals $nhq_j, q_j io^N_{j=1}$, and reintermediation quantities $1Q_j l^N_{j=1}$ such that

1. Sellers (myopically) maximize utility: for all i 2 □,

$$i\ 2\ hq_j, q_j i,\ c_j(x_i) \square 1+\gamma x_i \text{ and } j=\min 1j_0 : c_{j_0}(x_i) \square 1+\gamma x_i l$$

2. Market makers maximize profits: for all market makers j = 1,...,N,

nq
j ,qj,Qj, 1cj (x)lo 2 arg max πj = / qj cj (x) f (x) dx κj 2 Ñ/ qj f (x) dx + Qjé2
qj ,qj ,Qj ,{cj (x)} qj qj
| {z } | {z }
spreads inventory costs

/ qj xf (x) dx +κj1 / Qj ↓↓F Äq j1ä F ↓q j1 + Qj1 q dq
qj 0
| {z } | {z }
info costs reintermediation fees

3. Intermediated sets are feasible: for all market makers j = 1,...,N,

hq
j ,qji [A, B] < [ nhq j0 ,qj0 ioj1 j0=1

F Äqjä F ↓q j Qj+1

Equilibrium in the speed hierarchy with reintermediation is intuitive. Because sellers are passive up to bargaining, their utility maximization entails only accepting the first incentive-compatible spread and rejecting spreads that are too large (condition 1). Condition 2 is standard profit maximization by market makers over their control variables of intermediation intervals hq j ,qji, reintermediation quantities Qj, and spreads 1cj (x)l. Condition 3 requires the optimal intermediation sets to be feasible, namely, that intermediation intervals are contained in the support filtered by faster market makers and have sufficient mass for the subsequent market maker to reintermediate his desired quantity Qj+1. Market clearing is an amalgam of all three conditions: condition 1 describes the set from which liquidity is demanded; condition 2 describes the set to which liquidity is supplied, where spreads are set to ensure demand does not exceed supply for any i; and condition 3 ensures this set is only supplied to at most once so that liquidity supply does not exceed demand for any i. Equilibrium without reintermediation is identical with the exception of fixing Qj = 0 for all j = 1,...,N.

Theorem 1 characterizes optimal intermediation activity in the speed hierarchy equilibrium under general conditions on informedness of fundamental traders and the risk capacities of intermediaries. It also describes maximal intermediable sets and settings in which new intermediation activity requires infinitely many, finitely many, or zero intermediaries to achieve the maximal feasible level of fundamental trading volume. All implications from the model follow as corollaries to the theorem.

Theorem 1 (Main Theorem). If f (x) is defined over a set [A, B] R+ with convex support,

f (x) < 1 and continuously differentiable 8x 2 [A, B], then c(x) = 1 + γx and

1. If A < 1 1 B, all market makers will intermediate positive quantities but intermedi-

ation activity will not cover the interval. Optimal policies follow q← j = q←j1 8j 1 (where q←0 □ A) and q←j implicitly defined by (1 γ) q←j = lκj ↓F Äq←jä F ↓q← j □ + Q←j□. With reintermediation, the quantity of reintermediated contracts is Q←j = ↓F Äq← j1ä F ↓q← j1□ + Q← j1□ □j □j1 ↓F Äq←jä F ↓q← j □ + Q←j□. Otherwise, Q←j = 0.

2. If A<B< 1 1 , all trades will be intermediated for N large enough and, for some n<N, all agents j n will not intermediate new fundamental volume. Intermediaries follow the optimal policies prescribed above until no longer feasible, in which case intermediaries take the entire available [Aj,Bj] interval. Q←j is positive for all intermediaries j > 1 so reintermediation activity continues indefinitely.

3. If A 1 1 , no intermediation will occur.

Proof. I give a formal proof in the Mathematical Appendix. I provide a heuristic proof here. The Lagrangian associated with the optimization problem of the jth intermediary in the model without reintermediation is

L ↓q j ,qj□ = / qj qj (1 + (γ 1) x) f (x) dx | {z } spreads net of info loss κj 2 Ñ/ qj qj f (x) dxé2 | {z } inventory risk cost

+λ1,1 ↓q j Aj□ + λ1,2 ↓Bj q j□

+λ2,1 ↓qj q j□ + λ2,2 ÄBj qjä

where [Aj,Bj] is the subinterval of [A, B] available to the jth market maker. Absent constraints, the optimization problem is straightforward. Taking first order conditions with respect to the informedness bounds obtains

∂L/∂q j = n↓1 (1 γ) q j□ κj ↓F Äqjä F ↓q j□□o f ↓q j□ = 0

∂L/∂qj = nÄ1 (1 γ) qjä κj ↓F Äqjä F ↓q j□□o f Äqjä = 0 (1)

If both conditions were to hold, the intermediation interval would be of measure zero and profits would be zero. In general, positive profits are possible so one of the boundary conditions will be binding. Much of the proof is devoted to handling these boundary conditions rigorously. The first order condition for q j will be violated in general because taking less informed contracts is always preferred. The least informed counterparty intermediated has

informedness of Aj = q j1 . Where feasible, the first order condition for qj holds with equality. Note qj is bounded above by 1 1 ; otherwise, the first derivative is negative and a smaller qj would be optimal. When the unconstrained optimum qj is not available, the intermediary intermediates as much order flow as is available, that is, the entire interval $[A_j,B_j]$.

With reintermediation, additional first order and boundary conditions apply. Risk sharing with positive reintermediated volume $Q_j > 0$ is optimal if risk aversion-weighted quantities differ between adjacent market makers. $Q_j$ is set to equate these quantities to share risk optimally

$$Q_j = \downarrow F\ \ddot{A}q_{j1\ddot{a}} \qquad F \downarrow q_{j1\square} + Q_{j1\square} \qquad \frac{\kappa_j}{\kappa_{j1}} \downarrow F\ \ddot{A}q_{j\ddot{a}} \qquad F \downarrow q_{j\square} + Q_{j\square} \tag{2}$$

Positive risk sharing is always optimal in the speed hierarchy. Slower intermediaries desire the less informed flow intermediated by faster ones, and faster intermediaries want slower ones to absorb their inventory risk. The two features of heterogenous access to order flow and subsequent risk sharing drive the intermediation and reintermediation decisions and, by extension, the results of the model.

Theorem 1's optimal intermediation strategies are straightforward. Intermediaries take the least toxic contracts available and continue absorbing order flow until the marginal benefit of intermediating more toxic contracts declines to meet the rising marginal cost of contributing to additional inventory risk. Reintermediating the previous market maker's trades becomes preferable when previously intermediated contracts are sufficiently attractive relative to new order flow. Optimal risk sharing pushes positions closer to equality than they would be without reintermediation and continues indefinitely, regardless of whether new order flow is available for intermediation.

Theorem 1 provides a complete description of intermediation activity in the class of economies described by the model. Equations (1) and (2) in particular are useful for establishing additional propositions governing the properties of the model. On a micro level, the theorem describes which market makers trade with whom and at what prices, as well as which trades go unintermediated. On a more macro level the theorem speaks to concepts such as aggregate liquidity provision.

# 5 Empirical Facts of High Frequency Intermediation

## 5.1 Intermediation Chains

Exchanges exist in part to allocate assets from sellers to buyers at prevailing prices. In this section, I shed light on the allocative process and find seller to buyer transfers frequently involve several intermediaries of differing speeds. The prevalence of intermediation "chains," or contract flows among market makers, suggests liquidity provision is more nuanced than binary decisions to intermediate. Additionally, the large excess volume generated by retrading among intermediaries delinks liquidity for fundamental traders from observed volume. These findings extend work from the inter-dealer trading literature by documenting the (varying) extent to which contracts can remain in intermediation for prolonged periods.

I track contract flows to construct chains of intermediaries separating fundamental sellers and fundamental buyers. For each commodity and trading date, I initialize all market participants' inventories at zero. I then iterate sequentially through trades. I break trades of size n into n trades of size 1. For each contract bought or sold in the direction of the existing position, I augment the trader's position count by 1. For contracts bought or sold against the direction of the existing position, the decumulation methodology can be important. I repeat calculations for three decumulation methods:

1. First in, first out (FIFO) - remove the oldest contract held by the intermediary
2. Last in, first out (LIFO) - remove the newest contract held by the intermediary
3. Uniform - remove a contract at random from those held by the intermediary, where all contracts have equal probability of removal

Each contract tracks its trading history by holder classification. At the conclusion of the trading day, I restrict consideration to the subset of contracts sold by fundamental sellers and later purchased by fundamental buyers. I then compute the time in seconds, trades, and intermediaries for each contract to migrate to a fundamental buyer. The trade time is computed as the number of trades required for a contract to transit from a seller to buyer conditional on being in this set. Clock time is the associated time in seconds. Intermediary transit times are the number of market makers that hold the contract during the complete seller-buyer transaction. Appendix D provides additional details for intermediation chain construction. In this analysis, I examine only the cross-section of trades within each trading date because the five day time series is short and relatively uneventful.

Figure 4 presents histograms describing seller-buyer chains for intermediated trades. It omits direct seller-buyer trades—approximately 34%, 26%, and 28% for gold, silver, and

copper, respectively—to provide resolution on the intermediation process. Panel A plots the distribution of seller-buyer clock times in each market. The modal chain duration is between 20 and 50 minutes, although some chains take as long as five hours to effect the transfer of contracts. As a result, short-lived liquidity demands can reduce intermediation capacity for a substantial fraction of the trading day; if intermediaries hold open positions for several hours to accommodate a seller-buyer chain, they may be less able to handle contract inflows in the same direction. This result is robust to decumulation methodology. Dispersion among decumulation methods is high only in the tails of the clock time distributions.

Panels B and C present frequency distributions for the number of high frequency traders (Panel B) and of all market makers (Panel C) involved in intermediation chains. Intermediated contracts typically feature at least one high frequency trader linking sellers to buyers, indicating their central role in liquidity provision. Moreover, some intermediation chains involve as many as five or six high frequency traders trading amongst themselves—even among fast market makers, retrading potentially results in large, and non-constant, intermediation multipliers of fundamental trading volume. Panel C reports chain lengths for all market makers. Fewer than five percent of intermediated transactions do not involve a market participant of the "market maker" category, suggesting intermediaries are classified correctly. Seller-buyer chains include as many as nine intermediaries. Extended chains imply some fundamental trader transactions get "stuck" in intermediation. Volume and liquidity are likely negatively related in these instances as contracts tie up intermediation capacity and are slow to migrate to fundamental buyers. Both panels strongly reject the canonical model of a single market maker. Although zero or one market maker are common, the slight majority of intermediation activity involves at least two market makers, and cross-sectional dispersion among chain lengths is high.

Table 6 presents these results in numeric form. I take averages for each value across dates, so some values in Panel B are not integers. Panel A provides quantiles for seller-buyer times in trades and seconds and breaks down the results by market and decumulation approach. Although 10% of trades exit intermediation in under two minutes, the median intermediation duration is roughly an hour and the longest 10% of intermediation durations is about five hours. The impact of liquidity demands thus decays relatively slowly, even in these liquid contracts under normal market conditions. Panel B replicates the plots of Figure 4. The distribution of intermediation chains in gold has slightly shorter tails and more direct seller-buyer transactions, perhaps signifying more readily available fundamental counterparties than in silver and copper. This availability may be the salient definition of liquidity for market makers, who seek to close positions quickly rather than bear inventory risk or pay premia to other market makers for bearing that risk.

Figure 5 decomposes intermediation chains into high frequency traders and other market

makers. To compare with the standard view of market making, I include non-intermediated contracts and take logs to illustrate the relative counts for the full range of observed seller-buyer contract flows. The bars at coordinates (0, 0), (0, 1), and (1, 0) represent canonical market making activity, in which zero or one intermediary facilitates transactions. 32.7%, 42.6%, and 48.7% of seller-buyer flows involve multi-party intermediation chains in gold, silver, and copper futures markets, respectively, demonstrating that the cross-section of intermediation activity is in fact quite nuanced. Long chains appear to be driven roughly equally by retrading among high frequency traders and among slower market makers.[12]

The empirical cross-section of intermediation chains contrasts with a view of anonymous markets as a setting for random, two-sided seller-buyer matching or of intermediation as a one market maker affair. Instead, there appears to be a potentially dense network of intermediating market makers. Panels B and C of Figure 4 are visually consistent with undirected trading among market makers with a fixed absorption probability by fundamental traders, which could suggest the network topology plays little role in trading activity. I show in Section 7.1 that directed risk sharing among market makers belies this result. In sum, the intermediation process in asset markets takes far longer in time and trading parties than existing models may predict for seemingly liquid centralized markets.

Three caveats bear mention here. First, tracing individual, fungible contracts requires strong assumptions on how accounts decumulate contracts. I apply three decumulation methods to show results are robust: first in, first out (FIFO); last in, first out (LIFO); and uniformly drawn from contract holdings. Trade times in trades and seconds are slightly sensitive to the decumulation methodology, but other results are unaffected. Endogenous variation in decumulation methodologies among traders and across time may affect these results in a highly nonlinear fashion, however. Second, transaction histories are reset at the beginning of each trading day. I may underestimate intermediation chain lengths, as 36% (gold) to 46% (copper) of fundamental seller contracts are not received by fundamental buyers by the end of the trading day and are not recorded. The associated chains may be truncated by the end of the trading day, in which case selection effects are minimal, or some may find medium-term holders in non-fundamental buyers. Medium-term market makers would complicate the intermediation picture, but in general, their presence would contribute to an underestimation of the time and number of intermediaries required to process fundamental liquidity demands. Finally, positions accumulated prior to each trading day could alter results

[12] In unreported results, I find no robust relation between chain composition and volume, volatility, or fundamental trader returns. However, this sample period is quiescent, and the distribution of chain lengths may covary with these characteristics in longer or more active time series. Understanding the interplay between chain length and liquidity would be a valuable extension to this work.

in either direction if trading closes previously accumulated positions. Because intermediary inventories cross zero often and fundamental traders frequently transition into the small trader or inactive categories, I do not anticipate initial positions to be large and opposing observed accumulation for the primary liquidity suppliers and demanders in these commodities.

Long reintermediation chains imply trade typically generates significantly higher volume than fundamental trading motives alone would suggest. As a general point, volume and liquidity—defined here as the ability for fundamental traders to have their orders intermediated—are delinked when intermediation chains are prevalent. Not accounting for the excess volume generated by intermediation chains and the duration of inventory handling can have catastrophic effects. Taking the "Flash Crash" in U.S. stock and futures markets as an example, the initiator of the large sell orders on May 6, 2010, employed a sell program targeting the share of trading volume rather than price impact or other liquidity measures.[13] In light of the preceding intermediation chain results, high volume among market makers likely signaled the absence of liquidity rather than its availability as futures prices plunged and order book thickness dissolved. Not differentiating reintermediation activity from liquidity caused the sell program to continue issuing large liquidity demands, which in turn may have triggered the biggest one-day point decline in the history of the Dow Jones Industrial Average.

## 5.2 Cross-Sectional Implications of Market Maker Speed Heterogeneity

The previous section paints a new picture of the role of market makers in transferring contracts from sellers to buyers. Speed differences among market makers determine how they contribute to the process of liquidity supply. This section relates observed cross-sectional outcomes to variation in market maker speeds. I find faster market makers charge lower spreads, intermediate higher volumes, and earn larger profits. These results are robust to a variety of functional form specifications for the relationship with speed. The model suggests mechanisms that generate these results and provide structure for empirical tests. Accordingly, Proposition 1 summarizes the primarily testable implications of the model.

Proposition 1 (Spreads, Volume, Informedness, and Profits). The following relations hold among intermediation tiers

1. Average spreads for newly intermediated trades are decreasing in speed (increasing in j).

2. Holding risk capacity constant, passive volumes are increasing in speed. Total risk aversion-weighted passive volumes are always increasing in speed.

[13] Preliminary Findings Regarding the Market Events of May 6, 2010.

3. Profits are increasing in speed if risk capacity is non-increasing in speed. If risk capacity increases fast enough, profits may be decreasing in speed.

Proof. Total risk aversion-weighted passive volume for market maker j is given by

$$\kappa_j \downarrow F\ \ddot{A} q_j \ddot{a} \qquad F \downarrow q_{j} \square + Q_j \square = 1 (1\ \gamma)\ q_j$$

which is monotone decreasing in j. All else equal, the fastest traders will intermediate the highest volumes. However, passive volume increases with intermediation tier if risk capacity $\kappa_{j1}$ increases faster than $q_j$, namely if

$$\kappa_{j1} \qquad \kappa_j > (1\ \gamma) \qquad \frac{q_j \qquad q_j}{F\,\text{Ä}q_j\text{ä} \qquad F \downarrow q_{j\square} + Q_j}$$

Average spreads for newly intermediated trades are given by

$$\frac{\int_{q_j}^{q_j} (1+\gamma x)\, f(x)\, dx}{\int_{q_j}^{q_j} f(x)\, dx} = 1+\gamma \frac{\int_{q_j}^{q_j} xf(x)\, dx}{F\,\text{Ä}q_j\text{ä} \qquad F \downarrow q_{j\square}}$$

Taking the minimal and maximal intermediated contract informedness x gives bounds for the average spreads

$$1+\gamma q_{j1} \quad = 1+\gamma q_{j\square} \quad \frac{\int_{q_j}^{q_j} (1+\gamma x)\, f(x)\, dx}{\int_{q_j}^{q_j} f(x)\, dx} \quad \square\ 1+\gamma q_j$$

where the inequalities are strict if $q_j > q_j$. Average spreads for non-market makers increase in the intermediary tier j because $q_j$ increases in j.

I prove profits decrease with j in Appendix B.

Proposition 1 supports anecdotal evidence that high frequency traders offer lower spreads yet make higher profits than do slower market makers. In the model, fast market makers select the least toxic trades to intermediate and suffer only weak information costs as a consequence. They scale small spreads per trade, $1+\gamma x_i$, by large volumes to attain outsized returns. Despite their effective monopoly power, HFTs cannot achieve higher rents per trade because larger spreads are not incentive-compatible for sellers. Note that results for volumes and profits are guaranteed to hold only when risk capacity increases with speed. Slow market makers with high risk tolerance suffer higher order flow toxicity than do HFTs, but greater risk-bearing ability can nonetheless generate higher total volumes and compensation. For

**Page 24**

this reason, I supplement univariate regressions on speed with bivariate regressions that incorporate a risk capacity proxy.

Proposition 1 describes observable quantities as a function of rank in a sort by speed. For empirical tests, in addition to running regressions using raw speed values as the independent variable, I run regressions on speed rank to align with the setup of the model.[14] The model suggests differences in speed matter only if they change ordering in the intermediation hierarchy, and that outcome variables are monotone in this ordering. Additionally, relations with respect to raw speed are contingent on the distribution of informedness of fundamental traders and are potentially nonlinear. For example, variation in speeds among HFTs may be very small, but the corresponding differences in profits can be quite large. By contrast, similar speed differences among slow intermediaries may generate no spread in profits. Using ranks rather than raw variables can smooth out such nonlinearities and also serves as a robustness check for regression model misspecification.

I take speed as exogenous for testing Proposition 1. I discuss on a case-by-case basis how the choice of speed could drive or affect interpretations of the results. The model is extended in Section 6.2 to endogenize speed as a function of risk capacity and of heterogenous costs of investment in speed.

### 5.2.1 Spreads

Table 7 provides empirical support for average spreads varying negatively with market maker speed. I define spreads as the price impact in basis points from the previous mid price associated with trading aggressively as a fundamental or small trader. Average spreads are the dollar-weighted value of spreads charged by intermediaries to fundamental order initiators. Panel A presents the average spread charged to fundamental traders by intermediaries of each type. I expressly do not control for quantities or other attributes of the order or counterparty; the rationale for variation in spreads is that fast intermediaries are able to cherry-pick the most attractive order types. Spreads decrease monotonically with speed as predicted in the proposition. Additionally, dispersion among average spreads appears to decrease with speed, as well. Lower dispersion could suggest fast traders exclusively intermediate for uninformed traders, whereas slower market makers are less able to discriminate among order flow and intermediate for a wider range of traders.

[14] I use the MATLAB tiedrank function to account for ties in ranks. Tiedrank takes average ranks where values are equal. For example, a set 13,7,4,4,6l would be ranked in ascending order as 11,5,2.5,2.5,4l.

Panel B of Table 7 regresses average spreads on log switching frequency

$$\text{average spread}_i = \alpha + \beta\,(\text{log switching frequency})_i + \in_i \qquad (3)$$

where I adopt the convention that switching frequency ranks are increasing in speed. The left (right) subpanel performs regressions for raw (rank) variables. To ensure relations are robust across the market maker speed spectrum, I regress both for the fastest 20% of market makers, as determined by the number of intraday zero crossings, and for the entire population of market makers.[15] Results are also replicated with a control for market maker risk capacity—

the maximal absolute position of the account intraday—to isolate the effects of speed from the primary potential confound of Proposition 1.

I estimate βs separately for each trading day t = 1,..., 5 and then average coefficient estimates across dates. This approach is equivalent to a Fama-MacBeth procedure without factor prices (see, e.g., Section 12.3 of Cochrane (2005)). I correct for time series correlation in the parameter estimates using a one-lag Bartlett kernel for the $\hat{\beta}$ covariance matrix when calculating standard errors. The Fama-MacBeth procedure allows for each day to be treated separately while correcting OLS standard errors for cross-sectional correlation. I take this estimation approach throughout this work. The advantage over a true panel approach is most apparent for extensions of this work. For example, the E-mini S&P 500 contract alone has approximately 3,000,000 observations per trading day, or roughly 750,000,000 observations per year. Treating each day as independent up to a coefficient lag structure easily surmounts numerical issues when considering intermediation effects in the time series.

Panel B presents a robust, negative slope of spreads with respect to speed. In regressions with raw variables, a one point increase in switching frequency decreases spreads by 0.0781 to 0.320 basis points. This relationship holds across the speed spectrum and is invariant to the use of the maximal position size control. Although statistically weak, in large part due to the small sample of T = 5 days, the point estimates are universally economically significant. For example, improving log switching frequency by one in gold is associated with a predicted decrease in spreads of 6.3% of the equal-weighted average spread that market makers charge. The comparable estimates in silver (11.7%) and copper (6.4%) are larger still. Rank regression results are similarly strong: an improvement in speed rank by one position is associated with a 0.0157-0.112 decrease in rank of spreads charged. Put differently, advancing nine ranks in silver is predicted to reduce the position in an ordering by spreads by one.

Differences in risk aversion in addition to differences in speeds may drive dispersion in

[15] I split market maker speeds at the 80th percentile because it strikes a balance between capturing qualitatively fast accounts and having sufficient sample size. The sample size for the HFT subgroup is too small to obtain precise point estimates.

spreads. Market makers may charge high spreads if they are especially sensitive to adverse selection or inventory risk. In the model, high speed implies lower adverse selection, and hence, lower spreads. Speed only partially compensates for the risks to the market maker, so $R^2$s are generally quite low. However, $R^2$s are much larger among the fastest market makers than among the market making sector as a whole, perhaps indicating speed is a more salient characteristic for determining the observed behavior of faster intermediaries.

### 5.2.2 Volumes

Table 8 confirms volumes strongly increase in speed. I take logs of dollar volumes to prevent results from being dominated by the activity of high frequency traders. In computing volumes, I restrict trading to fundamental trader or faster market maker counterparties, as defined by zero crossing quantile. I impose this restriction for two reasons. First, it prevents double-counting of market maker volume. Second, it aligns with the spirit of passive trading in the model. Strictly speaking, Proposition 1 proves volumes from fundamental trader and faster market maker counterparties are increasing in speed rank and does not make predictions for total volumes. This restriction understates the contribution of fast market makers to volume as their offloading of risky positions to slower traders is ignored. As a result, estimates of slopes of volume with respect to speed are downwardly biased relative to the alternative in which reintermediation volume is included twice.

Panel A shows log trading volume decreases monotonically from 18.2-18.9 among HFTs to 15.1-15.7 among the slowest 50% of market makers. These differences in logs translate into an impressive factor of 25-33 difference in levels. On a per account basis, HFTs dominate market making activity.

Panel B finds volumes increase strongly in speeds. The regression specification parallels that of Equation (3):

$$\text{log volume}_i = \alpha + \beta\ (\text{log switching frequency})_i + \in_i$$

An increase in log switching frequency by one is associated with a log dollar volume increase of 0.464-0.608, or more than 60% in levels. Given the dispersion in the right-hand side variable from roughly 0 to 4, speed differences alone can explain a factor of 7-11 difference in volumes among market makers. Rank and rank regressions also show volumes are nearly monotone in speed. A one rank improvement in speed is associated with a 0.348-0.459 rank improvement in volumes. Higher slopes and stronger ordering of volumes by speed among the fastest accounts again suggests speed may be more salient for the operations of the fastest market makers.

The large t-statistics and $R^2$s of Panel B are cause for concern. The switching frequency measure is intimately related to volume. Consider the account of a slow market maker with several subaccounts representing the activity of a large firm. Switching between buy and sell orders may occur quickly because different subaccounts intermediate independently rather than because a single entity generates new opposing orders quickly. The mechanical linkage between speed and volume is a known issue for assessing the activity of high frequency traders; as an example, Kirilenko et al. (2011) do not attempt to distinguish the concepts and define HFTs simply as high volume market makers. Ideally I would use a volume independent measure of speed, such as latency with the exchange, but these data are not available.

Controlling for maximal net position partly addresses this concern. Unconstrained market makers with independent subaccounts will have higher variance in positions and, consequently, greater maximal positions. Coefficient estimates are also potentially subject to omitted variable bias with respect to risk capacities, for which I use maximal position size as a proxy. Accounts with large risk capacities may be faster and more active, biasing βs upward. Like volume, risk capacity and speed are difficult to disentangle. Indeed, for both reasons, it is unsurprising coefficients decline in all regression specifications when including this control. Nevertheless, the residual regression results still strongly link speed and volume statistically and economically.

### 5.2.3 Profits

Proposition 1 implies profits net of inventory risk are increasing in speed. If faster traders take on more inventory risk, as Table 10 suggests, then higher profits net of inventory risk also imply higher gross profits. Table 9 presents evidence that (gross) profits are increasing with speed. I take the signed log of profits, $\text{sign}(\pi) \rightarrow \log(1 + |\pi|)$, to address substantial right skew in profits. Estimating profitability over a five day period is a dangerous exercise as it is prone to noisy price dynamics outside of the model. As such the signed log of profits given in Panel A is highly variable and the averages themselves are not necessarily monotone in speed. Moreover, average log profitability is close to zero within each type classification.

Noise notwithstanding, slope results in Panel B testify to a strong relationship between profits and speed. Panel B regresses profits on switching frequency

$$\text{signed log profits}_i = \alpha + \beta\,(\text{log switching frequency})_i + \in_i$$

Slope coefficients for the raw on raw regressions are typically statistically significant at the 5% or 10% level and highly economically significant. An increase in log switching frequency by one is associated with roughly a 0.5 log increase in profits in all commodities, or in levels,

a 65% profit improvement. In gold and silver, the profit-speed relationship is particularly pronounced among the fastest market makers, where a comparable increase in absolute speeds (if feasible) is associated with more than a 140% increase in profits. Rank on rank regression coefficients are also universally positive, but the point estimates are less notable. Advancing 20 speed ranks predicts advancement of one profit rank. $R_2$'s are low throughout, likely due to the dominance on the intraday frequency of stochastic price movements over differing drift terms among market makers.

Proposition 1 implies the profit-speed relationship may be weakened (strengthened) if risk capacity decreases (increases) in speed. Slow market makers with high risk capacities could potentially generate higher profits than faster market makers. Controlling for the risk capacity proxy of maximal position, I find coefficient estimates for raw variables to be roughly the same across the spectrum of market maker speeds. Speed rank results are similar for slow market makers, but slopes are markedly reduced for the 20% of market makers with the most zero crossings. These results suggest risk capacity and speed are positively correlated and related to profits for the fastest accounts. Indeed, endogenizing speed as a function of risk capacity generates such an alignment, as in Section 6.2.

### 5.2.4 Summary

This section empirically establishes that spreads, volumes, and profits differ substantially across market makers and are closely linked to market maker speed. These results support the intuition of the model: fast intermediaries select the most advantageous order flow and accordingly can offer lower spreads and intermediate higher volumes. Speed's effects on all quantities are economically large, despite low $R_2$s. Given this cross-sectional variation within the intermediation sector, I now address how aggregate variables, such as liquidity, vary as a function of this dispersion and whether speed differences are sustainable. These questions are the subject of the following section.

# 6 Additional Implications of the Model

The speed hierarchy model of Section 4 has additional theoretical implications for short- and long-run liquidity provision by market makers. First, I show speed heterogeneity among market makers can be desirable in the sense of improving liquidity for fundamental parties seeking to trade. This result maintains the assumption of fixed market maker speeds. Second, I endogenize speed and show no equilibrium exists in absolute speeds. An arms race ensues even if fundamental traders are unwilling to compensate execution time improvements. Moreover,

if differences in speed costs drive the arms race, aggregate intermediation capacity can fall as risk tolerant market makers are marginalized by faster market makers.

## 6.1 Liquidity Benefits of Speed Heterogeneity

In Section 5.1, I show observed volume may be a poor proxy for fundamental trader volume. A sufficient statistic for liquidity in the model is aggregate volume net of reintermediation activity: orders are intermediated in a fixed sequence of increasing fundamental trader surplus, so the set of orders intermediated is increasing in fundamental volume. I can compare analytically the fundamental volume of economies with (tiered) and without (flat) speed heterogeneity. I find a speed hierarchy generates more total intermediation volume than a flat structure, and if risk capacities increase sufficiently with speed, more fundamental volume, as well. Proposition 2 works through the comparison of these intermediation structures.

Proposition 2 (Flat and Hierarchical Intermediation). The tiered economy without retrading described in Section 4.1 facilitates greater (new) intermediation capacity than a comparable "flat" structure in which intermediaries choose to intermediate simultaneously and symmetrically. Tiered economies with or without reintermediation also generate more volume overall.

Proof. I assume in the flat structure that there is no price competition and intermediaries are symmetric.[16] Further, for simplicity and to maintain symmetry among intermediaries, I take risk aversion κ as fixed and identical across market makers. By symmetry, all intermediaries act as though f (x) is divided by N and intermediation bounds are $q_1$ and $q_1$ = A. Aggregate intermediation volume is then given by

$$F\ (q_1)\ F\ (A) = \frac{N}{\kappa}\ [1\,(1\ \gamma)\ q_1]\ ,\ q_1 = \frac{1\ \ \square_N\ (F\ (q_1)\ F\ (A))}{1\ \gamma}$$

In the comparable tiered structure without reintermediation, intermediation volume for the jth market maker is

$$F\ \text{Ä}q^{\text{ä}}_{\text{tiered}_j}\quad F\ \text{Ä}q_{\text{tiered}_{j1}}\ \text{ä} = \frac{1}{\kappa\ \text{î}1\ q_{\text{tiered}_j}}\ (1\ \gamma)\text{ó}\ ,\ q_{\text{tiered}_j} = \frac{1\,\kappa\ \text{ÄF Ä}q_{\text{tiered}_j}\ ^{\text{ä}}\quad F\ \text{Ä}q_{\text{tiered}_{j1}}\ \text{ää}}{1\ \gamma}$$

[16]This equilibrium is not unique, but the symmetric action equilibrium is presented as a point of comparison.

Total intermediation volume sums individual F $\text{Ä}q_j\text{ä}$ F $\downarrow q_{j}\square$

$$F\ \text{Ä}q^{\text{ä}}_{\text{tiered}_N}\quad F\ (A) = \sum_{j=1}^{N}{}^{\text{ä}}\ \text{ÄF Ä}q_{\text{tiered}_j}\quad F\ \text{Ä}q_{\text{tiered}_{j1}}\ \text{ää}$$
$$= \frac{1}{\kappa}\sum_{j=1}^{N}\text{î}1\ (1\ \gamma)\ q_{\text{tiered}_j}\ \text{ó}\quad \frac{N}{\kappa}\ \text{î}1\ (1\ \gamma)\ q_{\text{tiered}_N}\ \text{ó} \qquad (4)$$

The right-hand side is identical in form across both market structures. The left-hand side is increasing in $q_{\text{tiered}_N}$ whereas the right-hand side is decreasing in $q_{\text{tiered}_N}$. Hence $q_{\text{tiered}_N}$ must be larger than $q_1$. Correspondingly, more volume must be intermediated with speed heterogeneity relative to the flat structure. Introducing price competition among market makers only strengthens this volume result: inventory costs are identical but profits per trade are reduced only in the flat structure, which in turn further reduces aggregate intermediation activity.

Extending the same argument to total intermediation volume in a reintermediation economy gives

$$F\left(q^{tiered}_N\right) - F(A) = \sum_{j=1}^{N} \left(F\left(q^{tiered}_j\right) - F\left(q^{tiered}_{j-1}\right)\right) = \frac{1}{\kappa} \sum_{j=1}^{N} \left[l - (1-\gamma)\, q^{tiered}_j - \kappa Q_j\right]$$

The inequality in (4) no longer holds because of the additional $\kappa Q_j$ term, so no statement can be made comparing fundamental intermediation volume between a reintermediation and a flat economy.

However, comparing total intermediation volume among economies is immediate. Including reintermediation volume gives

$$\begin{aligned} F\left(q^{tiered}_N\right) - F(A) + \sum_{j=1}^{N} Q_j &= \sum_{j=1}^{N} \left(F\left(q^{tiered}_j\right) - F\left(q^{tiered}_{j-1}\right) + Q_j\right) \\ &= \frac{1}{\kappa} \sum_{j=1}^{N} \left[l - (1-\gamma)\, q^{tiered}_j\right] - \frac{N}{\kappa} \left[l - (1-\gamma)\, q^{tiered}_N\right] \end{aligned}$$

Because each $Q_j$ is positive, the $q^{tiered}_N$ that equates this expression must be smaller than that in the no-reintermediation case. However, the total quantity on the left must be larger, because the right-hand side is decreasing in $q^{tiered}_N$. Therefore, both economies with speed heterogeneity have higher volumes than a flat intermediation economy, all else equal.

Tiered intermediation forces slow market makers to take relatively informed contracts or none at all. These trades are only (marginally) profitable when inventories are low. In an

Page 31

economy with equal speeds, even the slowest market makers have moderately large positions, so marginal inventory risk is too great to intermediate highly informed orders. Traders are then better off under speed heterogeneity in the narrow sense of more intermediation of privately desirable and especially well-informed trades. Outside the model, impounding the information from this highly informed order flow contributes to socially valuable price discovery as well. In a richer model, competition among market makers may drive down spreads to increase aggregate fundamental trader surplus, so I can make no statements on fundamental trader welfare.

From the perspective of fundamental traders, speed heterogeneity economies can facilitate more and more highly informed volume and hence offer liquidity for trades that would be otherwise unintermediated. However, risk sharing among intermediaries allows slower market makers to substitute new intermediation volume with reintermediation of less toxic order flow. The net effect on new intermediation volumes is then ambiguous. Liquidity is enhanced on net if risk capacities align sufficiently with speeds so that reintermediation volume is relatively low. Table 10 suggests such alignment occurs among fast market makers, as speed rank is strongly increasing in risk capacity proxies. The next section endogenizes speed, shows theoretically why such alignment should be expected, and considers this empirical fact in detail.

## 6.2 Endogenous Speed and the HFT Arms Race

The liquidity improvement result of the previous section assumes speeds and risk capacities are held fixed. However, variation in intermediary speeds can be socially costly, particularly when traders choose and compete on speed. I now augment the model with cost functions for speed. Traders can accelerate ever faster to leapfrog their competition in the intermediation hierarchy. In this section I show which traders one should expect ex ante to be fastest. Intriguingly, the identity of the fastest firms may shift simply as a function of technological upper bounds to speed.

Adding in reasonable cost of speed specifications generates an arms race in which absolute speeds s are permanently in disequilibrium, but a pure strategy equilibrium ordering can nevertheless be established. I formalize this intuition in Proposition 3 and the subsequent text.

Proposition 3 (Arms Race). If the speed cost function $_i$ satisfies $_{0i} > 0$ and $_{00i} > 0$ for each market maker i, then no Nash equilibrium in pure strategies exists in absolute speeds in the generic case.

Proof. Conjecture there exists an equilibrium set of (absolute) speeds chosen by market

**Page 32**

makers $s_1,...,s_N$. Without loss of generality, let the speeds be ordered $s_1$ $s_2$ ... $s_N$.
Denote the set of risk aversions as $□_1$, $□_2$,..., $□_N$. Let the profits associated with being the jth intermediary be given by ⇡ (j; $□_i$). Suppose that the speed cost functions $_i$(s) have support on the positive reals and are increasing in s. For the participation constraint to be satisfied, ⇡ (j ($s_1$,$s_2$,...,$s_N$); $□_i$) $_i$($s_i$) for i = 1,...,N.

If $s_i$ 6= $s_{i+1}$ for any i, the set of speeds and net profits 1($s_i$, ⇡ (j (·); $□_i$) $_i$($s_i$))l $^N_{i=1}$ cannot be an equilibrium. Intermediary i can unilaterally deviate to $s_{0i} = s_{i+1}$ + □, □ = $s_i$ $\frac{s_{i+1}}{2}$, to obtain net profits

$$⇡ (j; □_i) \quad _i(s_{0i}) > ⇡ (j; □_i) \quad _i(s_i)$$

Note that placing a lower bound s on speed does not affect this result because $s_{0i} > s_{i+1}$ s.
The Nth market maker may push this boundary, but it does not repair the equilibrium for any other intermediary.

Equilibrium also cannot exist if $s_i = s_{i+1}$ for any i. Because the cost function is differentiable and hence continuous, for any > 0, there exists an □ > 0 such that

$$_i(s_{i+1} + \square) \qquad _i(s_{i+1}) <$$

Let = ⇡ (j (·); □i) ⇡ (j (·) + 1; □i) > 0, where positivity follows from profits increasing in speed given fixed risk aversion. Then increasing speed from $s_{i+1}$ to $s_{i+1} + \square$ generates an increase in net profits of

$$[\uparrow (j; \square_i) \qquad _i(s_{j+1} + \square)]\ [\uparrow (j + 1; \square_i) \qquad _i(s_{i+1})] > 0$$

contradicting the conjecture of an equilibrium in this case. Even if the participation constraint were to bind for the j + 1st agent such that ⇡ (j, □i) < $_i(s_{i+1} + \square)$, there must be an $\square_0 = \min(\square, \uparrow (j, \square_i)\ \uparrow (j + 1, \square_i))$ such that leapfrogging is possible. Heterogeneity in cost functions or risk aversions is therefore insufficient to generate an equilibrium in absolute speeds.

Suppose now that none of the participation constraints bind for any i and j. An equilibrium does exist in the very special case in which an upper bound on speed s is preferred 8$s_0$ < s for all intermediaries i conditional on s $_i$ = s, namely, when

$$[\uparrow (1; \square_i) \qquad _i(s)]\ [\uparrow (N; \square_i) \qquad _i(s)] > 0$$

Since □ is bounded above by 0, there does not exist □ > 0 such that leapfrogging other market

makers is possible. One equilibrium is then given by

$$s_1 = s_2 = \ldots = s_N = s \qquad (5)$$

that is, all market makers achieve the fastest speed technologically available.

The non-existence of Nash equilibria in pure strategies outside of the special case follows from the absence of better-reply security (Reny (1999)). The special case is distinct in that slight deviations in s $_i$ do not alter $s_{-i}$. An immediate consequence of Proposition 3 is that an arms race in absolute speeds is possible and socially costly if small improvements to execution times are not valuable to fundamental traders. Investments in speed that do not change speed rankings are entirely wasteful in this setting.

Under mild conditions on the informedness distribution and costs of speed $_i$, a Nash equilibrium in relative speeds (orderings) exists under both heterogeneity in risk aversion and heterogeneity in underlying costs of speed. Orderings provide better-reply security, and under additional mild conditions, satisfy the requirements of Reny (1999) Theorem 3.1 to guarantee existence of an equilibrium. The incentive-compatibility condition necessary to support such an equilibrium is

$$\uparrow (j; \square_i) \quad {}_i(j) \quad \uparrow (j_0; \square_i) \quad {}_i(j_0)\ 8j_0\ 6{=}\ j$$

where I now define ${}_i(j)$ as the cost of speed for intermediary i to be jth in the intermediation hierarchy. The position j of intermediary i net of costs must be the most profitable among potential positions given the positions of other intermediaries. Market makers with particularly steep profit increases with respect to position or low costs of becoming faster satisfy increasing differences and will separate at the top of the speed hierarchy. A sufficient condition for equilibrium to obtain is that, for all i, there are speeds for which the jth intermediary in the hierarchy prefers to be at least that fast because of higher profits or lower costs and the j + 1 through Nth intermediaries prefer to be (weakly) slower. I show in the following proposition that this condition holds in the fairly general case in which profits are convex in speed rank and risk aversion varies among market makers.[17]

Proposition 4 (Equilibrium Ordering in Speed). A Nash equilibrium ordering of market makers exists if $f_0(x) < 0$ and profits are convex in speed rank.

Proof. I show existence formally in Appendix B. Heuristically, the proof proceeds in two parts. First, using $f_0(x) < 0$, I show increasing differences—higher risk capacity implies

[17] Equilibrium in pure strategies exists under more general conditions, i.e., quasiconcavity of payoffs in j.

improvements in speed rank generate larger profits. The heterogeneity in benefits to speed enables high risk aversion market makers to sustain a separating equilibrium if the cost of speed function is well-behaved. The remainder of the proof is dedicated to showing the necessary incentive-compatibility conditions indeed hold.

Table 10 offers empirical support for speed sorting by risk capacity, as also theorized by Biais et al. (2011). Panel A shows a roughly monotone relationship between market maker speed and log maximal position. Panel B regresses log switching frequency on risk capacity and finds positive, highly significant slopes throughout. The interpretation associated with the sorting on risk capacity differs between measures of risk capacity, however. If maximal position is the appropriate measure of risk capacity, fast market makers appear to separate, whereas slower market makers are closely bunched. Conversely, separation appears to occur over the entire range of market maker speeds if maximal flow rate is the correct measure of risk capacity. Both sets of results are consistent with the heterogenous risk capacity equilibrium presented above. However, if capital flows to more profitable positions in the intermediation hierarchy, risk capacity itself will be endogenous.

### 6.2.1 Equilibrium with Speed Cost Heterogeneity

I now consider a catch-all factor for determining speed rank: comparative advantage in speed. Suppose heterogeneity exists only in the cost functions for speed corresponding to an innate advantage in applying the human and physical capital required of fast market making. In particular, assume cost functions are parameterized by ↵, @(·; ↵) /@s > @ (·, ↵$_0$) /@s for all ↵ > ↵$_0$ and all j. Furthermore, suppose (N, ↵) = (N, ↵$_0$) = c > 0 for all ↵ and ↵$_0$. This parameterization captures roughly equal costs of very slow speeds—voice trading or online brokerages require minimal infrastructure for anyone—but increasing and diverging costs of applying the resources necessary to achieve higher speeds. I work through this case in detail because separation along the speed cost dimension is new to the HFT literature and has important implications for long-term liquidity provision.

In the special case of equal risk capacities, □$_i$ = □, separation will occur only when speed costs differ sufficiently so that incentive-compatibility conditions for remaining in place are met. In the three agent case, sufficient conditions are

↵$_1$ (1) ↵$_1$ (2) < ⇡ (1; □) ⇡ (2; □) < ↵$_2$ (1) ↵$_2$ (2) <

↵$_2$ (2) ↵$_2$ (3) < ⇡ (2; □) ⇡ (3; □) < ↵$_3$ (2) ↵$_3$ (3)

where all indexes for the profit functions are adjacent, and

↵$_1$ (1) ↵$_1$ (3) < ⇡ (1; □) ⇡ (3; □) < ↵$_3$ (1) ↵$_3$ (3)

otherwise. These expressions are generalizable beyond three agents analogously with the risk capacity heterogeneity case described in Appendix B. Here the profit differences within the hierarchy are given but the cost functions vary across agents. An equilibrium obtains if the set of cost functions is such that each is convex in speed rank and satisfies the assumed condition on partial derivatives to generate increasing differences. In Figure 6, I construct one such equilibrium with cost heterogeneity in which cost functions are of the form ↵$_i$s$_2$, ↵$_i$ = 1.05↵$_{i+1}$.

Figure 6's top subplot depicts the distribution of net profits and relative speeds by intermediation tier with separation by costs of speed. By Proposition 4 these absolute speeds are not pinned down, but they suffice for illustrating the ordering of market makers in the speed hierarchy. I normalize the net profit to being the Nth intermediary to a number □ > 0 to ensure all participation constraints are met. Net profits are again convex by assumption so that market makers with slight cost advantages invest most in speed and achieve far greater profits than do their competition. The bottom subplot shows signed logs of the incentive-compatibility conditions in this equilibrium

$$IC_{j,j}\,(\text{blue}) : (\uparrow\ (j)\ \uparrow\ (j+1))\ (_{j}\,(j) \qquad {}_{j}\,(j+1)) \tag{6}$$

$$IC_{j+1,j}\,(\text{green}) : (\uparrow\ (j)\ \uparrow\ (j+1))\ (_{j+1}\,(j) \qquad {}_{j+1}\,(j+1)) \tag{7}$$

The first condition, when positive, implies the jth intermediary will not deviate and slow to become the j + 1st in the hierarchy for any j. The second condition, when negative, implies the j + 1st intermediary will not deviate to advance a place in the hierarchy for any j. As speeds increase, these conditions will be enforced more rigidly as speed advantages—for example, in deployment in human or physical capital—are further amplified. Equilibrium in speed ranks exists because increasing differences with respect to costs of speed satisfy all incentive-compatibility conditions.

Heterogeneity in speed costs grows with s, whereas heterogeneity in risk aversion generates only fixed returns to advancing in the intermediation hierarchy. Hence, for s large enough, speed cost differences can dominate risk capacity profit differences. If s binds for the fastest market makers, technological advances that increase s will exacerbate cost heterogeneity—equilibrium at the former s disappears—but leave profit differences due to risk capacity heterogeneity unchanged. Consequently, we would expect traders with comparative speed advantages to supplant traditional market makers as trading technology improves over time.

All else equal, this transition makes worse off those with high risk capacities that had previously dominated the market.

The entry of new, low-cost market makers crowds out other intermediaries from the market if falling back in the speed hierarchy reduces their profits below fixed costs of operation. High risk capacity firms are especially likely to be displaced because the negative effect of small increases in the toxicity of order flow roughly scales with trading volume. Aggregate intermediation capacity decreases with the exit of these firms if high fixed costs and low intermediation revenues likewise deter potential entrants. Hence HFT entry has an adverse long-run impact on liquidity if speed costs are convex, profits for slow market makers are close to fixed costs of operation, and risk capacity is initially negatively correlated with speed.

Section 6.1 shows speed heterogeneity can improve liquidity in the short run with speeds held fixed. This section demonstrates that in the long run, the high frequency arms race can reduce intermediation capacity if capital does not move quickly to align with changing intermediary speed orderings. As an extreme example, anecdotal evidence suggests the fastest intermediaries in the E-Mini S&P 500 futures market suffered from low risk capacity during the Flash Crash, as most high frequency traders exited the market rather than intermediate at fire sale spreads (Kirilenko et al. (2011)). The worrisome failure of the intermediation sector to stabilize prices could be among the many consequences of an emerging misalignment of market maker speed and risk capacity.

# 7 Testing Assumptions of the Speed Hierarchy Model

Tables 7, 8, and 9 provide empirical support for Proposition 1. Faster market makers offer lower spreads, intermediate higher volumes, and achieve higher profits. In this section, I consider the model's assumptions of risk sharing between fast and slow market makers and order flow informedness predictability. I find additional evidence in support of the underlying mechanisms of the model in that slow market makers appear to close the positions of faster ones, and informedness of order flow is predictable out of sample.

## 7.1 Risk Sharing

I take two approaches to testing for risk sharing. First, I show faster intermediaries tend to close their positions when initiating orders against slower traders, and slow traders tend to open positions when initiating orders against faster traders. The asymmetric fast-slow contract flow suggests slower market makers indeed take on the risky positions of faster ones. Second, I show faster intermediaries do not profit directly from trading with slower market

makers, contrary to a common prior (e.g., that of Cartea and Penalva (2012)) that high frequency traders harm slower market makers directly.

I do not test risk sharing by evaluating whether intermediaries with large positions transfer contracts to intermediaries with smaller positions. Risk sharing in the model does not require market makers to attempt to equate inventories ex post. Some intermediaries are better able to tolerate risk than others, so the observed flow could instead go in the opposite direction under this alternative test. By contrast, in the model, reintermediated volume always flows from fast to slow market makers. Only if risk capacities are weakly increasing in speed would both models predict flows from accounts with large positions to those with small positions.

Table 11 presents these risk sharing results. Panel A calculates the share of volume that closes open positions when the intermediary with the higher switching frequency is the aggressive (top) and passive (bottom) party. In this context, I interpret the aggressive party as the one that motivates the trade. A share of 0.5 indicates an equal propensity to close and open positions. I find minimal indication of risk sharing among the fastest 20% of market makers. The probability of these market makers closing positions is roughly equal, implying inventory does not consistently flow in one direction. By contrast, for the full sample of market makers, values for faster (slower) trades are typically above (below) 0.5 and, in gold and copper, the propensity to close positions is significantly different at the 10% level.

The 1.5%-2.4% difference in opening propensities is economically large. Accumulating net positions equal to this share of HFT volume would swamp slower market makers if fundamental counterparties cannot be located and positions were opened in the same direction. Using values for gold from Tables 4 and 5 as an example, the 2.4% open-close discrepancy scaled by $14.9 billion HFT volume is $357.6 million, roughly half of the summed maximal observed position of the slowest 50% of market makers. High frequency trader pass-through therefore could potentially saturate slower market makers' intermediation capacity in the face of a Flash Crash-scale order flow imbalance.[18] Panel B presents a non-result: bilateral profits generated from trading aggressively or passively with faster intermediaries are universally indistinguishable from zero. The model predicts weak profits by slower market makers. However, given the prior that HFTs take considerable advantage of slower market makers, zero profits are a considerable step in the predicted direction. I find no sense in which fast traders directly take advantage of slower ones, although profits are generally decreasing in speed rank.

The risk-sharing mechanism in the model rationalizes the potentially surprising finding that high frequency traders do not appear to profit from trading with slower market makers.

[18] Of course, the counterfactual scenario of a flash crash in which HFTs do not intermediate and risk share may well be worse for slower market makers.

Proposition 1 suggests HFTs hurt slower market makers by leaving only marginally profitable order flow for them to intermediate. The harm is found in the absence of otherwise profitable trades rather than in the presence of unprofitable realized trades, as borne out by the negative result in Panel B.

Panel A of Table 11 suggests that as in the model, fast market makers primarily intermediate new volume from fundamental traders, whereas slow intermediaries largely serve to absorb the inventory risk of faster ones. Table 12 provides additional support for this interpretation. Entries in the table report the share of each counterparty type in opening positions held by market makers, where values are averaged across dates. The fastest market makers indeed have much larger shares of fundamental and small trader volume than other intermediaries, who in turn trade much more with HFTs than with fundamental counterparties. Furthermore, these relationships are approximately monotone with market maker speed. Fundamental counterparty volume shares increase with market maker speed and HFT counterparty shares decrease with it. These relations suggest fast market makers indeed primarily intermediate fundamental volume, whereas slower market makers largely absorb positions of faster ones. However, in isolation, these relationships are also consistent with several alternative stories, such as HFTs consciously avoiding each other or preying on fundamental traders, but it is nonetheless telling in juxtaposition to the position-closing result.

## 7.2 Order Flow Informedness

Most existing measures of trade informedness are valid only for single transactions or are evaluated in the aggregate.[19] They do not provide the resolution necessary to identify the informedness associated with individual trades. I proxy informedness on a trade-by-trade-basis as the time required for an aggressive buy (sell) price to return to its pre-trade best bid (offer) price. An extended sojourn time in the direction of a trade is more likely if the trade has a permanent price impact. Of course, not all trades associated with extended sojourns are informed. Trades with large temporary price impacts or coincident with others' impounding of information are misclassified by this measure. However, large orders that might contribute to large temporary impacts are quite rare, and the noise term in the regression should absorb the serendipitous trades.

I construct the sojourn time using future data from the perspective of agents at time t. As such, whereas the measure works for analyzing counterparty informedness ex post,

[19] Much related literature for permanent price impacts from information effects focuses on (single) block trades (e.g., Kraus and Stoll (1972), Loeb (1983), and Keim and Madhavan (1996)). Hasbrouck (1991a,b) considers a bivariate trade and quote innovation system for estimating trade informedness in the aggregate.

it is questionable for testing for contemporaneous predictability. Because I do not have a real time estimate of order flow informedness, I can only indirectly test for predictability by supplementing this imperfect measure with additional related order flow predictability results.

I first show some market maker relevant variables are highly contemporaneously predictable. In particular, Panel A of Table 13 presents coefficients and out-of-sample $R^2$s for a rolling forecast of the direction of trade of the trade's initiator, or aggressor. I denote an aggressive sell with 0 and an aggressive buy with 1. For each commodity and trading date, I perform the rolling regression for each transaction T of aggressiveness on lagged aggressiveness

$$\text{aggressiveness}_t = \alpha + \sum_{j=1}^{5} \beta_j \, \text{aggressiveness}_{t-j} + \epsilon_t \qquad (8)$$

for trades t = 1,...,T. I impose an AR(5) specification to capture lagged dependence or runs as in Hasbrouck and Saar (2010). I then forecast aggressiveness one (two) periods out for the left (right) panel of the table.

I update coefficients using recursive least squares to make the transaction-by-transaction forecasts computationally feasible. Results are identical across a wide range of initial (prior) parameters due to large intraday sample sizes. Σ is the estimated dynamic causal multiplier, or the sum of full sample coefficients across lags. I calculate pseudo out-of-sample (POOS) $R^2$s by taking one minus the variance of the forecast errors divided by the variance of aggressiveness.

Panel A finds aggressiveness to be highly predictable in real time. All else equal, aggressive buy trades cumulatively increase the probability of future aggressive buy trades by 0.657. This strong and positive dependence makes aggressive order flow prone to streaks of buyer- or seller-initiated orders. POOS $R_2$s are correspondingly large at between 0.2 and 0.3. I show coefficients and POOS $R_2$s for two-period-ahead dependence to account for a possible lag in receiving and processing transaction feeds. Out-of-sample forecasting power two periods ahead is still substantial but is reduced by roughly half relative to the one-period-ahead forecast. These results confirm the equity markets finding of Hasbrouck and Ho (1987) and Biais et al. (1995) that order flow aggressiveness is highly predictable.

I now repeat this analysis using estimated counterparty informedness in place of aggressiveness. I assume the log time to first return serves as a proxy for a true real-time measure of informedness in intermediaries' information sets. Unfortunately, this proxy may be the best available: no public, contemporaneous, trade-by-trade informedness measures exist to the best of my knowledge.

Panel B of Table 13 presents results. The estimated coefficient is again economically

and statistically large. All else equal, an increase of one log second until return implies the next five prices (of the same aggressor parity) have a total 0.32-0.36 increase in log seconds until return to the previous price. Permanent price impacts tend to come in runs, although these runs are much noisier than those for aggressiveness. The POOS $R_2$s are smaller as a result. I also find two-trade-ahead predictability with point estimates at roughly half the one-trade-ahead level, but $R_2$s are much closer to zero.

Small $R_2$s in Panel B notwithstanding, Panel C shows informedness predictability has appreciable economic significance. I combine the predictor variables in Panels A and B to generate signed order flow informedness, that is, informedness multiplied by the direction of the aggressor (+1 for buys, -1 for sells). I then forecast signed informedness as in Equation (8).[20]

I construct a simple front-running strategy based on forecasting signed informedness. When the forecast informedness is greater than two log trades in the positive (negative) direction, the trader buys (sells) at the current price and sells (buys) one or two periods ahead for the left and right subpanels, respectively. This strategy assumes the trader can analyze the order book in real time and execute just prior to the forecasted trades without altering the strategy of the next trader. As such, only high frequency traders are likely to be able to implement this strategy. Averaged across the five days, order flow predictability employed in this fashion generates profits of thousands or tens of thousands of dollars depending on the forecast horizon and commodity market. Profits per trade are small but consistent, yielding annualized Sharpe ratios in excess of 30 even for two-trade-ahead forecasts. However, orders closing the positions opened by this strategy must be placed passively for profits to survive transactions costs, again suggesting this strategy is available only to fast market makers.

## 8 Conclusion

The canonical view of market makers as a single entity or homogenous group is incorrect. Considerable speed heterogeneity is a defining attribute of the intermediation sector in asset markets. Variation among speeds generates lengthy chains of market makers separating fundamental sellers and buyers of assets. These chains challenge traditional conceptions of liquidity. First, most intermediated transactions involve several market makers of varying speeds. This finding suggests the concept of market liquidity can be meaningfully decomposed into the ease of trade for new orders and the ease of placement for orders in the intermediation system. Second, the high volume associated with inter-market maker risk sharing may be symptomatic of the use of intermediation capacity rather than of available liquidity.

Non-uniformity in speed also explains variation in market maker outcomes. Higher speed is associated with improved spreads, larger volumes, and greater profits. On the latter two dimensions, higher speed has such large benefits to market makers that the coexistence of intermediaries of differing speeds is surprising. I find empirical support for slow market makers' conjectured risk-sharing function: slow intermediaries close open positions of faster intermediaries significantly more than the reverse. This flow of contracts along the speed gradient may indicate a symbiotic relationship between fast market makers with short horizons and slower market makers that are more patient. Understanding the interactions between these groups and fast versus slow liquidity provision more broadly is a critical frontier for assessing market depth and stability.

In addition to describing the cross-section of market maker outcomes, the speed hierarchy model also resolves the puzzle of how high frequency traders can appear highly compensated for millisecond execution time improvements even if the value of such improvements is zero from the perspective of fundamental traders. In the model, high frequency traders are not compensated by a fixed set of fundamental traders for these improvements. Instead, speed increases enable intermediaries to overtake faster market makers for access to more desirable, less informed clientele. Advancing a rank in the speed hierarchy thus generates large profits regardless of the absolute change in speed associated with that advancement or its utility to fundamental traders. Selection effects are important even in ostensibly anonymous markets.

The model also speaks to aggregate variables. I show theoretically that speed heterogeneity can improve liquidity in the short run by facilitating trades that might otherwise be deemed

[20] I provide additional discussion of these regressions and of the associated trading strategy in Appendix D.

too toxic to intermediate. Conversely, the process that gives rise to speed heterogeneity may be socially costly. Separating equilibria in relative speeds exist, but tâtonnement can require continual investment in speed infrastructure without any benefits to fundamental traders or even to the investing intermediaries themselves. The resulting arms race creates a sheer welfare loss to market makers and potentially to liquidity consumers, as well.

The HFT arms race can also fundamentally change risk flows in the economy and aggregate intermediation capacity. If high frequency traders initially have low capital, their emergence at the front of the intermediation hierarchy depresses fundamental volumes in favor of large reintermediation volumes as slower market makers substitute risk sharing for adversely selected new order flow. If capital moves slowly within the intermediation sector, high frequency traders can adversely select slow intermediaries out of the market without offering new risk capacity in their place. As a result, the introduction of fast, low-capital intermediaries can render markets less able to bear large liquidity demand shocks. The sudden prevalence of flash crashes—Nanex, a market data feed provider, estimates more than 1,800 miniature flash

crashes occurred in 2010 alone—is not surprising when viewed from this perspective.

Speed heterogeneity among market makers has likely long been a salient feature of intermediation in asset markets. This study offers a first step to understanding its hallmarks and aggregate effects. Much additional work is needed to analyze market maker heterogeneity's welfare consequences and means to stabilize liquidity provision during periods of rapid innovation in market making technology. Incorporating realistic price and order book dynamics and explicitly including fundamental buyers in the speed hierarchy remain as important extensions to this work.

# References

Biais, Bruno, Pierre Hillion, and Chester Spatt, “An Empirical Analysis of the Limit Order Book and the Order Flow in the Paris Bourse,” Journal of Finance, December 1995, 50 (5), 1655–1689.

, Thierry Foucault, and Sophie Moinas, “Equilibrium High Frequency Trading,” Working Paper, SSRN, http://ssrn.com/abstract=1834344 September 2011.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, “High Frequency Trading and Price Discovery,” Working Paper, SSRN, http://ssrn.com/abstract=1928510 July 2012.

Cartea, ´Alvaro and José Penalva, “Where is the Value in High Frequency Trading?,” Working Paper, SSRN, http://ssrn.com/abstract=1712765 February 2012.

Chaboud, Alain, Benjamin Chiquoine, Erik Hjalmarsson, and Clara Vega, “Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market,” FRB International Finance Discussion Paper 980, SSRN, http://ssrn.com/abstract=1501135 June 2011.

Cochrane, John H., Asset Pricing - Revised Edition, Princeton University Press, 2005.

Grossman, Sanford J. and Merton H. Miller, “Liquidity and Market Structure,”Journal of Finance, July 1988, 43 (3), 617–633.

Hansch, Oliver, Narayan Y. Naik, and S. Viswanathan, “Do Inventories Matter in Dealership Markets? Evidence from the London Stock Exchange,” Journal of Finance, October 1998, 53 (5), 1623–1656.

Hasbrouck, Joel, “Measuring the Information Content of Stock Trades,”Journal of Finance, 1991, 46 (1), 179–207.

, “The Summary Informativeness of Stock Trades: An Econometric Analysis,” Review of Financial Studies, 1991, 4 (3), 571–595.

and Gideon Saar, “Low-Latency Trading,” Working Paper October 2010.

and Thomas S. Y. Ho, “Order Arrival, Quote Behavior, and the Return-Generating Process,” Journal of Finance, September 1987, 42 (4), 1035–1048.

Hendershott, Terrence and Ryan Riordan, “Algorithmic Trading and Information,” Working Paper 2011.

, Charles M. Jones, and Albert J. Menkveld, “Does Algorithmic Trading Improve Liquidity?,” Journal of Finance, February 2011, 66 (1), 1–33.

Ho, Thomas and Hans R. Stoll, “The Dynamics of Dealer Markets Under Competition,” Journal of Finance, September 1983, 38 (4), 1053–1074.

Keim, D. B. and Ananth Madhavan, “The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects,” Review of Financial Studies, 1996, 9, 1–36.

Kirilenko, Andrei, Albert (Pete) Kyle, Samadi Mehrdad, and Tugkan Tuzun, “The Flash Crash: The Impact of High Frequency Trading on an Electronic Market,” Working Paper, SSRN, http://ssrn.com/abstract=1686004 May 2011.

Kraus, A. and H. Stoll, “Price Impacts of Block Trading on the New York Stock Exchange,” Journal of Finance, 1972, 27, 569–588.

Locke, Peter R. and Pattarake Sarajoti, “Interdealer Trading in Futures Markets,” Working Paper, SSRN, http://ssrn.com/abstract=265932 April 2001.

Loeb, T. F., “Trading Cost: The Critical Link Between Investment Information and Results,” Financial Analysts Journal, 1983, 39, 39–43.

Lyons, Richard K., “A Simultaneous Trade Model of the Foreign Exchange Hot Potato,” Journal of International Economics, May 1997, pp. 275–298.

Reiss, Peter C. and Ingrid M. Werner, “Does Risk Sharing Motivate Interdealer Trading?,” Journal of Finance, October 1998, 53 (5), 1657–1703.

Reny, Philip J., "On the Existence of Pure and Mixed Strategy Nash Equilibria in Discontinuous Games," Econometrica, September 1999, 67 (5), 1029–1056.

U.S. Commodity Futures Trading Commission and U.S. Securities & Exchange Commission, "Preliminary Findings Regarding the Market Events of May 6, 2010: Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues," Technical Report May 2010.

Vogler, Karl-Hubert, "Risk Allocation and Inter-Dealer Trading," European Economic Review, 1997, 41, 1615–1634.

**Page 45**

# Guide to Notation

## Fundamental Trader Parameters

l> 0 The common liquidity motive for fundamental trading.

x The payoff (informedness) associated with trading one contract from the fundamental traders' perspective. Also the cost to the market maker of facilitating trade of one contract.

f (x) ,F (x) The density and cumulative distribution functions for informedness of fundamental traders in the economy. In general f (x) is not constrained to be a proper probability density.

[A, B] The interval denoting the bounds for fundamental trader informedness and for which f (x) > 0.

2 (0, 1) The bargaining power held by intermediaries in negotiation over informational rents from trading. = 1 implies all surplus goes to market makers.

## Market Maker (Intermediary) Parameters

□j The risk aversion of market maker j. Also, the inverse of the intermediary's risk capacity.

N The number of (discrete) intermediaries in the economy.

## Trading Variables

c(x) The spread or fee charged by the market maker to fundamental traders for intermediating a contract of informedness x.

$q_j$ ,$q_j$ The lower and upper limits on the informedness of trades intermediated by market maker j.

F Äq$_j$ä F ↓q $_j$□ The new intermediation volume of market maker j.

$Q_j$ The mass of contracts held by market maker j 1 that are reintermediated by market maker j.

↑$_j$ The profits net of inventory and informational risk for market maker j.

(j) The cost associated with the level of speed required to be in position j in the intermediation hierarchy.

# A Tables and Figures

## Table 1 – Contract and Trading Information for CME Metal Futures

This table presents contract information for each of the Chicago Mercantile Exchange (CME) futures contracts analyzed. Symbol refers to contracts' CME commodity codes. The most active contract is the contract month with the most volume during the week of December 12-16, 2011. Contract sizes are the amount in units of weight of the futures' underlying physical products. Dollar contract sizes multiply this weight by the 2011 average prices in dollars. The contract share of CME futures volume is the volume in the most active contract relative to total volume in all futures contracts on the underlying commodity at the CME. CME share of all futures volume denotes the total volume by weight traded in the specified CME futures contract relative to that traded in all futures markets in the underlying metal in the January-December 2011 period. CME (futures) share of futures and options incorporates all options market activity, as well. Ranks among global exchanges are given in parentheses.

| | Gold | Silver | Copper |
|---|---|---|---|
| Symbol | GC | SI | HG |
| Most Active Contract | February 2012 | March 2012 | March 2012 |
| Contract Size (weight) | 100 ozt. | 5,000 ozt. | 25,000 lbs. |
| Contract Size (dollars) | $157,000 | $176,000 | $100,000 |
| Contract Share of CME Futures Volume | 92.0% | 92.7% | 87.3% |
| CME Futures Volume Share | 77.5% (#1) | 71.6% (#1) | 10.9% (#3) |
| CME Futures + Options Volume Share | 66.8% (#1) | 66.5% (#1) | 10.3% (#3) |

## Table 2 – Trading Characteristics by Commodity

This table presents cross-day averages for trading characteristics by commodity market. The number of accounts denotes the number of unique CME customer account identifiers active in the contract. Average trade and order sizes are the average across dates of the mean values for the size in contracts of each transaction and each order for each date. Tick sizes are set by the Chicago Mercantile Exchange and reported publicly at http://www.cmegroup.com/trading/metals/.

| | Gold | Silver | Copper |
|---|---|---|---|
| Number of Trades (thousands) | 127 | 34.9 | 30.2 |
| Contract Volume (thousands) | 169 | 42.3 | 39.0 |
| Dollar Volume (billions) | 27.4 | 6.33 | 3.28 |
| Number of Accounts | 3360 | 1410 | 1100 |
| Average Trade Size (contracts) | 1.33 | 1.21 | 1.29 |
| Average Order Size (contracts) | 2.37 | 1.86 | 2.09 |
| Tick Size (dollars) | 0.1/ozt. | 0.005/ozt. | 0.0005/lb. |

# Table 3 – Classification Methodology and Robustness

Panel 3a summarizes the classification methodology for accounts in the CME-CFTC data set. Accounts that satisfy multiple criteria are classified by order of precedence. For example, a trader that buys nine contracts will be designated a small trader. Panel 3b presents average transition matrices among classifications across days and commodities. The first subtable computes observed transitions among major subgroups. The second subtable focuses on transitions among quantiles within the intermediary category of the number of times positions cross zero. The (i, j) entry corresponds to the empirical probability that an account classified as type i on day t will be classified as of type j on t + 1, where this probability is averaged across days in the sample and across commodities.

(a) Classification Methodology

| | Classifiers (in order of classification precedence) |
|---|---|
| Small Trader | Trades fewer than 10 contracts |
| Fundamental Trader | Absolute end-of-day position greater than 20% of volume |
| Market Maker | Intraday position crosses zero at least twice |
| Opportunistic Trader | Active but does not fall under other categories |

(b) Classification Robustness

| | MM | FT | ST | OT | NT |
|---|---|---|---|---|---|
| Market Maker | 0.549 | 0.100 | 0.109 | 0.156 | 0.0858 |
| Fundamental Trader | 0.0792 | 0.438 | 0.152 | 0.0987 | 0.233 |
| Small Trader | 0.0325 | 0.0408 | 0.409 | 0.039 | 0.479 |
| Opportunistic Trader | 0.218 | 0.130 | 0.205 | 0.209 | 0.238 |
| Not Trading | 0.00292 | 0.00807 | 0.0578 | 0.0059 | 0.925 |

|  | 95-100 | 90-95 | 80-90 | 50-80 | 0-50 | NMM |
|---|---|---|---|---|---|---|
| Market Maker - 95-100% | 0.610 | 0.203 | 0.0939 | 0.068 | 0.0110 | 0.0140 |
| - 90-95% | 0.250 | 0.239 | 0.212 | 0.186 | 0.0309 | 0.0816 |
| - 80-90% | 0.0431 | 0.159 | 0.250 | 0.254 | 0.119 | 0.175 |
| - 50-80% | 0.00761 | 0.0248 | 0.0942 | 0.281 | 0.175 | 0.417 |
| - 0-50% | 0.00107 | 0.00501 | 0.0334 | 0.150 | 0.184 | 0.626 |
| Not Market Making | 0.0000387 | 0.000116 | 0.000833 | 0.00461 | 0.00827 | 0.986 |

Page 48

## Table 4 – Participant Characteristics by Type: Aggregates

Panel A presents the average number of participants in each commodity by type across days in the sample. Panel B gives the total trading volume in dollars made by traders of a given type averaged across days. Panel C breaks this volume into aggressive and passive components, where aggression denotes initiating a trade by meeting an existing passive limit order in the limit order book. In all panels, splits among market makers are along quantiles of the number of times positions cross zero.

|  | Gold | Silver | Copper |
|---|---|---|---|
|  |  | Panel A: Number of Participants |  |
| Market Makers - 95-100% | 29.8 | 10.0 | 6.80 |
| - 90-95% | 29.4 | 10.4 | 7.40 |
| - 80-90% | 64.2 | 23.4 | 15.4 |
| - 50-80% | 208 | 70.4 | 46.6 |
| - 0-50% | 251 | 87.2 | 64.8 |
| Fundamental Traders | 505 | 212 | 188 |
| Small Traders | 1830 | 826 | 638 |
| Opportunistic Traders | 442 | 169 | 134 |
|  |  | Panel B: Trading Volume (billions of $) |  |
| Market Makers - 95-100% | 14.9 | 3.60 | 1.53 |
| - 90-95% | 2.47 | 0.986 | 1.02 |
| - 80-90% | 4.19 | 1.12 | 0.805 |
| - 50-80% | 6.86 | 1.53 | 0.688 |
| - 0-50% | 4.95 | 0.953 | 0.441 |
| Fundamental Traders | 14.3 | 2.48 | 1.15 |
| Small Traders | 1.13 | 0.466 | 0.200 |
| Opportunistic Traders | 5.99 | 1.52 | 0.729 |
|  |  | Panel C: Aggressiveness (% of $ volume) |  |
| Market Makers - 95-100% | 51.6 | 56.9 | 60.5 |
| - 90-95% | 50.6 | 60.5 | 52.7 |
| - 80-90% | 53.7 | 50.2 | 52.6 |

| | | | |
|---|---|---|---|
| - 50-80% | 53.2 | 49.9 | 42.2 |
| - 0-50% | 50.2 | 41.3 | 45.8 |
| Fundamental Traders | 44.7 | 40.4 | 43.4 |
| Small Traders | 58.1 | 51.9 | 52.0 |
| Opportunistic Traders | 47.6 | 43.4 | 43.1 |

Note: The number of participants in Panel A denotes the average number of accounts trading in each category per trading day. The total number of unique accounts over which averages are computed are often substantially larger. For example, the 10.0 and 6.80 values in the first row correspond respectively with 20 and 12 unique accounts that hold the 95-100% designation over the observation period.

Page 49

## Table 5 – Participant Characteristics by Type: Means

Panel A shows the average across days of the mean maximal net position by type. Panel B computes inter-day averages of means by type for the inverse of one second plus the 10th percentile of durations between submission of orders of opposite signs, i.e., a buy order followed by a sell order or the reverse. Panel C reports the same statistic for the number of times a trader's net position crosses zero. Panel D reports the Spearman rank correlation of the speed measures used for empirical work, the switch frequency of Panel B and the number of zero crossings of Panel C. Correlations are averaged across dates. In all panels, splits among market makers are along quantiles of the number of times positions cross zero.

| | Gold | Silver | Copper |
|---|---|---|---|
| | Panel A: Maximal Net Positions (millions of $) | | |
| Market Makers - 95-100% | 4.28 | 3.32 | 2.16 |
| - 90-95% | 2.82 | 2.09 | 2.13 |
| - 80-90% | 3.85 | 1.85 | 2.35 |
| - 50-80% | 3.41 | 2.02 | 1.39 |
| - 0-50% | 3.01 | 1.78 | 1.17 |
| Fundamental Traders | 16.6 | 7.38 | 4.25 |
| Small Traders | 0.345 | 0.329 | 0.196 |
| Opportunistic Traders | 3.26 | 2.28 | 1.41 |
| | Panel B: Order Direction Switch Frequencies (1/minutes) | | |
| Market Makers - 95-100% | 32.5 | 30.3 | 23.5 |
| - 90-95% | 14.2 | 23.8 | 20.7 |
| - 80-90% | 11.9 | 10.5 | 18.0 |
| - 50-80% | 7.16 | 8.51 | 7.04 |
| - 0-50% | 5.05 | 5.10 | 3.85 |
| Fundamental Traders | 2.95 | 1.74 | 2.01 |
| Small Traders | 1.97 | 2.03 | 1.49 |
| Opportunistic Traders | 3.49 | 3.42 | 3.06 |
| | Panel C: Net Position Zero Crossings | | |
| Market Makers - 95-100% | 95.9 | 130 | 135 |
| - 90-95% | 25.8 | 33.9 | 45.4 |

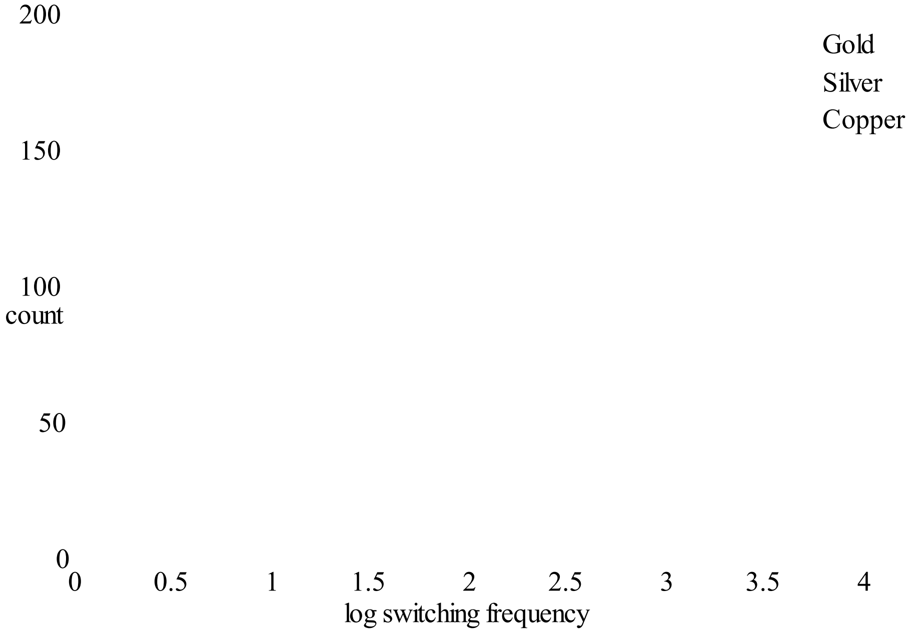
| | | | | |
|---|---|---|---|---|
| | - 80-90% | 13.7 | 14.6 | 18.9 |
| | - 50-80% | 5.96 | 6.10 | 6.86 |
| | - 0-50% | 2.42 | 2.56 | 2.62 |
| Fundamental Traders | | 0.472 | 0.405 | 0.331 |
| Small Traders | | 0.138 | 0.137 | 0.109 |
| Opportunistic Traders | | 0.313 | 0.328 | 0.349 |
| | | Panel D: Switching Time and Zero Crossing Rank Correlation | | |
| All Traders | | 0.374 | 0.396 | 0.483 |

Page 50

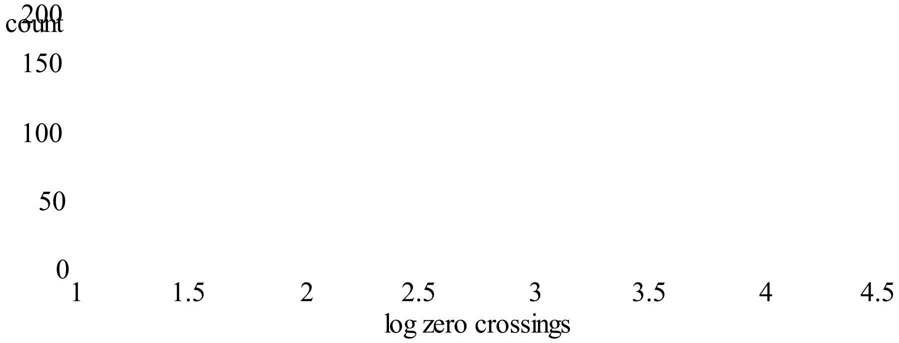
## Figure 1 – Market Maker Speeds in Gold, Silver, and Copper Futures

The top figure presents the empirical distribution of market maker log switching frequencies in gold, silver, and copper futures, where colors represent the respective commodities. Histogram counts are averaged across dates in the sample. The bottom figure presents the analogous empirical distribution for the number of times market maker positions cross zero intraday. The dashed black line denotes truncation on the right to preserve anonymity of individual accounts. Counts associated with truncated values are added to the rightmost visible stack.

(a) Order Direction Switch Frequencies (1/minutes)

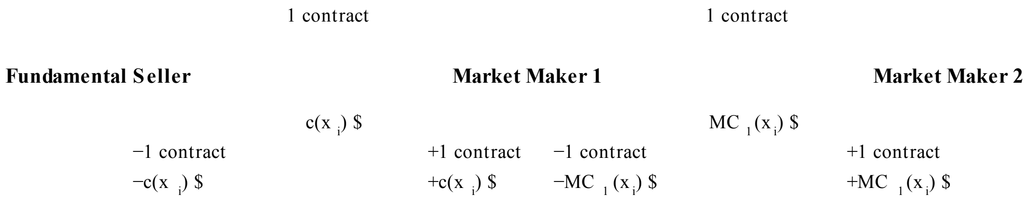


(b) Net Position Zero Crossings



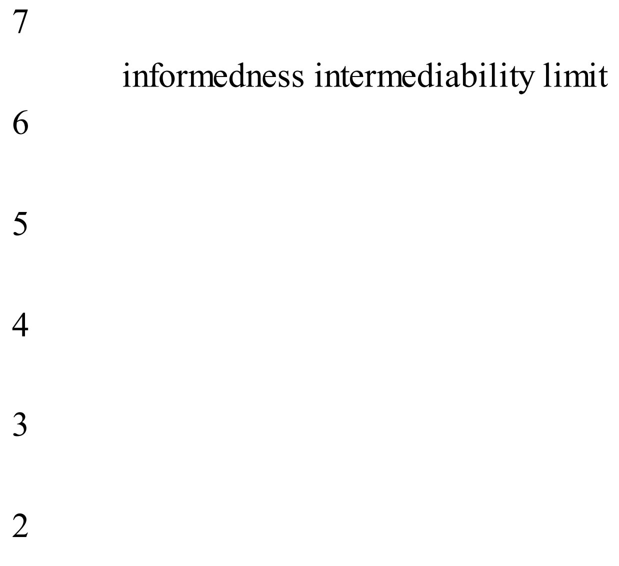
## Figure 2 – Trade in the Economy

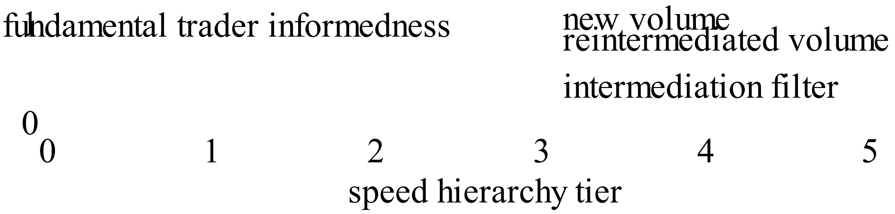
This figure depicts the flow of contracts in the model. Traders sell a price 0 contract to market makers and are charged a spread c(xi). Market makers (potentially) close faster market makers' positions at the faster intermediary's marginal cost of holding those contracts.



## Figure 3 – Intermediaries as Sequential Monopolists for Market Making

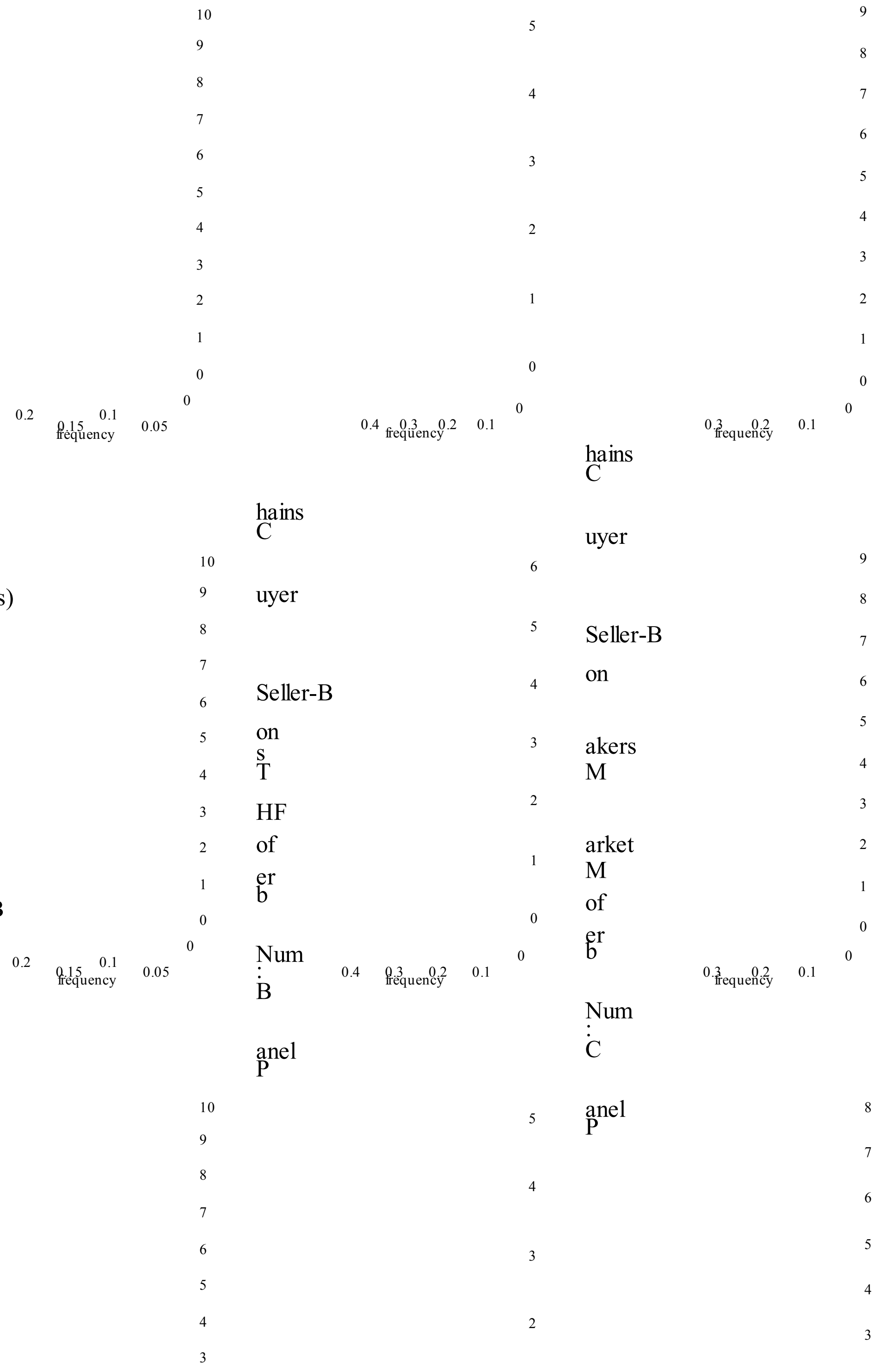
This figure presents order flow in a speed hierarchy in which risk tolerance increases with tier. Theorem 1 implies that the first market maker filters order flow and selects the least toxic contracts. Subsequent market makers filter the residual order flow and reintermediate contracts from the previous market maker. New volumes tend to decrease with intermediation tier. Contracts with informedness above the dashed line are too toxic to be intermediated.





Page 52

the
dates to denote
duration
ond bars
termediation across
in r e
trace corresp uency pp o C
buyer I averaged freq ectively.
to
buyer, olorings hue resp C B
seller tal uencies
for freq
tracts overlaps,
omitted. con
parties fundamen are of
a relative
to
the hich w
hains C
distribution Seconds)
ulation o
seller plot og
termediating tw (L
in tal
decum and
of for r e imes T
er b one
transactions,
um n fundamen Histograms to Silv uyer
a
ond
Intermediation the path
– from assumptions
4 and buyer-seller Seller-B
corresp A:
draw red
ure ig F times migrates direct
ck seller-buyer to anel P
clo that the and
log onds uniformly
of tract along G
and
con corresp O
each makers IF L
or ,
distribution residual O IF distributions.
the y. arket m he T F G



ts of three
economy. under all
presen
the num quan
in from
figure the each
his
This chains under for distribution overlaps

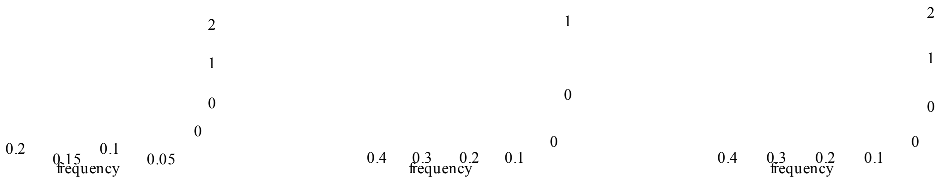


Table 6 – Intermediation Chains

| Decumulation Method | | OIF | OIF | | OIF | OIF | | OIF | OIF | | OIF | OIF | | OIF | OIF | | OIF | OIF | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Times (Log Seconds) | 90% | 9.89 | 9.66 | 9.84 | 9.95 | 9.75 | 9.93 | 9.98 | 9.75 | 9.92 | Total Number of Market Makers | 5 | 5.2 | 5.2 | 5.8 | 5.8 | 5.8 | 6.4 | 6.6 | 6.6 |
| | 70% | 8.98 | 8.36 | 8.81 | 9.06 | 8.52 | 8.91 | 9.14 | 8.55 | 8.95 | | 2.8 | 3 | 2.8 | 3 | 3 | 3 | 3.2 | 3.2 | 3.2 |
| | 50% | 8.02 | 7.12 | 7.78 | 7.99 | 7.35 | 7.86 | 8.17 | 7.34 | 7.92 | | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 1.8 | 2 | 2 | 2 |
| | 30% | 6.73 | 5.42 | 6.41 | 6.69 | 5.72 | 6.52 | 6.99 | 5.85 | 6.73 | | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 10% | 4.73 | 3.08 | 4.25 | 4.96 | 3.40 | 4.53 | 5.36 | 3.60 | 4.76 | | 0.8 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Times (Log Trades) | 90% | 10.7 | 10.5 | 10.7 | 9.59 | 9.37 | 9.54 | 9.46 | 9.17 | 9.38 | Number of HFTs | 2.8 | 2.8 | 2.8 | 3 | 3.2 | 3 | 2.8 | 2.8 | 2.8 |
| | 70% | 9.78 | 9.20 | 9.64 | 8.58 | 8.10 | 8.50 | 8.55 | 8.00 | 8.35 | | 1.2 | 1.2 | 1.2 | 1.2 | 1 | 1 | 1.2 | 1.2 | 1.2 |
| | 50% | 8.83 | 7.96 | 8.62 | 7.59 | 6.91 | 7.44 | 7.59 | 6.77 | 7.38 | | 1 | 1 | 1 | 1 | 1 | 1 | 0.8 | 0.8 | 0.8 |
| | 30% | 7.61 | 6.41 | 7.32 | 6.34 | 5.45 | 6.16 | 6.50 | 5.32 | 6.22 | | 0.4 | 0.4 | 0.4 | 0.2 | 0.2 | 0.2 | 0 | 0 | 0 |
| | 10% | 5.75 | 4.16 | 5.27 | 4.78 | 3.28 | 4.23 | 4.94 | 3.31 | 4.32 | | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Times | | | | | | | | | | | Counterparts | | | | | | | | | |
| | | | | | | | | | | | Intermediary | | | | | | | | | |

buyer to trace it clock
seller for buyer, they augmented (25-35%) as chains (in I. contracts
parties fundamental a to transactions tracking for dequeuing
intermediating seller term in of number the from path. each values
and migrates times that seller-buyer dates the within all
contract along and each logs uniform across decumulation keep to
trade log or FIFO. makers market of economy. averaged are tiles percentiles
the in number distribution the ts panel consider the chains each holdings.
seller-buyer direct assumptions fundamental common well-defined.
LIFO FIFO y. before b 1 y
times trade and taking

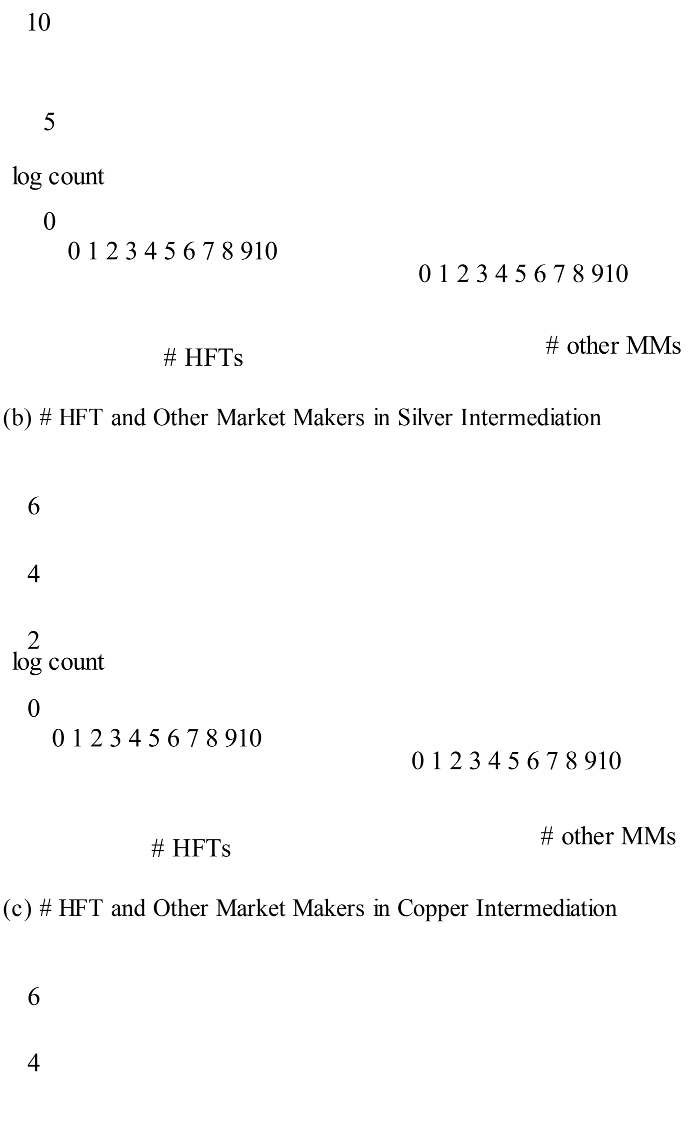
andE 's D termediation Fi L Uniform F L Uniform F L Uniform term F L Uniform F L Uniform F L Uniform
presen ed. second In In
1
y dity A: :
b B
table r r r r
duration e e e e
dropp anel ld pp anel ld pp
his termediation termediary ommop o P o
T in the are in times C P Go Silv C Go Silv C

Page 54

## Figure 5 – Distribution of Market Makers on Intermediation Chains

These figures present the empirical distribution of market makers on intermediation chains. For each contract passing from a fundamental seller to a fundamental buyer, I tally the number of high frequency traders and other market makers on the intermediation chain. Direct transactions are included as zeros and values are top-coded at 10. I average counts for each HFT-market maker pair across days and show the log of one plus the average count for gold, silver, and copper futures markets, respectively.

(a) # HFT and Other Market Makers in Gold Intermediation



(b) # HFT and Other Market Makers in Silver Intermediation

(c) # HFT and Other Market Makers in Copper Intermediation

log count
2
0
0 1 2 3 4 5 6 7 8 9 10
0 1 2 3 4 5 6 7 8 9 10

# HFTs # other MMs

Page 55

y b price trader across efined uencies. r e 0% 2 % 0 2 % 0 2 os. P

tal made trade freq are pp o C 0.54 1.10 1.82 2.10 2.65 2.38 -0.0322 -0.720 0.0243 109 st ste a F -0.265 -9.16 0.113 37 st ste a F -0.164 -4.60 0.112 22.6 ax M

ns nk a R astest F

the regression tio ia v e D

of initiated estimates itching sw positions p r e

fundamen ts) y b for Sp s. v All -0.0226 -1.21 -0.0167 441 ll A -0.119 -6.66 0.0277 156 ll A -0.0747 -2.04 0.0601 103 os. P ax M

oin p orders imal rd Silv 1.31 1.24 1.99 2.21 2.65 2.34

made for parameter maker M in nda nk a R 0% 2 % 0 2 % 0 2

basis t-statistics Sta

acity.

(in orders charged and □ i 's + ) i market of traday in cap ld o G 0.54 0.556 0.990 1.08 1.51 1.22 d a astest F -0.000855 -0.0903 0.0107 109 st ste a F -0.218 -5.51 0.0808 37 st ste a F -0.131 -2.90 0.0474 22.6 None

for erence correlation risk

diff spread each order zero for y Spre

charged the acB uency serial rank cross ll A -0.0157 -0.853 0.00232 441 ll A -0.112 -6.41 0.0225 156 ll A -0.0458 -1.27 0.0105 103 None

Spreads : 1 as mean freq for prox a r e tiles.

the ama-M F pp o C 1.51 1.38 2.20 2.63 2.23 2.31 %

spreads are ts itching sw I correct positions are p ascending and 20 % 0 2 % 0 2 ercen p

osition average defined maker B presen (log eed. sp the is times of er b d e e -0.0323 -1.08 0.0239 109 st ste a F -0.560 -5.36 0.132 37 st ste a F -0.562 -3.66 0.180 22.6 os. P ax M 99th d an ,

rop P the + ↵ for rank maker ns a e M r e astest F

– 7 of Spreads anel = i market y eed um n Sp market Silv 1.58 2.07 2.61 2.70 3.06 2.74 Sp w a R ll A -0.0877 -1.71 0.0155 441 ll A -0.333 -7.89 0.0326 156 ll A -0.210 -6.22 0.0304 103 os. P ax M 95th ,

ble a T iation for dates. party. prox a spread is the each on s. v 90th e

dev errors. for sort a ds a % 20 % 0 2 % 0 2 th at

across spreads passive uency average y b ld o G 0.923 0.943 1.16 1.22 1.36 1.24 ce

standard the is freq standard position p Spre -0.0185 -0.593 0.0100 109 st ste a F -0.503 -4.08 0.110 37 st ste a F -0.467 -3.02 0.0833 22.6 None can ifi

and averaged Average est given w a R astest F sign

are price hich w. itching sw W ey are traday in W) (E ll A -0.0781 -1.62 0.00879 441 ll A -0.320 -7.31 0.0302 156 ll A -0.148 -6.54 0.00733 103 None

mean in tiles statistical

the values mid New

ts order of lag one largest ercen p the of ls e v e L 95- - 90- - 80- - 50- - 40- - 20% - Average s e p ote en d

oth ious B traders

prev tal log

aker aker Slo

presen the the using maker log A l M M B l r e r e d an

anel P aggressors fundamen here w dates arket M as the ne a P arket M arket M ne a P ld Go ) ( t 2 R N Silv ) ( t 2 R N pp o C ) ( t 2 R N trols on C

faster averaged and are across efined uencies. freq are r e pp o C 0.898 1.10 1.02 1.07 0.978 1.38 0% 2

ns tio ia v e D nk a R astest F

traders values tal estimates itching positions sw p r e

oth B imal rd Silv 0.968 0.895 0.933 1.02 0.946 1.25 d e e Sp All

parameter maker in M nda s. v

fundamen trader. ith of w Sta nk a R 0% 2

market acity. of traday cap in ld o G e

risk correlation order zero 1.22 0.912 1.06 1.17 1.10 1.35 lum o V astest F

□ i + ) i makers category for y

serial rank cross ll A

olume V : 1 each y b market regression uency for freq the prox a r e

for pp o C 18.2 17.7 17.0 15.8 15.1 15.8 %

positions are corresponding ascending 20

volume traded itching and

osition sw the is times

eed of d e astest

rop P – 8 average t-statistics volume (log sp rank er b maker ns a e r e Sp F

the of and 's dollar + ↵ = i for y eed um n M Silv 18.9 17.5 16.7 16.1 15.6 16.2 w a R ll A

ble a T the iation of eth prox a Sp the market s. v

dev log acB is volume on errors. each for s e %

sort ld o G 18.8 17.4 17.0 16.2 15.7 16.3 20

the log a uency y lum o V

as F ama-M b osition

standard ts freq standard p w a astest F

est given R

and itching W are traday W) ll

measured presen B sw ey in (E A

mean is tiles 0.476

New

the anel P order lag largest ercen ls e 95-100% 90-95% 80-90% 50-80% 0-50% s e p

ts olume V of p the v - - - - - Average

one e L

log of aker aker Slo

presen dates. makers. the maker log : A M M : B

A using l l ld
the ne ne
anel here arket a arket arket a
P acrossmarket w datesM as P M M P Go