Relative Spread and Price Discovery

Eric M. Aldrich\textsuperscript{a}, Seung Lee\textsuperscript{b}

\textsuperscript{a}Department of Economics, University of California, Santa Cruz, ealdrich@ucsc.edu
\textsuperscript{b}Department of Economics, University of California, Santa Cruz, slee126@ucsc.edu

May 2016

Abstract

We develop a theoretical model to highlight a previously unexplored mechanism of price discovery: relative minimum price increments for equivalent assets trading on distinct financial exchanges. Although conventional wisdom dictates that futures market assets lead equities equivalents in terms of price formation, our model predicts that the opposite should be true when particular relative price conditions hold for the bids and offers of each asset. We develop a new empirical measure of price discovery which is suited to asynchronous, high-frequency transaction and quotation data, and apply it to the highly liquid E-mini/SPY pair in order to test the predictions of the model. Empirical evidence strongly supports the model and further demonstrates that relative minimum contract size plays an additional role in the formation of prices.

Keywords: Market microstructure, market design, high-frequency trading, entropy.
JEL Codes: G12, G14.

About CAFIN

The Center for Analytical Finance (CAFIN) includes a global network of researchers whose aim is to produce cutting edge research with practical applications in the area of finance and financial markets. CAFIN focuses primarily on three critical areas:

- Market Design
- Systemic Risk
- Financial Access

Seed funding for CAFIN has been provided by Dean Sheldon Kamieniecki of the Division of Social Sciences at the University of California, Santa Cruz.
Relative Spread and Price Discovery

Eric M. Aldrich∗
Department of Economics
University of California, Santa Cruz

Seung Lee†
Department of Economics
University of California, Santa Cruz

May 30, 2016

Abstract

We develop a theoretical model to highlight a previously unexplored mechanism of price discovery: relative minimum price increments for equivalent assets trading on distinct financial exchanges. Although conventional wisdom dictates that futures market assets lead equities equivalents in terms of price formation, our model predicts that the opposite should be true when particular relative price conditions hold for the bids and offers of each asset. We develop a new empirical measure of price discovery which is suited to asynchronous, high-frequency transaction and quotation data, and apply it to the highly liquid E-mini/SPY pair in order to test the predictions of the model. Empirical evidence strongly supports the model and further demonstrates that relative minimum contract size plays an additional role in the formation of prices.

Keywords: Market microstructure, market design, high-frequency trading, entropy.

JEL Classification: G12, G14.

∗Email: ealdrich@ucsc.edu.
†Email: slee126@ucsc.edu
1 Introduction

On May 10, 2010, four days after the Flash Crash, the CME Group issued a staff report explaining precautions taken at the Chicago Mercantile Exchange (CME) during the volatile events of May 6, 2010. According to the staff report, “The primary purposes of futures markets are to provide an efficient mechanism for price discovery and risk management” and “...stock index futures frequently represent the venue in which price information is revealed first, generally followed closely by spot markets” (Labuszewski and Co, 2010). Indeed, conventional trading wisdom dictates that “futures markets lead cash markets”.

Academic work has largely supported the statements of the CME Group staff report. Kawaller et al. (1987) and Stoll and Whaley (1990) are early examples, demonstrating a leading relationship between S&P 500 index futures and the S&P 500 index itself. Hasbrouck (2003) presents similar findings for small-denomination S&P 500 index futures contracts (E-minis) and equities market exchange traded funds (ETF) that track the S&P 500 index. Futures and cash markets for Canadian bonds are shown by Campbell and Hendry (2007) to behave in like manner. More recently, Laughlin et al. (2014) and Aldrich et al. (2016) use econometric methodology similar to that of this paper to articulate the tightly coupled relationship of messaging traffic between futures and equities exchanges.

While the “futures-leads-cash” relationship is widely considered a market standard, counterexamples exists. Stephan and Whaley (1990), Easley et al. (1998) and Chakravarty et al. (2004) document a reversal of informed trading for single stock options: equities lead the derivatives market. Easley et al. (1998), however, show that this reversal is an unconditional fact and is not true under specific put/call trade conditions. Yang (2009) shows the same reversal of informed trading to be true of currency markets.

The primary objective of this work is to contribute to current knowledge on the determinants of price discovery. In particular, we seek to understand why price discovery often occurs primarily in a single venue, despite the fact that a single asset, or its equivalents, trade in diverse venues. A standard explanation for the primacy of derivatives markets for price discovery is the ability of informed traders to exploit greater leverage. Fleming et al. (1996) argues that price discovery should occur in the market with lower trading costs as measured
by bid/offer spread, broker commissions and market impact of large orders. Following this logic, a wider minimum price increment (higher cost) would be associated with diminished price discovery.

Although contract specifications are uniform for stocks that trade on diverse exchanges in the National Market System (NMS), the same is not true across derivatives and cash (equities or currency spot) markets. After a simple basis adjustment (to account for dividends and interest), many pairs of futures and cash instruments may be considered identical, despite the fact that they trade in distinct locations. Further, because futures and cash instruments are regulated by different entities, they frequently differ with respect to contract specifications, such as minimum price increment, minimum contract size, notional value, etc.

A primary contribution of this paper is a model that highlights the relationship of relative minimum price increments to price discovery. Contrary to the reasoning above, that larger price increments should be associated with diminished information share, our model suggests a countervailing effect when the minimum increment in one market is larger than that of another. The mechanism relies on simple deterministic arbitrage among simultaneously posted bids and offers in each market. According to these arbitrage arguments, the model makes several specific predictions about the direction of informed trade and under what conditions price discovery reversals should occur.

Our second contribution is to refine the econometric methodology for detecting lead/lag behavior between financial exchanges. The current standard for measuring price discovery is developed in Hasbrouck (1995), which expresses fragmented market prices as a cointegrated system and which formally interprets market information share as the fraction of price variation attributed to the permanent innovation in each market. While this methodology is intuitively appealing when price discovery is viewed through the lens of variance contribution, it has a number of weaknesses. First, it requires synchronization of price observations across the cointegrated series. This is typically done by measuring prices in clock time. As demonstrated by Ane and Geman (2000) and Aldrich et al. (2015), transaction-time models are typically superior to clock-time models, especially for fine-grained, high-frequency applications. Second, VAR methodology imposes a parametric model. Third, identification issues often arise when incorporating sufficient lag information in the system. Hasbrouck (1995)
deals with this by imposing polynomial lag restrictions on the system coefficients. Finally, when adapting the method to recent, high-frequency data, the identification and estimation issues are compounded.

We introduce a method that is model free and well suited for high-frequency data. While the computational burden is not trivial, it does not suffer the identification issues that arise in richly parameterized vector autoregression systems. Further, it is not narrowly interpreted as a measure of variance decomposition, but is a direct measure of lead/lag relationships between transactions at distinct trading centers. It is specifically designed to deal with asynchronous data and implicitly tests a null hypothesis of no leading information in the transactions of a particular market.

Our final contribution is to apply our econometric methodology to test the model predictions on the liquid E-mini S&P 500 index futures/SPDR ETF pair (tickers ES and SPY, respectively). Our empirical work is especially careful with the issue of simultaneity, accounting for speed-of-light transmission latency between the CME Globex matching engine in Aurora Illinois, and the equities exchanges in New Jersey (we focus on Nasdaq OMX in Carteret). We find strong evidence in favor of the model, suggesting that relative minimum increments are an extremely important factor in terms of determining the share of informed trading and price discovery within a market. We further highlight the secondary effect of relative contract size, which accentuates/attenuates the share of price discovery.

Our results are supportive of recent work by Hagströmer and Menkveld (2016), who extend the methods of Hasbrouck (1995) to build a network map of information percolation, estimating not only the relative weights of market centers and participants in the FX market, but also the relative flow of information between them. While our approach is somewhat distinct, we likewise find that intermediaries are a fundamental component of price cointegration. Specifically, although price cointegration among fragmented markets is the result of a variety of factors, arbitrage opportunities within an information network of prices causes intermediaries to be a critical vehicle for maintaining price uniformity.

Fundamentally, rather than questioning how markets are organized and how information flows between participants and exchanges, our work questions why those flows are observed. Most importantly, we view our results as being of primary relevance to market regulators.
In October 2016, the Financial Industry Regulatory Authority (FINRA) will implement the Tick Size Pilot Program, as ordered by the Securities and Exchange Commission (SEC). Under the program, several test groups of small capitalization equities will be required to quote, and potentially transact, at wide price increments of $0.05, rather than the standard $0.01. Understanding the role of such price increments both within and across markets is important for regulators such as the SEC and Commodity Futures Trading Commission (CFTC), for whom policy coordination may be necessary in order to promote stable and well-functioning markets.

2 Model

2.1 Environment

The environment is comprised of a single asset, $S$, and a futures contract on the asset, $F$. The asset and futures contract trade in distinct markets separated by communication latency $\tau$. The markets are populated by three agents, who are distinguished by the following types: a market maker for $F$, a market maker for $S$ and an investor, who takes the role of an informed trader. All agents maximize linear utility

$$U(\mu, \sigma) = \mu - \frac{1}{2} \gamma \sigma^2. \quad (1)$$

In each market, $X \in \{F, S\}$, at time $t$, the market maker offers $q_{m,t}^{X,o}$ units of the asset for sale at price $p_{m,t}^{X,o}$ and bids to purchase $q_{m,t}^{X,b}$ at price $p_{m,t}^{X,b}$. She assesses the fair value of the asset to be $p_{m,t}^{X,f}$. The investor demands $q_{i,t}^{X,o}$ units of the asset at the market maker offer price and supplies $q_{i,t}^{X,b}$ units at the market maker bid. The difference between bid and offer prices is known as the spread and is denoted $2\xi_t^X$ (i.e. $\xi_t^X$ is the half spread). We make the following assumptions:

Assumption 2.1

*Market maker assessment of the fair price in market $X \in \{F, S\}$ is an increasing function of investor demand to buy in that market, $\partial p_{m,t}^{X,f}/\partial q_{i,t}^{X,o} > 0$, and a decreasing function of investor supply to sell in that market, $\partial p_{m,t}^{X,f}/\partial q_{i,t}^{X,b} < 0.$*
Assumption 2.2
Investor demand is a function of the bid/offer spread, with quantities demanded at the bid and offer decreasing with the size of the spread: $\frac{\partial q^{X,o}_{i,t}}{\partial \xi_t^X} < 0$ and $\frac{\partial q^{X,b}_{i,t}}{\partial \xi_t^X} < 0$ for $X \in \{F, S\}$.

Assumption 1 is a reflection of adverse selection: uninformed market makers adjust their assessment of fair market valuation with informed investor order flow. Assumption 2 is a reflection of trading costs: as costs increase, the informed investor demands less. The following proposition will allow us to simplify notation:

Proposition 2.3
The fair price for the market maker is equidistant between the posted bid and offer in each market: $p^{X,f}_{m,t} = p^{X,o}_{m,t} - \xi_t^X = p^{X,b}_{m,t} + \xi_t^X$, $X \in \{F, S\}$.

Proof By symmetry of the market maker problem outlined in Section 2.3, and the fact that they are uninformed about the direction of investor order flow, market makers achieve optimality by symmetrically placing bids and offers around their fair valuation of the assets, $F$ and $S$.

To ease notation in the sequel, we will only outline optimization problems for transactions at the best offer: a purchase by an investor and a sale by a market maker. The analogous problems for transactions on the bid are symmetric. We thus reduce notation in the following manner: $q^X_{m,t} = q^{X,o}_{m,t}$, $q^X_{i,t} = q^{X,o}_{i,t}$, and $p^X_{m,t} = p^{X,f}_{m,t}$, for $X \in \{F, S\}$.

2.2 Investor

The investor earns returns in market $X \in \{F, S\}$ by purchasing the asset at the current offer price, $p^X_{m,t} + \xi_t^X$, and liquidating at the subsequent bid price, $p^X_{m,t+1} - \xi_t^X$. We denote the realized return and its first two moments as

$$ r^X_{i,t+1} = p^X_{m,t+1} - \xi_t^X - p^X_{m,t} - \xi_t^X $$  \hspace{1cm} (2)

$$ \mu^X_{i,t} = E_i \left[ p^X_{m,t+1} - \xi_t^X \right] - p^X_{m,t} - \xi_t^X $$  \hspace{1cm} (3)

$$ \sigma^2_{i,t}^X = \text{Var}_i \left( r^X_{i,t+1} \right) $$  \hspace{1cm} (4)
Holding period returns over longer horizons are simply the sum of single-period realized returns. We make the assumption that assets $F$ and $S$ are perfectly correlated under the investor’s subjective assessment:

**Assumption 2.4**

\[
\text{Cov}_i(r_{i,t+1}, r_{i,t+1}^S) = \sigma_{i,t}^F \sigma_{i,t}^S,
\]

Given a budget, $B$, the investor allocates resources through share purchases, $q_{i,t}^F$, $q_{i,t}^S$:

\[
\max_{q_{i,t}^F, q_{i,t}^S} \mu_{i,t} - \frac{1}{2} \gamma \sigma_{i,t}^2
\]

subject to

\[
\mu_{i,t} = (\mathbb{E}_i [p_{m,t+1}^F - \xi_{t+1}^F] - p_{m,t}^F - \xi_t^F) q_{i,t}^F + (\mathbb{E}_i [p_{m,t+1}^S - \xi_{t+1}^S] - p_{m,t}^S - \xi_t^S) q_{i,t}^S
\]

\[
\sigma_{i,t}^2 = \left(q_{i,t}^F \sigma_{i,t}^F \right)^2 + \left(q_{i,t}^S \sigma_{i,t}^S \right)^2 + 2q_{i,t}^F q_{i,t}^S \sigma_{i,t}^F \sigma_{i,t}^S,
\]

\[
q_{i,t}^F (p_{m,t}^F + \xi_t^F) + q_{i,t}^S (p_{m,t}^S + \xi_t^S) \leq B.
\]

Equations (5b) and (5c) are the expected return and variance of the portfolio. In addition to the budget constraint, (5d), bounding inequality constraints exist for the control variables, $q_{i,t}^F$ and $q_{i,t}^S$; we have intentionally ignored these constraints as they enter the first-order conditions as constants and add little value to the economic digression of the model. The first-order conditions are

\[
\mathbb{E}_i [p_{m,t+1}^F - \xi_{t+1}^F] - p_{m,t}^F - \xi_t^F - (1 + \lambda_{i,t}^B) \frac{\partial p_{m,t}^F}{\partial q_{i,t}^F} q_{i,t}^F

- \gamma \sigma_{i,t}^F \left(q_{i,t}^F \sigma_{i,t}^F + q_{i,t}^S \sigma_{i,t}^S \right) - \lambda_{i,t}^B (p_{m,t}^F + \xi_t^F) = 0
\]

\[
\mathbb{E}_i [p_{m,t+1}^S - \xi_{t+1}^S] - p_{m,t}^S - \xi_t^S - (1 + \lambda_{i,t}^B) \frac{\partial p_{m,t}^S}{\partial q_{i,t}^S} q_{i,t}^S

- \gamma \sigma_{i,t}^S \left(q_{i,t}^S \sigma_{i,t}^S + q_{i,t}^F \sigma_{i,t}^F \right) - \lambda_{i,t}^B (p_{m,t}^S + \xi_t^S) = 0,
\]

7
where $\lambda_{i,t}^B$ is the Lagrange multiplier for constraint (5d). Equations (6) and (7) result in the following solutions:

$$q_{F,i,t} = \left( E_i \left[ p_{m,t+1}^F - \xi_{t+1}^F \right] - p_{m,t}^F - \xi_t^F \right) - \gamma q_{i,t}^F \sigma_{i,t}^F - \lambda_{i,t}^B (p_{m,t}^F + \xi_t^F) \times \left( \gamma \sigma_{i,t}^2 + (1 + \lambda_{i,t}^B) \frac{\partial p_{m,t}^F}{\partial q_{i,t}^F} \right)^{-1} \quad (8)$$

$$q_{S,i,t} = \left( E_i \left[ p_{m,t+1}^S - \xi_{t+1}^S \right] - p_{m,t}^S - \xi_t^S \right) - \gamma q_{i,t}^S \sigma_{i,t}^S - \lambda_{i,t}^B (p_{m,t}^S + \xi_t^S) \times \left( \gamma \sigma_{i,t}^2 + (1 + \lambda_{i,t}^B) \frac{\partial p_{m,t}^S}{\partial q_{i,t}^S} \right)^{-1}. \quad (9)$$

Equations (8) and (9) show that investor demand increases with expected return in the respective assets, decreases in the volatility of both assets (both own and cross volatility), decreases with the elasticity of price to investor demand ($\frac{\partial p_{m,t}^X}{\partial q_{i,t}^X}$), and decreases with demand of the other asset.

### 2.3 Market Maker

Market makers are required to continually post bids and offers in the $F$ and $S$ markets. They are compensated for their services via the bid/offer spreads, $\xi^F$ and $\xi^S$, which are their control variables. At the time of a passive sale in market $X \in \{F, S\}$ at the current offer price, $p_{m,t}^X + \xi_t^X$, the market maker earns returns via two mechanisms: (1) retaining some fraction of the sale quantity for a passive purchase (with an aggressive investor) in her own market in the subsequent period, and (2) aggressively purchasing some fraction of the sale quantity at the offer price in the other market, $X^c$, at time $t + \tau$, $p_{m,t+\tau}^{X^c} + \xi_{t+\tau}^{X^c}$. The time shift $\tau$ accounts for communication latency between the two markets. Since the market maker’s own-market transactions are passive, the expected repurchase price is her expected fair valuation, which is her long-term, passive transaction price. This is equivalent to saying that with equal probability, she will transact at the bid, $E_m \left[ p_{m,t+1}^X - \xi_{t+1}^X \right]$, and offer, $E_m \left[ p_{m,t+1}^X + \xi_{t+1}^X \right]$, in the subsequent period, which nets out to an expected repurchase price of $E_m \left[ p_{m,t+1}^X \right]$. We make the following assumption.
Assumption 2.5

Under the subjective assessment of the market maker, the fair price of each asset is a martingale: $E_m [p_{m,t+1}^X] = p_{m,t}^X$, $X \in \{F, S\}$.

As uninformed market participants, Assumption 2.5 is natural for the market makers. We denote the realized returns and their first two moments as

\begin{align*}
    r_{m,t+1}^X &= p_{m,t}^X + \xi_t^X - p_{m,t+1}^X \quad \text{(10a)} \\
    \mu_{m,t+1}^X &= p_{m,t}^X + \xi_t^X - E_m [p_{m,t+1}^X] = \xi_t^X \quad \text{(10b)} \\
    \sigma_{m,t+1}^{2,X} &= \text{Var}_m (r_{m,t+1}^X) \quad \text{(10c)}
\end{align*}

\begin{align*}
    r_{m,t+\tau}^F &= p_{m,t}^F + \xi_t^F - p_{m,t+\tau}^F - \xi_{t+\tau}^F \quad \text{(10d)} \\
    \mu_{m,t+\tau}^F &= E_m [r_{m,t+\tau}^F] = p_{m,t}^F + \xi_t^F - E_m [p_{m,t+\tau}^F + \xi_{t+\tau}^F] \quad \text{(10e)} \\
    \sigma_{m,t+\tau}^{2,F} &= \text{Var}_m (r_{m,t+\tau}^F) \quad \text{(10f)}
\end{align*}

Conditional on a passive sale in each market, makers oversee a control problem, which involves the choice of bid/offer spread and a portfolio of repurchases. As the two problems are symmetric, we focus on the case of a sale in the market for $F$. At the moment of a sale, the market maker is filled at an exogenously determined quantity, $q_{F,i}^F$, chosen by the investor. In addition to her choice of bid/offer spread, $\xi_t^F$, she also chooses an optimal portfolio of repurchases via the variable $q_{m,t+\tau}^S$. The resulting optimality problem is:

\begin{align*}
    \max_{q_{m,t+\tau}^S} & \quad \mu_{m,t+1,t+\tau}^F \cdot q_{m,t+1,t+\tau}^F - \frac{1}{2} \gamma \sigma_{m,t+1,t+\tau}^{2,F} \quad \text{(11a)} \\
\text{subject to} & \quad \mu_{m,t+1,t+\tau}^F = \xi_t^F (q_{F,i}^F - q_{m,t+\tau}^S) + (p_{m,t}^F + \xi_t^F - E_m [p_{m,t+\tau}^S + \xi_{t+\tau}^S]) q_{m,t+\tau}^S \quad \text{(11b)} \\
    & \quad \sigma_{m,t,t+\tau}^F = ((q_{F,i}^F - q_{m,t+\tau}^S) \sigma_{m,t+1}^F + (q_{m,t+\tau}^S \sigma_{m,t+\tau}^F)^2) \\
    & \quad \quad \quad + (q_{F,i}^F - q_{m,t+\tau}^S) q_{m,t+\tau}^S \sigma_{m,t+1}^F + (q_{m,t+\tau}^S \sigma_{m,t+\tau}^F) \quad \text{(11c)}
\end{align*}

Equation (11b) says after a passive fill at the $F$ offer, the market maker lays off her risk by retaining $q_{F,i}^F - q_{m,t}^S$ shares of $F$ in her own market at time $t + 1$ and purchasing $q_{m,t}^S$ shares of $S$ at time $t + \tau$. Her motive for laying off risk in the $S$ market is that the anticipated cross-market return, $p_{m,t}^F + \xi_t^F - E_m [p_{m,t+\tau}^S + \xi_{t+\tau}^S]$, may be larger than the anticipated
within-market return, $\xi_t^F$. The cross-market hedge can be thought of as latency arbitrage profits earned by high-frequency market makers. In addition to the spreads they choose in the separate markets, it is part of their total compensation package.

The first-order conditions of the market maker’s problem are:

$$p_{m,t}^F - \mathbb{E}_m \left[p_{m,t+\tau}^S + \xi_{t+\tau}^S + \frac{\partial p_{m,t+\tau}^S}{\partial q_{m,t+\tau}^S} q_{m,t+\tau}^S\right] + \gamma \left(q_{i,t}^F - q_{m,t+\tau}^S\right) \sigma_{m,t+\tau}^2$$

$$- \gamma q_{m,t+\tau}^S \sigma_{m,t+\tau}^2 - \gamma \left(q_{i,t}^F - 2q_{m,t+\tau}^S\right) \sigma_{m,t+\tau}^2 = 0$$  \hspace{2cm} (12)

$$q_{i,t}^F + \xi_{t}^F \frac{\partial q_{i,t}^F}{\partial \xi_t^F} - \gamma \left(q_{i,t}^F - q_{m,t+\tau}^S\right) \sigma_{m,t+\tau}^2 \frac{\partial q_{i,t}^F}{\partial \xi_t^F}$$

$$- \gamma q_{m,t+\tau}^S \sigma_{m,t+\tau}^2 = 0,$$  \hspace{2cm} (13)

where we have allowed the $S$ maker’s fair price at $t+\tau$, $p_{m,t+\tau}^S$, to be sensitive to the $F$ maker’s order flow, $q_{m,t+\tau}^S$, since the $F$ maker takes the role of an aggressive investor. Solving for the control variables in Equations (12) and (13):

$$q_{m,t+\tau}^S = \left(p_{m,t}^F - \mathbb{E}_m \left[p_{m,t+\tau}^S + \xi_{t+\tau}^S\right] + \gamma q_{i,t}^F \sigma_{m,t+1}^F \left(\sigma_{m,t+1}^F - \sigma_{m,t+\tau}^S\right)\right) D^{-1}$$  \hspace{2cm} (14)

$$\xi_t^F = \gamma q_{i,t}^F \sigma_{m,t+1}^F - \gamma q_{m,t+\tau}^S \left(\sigma_{m,t+1}^F - \sigma_{m,t+\tau}^S\right) - q_{i,t}^F \left(\frac{\partial q_{i,t}^F}{\partial \xi_t^F}\right)^{-1}$$  \hspace{2cm} (15)

where

$$D = \gamma \left(\sigma_{m,t+1}^F - \sigma_{m,t+\tau}^S\right)^2 + \frac{\partial p_{m,t+\tau}^S}{\partial q_{m,t+\tau}^S}.$$

Substituting Equation (14) into (15):

$$\xi_t^F = q_{i,t}^F \left(\gamma \sigma_{m,t+1}^F - \gamma^2 \sigma_{m,t+1}^F \left(\sigma_{m,t+1}^F - \sigma_{m,t+\tau}^S\right)^2 D^{-1} - \left(\frac{\partial q_{i,t}^F}{\partial \xi_t^F}\right)^{-1}\right)$$

$$- \gamma \left(p_{m,t}^F - \mathbb{E}_m \left[p_{m,t+\tau}^S + \xi_{t+\tau}^S\right]\right) \sigma_{m,t+1}^F \left(\sigma_{m,t+1}^F - \sigma_{m,t+\tau}^S\right) D^{-1}.$$  \hspace{2cm} (17)

Equation (14) shows that the $F$ market maker’s optimal cross-market arbitrage quantity is determined by three primary components: (1) size of the expected arbitrage profit, (2) $S$ market price elasticity to the cross-market arbitrage order size and (3) the relative volatilities in the two markets. The arbitrage profit is captured by the first term, $p_{m,t}^F - \mathbb{E}_m \left[p_{m,t+\tau}^S + \xi_{t+\tau}^S\right]$, which represents the difference between the expected repurchase prices in the two
markets. The price elasticity, which according to Assumption 2.1 is positive, enters in the denominator, and highlights how increasing market impact in the $S$ market (high elasticity) reduces the incentive of the $F$ market maker to route orders to that market. The relative market volatilities have several interesting effects. First, note from the profit condition that if $\sigma_{F,m,t+1} > \sigma_{S,m,t+\tau}$, the arbitrage quantity, $q_{S,m,t+\tau}$, is only positive if the expected $S$ offer price at $t+\tau$ is below the expected $F$ mid price at $t$. Further, the only term in the denominator, which influences the size, not the sign, of the order, is the price elasticity. When $\sigma_{F,m,t+1} > \sigma_{S,m,t+\tau}$, however, the second term of Equation (14) has a positive effect, which weakens the profit requirements of the $F$ market maker: the arbitrage order quantity can be positive even when the expected $S$ offer price is above the expected $F$ mid price. At the same time, the denominator increases because of the additional positive term related to volatility. The net effect on $q_{S,m,t+\tau}$ may be positive or negative. The higher relative volatility in the $F$ market increases the potential upside gains of expected arbitrage profits, but also increases the potential downside losses. Clearly, risk aversion, captured by $\gamma$, plays an important role in determining the net effect. When $\sigma_{m,t+1} < \sigma_{m,t+\tau}$, the effect on $q_{m,t+\tau}$ (relative to the equal volatility case) is unambiguously negative: the second term in Equation (14) is negative, which effectively states that the $F$ maker demands a more strict profit condition, and the first term of the denominator also increases. This latter effect is natural, as the increased $S$ volatility, and the maker’s risk aversion to it, decrease the utility from potential arbitrage gains. Additionally, we note that $q_{m,t+\tau}$ is increasing in investor order flow to the $F$ market maker, $q_{F,i,t}$, and that when $\sigma_{F,m,t+1} = \sigma_{S,m,t+\tau}$, $q_{S,m,t+\tau} \to \infty$ as $\partial p_{m,t+\tau}/\partial q_{S,m,t+\tau} \to 0$. The latter statement highlights the fact that when $S$ market impact of the $F$ maker’s aggressive order flow diminishes to zero, her desire to route arbitrage orders grows unboundedly. In practice, since $q_{m,t+\tau} \leq q_{i,t}$, there is a small enough price elasticity to cause the market maker’s order flow to achieve the constraint.

Equation (17) highlights the mechanisms that determine the market maker’s optimal spread. First, the term multiplying $q_{i,t}$ is always positive, since the spread elasticity of investor demand is negative, according to Assumption 2.2. The result is that investor demand induces the market maker to increase the $F$ spread, all else equal. The magnitude of the effect is positively related to $\partial p_{m,t+\tau}/\partial q_{m,t+\tau}$ (through the denominator $D$), and negatively
related to $\partial q_{i,t}^F / \partial \xi_t^F$. Intuitively, the market maker’s spread response to increased investor demand increases with her market impact in the $S$ and decreases with investor sensitivity to spread. The second term of the equation shows that the relationship of $\xi_t^F$ to $\xi_t^S$ depends on the relative market volatilities: it is positive when $\sigma_{m,t+1}^F > \sigma_{m,t+\tau}^S$, and negative when $\sigma_{m,t+1}^F < \sigma_{m,t+\tau}^S$. This captures the market maker’s desire to both increase arbitrage profits (the positive relationship) while also reducing portfolio volatility (the potentially negative relationship). As with the sensitivity to investor demand, the magnitude of the spread relationship is determined by the risk aversion parameter and the relative market volatilities.

To reconcile the empirical relationship of spreads in the $F$ and $S$ markets, we make the following assumption and proposition.

**Assumption 2.6**

The spread elasticity of investor $F$ demand is smaller (in magnitude) than that of $S$ demand:

$$\left| \frac{\partial q_{i,t}^F}{\partial \xi_t^F} \right| < \left| \frac{\partial q_{i,t}^S}{\partial \xi_t^S} \right|.$$  

**Proposition 2.7**

Conditional on constant and equal market volatilities

$$\sigma_{m,t}^F = \sigma_{m,t+1}^F = \sigma_{m,t+\tau}^S = \sigma_{m,t}^S,$$

and conditional on equal investor order flow arriving at each market, $q_{i,t}^F = q_{i,t}^S$, the spread for $F$ is greater than that of $S$: $\xi_t^F > \xi_t^S$.

**Proof** We denote the common, constant market volatility as $\sigma_{m,t}$ and the common investor flow as $q_{i,t}$. According to Equation (15), and exploiting the symmetry of the $S$ market maker control problem, the equilibrium spreads are

$$\xi_t^F = \gamma q_{i,t} \sigma_{m,t}^2 - q_{i,t} \left( \frac{\partial q_{i,t}}{\partial \xi_t^F} \right)^{-1},$$

$$\xi_t^S = \gamma q_{i,t} \sigma_{m,t}^2 - q_{i,t} \left( \frac{\partial q_{i,t}}{\partial \xi_t^S} \right)^{-1}.$$  

The result directly follows from Assumption 2.6. □

Assumption 2.6 states that investor order flow in the $F$ market is less sensitive to changes in the bid/offer spread than investor order flow in the $S$ market. Such an assumption is reasonable from a practical perspective, as it captures incentives, such as increased leverage,
for investors to direct informed flow to the futures market, which we have not included in the model. The result is that under a reasonable assumption of equal or similar market volatilities in the two assets, the equilibrium spread for $F$ will be larger than that of $S$. This is observed in practice (see Section 4.1).

Combining Assumption 2.6 with the preceding model results, we see that the relationship of equilibrium market spreads is determined by three primary mechanisms:

1. Expected arbitrage profits (positive effect).

2. A desire to attract more investor order flow (negative effect, through the quantity elasticity).

3. Non-modeled effects, such as leverage ($\xi^F > \xi^S$).

We conclude by summarizing the important implications of our model in a proposition. The proposition is stated only for the case of equal market volatilities, but a similar version holds when volatilities are not equal, and the market maker assessment of cross-market arbitrage profits includes a risk adjustment (the second term of Equation (14)).

**Proposition 2.8**

*Conditional on constant and equal market volatilities*

\[
\sigma^F_{m,t} = \sigma^F_{m,t+1} = \sigma^S_{m,t+\tau} = \sigma^S_{m,t},
\]

(21)

the following statements hold:

1. *When the $F$ market maker passively transacts at the $F$ offer at time $t$, she will seek to aggressively hedge herself by transacting some quantity at the $S$ offer at time $t + \tau$ if the expected $S$ offer price at $t + \tau$ is below the $F$ mid price at $t$: $\mathbb{E} \left[ p^S_{m,t+\tau} + \xi^S_{t+\tau} \right] < p^F_{m,t}$.*

2. *When the $F$ market maker passively transacts at the $F$ bid at time $t$, she will seek to aggressively hedge herself by transacting some quantity at the $S$ bid at time $t + \tau$ if the expected $S$ bid price at $t + \tau$ is above the $F$ mid price at $t$: $\mathbb{E} \left[ p^S_{m,t+\tau} - \xi^S_{t+\tau} \right] > p^F_{m,t}$.*

3. *When the $S$ market maker passively transacts at the $S$ offer at time $t$, she will seek to aggressively hedge herself by transacting some quantity at the $F$ offer at time $t + \tau$ if the expected $F$ offer price at $t + \tau$ is below the $S$ mid price at $t$: $\mathbb{E} \left[ p^F_{m,t+\tau} + \xi^F_{t+\tau} \right] < p^S_{m,t+1}$.*
4. When the $S$ market maker passively transacts at the $S$ bid at time $t$, she will seek to aggressively hedge herself by transacting some quantity at the $F$ bid at time $t + \tau$ if the expected $F$ bid price at $t + \tau$ is above the $S$ mid price at $t$: $E\left[p_{m,t+\tau}^S - \xi_t^{F}\right] > p_{m,t+1}^S$.

**Proof** When $F$ and $S$ market volatilities are equal, Equation (14) and its analog for the $S$ market maker, reduce to

$$q_{m,t+\tau}^S = \left(p_{m,t}^F - E_m\left[p_{m,t+\tau}^S + \xi_t^{S}\right]\right)\left(\frac{\partial p_{m,t+\tau}^S}{\partial q_{m,t+\tau}^S}\right)^{-1}$$

(22)

$$q_{m,t+\tau}^F = \left(p_{m,t}^S - E_m\left[p_{m,t+\tau}^F + \xi_t^{F}\right]\right)\left(\frac{\partial p_{m,t+\tau}^F}{\partial q_{m,t+\tau}^F}\right)^{-1}.$$  (23)

By inspection, we observe that $q_{m,t+\tau}^S$ and $q_{m,t+\tau}^F$ are only positive when the conditions of the proposition are met. 

In the foregoing development of the model, we have neglected the treatment of rebates and transactions costs. Since fee structures vary widely across exchanges and markets, it is difficult to make uniform statements as to how they would affect our stated results. In the markets used for our analysis in Section 4.3, both market makers and aggressive takers pays fees to transact futures contracts at the CME, while market makers receive rebates and aggressive takers pay fees to trade equities at the Nasdaq. Let $\phi^F$ denote the fee to transact in the futures market and $\phi^S$ and $\mu^S$ the taker fee and maker rebates in the spot market, respectively. For the investor, $-2\phi^F$ and $-2\phi^S$ are added to numerators of Equations (8) and (9), respectively. For the $F$ maker, who pays fees to transact in both markets, $-\phi^F - \phi^S$ is added to the numerator of Equation (14), which serves to tighten the arbitrage condition: the $F$ market maker will demand an even lower offer price in the $S$ market in order to exploit a potential cross-market arbitrage. On the other hand, for the $S$ market maker, $\mu^S - \phi^F$ is added to the numerator of Equation (14), since the $S$ maker earns a rebate for passively providing liquidity in the $S$ market and pays a fee to exploit an arbitrage profit in the $F$ market. Depending on the net effect ($\mu^S > \phi^F$ or $\mu^S < \phi^F$) the fees may serve to relax or tighten the required difference in prices across market to make a profitable arbitrage. For the CME and Nasdaq in particular, this difference is likely to be positive, where the fee to trade a single ES futures contract is $0.25$ (for the most active market makers, who have access
to lowest latency communications technology) and the rebate to trade 500 SPY shares (the equivalent of one ES contract – see Section 4) is roughly $1.05. The net difference, $0.80, however, is almost five times smaller than the difference between the two half spreads, $3.75, and so we do not deal with it here.

3 Econometric Methodology

To test the implications of our model, we need an econometric measure of information flow and price responsiveness across markets. Hasbrouck (1995) introduced an econometric methodology which has become the backbone for measuring information shares of distinct markets trading a single asset. Formally, given \( N \) markets and an \( N \times 1 \) vector of prices in those markets, \( \mathbf{p}_t \), Hasbrouck (1995) assumes that the prices are individually nonstationary and that the vector is cointegrated of order \( N - 1 \). Expressing the cointegrated system with the common trends representation of Stock and Watson (1988),

\[
\mathbf{p}_t = \mathbf{p}_0 + \mathbf{\psi} \left( \sum_{s=1}^{t} \mathbf{e}_s \right) \mathbf{\iota} + \mathbf{\Psi}^* (L) \mathbf{e}_t,
\]

Hasbrouck (1995) defines the information share of market \( j \) to be

\[
S_j = \frac{\psi_{jj}^2 \Omega_{jj}}{\mathbf{\psi} \Omega \mathbf{\psi}^*},
\]

where \( \mathbf{e}_t \) is a vector of price innovations to each of the markets, \( \sum_{s=1}^{t} \mathbf{e}_s \) is the common random walk component to prices, \( \mathbf{\psi} \) is a row vector which measures the long-run impact of innovations \( \mathbf{e}_t \) on prices (which is taken from the moving average representation of price differences), and \( \Omega \) is the covariance matrix of innovations. In essence, this measure of information share is variance decomposition: Equation (25) equates the share of information in market \( j \) to the total variance contribution of its price innovation. While Equation (25) assumes a diagonal covariance matrix, \( \Omega \), Hasbrouck (1995) also computes bounds on the variance decomposition by using a Cholesky factorization for non-diagonal covariance matrix.

The aforementioned measure of information share has been widely adopted because of its elegance of interpretation and its parsimonious representation as a cointegrated time series model. However, it suffers from several weaknesses. First, to maintain synchronicity across
elements of the price vector, prices must be observed at identical times. In practice, this means that they must be observed in clock time. Brada et al. (1966), Mandelbrot and Taylor (1967), Clark (1973), and more recently Ane and Geman (2000) and Aldrich et al. (2015) all demonstrate the advantages of expressing prices at granular time intervals using some measure of (subordinated) transaction time. In particular, Aldrich et al. (2015) highlights the importance of expressing price evolution through a dynamic model of transaction arrival. Second, it requires a parametric, linear time series model and a set of assumed cointegrating relationships. Third, as high-frequency time series data becomes increasingly rich, the time series model becomes increasingly large and complex. To deal with overparameterization in the model, it is necessary to impose parameter restrictions, such as the polynomial smoothing of coefficients performed by Hasbrouck (1995). Finally, the methodology only utilizes price information and neglects a second piece of information that is attributed to each trade and quote: size. Incorporating volume and order flow (defined as the difference between volume done on the best offer and volume done on the best bid) is essential to understanding total market share of information.

We propose a measure of information flow that is non-parametric and designed explicitly for high-frequency, transaction-time data. It is also well suited to measuring the model implications of Proposition 2.8. This econometric measure correlates events (e.g., transactions or changes in displayed liquidity) in two markets and measures deviations from a null hypothesis of the events in one market, under specific conditions, having no informative impact on concurrent or subsequent activity in another market. The measure inherently accounts for volume and order flow and also flexibly accounts for communication latency between markets.

Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be the sets of all possible events that can occur in a conditioning market, $X$, and a responding market, $Y$, at any time $t$. Suppose we are interested in understanding how the events in a subset $\mathcal{X}^* \subseteq \mathcal{X}$ affect events in a subset $\mathcal{Y}^* \subseteq \mathcal{Y}$. To do so, we define,

$$\delta_s = \mathbb{E}(y_{t+s}|x_t) - \mathbb{E}(y_{t+s}), \text{ for } s = 0, 1, \ldots, S;$$

for $x_t \in \mathcal{X}^*$ and $y_{t+s} \in \mathcal{Y}^*$ for $s = 0, 1, \ldots, S$. That is, $\{\delta_s\}_{s=0}^{S}$ simply measures the
conditional effect of events in $\mathcal{X}^*$ on events in $\mathcal{Y}^*$ over horizon $S$. We estimate $\{\delta_s\}_{s=0}^S$ via simple frequency counts:

$$
\hat{\delta}_s = \frac{1}{N_{x^*}} \sum_{i=1}^{N_{x^*}} \mathbb{1}(y_{t_i+s}|x_{t_i}) - \mathbb{E}(y_{t+s}),
$$

(27)

where $N_{x^*}$ is the number of events in $\mathcal{X}^*$ that occurs during the period of interest and where we have assumed that the unconditional expectation of $y_{t+s}$ is known a priori. In the case that the unconditional expectation of $y_{t+s}$ is unknown, it may also be estimated through a similar frequency count, by sampling an identical number, $N_{x^*}$, of events during the period of interest. The estimator of Equation (27) amounts to computing the difference of two histograms.

In Section 4, we will utilize the estimator $\{\hat{\delta}_s\}_{s=0}^S$ in the following manner. We will separately consider two cases for the originating market, $X \in \{ES, SPY\}$, with the responding market, $Y$, being the other. Given a choice of $X$, we will focus on distinct subsets of market events, $\mathcal{X}^*$: all possible transactions (sizes) on one side of the order book (bid or offer), where the transacted price at time $t$ is either above or below the displayed quotation on the same side of the order book in the responding market at time $t+s$, for $s \leq S$. For example, with $X = ES$, one of the four subsets would be transactions on the ES bid, when that bid price is below the quoted bid in the SPY market 7 ms later. Note that as we have defined it, the number of events in the conditioning set, $N_{x^*}$, is dependent on lag $s$. In this case, $\{x_{ti}\}_{i=1}^{N_{x^*}}$ represents the observed bid transactions in the ES during a single day that abide by the aforementioned restriction. The result is eight different collections of estimators: $\{\delta_s^{X,p,d}\}_{s=0}^S$, where $X \in \{ES, SPY\}$, $p \in \{b, o\}$ (bid or offer) and $d \in \{\downarrow, \uparrow\}$ (below or above). $\mathcal{Y}^*$ will represent the set of all possible values of order flow, or differences (in size) of transactions done on the offer and bid, in the responding market per unit of volume in the originating market, and $\{\{y_{ti+s}\}_{i=1}^{N_{x^*}}\}_{s=0}^S$ will represent the observed differences in sizes (per unit $\{x_{ti}\}_{i=1}^{N_{x^*}}$) following each observed event in the originating market. We make the assumption that $\mathbb{E}[y_t] = 0 \forall t$, which states that unconditional expectation of transacted size on the offer is equal to that on the bid. A simple empirical check has shown this to be a very reasonable assumption.

We maintain the following null hypothesis:
Hypothesis 3.1

Events in the originating market do not informatively affect the distribution of subsequent events in the responding market: \( \delta_{s}^{X,p,d} = 0 \) for all possible choices of \((X, p, d)\).

As discussed in Section 1, conventional wisdom suggests that events in the ES market subsequently inform events in the SPY market. Thus, a naive assumption would be that \( \delta_{s}^{ES} \neq 0 \) for some \( s \in \{0, \ldots, S\} \), but that \( \delta_{s}^{SPY} = 0 \ \forall s \) (note that we have dropped the superscripts \( b \) and \( d \) since the stated assumption would hold for all values). Proposition 2.8, however, provides a set of testable predictions, which are sometimes incongruent with the stated conventional wisdom. We consolidate these testable predictions in the following proposition.

Proposition 3.2

1. \(-1 \times \delta_{s}^{ES,b,\downarrow} > 0\) for some \( s \).

2. \( \delta_{s}^{ES,b,\uparrow} = 0 \) for all \( s \).

3. \( \delta_{s}^{ES,o,\uparrow} > 0 \) for some \( s \).

4. \( \delta_{s}^{ES,o,\downarrow} = 0 \) for all \( s \).

5. \(-1 \times \delta_{s}^{SPY,b,\downarrow} > 0\) for some \( s \).

6. \( \delta_{s}^{SPY,b,\uparrow} = 0 \) for all \( s \).

7. \( \delta_{s}^{SPY,o,\uparrow} > 0 \) for some \( s \).

8. \( \delta_{s}^{SPY,o,\downarrow} = 0 \) for all \( s \).

Note that in two cases we have reversed the sign of the estimator in order to consistently focus on positive deviations from the null hypothesis. This will assist visualization and aggregation of our empirical results in Section 4.

Figure 1 displays a stylized representation of \( \delta_{s} \) in statements 1,3,5 and 7 of Proposition 3.2. If the figure conformed exactly to the model in Section 2, \( \delta_{s} \) would be a Dirac function at \( s = \tau \), the inter-market communication latency, with height \( q_{m}^{X} \), the number of shares that the market maker chooses to route to the responding market. Instead, we have...
depicted the response function as we might expect it to appear in a market with a heterogeneous group of market makers who have access to different communication technology and different ability to assess and compute the relative information across markets. In the next section, we will empirically estimate $\delta_s$ for each of the subcases considered in Proposition 3.2.

Figure 1: Stylized representation of information response measure.

4 Empirical Application

We now apply the econometric methodology of the last section to empirically test the model implications of Section 2. We begin by describing our data and the necessary adjustments we make in order to equate the prices of futures and equity assets.
4.1 Data

Our data comprise all transactions for the E-mini S&P 500 futures contract (CME Group symbol ES, commonly known as the E-mini) and the State Street Global Advisers SPDR S&P 500 exchange traded fund (NYSE Arca symbol SPY) between Jun 16, 2014 and Sep 11, 2014. The E-mini trades exclusively at the CME Group Globex matching engine in Aurora, Illinois, longitude -88.24° W, latitude 41.80° N. The market consists of five listed contracts at all times, expiring on the March quarterly cycle, with expiry occurring on the third Friday of the designated month. The bulk of trading interest resides in the near-month contract, although common practice dictates that interest shifts to the second-month contract exactly one week prior to expiry. This transition, on the second Friday of the expiry month, is known as the roll date. The sample period for our data was selected to coincide with all trading days between the two roll dates for the September 2014 ES contract (symbol ESU4). The E-mini market is open nearly continuously each week from Sunday, 6:00 p.m. ET to Friday, 5:00 p.m. ET, aside from a daily trading halt from 4:15 – 4:30 p.m. ET and a daily maintenance period from 5:00 – 6:00 p.m. ET. Although trading occurs during all market hours, the majority of activity coincides with U.S. equities market hours, 9:30 a.m. – 4:00 p.m. ET, Monday through Friday. For this reason, and because we are interested in correlating with activity in the equities market, we restrict attention to U.S. equities market hours. The contract is quoted in S&P 500 index points, with a minimum spread of 0.25 index points, although the notional value is $50 \times$ the index. Panel (a) of Table 1 reports basic price and transaction size summary statistics for the period we consider; for example, the average price of the contract was 1959.40 index points, corresponding to a notional value of $97,970.00.

The SPY exchange traded fund (ETF) is listed with NYSE Arca and trades on U.S. equities exchanges. We confine attention to direct feed data obtained from the Nasdaq matching engine in Carteret, New Jersey (longitude -74.25° W, latitude 40.58° N). While it would be desirable to obtain a transaction record for SPY activity across all exchanges (such as from the consolidated tape), we emphasize the importance of using direct feed data, which has more accurate time stamps and which does not include an additional reporting delay to the Securities Information Processor (SIP). Consolidated tape data (e.g. NYSE
(a) Price and Size Summary Statistics

<table>
<thead>
<tr>
<th>Asset</th>
<th># Trades</th>
<th>$P_{\text{min}}$</th>
<th>$P_{\text{max}}$</th>
<th>$\bar{P}$</th>
<th>$S_{\text{min}}$</th>
<th>$S_{\text{max}}$</th>
<th>$\bar{S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>16,448,563</td>
<td>1899.75</td>
<td>2010.00</td>
<td>1959.40</td>
<td>1</td>
<td>1792</td>
<td>3.826</td>
</tr>
<tr>
<td>SPY</td>
<td>2,781,812</td>
<td>190.55</td>
<td>201.58</td>
<td>196.43</td>
<td>1</td>
<td>92,600</td>
<td>309.57</td>
</tr>
</tbody>
</table>

(b) Bid/Offer Summary Statistics

<table>
<thead>
<tr>
<th>Asset</th>
<th># Bid Trades</th>
<th>$P_{b\text{below}}$</th>
<th>$P_{b\text{above}}$</th>
<th># Offer Trades</th>
<th>$P_{a\text{below}}$</th>
<th>$P_{a\text{above}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>1,829,316</td>
<td>1,159,262</td>
<td>670,054</td>
<td>1,815,467</td>
<td>692,672</td>
<td>1,122,795</td>
</tr>
<tr>
<td>SPY</td>
<td>615,919</td>
<td>137,516</td>
<td>478,403</td>
<td>517,517</td>
<td>411,338</td>
<td>106,179</td>
</tr>
</tbody>
</table>

c) Integrated Order Flow Response (units = SPY shares)

<table>
<thead>
<tr>
<th>Market</th>
<th>$s = 0$</th>
<th>$s = 5$</th>
<th>$s = 10$</th>
<th>$s = 15$</th>
<th>$s = 20$</th>
<th>$s = 25$</th>
<th>$s = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>0.04564</td>
<td>4.811</td>
<td>9.532</td>
<td>10.83</td>
<td>11.88</td>
<td>12.72</td>
<td>13.21</td>
</tr>
<tr>
<td>SPY</td>
<td>1.888</td>
<td>89.01</td>
<td>640.1</td>
<td>1135</td>
<td>1260</td>
<td>1335</td>
<td>1397</td>
</tr>
</tbody>
</table>

d) Integrated Order Flow Response Value (dollars)

<table>
<thead>
<tr>
<th>Market</th>
<th>$s = 0$</th>
<th>$s = 5$</th>
<th>$s = 10$</th>
<th>$s = 15$</th>
<th>$s = 20$</th>
<th>$s = 25$</th>
<th>$s = 30$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES</td>
<td>633.31</td>
<td>66,760.18</td>
<td>132,260.25</td>
<td>150,301.39</td>
<td>164,804.12</td>
<td>176,503.71</td>
<td>183,258.99</td>
</tr>
<tr>
<td>SPY</td>
<td>230.71</td>
<td>10,877.46</td>
<td>78,223.08</td>
<td>138,747.88</td>
<td>153,994.15</td>
<td>163,184.48</td>
<td>170,688.22</td>
</tr>
</tbody>
</table>

Table 1: Panel (a): Summary statistics for ES and SPY transaction prices ($P$) and sizes ($S$) between Jun 16 and Sep 11, 2014. Panel (b): Summary statistics for ES and SPY bid/offer transaction prices relative to bid/offer quotes in the other market. Panel (c): Integrated order flow responses for bids and offers in each originating market. Panel (d): Valuation of integrated order flow responses for bids and offers in each originating market.

Daily TAQ) includes the SIP delay and would introduce ambiguity in our cross-market latency analysis. Table 2 lists the average market share (and standard deviation) of SPY transactions conducted on each equities exchange for all trading days in Apr, 2014, the most recent month for which we have consolidated tape data. The table shows that Nasdaq represents nearly 25% of all SPY market volume and is the largest exchange in terms of SPY market share. As noted above, equities markets are open from 9:30 a.m. – 4:00 p.m., Monday through Friday. Although some before- and after-market trading occurs for liquid equities, the vast majority of trading occurs during regular market hours, and for this reason we confine attention to those hours. The SPY ETF is priced at 1/10th the value of the...
S&P 500 index, and per Rule 612 of Reg NMS ("the rule"), market makers are prohibited from displaying or accepting quotations priced in an increment smaller than $0.01. Panel (a) of Table 1 reports basic price and transaction size summary statistics for SPY during our sample period.

<table>
<thead>
<tr>
<th>Exchange</th>
<th>Mean</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nasdaq</td>
<td>0.234</td>
<td>0.031</td>
</tr>
<tr>
<td>NYSE Arca</td>
<td>0.210736</td>
<td>0.018592</td>
</tr>
<tr>
<td>FINRA</td>
<td>0.163565</td>
<td>0.015833</td>
</tr>
<tr>
<td>BATS</td>
<td>0.157954</td>
<td>0.013971</td>
</tr>
<tr>
<td>DirectEdge X</td>
<td>0.061514</td>
<td>0.005623</td>
</tr>
<tr>
<td>BATS Y</td>
<td>0.056861</td>
<td>0.006155</td>
</tr>
<tr>
<td>Nasdaq BX</td>
<td>0.054256</td>
<td>0.025917</td>
</tr>
<tr>
<td>DirectEdge A</td>
<td>0.043766</td>
<td>0.004578</td>
</tr>
<tr>
<td>Nasdaq PSX</td>
<td>0.013788</td>
<td>0.001938</td>
</tr>
<tr>
<td>NSX</td>
<td>0.001936</td>
<td>0.000586</td>
</tr>
<tr>
<td>CSE</td>
<td>0.001307</td>
<td>0.000401</td>
</tr>
<tr>
<td>CBOE</td>
<td>0.000022</td>
<td>0.000001</td>
</tr>
</tbody>
</table>

Table 2: Market share (by number of transactions) for all participating equities exchanges during Apr, 2014.

Given that the E-mini and SPY ETF trade the same underlying quantity, it is expected that their markets are tightly linked. Laughlin et al. (2014) highlights the correlation of shifting liquidity from E-mini futures to equities markets, using an empirical measure similar to what we propose in Section 3. Their work further documents that microwave communication infrastructure has reduced the latency of information transmission to just over 4 milliseconds (ms), only slightly more than the 3.93 ms theoretical minimum to transmit messages in a frictionless environment on the great circle between Aurora and Carteret. An important distinction to make between the two assets is their relative notional values: since the E-mini
is valued at $50 \times$ the index and the SPY is valued at $1/10$th the index, one contract of the former is worth roughly 500 shares of the latter. Further, after accounting for the order of magnitude difference in price quotations between the two assets, the minimum increment of 0.25 index points for the E-mini is 2.5 times larger than the $0.01$, or 0.10 index point, minimum increment for the SPY. This latter relationship is of crucial importance to our empirical application, as it represents the channel of information transmission that we highlighted in Section 2.

4.2 Basis Adjustment

In order to empirically test the statements of Propositions 2.8 and 3.2, we need to adjust the prices of the ES and SPY contracts so that they are comparable. The standard futures pricing equation is

\[
F_t = e^{(r_f - \rho)(T-t)}S_t,
\]

where \( F_t \) is the futures price, \( S_t \) is the spot price of the underlying asset (in this case the S&P 500 index), \( r_f \) is the risk-free interest rate, \( \rho \) is the dividend rate and \( T - t \) is the time until expiry of the futures contract. Equation (28) shows that differences in futures and spot prices arise from stochastic variations in \( r_f \) and \( \rho \), as well as the deterministic movement of time, \( T - t \). The conventional way to quantify the wedge between prices is via the basis,

\[
b_t = F_t - S_t = \left( e^{(r_f - \rho)(T-t)} - 1 \right) S_t,
\]

which can be accurately estimated since the stochastic fluctuations in \( r_f \) and \( \rho \) are typically quite small over short time horizons. Rather than working in a forward fashion, estimating interest and dividend rates at a daily frequency, we introduce a methodology to infer what the basis must be from the empirical distribution functions of ES/SPY bid and offer differences. Figure 2 displays histograms of ES/SPY best bid differences (blue) and best offer differences (red) at the times of all transactions in the SPY market on Aug 4, 2014. The histogram bars are centered on the unique values of the differences (i.e. the bars do not aggregate values on the x-axis) and demonstrate that there are very few unique differences on Aug 4, 2014. This latter fact is owing to the price discreteness of both assets.
To infer the basis for a specific day, we must first understand the nature of adjusted prices. We begin with the following assumption.

**Assumption 4.1**

*On average, basis-adjusted bids and offers for the ES and SPY are symmetrically quoted around a latent, fair price for the S&P 500 index.*

Combining Assumption 4.1 with the fact that the ES price increment is 2.5 times larger than that of the SPY, we conclude that the SPY bid and offer will typically sit completely inside of the ES bid and offer. Panel (b) of Table 1 separates all ES and SPY transactions according to whether they were done on the bid or offer, and reports the total counts for which basis-adjusted (described below) bid/offer transactions in each market are above or below the best bid/offer quote in the other market. The data corroborate our conclusion: at the time of an ES transaction, roughly two-thirds of ES prices sit outside the best quotes in the SPY market and at the time of a SPY transaction, roughly 80% of SPY prices sit inside
the best quotes in the ES market. Although the values reported in Table 1 do not account for inter-market communication latency, we have computed them for time shifts of $\tau \leq 5$ ms and verified that the aggregate numbers change very little. We also note that the number of ES and SPY transactions in panel (b) do not equal the totals in panel (a); this is due to the fact that many transactions occur in the same millisecond (a result of order splitting) and because panel (b) only considers unique transaction times.

Figure 3 depicts several scenarios when the basis-adjusted ES bid and offer are quoted at 200.00 and 200.025. While our choice of price grid in the figure only allows the ES/SPY bids to perfectly align, a symmetric example exists in which the offers perfectly align (e.g. when the ES is quoted at 200.025 and 200.05). Assumption 4.1 states that Case 3 (and its symmetric analog) in the diagram is most common: with highest probability, the SPY bid is either $0.005$ or $0.01$ greater than the ES bid, and, respectively, the SPY offer is $0.01$ or $0.005$ less than the ES offer. These two scenarios correspond to the histogram bars labeled ‘1’ and ‘2’ in Figure 2. The next most probable cases (Case 2 in Figure 3 and its analog) occur when bids or offers align, which correspond to the histogram bars labeled ‘3’ and ‘4’. The central overlapping bars (labeled ‘3,4’) identify the basis: this is the price shift which causes the bids and offers to exactly align. For Aug 4, 2014, we conclude that the basis was -0.55. Repeating this procedure for each day in our dataset, we construct an inferred basis series, which is used to align prices for our econometric test. We view the resulting aligned prices as reasonable inputs for cross-market comparison since the intra-day basis market is very active for traders that participate in the ES and SPY markets.

In practice, a single basis does not exist since hedging across markets and products entails a net lending or borrowing position; a small interest rate spread applies to borrowers (traders that are short ES and long SPY) and lenders (traders that are long ES and short SPY). Although the spread is trader specific and dictated by the market for the basis, we take a commonly reported series of basis spreads (ranging from 0.065 on Jun 16, 2014 to 0.001 on Sep 11, 2014) from Bloomberg and further adjust ES prices upward to account for positions that are net borrowing (a passive fill on the ES offer and aggressive hedge on SPY offer or a passive fill on SPY bid and aggressive hedge on ES bid) and downward to account for net lending (a passive fill on the SPY offer and aggressive hedge on ES offer or a passive
<table>
<thead>
<tr>
<th>Case1</th>
<th>199.99</th>
<th>200.00</th>
<th>200.01</th>
<th>200.02</th>
<th>200.025</th>
<th>200.03</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ES^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case2</th>
<th>199.99</th>
<th>200.00</th>
<th>200.01</th>
<th>200.02</th>
<th>200.025</th>
<th>200.03</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ES^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case3</th>
<th>199.99</th>
<th>200.00</th>
<th>200.01</th>
<th>200.02</th>
<th>200.025</th>
<th>200.03</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ES^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case4</th>
<th>199.99</th>
<th>200.00</th>
<th>200.01</th>
<th>200.02</th>
<th>200.025</th>
<th>200.03</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ES^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$ES^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^b$</td>
<td>▲</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SPY^o$</td>
<td>▼</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 3:** Examples of adjusted relative ES and SPY prices.
fill on ES bid and aggressive hedge on SPY bid).

4.3 Results

To test the implications of our model, we use basis-adjusted prices to determine the relationship of ES/SPY bids and offers at the times of all transactions in each market. Since the data are not contiguous across days, we separately estimate the values of $\delta_{X,p,d}^{X,p,d, s} = \{\delta_{s}^{X,p,d}\}_{s=0}^{S}$ in Proposition 3.2 for each day in our dataset. Recall that $\delta_{s}^{X,p,d}$ represents the difference in number of shares transacted at the offer and bid in the responding market, per unit of volume transacted in the originating market (on bid or offer, depending on $p \in \{b, o\}$) when the prices in the two markets abide by restriction $d$ at time lag $s$. We emphasize the importance of comparing prices at distinct time lags, $s$: it is essential for estimation purposes to account for hypothetical communication latency when comparing bids and offers across markets. For example, conditional on a transaction at the ES bid at time $t$, $\delta_{s}^{ES,b, \downarrow}$ must be estimated only for SPY bid transactions at time $t + s$ when $P_{t}^{ES,b} < P_{t+s}^{SPY,b}$ and not necessarily when $P_{t}^{ES,b} < P_{t+s}^{SPY,b}$. In doing this, we do not rigidly assume a value for communication latency, $\tau$, but instead allow the communication latency to be observed in the econometric estimates. As noted in Section 4.1, we expect $\tau \approx 4$ ms, which would imply $\delta_{s}^{X,p,d} \approx 0$ for $s < 4$ ms in all cases outlined in Proposition 3.2.

We make two more refinements on the sets of conditioning events. In order to ensure that the estimated order flow responses, $\hat{\delta}_{s}^{X,p,d}$, are not partially attributed to information prior to their conditioning events, we only consider transactions in each market that are not preceded by transactions in the 7 ms immediately prior. This threshold is enough to ensure that all information from both markets has been impounded into the conditioning event price, and that the estimated responses are not an artifact of information that was available before the event. In addition, for the sets of SPY conditioning transactions, we only consider transactions of 500 shares or more. Since a single ES contract is equivalent to 500 SPY shares, a SPY market maker that is filled on bid or offer for less than 500 shares would have to reverse her net position in the ES market by taking advantage of a potential arbitrage opportunity. These two refinements reduce the total sample size from 1,829,316 ES events and 615,919 SPY events during our sample period (reported in panel (b) of Table 1).
to 1,145,481 and 112,874 events, respectively.

Figures 4 and 5 show daily estimates of the eight order flow response estimators listed in Proposition 3.2. Specifically panels (a), (b), (c) and (d) of Figure 4 show bid responses, $\delta^{ES,b,\downarrow}$, $\delta^{ES,b,\uparrow}$, $\delta^{SPY,b,\downarrow}$ and $\delta^{SPY,b,\uparrow}$, respectively, and panels (a), (b), (c) and (d) of Figure 5 show offer responses, $\delta^{ES,o,\uparrow}$, $\delta^{ES,o,\downarrow}$, $\delta^{SPY,o,\uparrow}$ and $\delta^{SPY,o,\downarrow}$, respectively. The bold, solid lines in each panel represent the median of daily response values and dotted lines correspond to 0.05 and 0.95 empirical quantiles of the same. All y-axis units have been scaled to units of SPY shares (i.e. ES responses have been scaled by 500). As described in Section 4, the order flow response estimators, $\delta^{X,p,d}$, are computed point by point, for each millisecond following a bid or offer transaction in an originating market. Specifically, the number of conditioning events $N_{x^*}$ is determined by the number of originating events, $N_x$, that satisfy specific relationships of relative bids and offers across markets, and changes for each millisecond time interval following an event in the originating market. In Figures 4 and 5 we have rescaled the all values by $N_{x^*}/N_x$, where $N_x$ is the common number of events in the originating market; e.g. ES bid events. Thus, prior to the rescaling, the order flow response for an ES bid transaction, when that bid is below a subsequent SPY bid, is interpreted as $\delta^{ES,b,\downarrow}_s = \mathbb{E}[SPY\text{ order flow}|ES^b_t < SPY^b_{t+s}]$, whereas after the rescaling it is interpreted as $\delta^{ES,b,\downarrow}_s = \mathbb{E}[SPY\text{ order flow}|ES^b_t < SPY^b_{t+s}] Pr(ES^b_t < SPY^b_{t+s})$. Given a symmetric interpretation of $\delta^{ES,b,\uparrow}_s$, the sum of the rescaled responses is interpreted as $\delta^{ES,b,\downarrow}_s + \delta^{ES,b,\uparrow}_s = \mathbb{E}[SPY\text{ order flow}|ES^b_t]$. The result is that the summation of response estimates in panels (a) and (b) or panels (c) and (d) of Figures 4 and 5 are estimates of order flow following bid or offer events in the origination markets, regardless of relative price conditions.

Panels (a) in both figures depict the expected, conventional response: when a trade occurs on the ES bid or offer, there is a subsequent preponderance of trading on the SPY bid and offer, respectively. Further, these panels clearly identify the cross-market communication latency to be approximately 4 or 5 ms, as the order flow response estimates do not appear to be statistically different from zero for $s \leq 3$ ms. Panels (b) in both figures, however, demonstrate that when a transaction occurs at the ES bid or offer under conditions that our model suggests are not profitable in terms of cross-market arbitrage ($P^{ES,b}_t \geq P^{SPY,b}_{t+s}$).
or $P_{i,t}^{ES,o} \leq P_{i,t+s}^{SPY,o}$) the statistical significance of the response disappears. This latter result directly contradicts conventional wisdom that information in the futures market precedes that of the spot market.

The lower rows of Figures 4 and 5 show the responses of the ES market to transactions on the bid and offer in SPY. Similar to panels (a), panels (c) depict a strong order flow response in the ES market when a trade occurs in the SPY market under favorable cross-market arbitrage conditions: $P_{SPY,b} < P_{ES,b}$ and $P_{SPY,o} > P_{SPY,o}$. Panels (d), however, show that when these conditions do not hold, the significance of the response disappears, as predicted by the model. Once again, while panels (d) are congruent with unconditional behavior across the ES/SPY market, panels (c) violate convention and strongly support our model.

An interesting feature of panels (c) is that the peak of the response occurs much later than that of panels (a) – closer to 10 ms than 4 ms. This is somewhat unexpected, as the microwave lines allow for full duplex transmission rates, both for the purpose of providing reverse information flow, as well as for the basic need to deliver fast message confirmations in the traditional, Aurora to New Jersey, direction. The delayed peak in the New Jersey to Aurora direction could arise because the bulk of traders exploiting the reverse cross-market arbitrage are less sophisticated, and are somewhat slower, or because they are more likely to use fiber optic infrastructure instead of microwave infrastructure. Despite this somewhat delayed response, it is clear that events in the SPY market have a significant impact on trading in the ES market when cross-market arbitrage conditions create profit opportunities for market makers.

The final important feature to note from Figures 4 and 5 is that the responses in panels (c) are almost two orders of magnitude larger than those of panel (a). This is partly related to the relative coarseness of ES/SPY contract sizes. When an ES market maker is passively filled on one or multiple contracts, the relative granularity of SPY share sizes gives her a high degree of flexibility in choosing her hedge quantity. In the reverse scenario, however, SPY market makers are forced to discretely round up or down to multiples of 500 SPY share equivalents. As a whole, the effect of relative contract size contributes to an overall larger impact of SPY events on subsequent ES trading. Thus, although the conditions under which
Figure 4: Order flow responses for ES and SPY bid transactions. Panel (a) shows SPY order flow response after ES bid event, when $P_{t}^{ES,b} \leq P_{t+s}^{SPY,b}$. Panel (b) shows the same response but for $P_{t}^{ES,b} \geq P_{t+s}^{SPY,b}$. Panels (c) and (d) display analogous responses of the ES (expressed in units of SPY shares) after SPY bid transactions for the (respective) cases of $P_{t}^{ES,b} \geq P_{t+s}^{SPY,b}$ and $P_{t}^{ES,b} \leq P_{t+s}^{SPY,b}$.

SPY events informatively lead ES transactions are relatively infrequent, they are individually more informative.
Figure 5: Order flow responses for ES and SPY offer transactions. Panel (a) shows SPY order flow response after an ES offer event, when $P_{t}^{ES,o} \geq P_{t+s}^{SPY,o}$. Panel (b) shows the same response but for $P_{t}^{ES,o} \leq P_{t+s}^{SPY,o}$. Panels (c) and (d) display analogous responses of the ES (expressed in units of SPY shares) after SPY offer transactions for the (respective) cases of $P_{t}^{ES,o} \leq P_{t+s}^{SPY,o}$ and $P_{t}^{ES,o} \geq P_{t+s}^{SPY,o}$.

We summarize the estimated order flow responses in panel (c) of Table 1 by integrating the curves generated by the medians depicted in Figures 4 and 5. Specifically, the integrated
order flow response for originating market $X \in \{ES, SPY\}$ is defined as

\[
\Delta^X_s = \int_0^s 0.5 \left( \delta^X_{s,b} + \delta^X_{b,s} \right) + 0.5 \left( \delta^X_{s,o} + \delta^X_{o,s} \right) ds
\]

(30)

Given our prior interpretation of $\delta^X_{s,b} + \delta^X_{b,s}$, $\Delta^X_s$ represents the total order flow response in the responding market over horizon $s$, for any transaction (bid or offer) in the originating market. The integrated response implicitly assumes that 50% of transactions in the originating market separately occur on bid and offer. We find this is a very accurate approximation.

To estimate $\Delta^ES_s$, we simply sum the median responses in panels (a) and (b) of Figure 4, average them with the sum of median responses in panels (a) and (b) of Figure 5 and integrate over horizon $s$. $\Delta^{SPY}_s$ is analogously estimated with panels (c) and (d) of Figures 4 and 5. Naturally, $\Delta^{ES}_s$ is reported in units of SPY shares, as it measures the SPY order flow (negative order flow in the case of bid transactions) at the Nasdaq exchange after an ES transaction at the CME. However, while the natural units of $\Delta^{SPY}_s$ would be ES contracts, we have scaled by 500 in order to maintain identical units across originating markets. The integrated responses exhibit the same order of magnitude effects that are depicted in the figures: SPY transactions have an effect that is up to 100 times larger than those of ES.

Further, the first row of panel (c) shows that 72% of the SPY response to ES occurs by 10 ms while the second row reports only 45% of the ES response to SPY occurs in the same time frame.

Panel (c) of Table 1 clearly demonstrates that SPY transactions have a relatively larger impact on the ES market than vice versa. However, given the relative frequency of cross-market arbitrage opportunities at the time of an ES transaction (and the relative paucity of opportunities at the time of a SPY transaction), we expect that the total profit from exploiting these discrepancies will be much more balanced. Panel (d) of Table 1 weights the integrated responses in panel (c) by the average number of transactions per day, and additionally scales by an approximated profit per opportunity. As reported in panel (a) of the table, there are $16,448,563 \times 3.826 = 62,932,202$ ES contracts traded during our 62-day sample period, resulting in an average $1,015,036$ contracts per day. For the SPY market, we subset just the transactions of 500 shares or more. During our 62-day sample period there were a total of 402,398 such transactions, with a median size of 1002 shares.
resulting in approximately 403,202,796 shares during the entire period, or 13,007 lots of 500 shares per day. We use the median to approximate the number of shares traded per day because of the large skew in the SPY volume distribution, attributed to a very small fraction of extremely large trade sizes, which are unlikely to exploit the cross-market mechanisms that we outline in this work. The final step in computing the integrated order flow profit is to approximate the profit per opportunity in each direction. These values are obtained on a daily basis by determining the relative frequency of each profitable opportunity and computing the associated weighted average of profits. This amounts to using the basis-adjusted histograms in Figure 2 as weights for the associated profit opportunities for specific bid and offer differences. Averaging across days we arrive at a profit of 1.367 cents per ES contract event and 0.9396 cents per 500 share SPY event. After scaling the integrated responses of panel (c) by the number of events and profit per event, we see that the integrated response profit is much more balanced – over the 30-millisecond horizon, ES events result in an expected profit of roughly $183,000 per day and SPY events result in roughly $171,000. We highlight that these values only represent the expected arbitrage profits between the CME and Nasdaq exchanges; given that Nasdaq accounts for approximately 25% of equities market trading, we anticipate the aggregate size of the trade to be about 4 times larger, per day, than the numbers we report. Our results demonstrate that although the per-event impact of SPY arbitrage opportunities is about two orders of magnitude larger than that of ES opportunities, the aggregate effect is ameliorated by the fact that there are almost two orders of magnitude fewer events in the SPY market that give rise to cross-market arbitrage.

While the literature on market information share typically attributes only a small weight to the ETF or cash market (typically around 10%), the foregoing results are striking because they demonstrate the relative information/profit importance of transactions in the ETF market. Further, our results are quite striking in that they resoundingly support the predictions of our model and highlight a previously unknown mechanism of cross-market information flow: relative price increment. The relevance of this mechanism is emphasized by the hundreds of billions of dollars of notional value traded daily in the ES/SPY market, and the trillion dollar size of the futures/spot markets as a whole. From a regulatory perspective, our results suggest that both price increment and contract size are not arbitrary
variables in contract and exchange design, but that regulators such as the SEC and CFTC should carefully consider the implications and consequences of contract relationships across equivalent or nearly equivalent assets.

5 Conclusion

The relationship of informed price movements, liquidity changes and order flow are important considerations in fragmented markets that trade equivalent, or very similar assets. These considerations become increasingly more important as markets become faster, and as market makers gain access to technology to intelligently traverse distinct market centers in very short intervals of time. Our work highlights the role of relative contract specifications, especially relative price increments, between two similar assets, in determining the direction of informed order flow between markets. We develop a model which demonstrates that market makers must balance competing needs to quote narrow spreads, in order to promote increased own-market order flow through reduced transactions costs, with the desire to quote wide spreads, in order to create more profitable cross-market arbitrage opportunities. Coupled with a leverage channel, which induces market makers in the futures market to quote wider spreads, our model makes several predictions regarding the direction of informed order flow between markets: when a market maker is passively filled on bid (offer) and the their transacted price is less (more) than the bid (offer) in the other market, they try to lay off some portion of their risk by directing aggressive order flow to the other market, while simultaneously earning an arbitrage profit. Despite preconceived notions of futures/cash market order flow relationships in practice, our model is agnostic of those observed relationships and makes predictions that information should move in both directions across markets, albeit under differing circumstances.

In our empirical study, we show that the contracts for futures and cash instruments on the S&P 500 index (the E-mini and SPY ETF, respectively) are designed so that the minimum price increment of the former is $2.5 \times$ larger than that of the latter, resulting in cross-market arbitrage opportunities that typically favor market makers in the E-mini. Indeed, conventional understanding of these two markets dictates that price or liquidity changes
in the E-mini almost always precede those of the SPY ETF by approximately the time it takes to transmit messages between the two trading venues – 4 ms. Using new econometric methodology, we show this standard direction of informed order flow is supported in the data: the SPY market typically reacts to the ES market, but only under conditions that our model suggests are favorable for market-maker arbitrage. Likewise, our empirical work shows that the conventional direction of informed order flow is reversed under the same conditions that favor cross-market arbitrage opportunities for SPY market makers. Our econometric method is simple, non-parametric and designed for asynchronous, high-frequency data and in all cases shows that the data strongly support the conclusions of our model.

We live in an era in which exchanges are proliferating, markets are becoming increasingly fragmented, trading technology is quickly changing and in which regulators are pressured to respond to perceived instability. As new exchanges, with differing contract specifications and order matching mechanisms are being added to a complex financial market network, the determinants of informed trading, and its direction within the system, become of vital importance to regulators. Our work demonstrates that a seemingly innocuous control variable, relative (not absolute) price increment, can have important implications for cross-market arbitrage, and hence direction of informed order flow. Understanding this channel, as well as others like it, will be necessary for regulators such as the SEC and CFTC and may guide them in further coordinating these distinct markets.
References


