Essays on Price Discovery Measure, Exchange-Traded Funds and Liquidity

Syed Galib Sultan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

Eric Zivot, Chair

Jonathan Brogaard

Yu-Chin Chen

Program Authorized to Offer Degree:

Department of Economics

University of Washington

Abstract

Essays on Price Discovery Measure, Exchange-traded Funds and Liquidity

Syed Galib Sultan

Chair of the Supervisory Committee:
Professor Eric Zivot
Department of Economics

Price Discovery is the process by which new information is impounded into asset prices through trading activity. A market is considered to contribute more to price discovery if it is the first to capture new information regarding the fundamental value of an asset. Hasbrouck's (1995) information share (IS) is the most widely used measure for price discovery contribution even though there is a well-documented concern with identification: its dependence on the ordering of the variable in the price vector and its non-uniqueness. In the first chapter, we propose a new measure, "Price Discovery Share" (PDS) that is closely related to IS and resolves the identification problems inherent in the IS method. PDS is motivated by a widely used method in risk management literature called the "risk-budgeting" or additive decomposition of portfolio volatility. Using simulated data based on different structural asset pricing models, we find that PDS measures the structural price discovery contribution more accurately than IS.

In the second chapter, we apply Price Discovery Share (PDS) to investigate the "duplication of Exchange-Traded Funds (ETFs)" phenomenon, a recent institutional trend in financial markets. We show that although there are multiple ETFs tracking the S&P 500 index, one specific S&P 500 ETF ('SPY') always contributes more to price discovery than the rest. We also find that PDS, unlike Information Share (IS), is robust to the use of intra-day market price data sampled at different frequencies.

In the third chapter, we study the effect of bond Exchange-traded funds (ETFs) and bond mutual funds on the liquidity of U.S. corporate bonds. Depending on the liquidity measure used, we find different statistically significant results. ETF ownership has a positive impact on their underlying corporate bonds liquidity when we only consider bonds that are already bought and held by ETFs. Bond mutual funds ownership is found to play a positive impact on the liquidity of high yield corporate bonds.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I would like to thank my thesis advisor Eric Zivot. I have been really fortunate to have a great mentor like Eric who was instrumental in every step of this journey. I would like to thank my dissertation committee members Jonathan Brogaard, Yu-Chin Chen and Ji Hyung Lee. I would also like to thank Jennifer Koski who served as my Graduate Student Representative. A special thanks goes to Fahad Khalil, the chair of Economics dept. at University of Washington Seattle for his constant support and guidance. Another special thanks goes to Yoshio Nozawa from Board of Governors of the Federal Reserve System (FRB) as well. I am also grateful for helpful comments and suggestions from the participants of Southern Economics Association Conference, Canadian Economics Association Conference, MTI brownbag at University of Washington Seattle, presentation in State Street Corporation and presentation in Texas A&M International University. I am also indebted to my two mentors from Washington University in Saint Louis- James Morley and Werner Ploberger.

I would like to express my heartfelt gratitude to my family- my mother Syeda Farida Parvin, my wife Nisha Noor and my younger brother Syed Sadaf Sultan. Their constant support is definitely one of the reasons why I could finish my PhD in timely manner. I also fondly remember the memory of my late father Mohammad Tawhidul Anwar who was a professor at the Department of Mass Communication and Journalism in University of Dhaka, Bangladesh. I am also grateful to my Father- and Mother-in-law for their constant prayers and support.

Lastly, I would like to thank my friends and colleagues for their help and support during different stages of my PhD- Wahid Abdallah, Nandini Abedin, Tanweer Akram, Shamma Adeeb Alam, Juan Equiza, Sohani Fatehin, Samuel Henly, Mudabbir Hussain, Minho Kim, Xuyang Ma, Orchi Modhurima, Romel Mostafa, Quyen Nguyen, Zaki Rahman, Mainul Raju, Jagori Saha and Bakhtiyar Uddin.

DEDICATION

To the three most important women in my life- my mother Syeda Farida Parvin, my wife Nisha Noor and my little angel Aleena Tehzeeb Syed.

CHAPTER 1

Price Discovery Share

1.1 Introduction

New market trends have made price discovery, the process by which new information is impounded into asset prices through trading activity, an important research agenda in the financial economics literature. Recent features of financial markets include the trading of identical stocks in multiple venues (market fragmentation) and the trading of closely related assets (e.g., derivatives, future and spot, ETFs tracking the same market index, etc.) in the same or different venues. Hasbrouck's (1995) information share (IS) is the most widely used empirical measure to identify and quantify the process of price discovery. [1] IS of a given market is defined as the share of that market to the innovation variance of the "efficient" [2] price or the random walk component of asset price. It is typically interpreted as identifying who moves first in the process of price adjustment when new trade related information or permanent shock is received.

IS has been used by numerous studies in different financial market related contexts. [3] However, it is also well-documented that IS has a potentially serious identification problem when idiosyncratic innovations to different market prices in Hasbrouck's (1995) model are contemporaneously correlated. When the correlation is significantly high, the IS measure, which is typically reported as a range, can become very wide and it does not clearly identify the price/information leader or the follower and their individual contributions to price discovery. This limitation has been referred to in previous literature as the "order-dependence problem of IS" because the upper and the lower bound of the range that IS reports depends on the order that the

[1] Special Issue on Price Discovery in the *Journal of Financial Markets* (2003) is an excellent source for a review of price discovery measures.

[2] According to Hasbrouck's (1995) model, intraday stock price at any time of the day is assumed to contain two components- a permanent component and a transient component. The permanent component is called the "efficient" or "fundamental" price of asset. It is assumed to have martingale property and to follow random walk. "Efficient" price is assumed to capture any change in price due to the arrival of new trade-related information into the stock market.

[3] IS has been applied to cross-listed stock to determine the information or price leadership among different stock exchanges (Hasbrouck, 1995; Huang, 2002; Harris et al.,2002) . IS has also been used to determine the information or price leadership between quotes and trade prices of stock (Hasbrouck, 2002), between stock options and underlying stocks (Chakravarty et al., 2004), between futures and their spots (Mizarch and Neely, 2008; Lien and Shrestha 2009), among Credit Default Swap (CDS), bonds and stocks (Grammig and Peter 2014) etc. Also, a brief summary of different studies which use IS can be found in pp- 78 of Putninš (2013).

prices enter into the vector of prices.[4] Numerous studies have proposed different solutions and measures (e.g. Hasbrouck, 1995; Baillie et al, 2002; Lien and Shrestha, 2009 and Grammig and Peter, 2014) to address this shortcoming of IS. However, no consensus has emerged so far because all of these approaches have been found to be either effective in particular context or to have their own identification issues. [5]

In this paper we address the identification problem of IS and propose a closely related measure for price discovery that is unique and order-invariant. Our measure of price discovery is motivated by a widely used method in portfolio risk management literature which additively decompose portfolio volatility into asset specific contribution.[6] A notable feature of Hasbrouck's (1995) model is that the volatility of the efficient price innovation (VEPI) is linearly homogeneous in the common factor weights of each market's innovation just as portfolio volatility is linearly homogenous in portfolio weights. We use this property and apply Euler's theorem to additively decompose the VEPI into components attributed to each market. Each of these components is defined as the contribution of each market to VEPI. Moreover, a key component of this decomposition is what we call the price discovery beta of a market. Price discovery beta of a market is the regression coefficient of a market's innovation on the efficient price innovation. We convert the calculated market contributions to the market shares by dividing these contributions by the VEPI. Our new measure of price discovery for each market is this contribution share which we call price discovery share (PDS).

PDS is applicable to the general n-assets or n-markets model. As a special case, we provide an analytical comparison between IS, IS-mean and PDS in a simple bi-variate case. We also

[4] An extensive discussion regarding this can be found in the section II of this paper.

[5] Hasbrouck (1995) suggests sampling the trade and quote prices at a high enough frequency such that contemporaneous correlation among the innovations becomes negligible. However, numerous studies including ours find that even with the use of bid-ask quotes sampled at a 1-second interval there is still enough residual correlation to produce a wide range for IS. Baillie et al. (2002) argue in support of using the mean or mid-point of the upper and lower bound of the range as a unique measure of IS. This approach, while intuitively appealing, is ad hoc. Lien and Shrestha (2009) correctly point out that the average of the two bounds of IS cannot be derived as a result of any particular factor structure. In addition, the calculated means or mid-points of the estimated individual price discovery contributions often do not add up to 100% in applications with more than two prices. Lien and Shrestha (2009) alternatively propose a modified information share (MIS) measure that is derived from the squared root of the eigenvalues of the innovation correlation matrix. A limitation of this approach is that it considers only the positive value of squared root of the eigenvalues in order to reach a unique result. If the negative values of the squared root of the eigenvalues are considered, then it would produce a different value for the MIS. Grammig and Peter (2014) propose another unique measure for IS which is derived by exploiting two properties of price changes - fat tails and tail-dependence.

[6] See Bruder and Roncalli (2012) for a nice description of risk budgeting and use of Euler theorem to determine individual risk contribution of assets in a portfolio.

compare IS and PDS using simulated market data. We generate simulated asset price data following four different structural asset pricing models and compare the sampling properties of IS and PDS. In every case, PDS is found to estimate the true structural price discovery contribution more accurately than IS.[7]

The remainder of the paper is organized as follows. In Section II we describe the reduced-form cointegration framework used by Hasbrouck (1995) for modeling price discovery in arbitrage linked market and also define our new measure of price discovery, PDS. In Section III, we compare PDS to IS using simulated market data generated from different structural models of asset prices.

1.2 Model Description

We use the arbitrage linked cointegration model approach of Hasbrouck (1995). Let $\boldsymbol{p}_t = \left(p_{1,t}, \ldots, p_{n,t}\right)'$ denote an $n \times 1$ vector of $I(1)$ log prices. In the price discovery literature, P_t either represents the vector of prices of a single asset that is traded in n market locations and linked by arbitrage or it represents the vector of prices of n similar or closely related and arbitrage linked assets that are being traded in the same market location.

It is assumed that there is a common stochastic component or fundamental value that drives all prices. As a result, there are n-1 cointegrating vectors $\boldsymbol{\theta}_i$ such that, $\boldsymbol{\theta}_i'\boldsymbol{p}_t \sim I(0)$. Furthermore, since the difference between any two prices in P_t is $I(0)$ it is convenient to use the following $(n-1) \times n$ matrix of rank n-1 as a basis for the cointegrating space:

$$\boldsymbol{\Theta}' = \begin{bmatrix} \boldsymbol{\theta}'_1 \\ \vdots \\ \vdots \\ \boldsymbol{\theta}'_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ 1 & 0 & \cdots & \cdots & -1 \end{bmatrix} = (\mathbf{1}_{n-1} \vdots -\mathbf{I}_{n-1}) \tag{1.1}$$

where $\mathbf{1}_{n-1}$ is an $(n\text{-}1) \times 1$ vector of ones and $\mathbf{I}_{n-1}$ is the identity matrix of dimension n-1. Since $\Delta\boldsymbol{p}_t$ is $I(0)$, it has a Wold representation:

$$\Delta\boldsymbol{P}_t = \boldsymbol{\Psi}(L)\boldsymbol{e}_t = \boldsymbol{e}_t + \boldsymbol{\Psi}_1\boldsymbol{e}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{e}_{t-2} + \ \ldots \tag{1.2}$$

[7] Due to the popularity of Baillie et al (2002) measure in different literature, we have decided to report this measure along with upper and lower bound of IS in our analysis. We call this measure "IS-mean".

where $\boldsymbol{\Psi}(L) = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k L^k$, $\boldsymbol{\Psi}_0 = \boldsymbol{I}_n$ and $\boldsymbol{e}_t \sim iid\ (0, \boldsymbol{\Sigma})$. It is assumed that elements of $\boldsymbol{\Psi}(L)$ are 1-summable and $\boldsymbol{\Psi}(1) \neq 0$. Also,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{bmatrix} \tag{1.3}$$

Using the Beveridge-Nelson decomposition (Beveridge and Nelson, 1981), we can write:

$$\boldsymbol{P}_t = \boldsymbol{P}_0 + \boldsymbol{\Psi}(1) \sum_{j=0}^{t} \boldsymbol{e}_j + \boldsymbol{\Psi}^*(L)\boldsymbol{e}_t \tag{1.4}$$

where $\boldsymbol{p}_0$ is n ×1 vector of initial values, $\boldsymbol{\Psi}(1) = \sum_{k=0}^{\infty} \boldsymbol{\Psi}_k$, $\boldsymbol{\Psi}^*(L) = \sum_{k=0}^{\infty} \boldsymbol{\Psi}^*{}_k$, $\boldsymbol{\Psi}^*{}_k = -\sum_{j=k+1}^{\infty} \boldsymbol{\Psi}_j$. Also, $\boldsymbol{\Psi}^*(L)e_t \sim I(0)$ and $\boldsymbol{\Theta}'\boldsymbol{\Psi}(1) = 0$. The restriction $\boldsymbol{\Theta}'\boldsymbol{\Psi}(1) = 0$ implies that the $n \times n$ matrix $\boldsymbol{\Psi}(1)$ has rank one and can be expressed as

$$\boldsymbol{\Psi}(1) = \boldsymbol{1_n}\boldsymbol{\psi}' = \begin{bmatrix} \psi_1 & \cdots & \psi_n \\ \vdots & \ddots & \vdots \\ \psi_1 & \cdots & \psi_n \end{bmatrix} \tag{1.5}$$

where $\boldsymbol{\psi} = (\psi_1, \dots, \psi_n)'$ is an $n \times 1$ vector. The matrix $\boldsymbol{\Psi}(1)$ contains the cumulative impacts of the innovation $\boldsymbol{e}_t$ on all future price movements, and thus measures the long-run impact of $\boldsymbol{e}_t$ on prices. Since the rows of $\boldsymbol{\Psi}(1)$ are identical, the long-run impact of $\boldsymbol{e}_t$ on each price is identical. Substituting (5) into (4) gives us

$$\boldsymbol{P}_t = \boldsymbol{P}_0 + \boldsymbol{1_n} \sum_{j=0}^{t} \eta_j^P + \widetilde{\boldsymbol{\varepsilon}}_t \tag{1.6}$$

Here, $\widetilde{\boldsymbol{\varepsilon}}_t = \boldsymbol{\Psi}^*(L)\boldsymbol{e}_t$ is an *I*(0) pricing error vector, $\sum_{j=0}^{t} \eta_j^P$ is the random walk component that is common to all prices. The following representation of equation (6) makes the role of η_t^P easier to interpret:

$$\boldsymbol{P}_t = \boldsymbol{P}_0 + \boldsymbol{1_n m_t} + \widetilde{\boldsymbol{\varepsilon}}_t \tag{1.7}$$

where, $m_t = m_{t-1} + \eta_t^P$. And $\eta_t^P = \boldsymbol{\psi}'\boldsymbol{e}_t = \sum_{k=1}^{n} \psi_k e_{kt}$ is therefore the innovation to the random walk component of the price. Hasbrouck (1995) described η_t^P as the component of the price change that is permanently impounded into the price that is due to new information. Transient pricing errors such as bid-ask bounces and inventory adjustments are absorbed by the *I*(0) component $\widetilde{\boldsymbol{\varepsilon}}_t$.

1.2.1. Hasbrouck's Information Share

Hasbrouck (1995) defines a market's price discovery contribution of i th market as its contribution to the permanent shock variance, var(η_t^p) $= \boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi}$.

In practice, IS_i (i=1,…,n) is computed from the estimated parameters of an empirical Vector Error Correction Model or VECM (K-1) for asset prices:

$$\Delta\boldsymbol{P}_t = \boldsymbol{A}(\boldsymbol{\Theta}'\boldsymbol{P}_{t-1} - \boldsymbol{\mu}) + \sum_{j=1}^{K-1}\boldsymbol{\Gamma}_j\Delta\boldsymbol{P}_{t-j} + \boldsymbol{e}_t \tag{1.8}$$

where $\boldsymbol{\Gamma}_j$ is an $n \times 1$ matrix. The lag length, K, is typically chosen by some model selection criterion such as BIC or AIC. Because the cointegration matrix $\boldsymbol{\Theta}'$ is known, equation (8) can be estimated by least squares equation by equation. The long-run impact matrix $\boldsymbol{\Psi}(1)$ can be computed directly using Johansen's factorization and the estimation coefficients (A, $\boldsymbol{\Theta}$ and $\boldsymbol{\Gamma}_j s$) from the VECM:

$$\boldsymbol{\Psi}(1) = \boldsymbol{\Theta}_\perp(\boldsymbol{A}'_\perp\boldsymbol{\Gamma}(1)\boldsymbol{\Theta}_\perp)^{-1}\boldsymbol{A}'_\perp \tag{1.9}$$

where $\boldsymbol{\Theta}_\perp$ and $\boldsymbol{A}_\perp$ are vectors satisfying $\boldsymbol{\Theta}'\boldsymbol{\Theta}_\perp = \boldsymbol{0}$ and $\boldsymbol{A}'\boldsymbol{A}_\perp = \boldsymbol{0}$, respectively, and $\boldsymbol{\Gamma}(1) = \boldsymbol{I}_n - \sum_{j=1}^{K-1}\boldsymbol{\Gamma}_j$. Information share (IS) measure for price discovery is defined as follows:

Case1: $\boldsymbol{\Sigma}$ is diagonal:

$$IS_i = \frac{(\psi_i\sigma_i)^2}{\boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi}} \quad , \ i=1,\ldots,n \tag{1.10}$$

Case 2: $\boldsymbol{\Sigma}$ is non-diagonal,

$$IS_i = \frac{([\boldsymbol{\psi}'\boldsymbol{F}]_i)^2}{\boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi}} \quad , \ i=1,\ldots,\mathrm{n} \tag{1.11}$$

where $[\boldsymbol{\psi}'\boldsymbol{F}]_i$ is the i-th element of $\boldsymbol{\psi}'\boldsymbol{F}$ and **F** is a lower triangular matrix (Cholesky factor) such that $\boldsymbol{F}\boldsymbol{F}' = \boldsymbol{\Sigma}$. The value of $\boldsymbol{F}$, and hence the value of IS_i, depends on the ordering in which individual prices enter into the vector of price, P_t. Therefore, when $\boldsymbol{\Sigma}$ is non-diagonal, Hasbrouck's approach can only provide upper and lower bounds for IS_i based on all possible orderings of prices. In particular, Baillie et al. (2002) showed that largest information share for a given market occurs when it is placed first in the price vector.

1.2.2. New Order Invariant Measure of Price Discovery: Price Discovery Share

Our new measure of price discovery is motivated by the additive decomposition of portfolio volatility that is widely used in risk management. Recall, the permanent shock is defined as a weighted average of individual market innovations $\eta_t^P = \boldsymbol{\psi}'\boldsymbol{e}_t = \sum_{k=1}^{n} \psi_k e_{kt}$. The volatility of the permanent shock is $\sigma_\eta(\boldsymbol{\psi}) = (\boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi})^{\frac{1}{2}}$. An interesting property of $\sigma_\eta(\boldsymbol{\psi})$ is that it is linearly homogenous in $\boldsymbol{\psi}$ since $\sigma_\eta(c \cdot \boldsymbol{\psi}) = c \cdot \sigma_\eta(\boldsymbol{\psi})$ for any constant *c*. We apply Euler's theorem and derive the following unique and additive decomposition of $\sigma_\eta(\boldsymbol{\psi})$:

$$\sigma_\eta(\boldsymbol{\psi}) = \boldsymbol{\psi}' \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} = \sum_{i=1}^{n} \psi_i \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_i} = \psi_1 \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_1} + \cdots + \psi_n \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_n} \quad (1.12)$$

Hence, the volatility of the permanent shock can be expressed as a weighted sum of marginal contributions from each asset (or market *i*). The *i*-th term on the right-hand side (1.12), $\psi_i \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_i}$, is asset *i*'s (or market *i*'s) contribution to the volatility of the permanent shock. In the spirit of Hasbrouck (1995)'s information share, $\psi_i \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_i}$ is a natural measure of an asset's (or market's) contribution to price discovery. Our new measure price discovery share of asset i (or market i), denoted by PDS_i, is its contribution divided by $\sigma_\eta(\boldsymbol{\psi})$:

$$PDS_i = \frac{\psi_i \frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\psi_i}}{\sigma_\eta(\boldsymbol{\psi})} \quad (1.13)$$

By construction $\sum_{i=1}^{n} PDS_i = 1$. A small amount of mathematical derivation gives us the following result:

$$\frac{\partial\sigma_\eta(\boldsymbol{\psi})}{\partial\boldsymbol{\psi}} = (\sigma_\eta(\boldsymbol{\psi}))^{-1}\boldsymbol{\Sigma}\boldsymbol{\psi} = \sigma_\eta(\boldsymbol{\psi})\boldsymbol{\beta}, \quad (1.14)$$

where $\boldsymbol{\beta} = (\beta_{1,\cdots,}\beta_n)' = \sigma_\eta(\boldsymbol{\psi})^{-2}\boldsymbol{\Sigma}\boldsymbol{\psi}$ with $\beta_i = \frac{cov(e_{it},\eta_t^P)}{var(\eta_t^P)} = \frac{\psi_i\sigma_i^2 + \sum_{j=1}^{n-1}\psi_{j\neq i}\sigma_{ij\neq i}}{\boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi}}$, we deduce the following analytic expression for PDS_i

$$PDS_i = \psi_i\beta_i = \frac{{\psi_i}^2\sigma_i^2 + \sum_{j=1}^{n-1}\psi_i\psi_{j\neq i}\sigma_{ij\neq i}}{\boldsymbol{\psi}'\boldsymbol{\Sigma}\boldsymbol{\psi}} \quad (1.15)$$

We denote β_i in equation (15) as the "*price discovery beta*" of asset *i* (or market *i*). The *price discovery beta* is the slope coefficient from the regression of η_t^P on e_{it} and summarizes the (normalized) covariance contributions of an asset's (or market's) innovation to the variance of the efficient price innovation. Hence, *PDS$_i$* is an asset *i*'s (or market *i*'s) contribution to η_t^P weighted by its price discovery beta.

1.2.3 Comparing PDS to IS in the case of Two Assets/Markets

We consider the case of *n* = 2, so that $\boldsymbol{P}_t = \left(p_{1,t}, p_{2,t}\right)'$. This allows us to analytically compare PDS to IS. Under the assumption of uncorrelated innovations ($\sigma_{12} = 0$) or diagonal $\boldsymbol{\Sigma}$ and using equation 1.10 and equation 1.15, we find that *IS$_i$* and *PDS$_i$* are identical:

$$IS_{i,diag} = \frac{\psi_i^2\sigma_i^2}{\psi\prime\Sigma\psi} = \frac{\psi_i^2\sigma_i^2}{(\psi_1^2\sigma_1^2+\psi_2^2\sigma_2^2)} = \psi_i\beta_i = PDS_{i,diag} \tag{1.16}$$

However, this is not the case when $\boldsymbol{\Sigma}$ is non-diagonal. Let, $\boldsymbol{\Sigma} = \boldsymbol{F}\boldsymbol{F}'$ where $\boldsymbol{F}$ is the 2×2 lower triangular matrix (Cholesky factor) which is defined as follows:

$$\boldsymbol{F} = \begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sigma_2(1-\rho^2)^{\frac{1}{2}} \end{bmatrix} \tag{1.17}$$

where $\rho^2 = \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}$. Then, using equation equation 1.11, *IS$_i$* is given by:

$$IS_{1,non-diag} = \frac{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2\rho^2 + 2\psi_1\psi_2\sigma_{12}}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \tag{1.18}$$

$$IS_{2,non-diag} = \frac{\psi_2^2\sigma_2^2 - \psi_2^2\sigma_2^2\rho^2}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \tag{1.19}$$

When the ordering of prices is reversed subscripts 1 and 2 get reversed in 1.18 and in 1.19. Inspection of equation 1.18 and equation 1.19 reveals that the highest (lowest) *IS$_i$* value occurs when asset *i* (or market *i*) is ordered first (last) in the vector of prices. This gives rise to upper and lower bounds for *IS$_i$* based on ordering of prices. To get a unique value for *IS$_i$*, Bailie et al. (2002) proposed to use the mean of the upper and lower bounds derived from 1.18 and 1.19

$$IS_{1,Bailie} = \frac{\psi_1^2\sigma_1^2 + (\psi_2^2 - \psi_1^2)\sigma_2^2\rho^2 + \psi_1\psi_2\sigma_{12}}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \quad (1.20)$$

$$IS_{2,Bailie} = \frac{\psi_2^2\sigma_2^2 + (\psi_1^2 - \psi_2^2)\sigma_2^2\rho^2 + \psi_1\psi_2\sigma_{12}}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \quad (1.21)$$

From equation 1.15, PCR_i for non-diagonal $\mathbf{\Sigma}$ is given by

$$PDS_{1,non-diag} = \frac{\psi_1^2\sigma_1^2 + \psi_1\psi_2\sigma_{12}}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \quad (1.22)$$

$$PDS_{2,non-diag} = \frac{\psi_2^2\sigma_2^2 + \psi_1\psi_2\sigma_{12}}{\psi_1^2\sigma_1^2 + \psi_2^2\sigma_2^2 + 2\psi_1\psi_2\sigma_{12}} \quad (1.23)$$

From equations 1.18 – 1.23, we make the following observations regarding IS, PDS and IS-mean when $\mathbf{\Sigma}$ is non-diagonal. First, PDS_i distributes the covariance contributions of each asset (or market) to the permanent shock variance, $\psi_1\psi_2\sigma_{12}$, evenly across assets (or markets). PDS_1 differs from PDS_2 only due to the difference between $\psi_1^2\sigma_1^2$ and $\psi_2^2\sigma_2^2$. Second, for $n > 2$, the calculation of the upper and lower bounds of IS_i requires recalculation of 1.11 for all the possible orderings of prices. For example, when $n = 5$ there are 120 possible ordering of prices which need to be considered. Also, for each ordering, we will get different values of IS from which we have to pick the highest and lowest value in order to define the range of IS. The calculation of PDS_i is invariant to the ordering of prices. Third, if $\psi_1 = \psi_2$ the mid-point IS_i (i = 1,2) in (20) and (21) is equal to PDS_i (i = 1,2) in equation 1.22 and 1.23. Fourth, it is possible for PDS_i to be negative. From (14), this can happen if ψ_i is negative and β_i is positive and vice-versa. It is unusual for either ψ_i or β_i to be negative. It can be shown (cf. Zivot and Yan, 2010) that $\boldsymbol{\psi} \propto \boldsymbol{\alpha}_\perp$ where $\boldsymbol{\alpha}_\perp$ is a 2×1 vector such that $\boldsymbol{\alpha}'_\perp\boldsymbol{\alpha} = \mathbf{0}$ and $\boldsymbol{\alpha}_\perp$ is the 2×1 vector of error correction coefficients from the VECM in equation (8) when n=2. In equation 8, we use the $n \times 1$ version of error correction matrix and express it as "A". In typical applications α_1 and α_2 have opposite signs so that ψ_1 and ψ_2 are both positive. However, it is possible to have a stable VECM with α_1 and α_2 having the same sign. In that case, ψ_1 and ψ_2 will have opposite signs. On the other hand, if ψ_1 and ψ_2 have the same sign then $\beta_i = cov(e_{it}, \eta_t^P) = \psi_1 \sigma_1^2 + \psi_2 \sigma_{12}$ can be negative if σ_{12} is a sufficiently large negative number.[8]

1.3 Application to Simulated Market Data

In this section we use simulated market data to provide comparison between IS and PDS. The simulated market data are generated from three different stylized structural models of asset prices described in Hasbrouck (2002). We also propose a modified version of one of these three models as an additional example. Using each of these simulated market data sets, we compute IS (upper and lower bound), IS-mean (average of these two bounds) and PDS. Using Monte Carlo simulations, we report mean, standard errors and 95% confidence intervals of these three price discovery measures.[9]

1.3.1 Applications to simulated market data: two-market "Roll" model

The first example is a model of high-frequency trade prices suggested by Roll (1984). Hasbrouck (2002) uses a simplified two-market version of this model and compare IS with another measure of price discovery called "component share". This model suggests that the trade price at any time during the trading hours has two components. One is the "efficient" or "fundamental" price which has the martingale property and follows a random walk. The other is the transient component which mostly arises from bid-ask bounce, inventory effects, discreteness etc. This model assumes that there are two markets where a cross-listed identical stock is being traded at price p_{1t} in Market 1 and at p_{2t} in Market 2. The common efficient price of this stock is defined as follows:

$$m_t = m_{t-1} + u_t, \tag{1.24}$$

where $u_t \sim N(0, \sigma_u^2)$ .There is a common and identical fixed cost per trade in each market denoted by "c". The bid-price at time t is $m_t - c$ and the ask-price at time t is $m_t + c$. The trade

[8] In the risk management context, an asset's contribution to portfolio volatility can be negative if it has a negative weight in the portfolio or if its beta with respect to the portfolio is negative (natural risk reducer). In the latter case the asset is negatively correlated with the portfolio.

[9] 95% Confidence interval is calculated as, $CI_{95} = mean \pm 2 * s.e.$

direction indicator is denoted by "q_t" which takes a value equal to 1 if the trader is buying and -1 if the trader is selling. Buys and sells are assumed to be equally likely and serially independent. The traders are also assumed to buy or sell independently of the innovation to efficient price (denoted by u_t). Therefore, trade direction and transaction price are defined as follows,

Trade direction: $q_{it} = \pm 1$, each with probability ½ for market, $i = 1{,}2$,

Transaction price: $p_{it} = m_t + cq_{it}$ for market, $= 1{,}2$,

From the set-up of the model, it is clear that two markets are structurally identical and therefore the true contribution of each market to price discovery is 50%. Following Hasbrouck (2002) we set, c = 1 and $\sigma_u = 1$. We generate 100,000 observations of transaction prices, p_{1t} and p_{2t} for each market. We use the prices to estimate the VECM as described in equation 8 and use the estimated parameters and covariance matrix to calculate IS, IS-mean and PDS. We repeat this process 1000 times and calculate the mean, standard error and 95% confidence interval of these 1000 estimated IS, IS-mean and PDS .Table 1 reports the simulation results for Market 1.

From Table 1.1 we find that the mean of PDS (50.1%) and IS-mean (50%) of Market 1 are very close to the structural price discovery contribution or 50%. However, the upper and lower bound of report a large range of price discovery contribution. The upper bound IS identifies Market 1 as the price leader (with estimated 78.9% contribution to price discovery) and the lower bound IS identifies Market 1 as the price follower (21.2% of estimated contribution to price discovery).

1.3.2 Applications to simulated market data: two markets with private information

We also take our second example from Hasbrouck (2002). In this model, it is assumed that all the informed trading happen in Market 1. This implies that changes in efficient price are driven only by trading activities in Market 1.The efficient price is defined as following:

$$m_t = m_{t-1} + \lambda q_{1t} \tag{1.25}$$

where λ is defined as liquidity parameter and assumed to be strictly positive. Trade directions (q_{it}) are defined in similar way as before. The transaction price of Market 2 depends on the lagged value of m_t and this assumption essentially defines the Market 2 as the price follower.

Trade direction: $q_{it} = \pm 1$, each with pr. ½ for $i = 1,2$,

Transaction price: $p_{1t} = m_t + cq_{1t}$ and $p_{2t} = m_{t-1} + cq_{2t}$

The structural model suggests that all the price discovery happens in Market 1 and therefore, the structural price discovery contribution of Market 1 is 100%. Following Hasbrouck (2002), we set $c = 1$ and $\lambda = 1$. We again generate 100,000 sample observation, calculate the IS, IS-mean and PDS and repeat 1000 times. Table 2 summarizes the moments and 95% confidence interval of each price discovery measure.

From Table 1.2, we find out that in Example 2 set-up, all the three price discovery measures estimate the contribution of Market 1 to price discovery as 99.9% which is almost equal to the structural contribution (100%). Therefore, in this setup, we do not find any difference among the three measures and all of them estimate the price discovery contribution with almost perfect accuracy.

1.3.3 Applications to simulated market data: two markets with private and public information

In Example 3, the efficient price, m_t contains a non-trade public information component (u_t) and a private information component (λq_{1t}) which is driven by Market 1's trade. Efficient price is defined as follows,

$$m_t = m_{t-1} + \lambda q_{1t} + u_t \quad (1.26)$$

where $u_t \sim N(0, \sigma_u^2)$. Trade direction is defined same as before. Trade cost for each market is now different from each other. The transaction price of Market 2 is again assumed to depend on lagged information regarding efficient price.

Trade direction: $q_{it} = \pm 1$, each with pr. ½ for $i = 1,2$,

Transaction price: $p_{1t} = m_t + c_1 q_{1t}$ and $p_{2t} = m_{t-1} + c_2 q_{2t}$

Market 1 is again the price leader in the structural model with 100% contribution to price discovery. The trading cost in Market 1 (c_1) is higher than Market 2 because the cost of market making is higher in Market 1, where all the informed traders are trading. Trades are done cheaply at stale prices in Market 2. For simulation, we set $c_1 = 1$, $c_2 = 0$, $\lambda = 1$ and $\sigma_u = 1$. We conduct the simulation in the same way as before.

Table 1.3 summarizes the result. The IS reports a upper bound of 98.4% and a lower bound of 90% for price discovery contribution of Market 1. The range in IS in this model is significantly larger than before and suffer from lack of identification. PDS estimate do not have any identification issue and reports a comparatively accurate estimate (96%) of actual price discovery contribution than that of IS-mean (94.2%). It should also be noted that the 95% confidence interval of lower bound of IS, IS-mean and PDS do not contain the true value of price discovery contribution.

1.3.4 Applications to simulated market data: Modified two-market "Roll" model

All the three models defined earlier are either 50-50 or all-or-nothing situation. In this example, we modify the first model ("Roll" model) in such a way so that it produces a structural price discovery contribution of 70% for Market 1 and 30% for Market 2.

We define a binary variable D such that D=1 with probability 0.7 and D=0 with probability 0.3. The assumption is that efficient price is driven by i.i.d. non-trade information which is revealed contemporaneously only to Market 1 70% of the time and the rest of the time to Market 2. We keep the model very simple by assuming no liquidity effect ($\lambda = 0$) and identical trade cost (c) for both of them. Efficient price is defined the same way as in first example (equation 1.24) and the transaction prices are defined as follows:

$$p_{1t} = Dm_t + (1 - \mathrm{D})m_{t-1} + cq_{1t} \tag{1.27}$$

$$p_{2t} = (1 - D)m_t + Dm_{t-1} + cq_{2t} \tag{1.28}$$

The structural price discovery shares are 70% for Market 1 and 30% for Market 2. We proceed with the simulation exercise in the same way as before.

Table 1.4 summarizes the simulation results. IS upper bound assigns a price discovery contribution of 80.5% to Market 1 (or identify Market 1 as the price leader) and lower bound assigns a contribution of 50.8% (no clear price leader or follower). PDS assigns a price discovery contribution of 67.5% to Market 1 which is very close to its true structural contribution (70%) and the 95% confidence interval of PDS contains the true contribution. IS-mean estimates a contribution of 65.6% contribution which is also close to actual contribution but its 95% confidence interval does not contain the true structural contribution.

1.4 Conclusion

This paper proposes a solution to the arbitrariness and lack of identification that come with Hasbrouck's (1995) IS. We propose a new method in quantifying the contribution to price discovery. We have shown analytically that our measure, PDS is identical to IS when there is zero contemporaneous correlation among the innovations to different market prices. When the correlation is non-zero, IS reports a range that can be excessively wide with strong correlation. On the other hand, PDS reports a unique measure under the similar condition. Our method also demonstrate a systematic way of distributing covariance contribution of each asset price to the permanent shock variance evenly across prices. We also demonstrated that PDS performs better compared to IS in identifying the price discovery contribution using simulated market data with known structural price discovery contribution.

Our expectation is that, our new order invariant measure of price discovery, PDS will be adopted widely in the future discourse on price discovery. We also hope that our study on the market of exchange-traded funds will also attract more attention in future to this class of assets and their price dynamics.

Table 1.1 Two-market 'Roll' model

Model 1 : structural price discovery share of market 1 = 0.50	**Hasbrouck (1995) Model: IS for market 1**			
	Upper bound	**Lower Bound**	**IS Mean**	**PDS**
Mean	0.789	0.212	0.500	0.501
Standard deviation	0.011	0.011	0.011	0.017
95% confidence interval	[0.766, 0.812]	[0.188, 0.235]	[0478, 0.522]	[0.466, 0.535]

Table 1.2: Two markets with private information

Model 2 : structural price discovery share of market 1 = 1.0	Hasbrouck (1995) Model: IS for market 1		IS Mean	PDS
	Upper bound	Lower Bound		
Mean	0.999	0.999	0.999	0.999
Standard deviation	0.0002	0.0002	0.0002	0.0002
95% confidence interval	[0.999,1.0]	[0.999,1.0]	[0.999,1.0]	[0.999, 1.0]

Table 1.3: Two markets with private and public information

Model 3 : structural price discovery share of market 1 = 1.0	**Hasbrouck (1995) Model: IS for market 1**		**IS Mean**	**PDS**
	Upper bound	**Lower Bound**		
Mean	0.984	0.900	0.942	0.960
Standard deviation	0.003	0.008	0.005	0.006
95% confidence interval	[0.978, 0.990]	[0.884,0.916]	[0.932, 0.952]	[0.948, 0.972]

Table 1.4: Modified Two-market "Roll" model

Model 4 : structural price discovery share of market 1 = 0.7	Hasbrouck (1995) Model: IS for market 1		IS Mean	PDS
	Upper bound	Lower Bound		
Mean	0.805	0.508	0.656	0.675
Standard deviation	0.011	0.014	0.012	0.014
95% confidence interval	[0.827, 0.783]	[0.536, 0.480]	[0.632, 0.680]	[0.703, 0.647]

Chapter 2

Price Discovery in S&P 500 Exchange-Traded Funds (ETFs)

2.1 Introduction

In this chapter, we apply our newly constructed measure of price discovery called 'Price Discovery Share' or PDS in an empirical investigation of price discovery in the market of exchange-traded funds (ETFs) that track the S&P 500 index. ETFs are defined as securities that track the performance of the market index, commodity and bonds like an index fund. ETFs can be traded like close-end mutual funds, that is, they can be traded throughout the day like a stock. ETFs are usually highly liquid assets with high tax-efficiency and very low expense ratio compared to mutual funds. All these favorable features have made ETFs extremely popular for investment risk management purposes.

After the "Flash Crash" of May 6, 2010 ETF trading has also caught the attention of regulators and academicians alike. Sharp price falls in a disproportionate number of ETFs during the Flash Crash have been deemed to be responsible for the abrupt market crash.[10] An investigative report by Borkovec, Domowitz, Serbin and Yegerman (2010) finds that price discovery failed dramatically for the class of ETF securities during the flash crash. In this backdrop, understanding the price discovery dynamics in the ETFs has become an important research agenda.

A recent trend in financial markets, which is relatively unexplored in the price discovery literature is the proliferation of multiple ETFs tracking the same index. In this paper, we denote this as "duplication of ETFs". The natural research question that rises from this trend, "What is the rationale behind the "duplication of ETFs" and what is its effect on price discovery?" In this study, we mainly focus on the effect of ETF duplication on price discovery. We select a particular index ETF, the S&P 500 ETF and study price discovery between two nearly identical and competing S&P 500 ETFs. The two S&P 500 ETFs we consider are SPY (issued by State Street) and IVV

[10] According to the joint SEC/CFTC flash Crash Report, *"ETFs accounted for 70% of all US-listed securities that declined by 60% or more during the May 6, 2010 Flash Crash".* See Borkovec, Domowitz, Serbin and Yegerman (2010)

(issued by iShare).[11] We investigate price discovery between these two competing ETFs during a regular trading week in 2012 and also, during two highly volatile trading days in recent years-the Flash Crash of May 6th, 2010 and stock market fall of August 8th, 2011 (two days after US lost its AAA credit rating). SPY is found to be the price leader in all instances and particularly, during the flash crash, SPY is found to be the absolute price leader with more than 95% share of price discovery.

We also pursue a comparative assessment of IS and PDS in two different empirical settings. First, we use our "duplication of ETF" application to show that even with one-second tick-by-tick quotes, IS can report very misleading results with wide ranges of price discovery share. By contrast, PDS always provides a clean decomposition and a unique value of price discovery contribution. Second, we use quotes of a single day (Dec 3rd, 2012) of the cross-listed ETF "SPY" to calculate price discovery share across two stock exchanges (NASDAQ and BATS). IS reports misleading results whereas, PDS results are consistent and also robust to the use of quotes data with higher time intervals.

The remainder of the chapter is organized as follows. Section II presents our empirical application in S&P 500 ETFs and evaluates the effect of "duplication of ETFs" on price discovery. In Section III, we compare the performance of IS and PDS in two different empirical settings. Finally, we conclude with a brief summary of the paper's findings and provide some guidelines for future research.

2.2 Empirical Application: Price Discovery in S&P 500 ETF Market

In this section, we apply PDS to quantify and analyze price discovery in the market for ETFs which track the S&P 500 index. In particular, we examine two competing S&P 500 ETFs (SPY and IVV) and discuss the effect of duplication of ETFs on price discovery. We first give an overview of the market for S&P 500 ETFs. We then review the existing literature on S&P 500 ETFs, describe our data and present our results and analysis.

[11] SPY is the largest, oldest and most popular ETF of its kind which was issued by SPDR State Street Global Advisors for public trading in 1993. IVV is the second most popular ETF that tracks S&P 500 index. It was first issued in 2000 by iShares More information regarding these ETFs can be found in section 4.1.

2.2.1 Market for S&P Exchange-Traded Funds (ETFs): An overview

In the past decade, the U.S. stock market has been characterized by a new market phenomena which we call "duplication of ETFs". More precisely, this refers to the proliferation of ETFs that track an identical index. For example, SPY (issued by SPDR of State Street), IVV (issued by iShares of BlackRock) and VOO (issued by Vanguard) track the S&P 500 index, IWM (issued by iShares of BlackRock), VTWO (issued by Vanguard) and TWOK (issued by SPDR of State Street) track the Russell 2000 index; and QQEW (issued by First Trust) and QQQE (issued by Direxion) track the NASDAQ-100 equal weighted index. Natural questions to ask are why this duplication is occurring and what effect does this dupli cation have on market prices. In our application, we address the second question from a price discovery perspective. The Duplication of ETFs shares similarity with "market fragmentation" as far as competition is concerned. Therefore, price discovery analysis of competing ETFs fits agreeably into this line of research.

Among the aforementioned index ETFs, we particularly focus on the S&P 500 ETFs because they are the most popular and highly traded. S&P 500 ETFs are also highly liquid assets with very low expense ratios and command a larger portion of market share among other index ETFs. There are currently three ETFs that track S&P 500 index- SPY, IVV and VOO.

Table 2.1 and 2.2 provide a brief comparison among SPY, IVV and VOO. SPY was introduced to the market first in 1993 and in fact, was the very first ETF of its kinds. IVV was issued next in 2000 and VOO was introduced very recently in 2010. Table 2.1 also reveals that they are similar to each other in terms of performance measures. Although VOO is the newest ETF, it is popular among traders due to its impressively low expense ratio (0.05%). However, VOO still hasn't managed to capture significant market share since SPY and IVV command almost 91.75% of the total market capitalization for S&P 500 ETF.

Analysis of Table 2.2 reveals the fact that in terms of top 10 holdings SPY and IVV are more similar to each other than VOO. This is also true when we see the sector-wise decomposition of these three ETFs. The market price data also reveals this fact in the sense that at any given time, prices of SPY and IVV are very close to each other whereas price of VOO is significantly different from the rest.

There are two additional features of S&P 500 ETFs that need to be discussed here in order to have a better understanding of the price dynamics of these ETFs. These are the tracking error of ETFs and the arbitrage opportunities in ETF trading.

The S&P 500 index uses a market capitalization weighting structure to construct its portfolio. However, S&P 500 ETFs cannot exactly replicate these portfolio weights for several reasons. First, there may be a copyright issue. And second, portfolio weights of the S&P 500 index are constantly changing depending on the change in market capitalization of the underlying assets. Instant portfolio rebalancing for an ETF is costly and therefore, their portfolio weights are not identical to the S&P 500 index. This also creates tracking error in the price of each S&P 500 ETF. Since a low tracking error of an ETF makes it more attractive for the ETF investors, all the ETFs issuers have strong incentives to reduce this tracking error as much as possible. Therefore, the tracking errors of each ETF are bounded.

Arbitrage opportunities can be created in ETF trading in two different ways. First, an arbitrage opportunity can be created between the ETF price and its Net Asset Value or NAV[12] (e.g. Ben-David, Franzoni, Moussawi, 2014[13]; Madhavan and Sobczyk, 2014[14]). Second, an arbitrage opportunity can be created between two similar ETFs which track the identical index (e.g. Marshall et al, 2013). Authorized Participants (APs[15]) can only take advantage of the first type of arbitrage. For example during the closing hour of trading, if the ETF price exceeds its NAV, the APs can buy the underlying securities of that ETF from the secondary market and submit them to the ETF issuers in exchange of new ETFs in the primary market. The APs can then sell the ETFs in the secondary market at a premium. The APs do exactly the opposite when the price of ETF falls below its NAV. They buy ETF at discount from the secondary market, redeem the ETFs into its underlying stocks in the primary market and then sell the underlying securities at a profit in the secondary market. In contrast, an arbitrage opportunity between two similar or nearly identical ETFs can occur at any time during the trading hours. Any ETF investors (retail or institutional)

[12] NAV per share is computed once a day based on the closing market prices of the underlying securities in the ETF's portfolio.
[13] Ben-David, Franzoni, Moussawi (2014) investigate the effect of ETFs on their underlying stocks and find that stocks owned by ETFs have significantly higher intra-day and daily volatility.
[14] Madhavan and Sobczyk (2014) utilize the arbitrage link between ETF closing price and its NAV and propose a state-space model of ETF price dynamics.
[15] APs are a class of institutional investors of ETFs who usually has legal contract with the ETF issuers.

can take advantage of this. Let us consider the case of SPY and IVV. If the price of SPY exceeds the price of IVV at any time during the day, there is a strong incentives for the investors who treat both of them as close substitute, to sell SPY at the high price and buy IVV at the low price. This would allow the investors to sell the newly bought IVV at a premium when the price of IVV finally catches up with SPY.

2.2.2 Price Discovery in S&P 500 ETFs (SPY and IVV): Previous Literature and Research Agenda

Marshal et al (2013) argue that even though SPY and IVV are not perfect substitutes but investors consider them as close substitutes and when mispricing is allowed, an arbitrage opportunity is created between them. They also point out that one possible source of these ETFs being mispriced is the difference in price discovery between these two assets. More importantly, their study finds that the prices of SPY and IVV do not diverge from each other and whenever there is significant dispersion between the two prices, arbitrage opportunities make the two prices converge to each other. As a result, the prices are co-integrated with co-integrating vector, $\theta = [1, -1]'$. This also makes them an ideal candidate to analyze their price discovery in Vector Error Correction Model (VECM) as done earlier by Hasbrouck (1995).

Fang and Sanger (2012) examine price discovery across SPY, IVV and the reconstructed price series for the S&P 500 index. The constructed price series of the S&P 500 index captures the second-by-second price movement in the underlying securities for S&P 500 ETFs. They find that SPY and IVV contribute half of the price discovery share compared to their underlying stocks. They also look into price discovery across SPY and IVV using Hasbrouck's (1995) IS and find that both of them contribute equally in price discovery. A major concern with their findings is the lack of interpretation in their reported results. For example, in Table 1 of their paper, they report that during 2006 Q4, SPY contributes 51.4%, IVV contributes 52.4% and underlying component assets contribute 29.5%. The contributions add up to more than 100%. Similarly, in Table 2, when they look into SPY and IVV, they report SPY contributes 85.5% and IVV contributes 82.4% of the price discovery in the whole sample period.

Our analysis is motivated by the results in Marshal et al (2012) and Fang and Sanger (2012). We use PDS to give a clean decomposition of price discovery between SPY and IVV on a selected of set of trading days across a number of exchanges. Our objective is to uncover the price discovery contributions of these two ETFs in normal and unusual trading environments. In other words, we seek to evaluate the relative importance of these two ETFs as a source of information for the traders under different market conditions.

2.2.3 *Price Discovery in S&P 500 ETFs (SPY and IVV): Interpretation of PDS Estimates*

When traders are buying or selling either SPY or IVV, they are essentially betting on their perceived forecast regarding future market performance. Any permanent change in S&P 500 ETF prices reflect the arrival and absorption of new information regarding the future movement of the market index. Price discovery across different S&P 500 ETFs describes this process.

To interpret our price discovery measure, it is important to define the set of information available to the traders of S&P 500 ETFs. According to Hasbrouck's (2003), the price vector P_t used in the cointegration analysis should be a comprehensive set of prices which serves as an information set that all the traders possess at time t and also 'a poor proxy for common public information'. The observable public information that a typical S&P 500 ETF trader has are the bid and ask price and trade price of the ETFs. We leave out the trade or transaction price from the information set of SPY and IVV in order to reduce the micro-structure noise (e.g. bid-ask bounce) in these prices. Following the previous literature, we only include the bid-ask mid-point of the two ETFs to construct the information set, so that $P_t = [P_{SPY,Bid-Ask\ mid,t}, P_{IVV,Bid-ask\ mid,t}]$. We use this price vector, P_t in the VECM in equation (9) and use the estimated parameters to derive PDS from equation (14).

Given our definition of P_t, it is straightforward to interpret the price discovery share of individual ETF. Suppose, SPY is estimated to have an X% price discovery share (according to IS or PDS), it can be interpreted in following ways:

1. SPY contributes X% of the volatility of the innovation to the common random walk efficient price.

`

2. If X > 50%, then SPY will be dominant in price discovery. In other words, SPY will be the price leader. Therefore, SPY price will be the first to adjust to a new information about the fundamental value. It will be considered as the more important source of information regarding future movement in market index or the overall market performance.

Based on the analysis in Yan and Zivot (2010), Putninš (2013) argues that an asset price which is first to adjust (the price leader) also has the potential to be more noisy than the follower. When the noise level differs significantly in two prices, IS may end up assigning higher price discovery share to the less noisy asset price even though it may not be the price leader. As a result, IS may identify the wrong asset as the price leader. Our measure, PDS shares this property as IS. However, in our application to S&P 500 ETFs the noise levels in the prices of SPY and IVV are found to be very similar and small in magnitude so that it is unlikely that IS or PDS will misidentify the price leader due to difference in noise level.

2.2.4 Price Discovery in S&P 500 ETFs (SPY and IVV): Data Description and Descriptive Statistics

We use the NYSE TAQ database as our source of high-frequency quotes for SPY and IVV. We choose two snapshots of data in two distinct trading environment. For a normal trading period, we choose intra-day quotes data from Dec 3rd to Dec 7th in 2012. For an extremely volatile trading period, we choose the quotes data of May 6th in 2010, the day of "Flash Crash" and Aug 8th, 2011, the day of the worst stock market fall in US since 2008.

For Dec 3rd - Dec 5th, 2012 we collect data from the following stock exchanges- BATS, Nasdaq, Arca, CBOE, NSX, Boston, Philadelphia and EDGE A. For May 6th, 2010 and Aug 8th, 2011 we collect dataset for BATS, Nasdaq and Arca. The reason for excluding the rest of the stock exchanges is either SPY or IVV are not traded in these stock exchanges or the trading frequencies are too low.

Table 2.3 reports the S&P volatility index, VIX sometimes called the "fear index" for the sample considered. The average of closing price of the VIX during the normal trading period is 16.53. In contrast, VIX on May 6th, 2010 was twice as large and on Aug 8th, 2011 it was three times as large (closing VIX=48) than those during normal period.

On May 6th, 2010 the US stock market experienced an abrupt crash. The abnormal plunge in the market index was first seen at 2:42 pm and the fall in prices continued for next 20 minutes. The Dow Jones Industrial Average experienced the biggest one-day point decline during that day.

August 11th, 2011 is considered to be the worst day in Wall Street since the crisis of 2008. All three major stock market indexes (S&P 500, Dow Jones Industrial Average, NASDAQ composite) fell sharply (between 5% to 7%) during that day. The day was also known for the wide-spread panic among the investors regarding the US losing its AAA credit rating on Aug 6th, 2011.

Tick-by-tick raw trade and quote data typically contain numerous types of data errors and need to be thoroughly cleaned prior to being analyzed. We use the data cleaning procedure recommended for the TAQ data described by Barndorff-Nielsen et al (2008) and implemented in the R package “highfrequency”. Data-cleaning steps for the bid-ask quotes involved the following-

1. Restrict data to exchange hours (9:30 am to 4:30 pm)
2. Delete entries with zero quotes.
3. Delete entries with negative spreads
4. Delete entries if spread > maximum*median daily spread
5. Delete entries for which the mid-quote is outlying with respect to surrounding entries
6. Restrict data to a specific exchange for analysis. After this step we get intra-day time-series dataset of SPY quotes and IVV quotes for each stock exchange separately.
7. For each stock exchange dataset, we delete entries with same time stamp and use median quotes.
8. In each stock exchange dataset, when there is a time-stamp with no ask/bid price reported for it, we use the last observed ask/bid price to replace the missing values. After cleaning for each ETF we have 25201 observations for a given stock exchange and for a given day.

Table 2.4 contains descriptive statistics for the intraday quotes of SPY and IVV in eight different stock exchanges during a normal trading week in December, 2012. We first calculate the average 1 second continuously compounded returns for each day in that week and then report the average of these calculated returns. We do the same for 1 second return volatility, bid-ask spreads and number of shares traded. Table 2.4 reveals that in terms of return and volatility, in every stock exchange except NSX, both ETFs were performing almost similarly. In the National Stock Exchange (NSX), IVV is found to be abnormally volatile compared to SPY and this irregularity is also captured in the spread of their average returns. NASDAQ, BATS and Arca are the exchanges with highest number of trades per day and also have the lowest average bid-ask spread. CBOE, which has the lowest trades per day, also has the highest average bid-ask spread. In each stock exchanges SPY is much more heavily traded than IVV. This was also true in May 6th, 2010 and August 8^{th}, 2011.

Table 2.5 reports descriptive statistics of the day of Flash Crash in 2010. Compared to our sample of a normal trading week, the average 1 second return volatility is almost 1000 times higher. The average 1 second returns are also 10 times lower than our previous sample. NASDAQ, BATS and Arca are again the stock exchanges with the highest number of shares traded and lowest average bid-ask spread. In these three exchanges the return volatility for IVV is much higher than that of SPY.

Table 2.6 reports the descriptive statistics of the day of stock market fall in August 8^{th} , 2011. The loss in 1-sec return is on average 100 times larger than that of regular trading period. The 1-sec return volatility is also high for both SPY and IVV in all the three stock exchanges. One interesting point here is that the bid-ask spread of both the ETFs were lower compared to May 6^{th}, 2010. This indicates that the market turmoil on August 8^{th} did not have any significant effect on the liquidity of SPY or IVV.

Figure 2.1-2.3 shows the intra-day bid-ask mid quotes of SPY and IVV in NASDAQ during three different days. It is very evident from these figures that both the ETF prices move in tandem throughout the day (i.e. highly cointegrated). The only exception was the short period in the afternoon during the Flash Crash when SPY price plummeted by a significant amount compared to IVV.

2.2.5 *Price Discovery in S&P 500 ETFs (SPY and IVV): Estimation and Results*

We estimate and report PDS for each exchange in Table 2.7 during the two selected periods. For the normal trading period (Dec 3^{rd} – Dec 7^{th}, 2012), we first estimate PDS of SPY and IVV in every day for a given stock exchanges and then report the daily average of PDS. We also do the same for abnormal periods- May 6^{th}, 2010 and Aug 8^{th}, 2011.

The third column of Table 2.7 reports PDS for the normal trading week in December, 2012 in eight stock exchanges. SPY leads IVV in price discovery in every stock exchange considered. On average, SPY contributes 61.25% of the price discovery compared to IVV across all stock exchanges. Interestingly, for the NASDAQ and Philadelphia, IVV price is found to be an almost equally important source of trade-related information as it contributes 47% of the price discovery. Although, the daily average PDS is indicating that SPY is leading IVV everywhere, there are days for which IVV beats SPY in price discovery shares. Among the 40 cases (eight stock exchanges in five days), IVV contributes more to price discovery than SPY in 9 occasions. The fourth column reports PDS across SPY and IVV on the day of Flash Crash in NASDAQ, BATS and Arca. We find that SPY leads IVV in every stock exchange by a large margin. On an average, in every stock exchange SPY contributed 95% of the price discovery. The fifth column reports the result during August 11, 2011. Here, we again find that in each market SPY contributes a major portion to price discovery compared to IVV. On an average, the contribution to price discovery of SPY in each stock exchange was 75%.

We interpret these results in the following way. Our goal is to check whether 'duplication' in S&P 500 ETFs provides the traders with a better source of trade related information. Looking at three snapshots in two different environments, we find the consistent result. Although there is more than one ETF tracking S&P 500 index, traders still consider SPY to be a better source of information. This dependence becomes extreme or significantly high in the event of abnormal jumps or falls in intra-day prices (Flash Crash, 2010 or August stock market fall in 2011). Particularly, on the day of Flash Crash, IVV price is found to be very less informative as far as capturing the new information in a timely manner is concerned.

A hypothesis that we think is feasible to explain this outcome is the attributes of the buyers of ETFs. The answer to the question that "Who buys S&P 500 ETFs and Why?" can give us a much better idea to explain the dominance of one particular S&P 500 ETF (namely SPY) in price discovery. If more informed traders choose to trade in one particular S&P 500 ETF most of the time, then by definition that ETF should contribute more to price discovery. Therefore, we think that the distribution of institutional and retail buyers of S&P 50 ETF and their objectives (eg. hedging against market makers) behind investing in these ETFs can help us understand the dominance of SPY over IVV. In previous literature, liquidity of assets has been found to be an important factor in explaining the price discovery contribution. However, in this case, we have observed that both ETFs are highly liquid and their liquidity measure like bid-ask spread are very close to each other in both normal and volatile conditions.

Trading volume is also a key factor behind the dominance of SPY. Analyzing the descriptive statistics tables (Table 2.4, 2.5 and 2.6), we find that IVV is comparatively highly traded in three stock exchanges- NASDAQ, BATS and Arca. However, even in these stock exchanges, during the normal period, the number of shares traded of IVV is only about 5% of that of SPY. During the volatile period, this ratio becomes less than 1%. This complete dominance of SPY over IVV in terms of trading volume can also explain the higher price discovery contribution of SPY.

2.3 Empirical Application: Comparative Assessment between PDS and IS

In this section we provide a comparative assessment between our measure, PDS and IS. First, we show their difference in a setting which is similar to first empirical application. That is, measuring price discovery across SPY and IVV in different stock exchange. But here we only consider one day from the normal trading period to show our result. Next, we demonstrate a more conventional application of price discovery which is analyzing price discovery of a cross listed asset. We pick SPY as the cross listed asset and measure price discovery between NASDAQ and BATS where it is most heavily traded. We report our result using different time intervals and compare the estimates of PDS and upper bound, lower bound and mean of IS.

2.3.1 *Comparative Assessment between PDS and IS: S&P 500 ETF*

Here we utilize a subset of our previous results to capture comparison between IS and PDS. A particular drawback of IS, as we have discussed earlier is that it can reports the contribution of a given asset price to the price discovery in a range and this range can be so wide that it becomes very difficult to make a meaningful inference. Hasbrouck (1995) recognized this lack of identification issue with IS and proposed the use of high-frequency data (e.g. quotes at every second) to get a tighter bound in IS. Here we look into the price discovery contributions between

SPY and IVV on December 3rd, 2012 which was a normal trading day. We find that even with high frequency 1-sec interval quotes data, IS produces a very large range.

We look into the bid-ask mid-quotes of these as before and estimating the VECM equation to derive the parameters necessary for the calculation of IS and PDS. While calculating IS, it is important note which of the two prices was placed first in the price vector. We first put bid-ask quotes SPY as the first element in the price vector and calculate the upper (lower) bound IS for SPY (IVV). Then we put bid-ask quotes of IVV as the first element and calculate the IS again. This is the lower (upper) bound IS for SPY (IVV). We also take the average of these two bounds and report it. We call it "IS-mean" which is an ad-hoc methodology proposed by Baillie et al (2002) for a unique measure of IS.

The 3rd and 4th column of Table 2.8 reports the upper and lower bound of IS. If the difference or the range between two bounds get larger, we face the identification problem in determining the price leader out of the two. A closer look at these two columns tell us that in the case NASDAQ, Arca, Boston, Philadelphia and EDGE A, the upper bound of IS for each ETF identifies itself as price "leader" and the lower bound of the same ETF identifies itself as a "follower". On the other hand, our measure, PDS gives a clean decomposition, reports a unique value for each ETF and more importantly, identifies the price "leader" correctly in every occasion. The IS-mean also identifies the same leaders as PDS does but the contribution reported by IS-mean is found to be under-estimated in every case. Particularly, in the case of BATS and CBOE, IS-mean under-estimates the price discovery contributions by more than 10%. We also estimate the standard error for both measures (IS and PDS) in every case by bootstrap method[16]. For both measure, the reported boot-strap standard errors are very low and we also do not find a significant difference between IS and PDS in terms of standard errors reported.

[16] Bootstrap method is done following Grammig et al (2005). First, VECM parameters are estimated and we derive the estimated residuals from the difference of actual data and fitted data. Estimated residuals are used to simulate price series using original price series as starting values. The price discovery measure is calculated from the simulated data and the whole process is repeated 1000 times to get the mean and standard error of IS and PDS.

2.3.2 Comparative Assessment between PDS and IS: Sampling at different frequencies

Another drawback of Hasbrouck's IS is that the gap between upper and lower bound gets extremely large if data at higher interval is used for analysis. Gramming et al. (2005), Theissen (2002) and Huang (2002) –all of these studies while studying quotes data with more than 1-sec interval found a substantial contemporaneous residual correlation and a wide upper and lower bounds for the IS.

Table 2.9 reports the results of an application where we focus on a single cross-listed stock (SPY) and the price discovery contribution between two competing stock exchanges (NASDAQ and BATS). We start our analysis with the tick-by-tick data (1 second interval) and then continue to calculate IS and PDS at lower frequency observations (data with 5 seconds-, 10 seconds-, 20 seconds-, 30 seconds, 40 seconds- , 50 seconds- , 1 minute- , 1 minute 30 seconds, 2 minutes-interval). We find that moving from high to lower frequency data, the gap between lower and upper bound Hasbrouck's (1995) IS becomes very large. We also report the IS-mean and find that for all the time intervals, IS-mean identifies both stock exchange contributing almost equally to the price discovery. On the hand, according to PDS, NASDAQ is the price "leader" in every sample with different time interval. This result proves that our measure PDS, unlike IS, is robust to the choice of time interval between two data-points in a data set and always gives a consistent result.

2.4 Conclusion

Our empirical application investigates the new financial market phenomena of "duplication of ETFs". Our paper contributes both into the literature of price discovery and the ETFs by showing presence of multiple near-identical ETFs do not result in creating equally important alternative sources of information. We have analyzed the price discovery across SPY and IVV, two popular and highly traded S&P 500 ETFs in two different trading conditions-normal and extremely volatile. SPY is found to be information leader in both cases. This indicates that despite the existence multiple S&P 500 ETFs, a stock market trader treats the quotes of SPY as a dominant source of information. This dominance is found to be absolutely leaning towards SPY during extreme volatile trading condition a.k.a. flash crash, 2010.

This application also provides a useful information for the ETF investors. As there are numerous ETFs that track the S&P 500 index, investors usually need to choose an ETF for their investment portfolio which works for them the best. Low tracking error, low expense ratio and low tax burdens are the common determinants behind this decision. ETF's price discovery contribution can also become important consideration for choosing an ETF for an investor's portfolio. In addition, if a clear pattern of price leadership is identified between two competing ETFs in a particular stock exchange, it may be possible for arbitragers to make profits by adopting pair-trading. In both cases, our new measure and our application can provide the investors with useful information regarding ETFs.

In our second application we provide comparative assessment between IS and PDS. We show that even using one second interval quotes, IS can report uninformative results. By contrast, PDS always gives clean decomposition of price discovery contribution and identifies the information leader correctly. In a separate application, we also demonstrate PDS's robustness to the use of quotes with high time intervals.

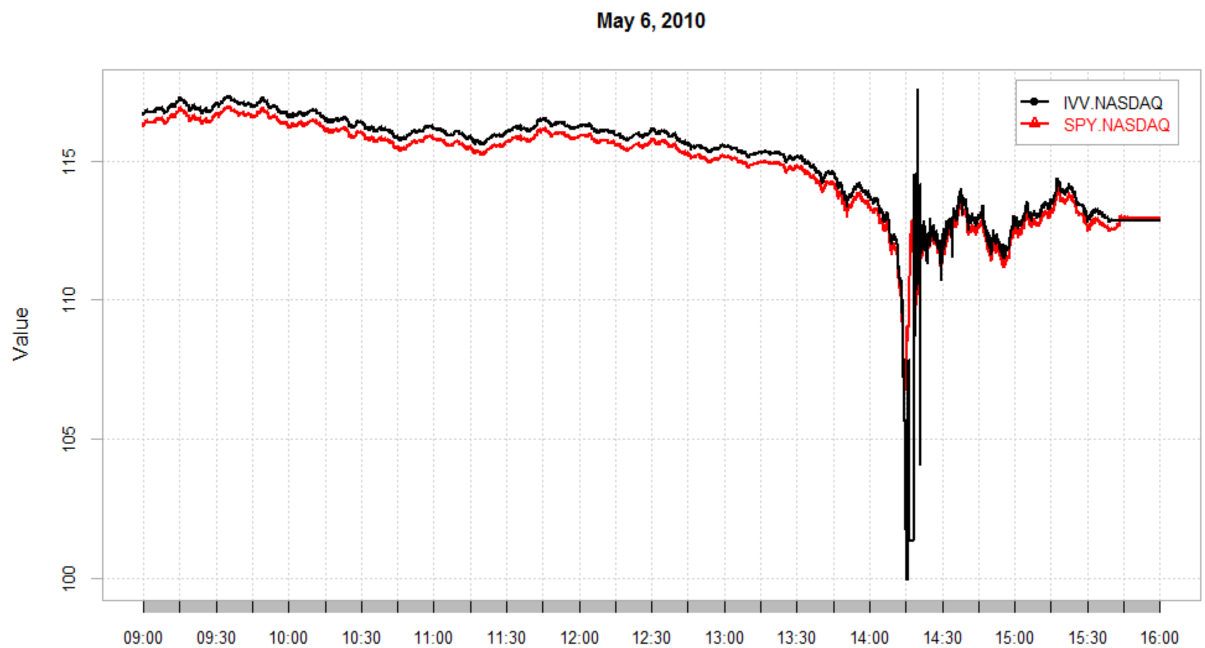

Figure 2.1: Mid-Quotes of SPY and IVV at NASDAQ on May 6, 2010 (from 9:30 am- 16:30 pm)

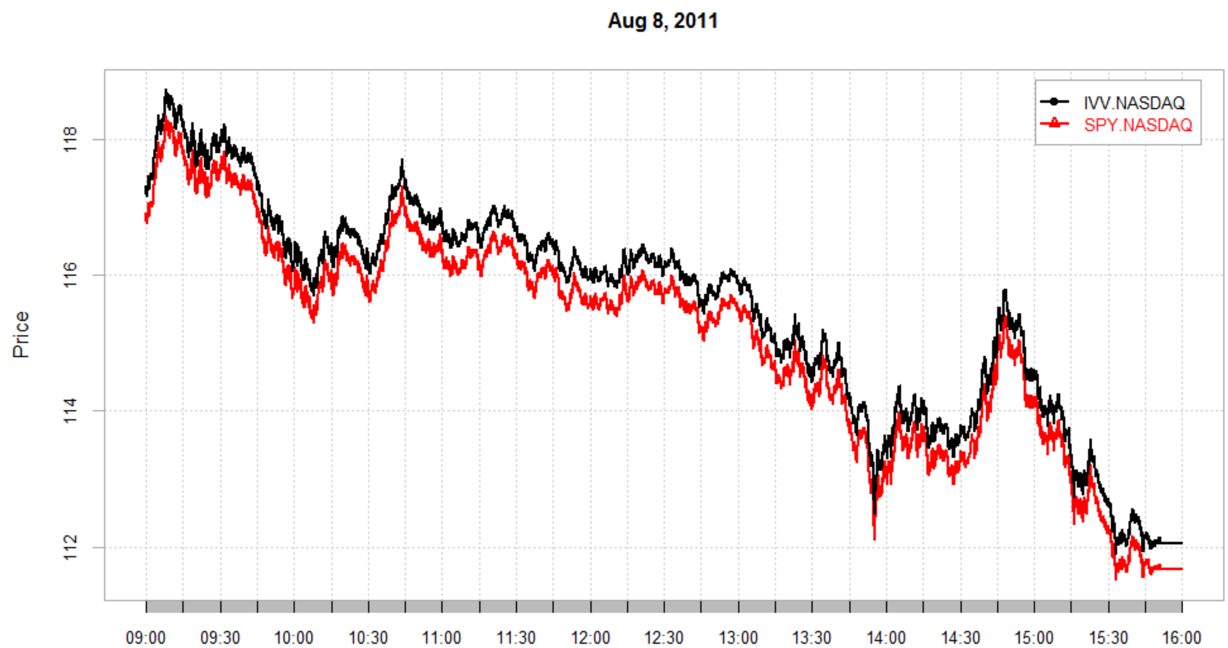

Figure 2.2: Mid-Quotes of SPY and IVV at NASDAQ on Aug 8, 2011 (from 9:30 am- 16:30 pm)

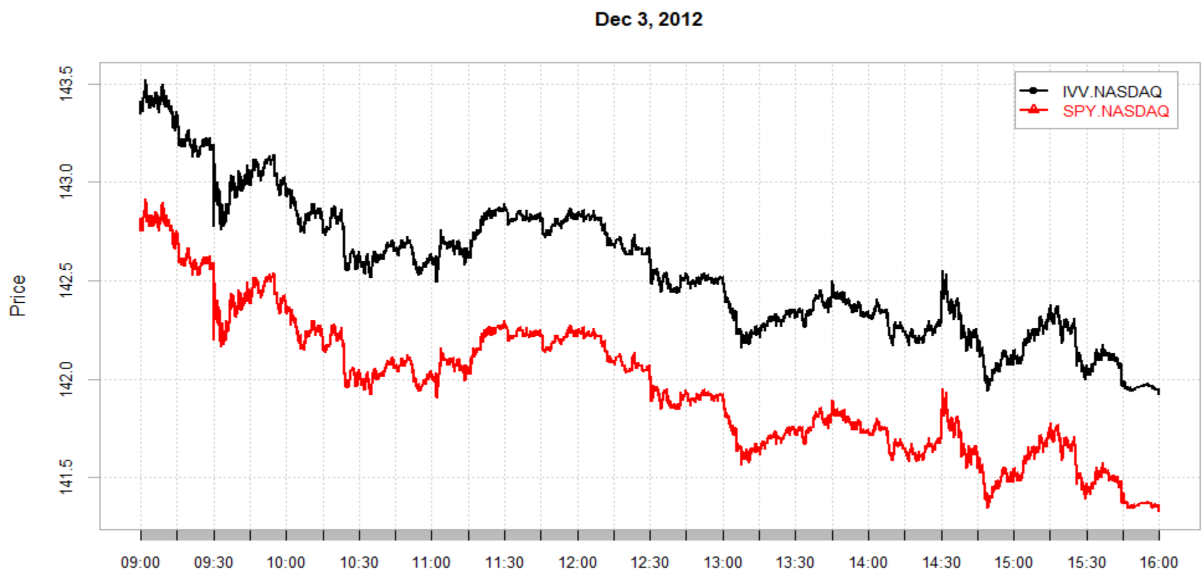

Figure 2.3: Mid-Quotes of SPY and IVV at NASDAQ on Dec 3, 2012 (from 9:30 am- 16:30 pm)

Table 2.1: Comparison between SPY and IVV (Overview and Performance)

Overview	SPY	IVV	VOO
Issuer	State Street SPDR	iShares	Vanguard
Inception	22,Jan-93	15,May-2000	7, Sept-2010
Asset Under Management	$165,308.6 M	$61,743.0 M	$21,794.6 M
Shares Outstanding	868.6 M	322.3 M	124.9 M
Expense Ratio	0.09%	0.07%	0.05%

Performance Comparison	SPY	IVV	VOO
1 Week Return	3.39%	3.36%	3.38%
4 Weeks Return	-1.99%	-2.01%	-1.99%
26 Weeks Return	4.27%	4.32%	4.33%
1 Year Return	13.44%	13.53%	13.48%
5 Year Return	96.65%	96.05%	n/a
Beta	1.00	0.99	0.99
P/E ratio	15.24	15.24	15.37
Annual Dividend Rate	$3.68	$3.64	$3.38
Annual Dividend Yield	1.93%	1.90%	1.94%
5 day Volatility	18.23%	18.56%	20.63%
200 day Volatility	10.93%	11.01%	11.20%

Source: ETF Database. Web link: http://etfdb.com/tool/etf-comparison/IVV-SPY/, All the results are reported on October 21st , 2014

Table 2.2: Comparison between SPY and IVV (Holdings and Sector Breakdown)

	SPY	IVV	VOO
Top 10 Holdings			
Apple Inc	3.43%	3.44%	3.46%
Exxon Mobil Corporation	2.28%	2.28%	2.39%
Microsoft Corporation	2.17%	2.18%	1.91%
Johnson & Johnson	1.71%	1.71%	1.64%
General Electric Co	1.46%	1.46%	1.46%
Berkshire Hathaway Inc Class B	1.43%	1.43%	1.27%
Wells Fargo & Co	1.40%	1.40%	1.38%
Procter & Gamble Co	1.29%	1.29%	1.26%
Chevron Corp	1.29%	1.29%	1.38%
JPMorgan Chase & Co	1.29%	1.29%	1.26%
Sector Breakdown			
Technology	17.94%	17.98%	17.64%
Financial Services	14.89%	14.92%	14.70%
Health Care	14.24%	14.28%	14.02%
Industrials	10.99%	10.99%	10.89%
Consumer Cyclical	10.17%	10.19%	10.37%
Energy	9.65%	9.67%	10.34%
Consumer Defensive	9.47%	9.49%	9.33%
Communication Services	4.01%	4.02%	3.96%
Basic Materials	3.28%	3.29%	3.35%
Utilities	2.97%	2.99%	3.00%
Real Estate	1.91%	1.92%	1.96%
Others	0.48%	0.26%	0.44%

Source: ETF Database. Web link: http://etfdb.com/tool/etf-comparison/IVV-SPY/ . All the results are reported on October 21st , 2014

Table 2.3: S&P 500 volatility index (VIX or the "fear index") in normal trading period (Dec 3-5, 2012) and in abnormal (volatile) trading period (May 6, 2010 and Aug 8, 2011)

Date	Open	High	Low	Close	Adj Close*
Normal Trading Period					
3-Dec,12	15.81	16.69	15.76	16.64	16.64
4-Dec,12	16.66	17.37	16.38	17.12	17.12
5-Dec,12	16.95	17.53	16.27	16.46	16.46
6-Dec,12	16.59	16.85	16.31	16.58	16.58
7-Dec,12	16.12	16.65	15.73	15.87	15.87
Average	16.426	17.018	16.09	16.534	16.534
Abnormal Trading Period					
6-May,10	25.88	40.71	24.43	32.8	32.8
8-Aug,11	36.9	48	35.29	48	48
Average	31.39	44.355	29.86	40.4	40.4

Source : Yahoo! Finance

Table 2.4: Descriptive Statistics of SPY and IVV in different stock exchanges on Dec 3-7, 2012

Stock Exchange	Vectors of Prices	average 1 sec return ($\times 10^{-8}$)	Average 1 sec return volatility ($\times 10^{-8}$)	Intra-day average number of shares traded	Intra-day average bid-ask spread
NASDAQ	SPY	-7.52	3.87	23,960,312	0.01
	IVV	-7.46	4.12	1,155,608	0.02
BATS	SPY	-7.48	4.05	18,935,201	0.01
	IVV	-7.55	4.03	964,979	0.02
Arca	SPY	-7.52	4.02	26,708,530	0.01
	IVV	-7.52	4.04	744,723	0.03
CBOE	SPY	-6.92	6.12	331,915	0.06
	IVV	-9.28	4.62	2980	0.06
NSX	SPY	5.97	7.62	235,790	0.02
	IVV	-9.49	36.80	4973	0.03
Boston	SPY	-7.58	3.99	3,018,644	0.02
	IVV	-7.57	4.23	107,219	0.04
Philadelphia	SPY	-7.78	4.33	1,963,004	0.02
	IVV	-7.49	4.12	123,952	0.03
EDGE A	SPY	-7.71	4.04	2,204,523	0.02
	IVV	-7.60	4.89	63,747	0.06

Table 2.5: Descriptive Statistics of SPY and IVV in different stock exchanges on May 6, 2010 (Flash-crash)

Stock Exchange	Vectors of Prices	Average 1 sec return ($\times 10^{-8}$) On May 6th, 2010	Average 1 sec return volatility, May 6th, 2010	Number of shares traded, May 6th, 2010	Intra-day average bid-ask spread, May 6th, 2010
Nasdaq	SPY	-58.0	0.0001	201,085,629	0.02
	IVV	-67.0	0.001	3,754,730	0.09
BATS	SPY	-59.0	0.0002	99,711,462	0.02
	IVV	-67.0	0.0005	3,712,119	0.07
Arca	SPY	-58.0	0.0001	145,771,969	0.02
	IVV	-59.0	0.0009	2,771,224	0.07
CBOE	SPY	-82.0	0.0002	5,257,700	0.11
	IVV	-21.0	0.0003	1700	2.28

Table 2.6: Descriptive Statistics of SPY and IVV in different stock exchanges on Aug 8, 2011

Stock Exchange	Vectors of Prices	Average 1 sec return ($\times 10^{-8}$) On Aug 8th, 2011	Average 1 sec return volatility ($\times 10^{-5}$), Aug 8th, 2011	Number of shares traded, Aug 8th, 2011	Intra-day average bid-ask spread, Aug 8th, 2011
Nasdaq	SPY	-180.0	14.0	151,577,895	0.01
	IVV	-181.0	13.0	4,760,661	0.04
BATS	SPY	-181.0	14.0	128,578,031	0.01
	IVV	-181.0	13.0	3,245,591	0.04
Arca	SPY	-181.0	14.0	140,039,777	0.01
	IVV	-181.0	13.0	3,660,737	0.04

Table 2.7: PDS between SPY and IVV in different stock exchanges

Stock Exchange	Vectors of Prices	Daily average of PDS on Dec 3rd -7th, 2012	PDS on May 6th, 2010 (Flash-Crash)	PDS on Aug 8th, 2011
NASDAQ	SPY	0.53	0.92 (0.002)	0.83 (0.009)
	IVV	0.47	008 (0.002)	0.17 (0.009)
BATS	SPY	0.59	0.99 (0.005)	0.62 (0.012)
	IVV	0.41	0.01 (0.005)	0.38 (0.012)
Arca	SPY	0.62	0.94 (0.005)	0.79 (0.016)
	IVV	0.38	0.06 (0.005)	0.21 (0.016)
CBOE	SPY	0.56		
	IVV	0.44		
NSX	SPY	0.69		
	IVV	0.31		
Boston	SPY	0.58		
	IVV	0.42		
Philadelphia	SPY	0.53		
	IVV	0.47		
EDGE A	SPY	0.80		
	IVV	0.20		

Table 2.8: IS and PDS across SPY and IVV in the same stock exchange on Dec 3, 2012.

Note: Bootstrap standard errors are reported in parenthesis

Vector of Prices	Stock Exchanges	Hasbrouck Information Share -Upper bound	Hasbrouck Information Share -Lower bound	Information Share- Mean	PDS
NASDAQ	SPY	0.9266	0.1325	0.5295	0.5813
		(0.023)	(0.023)	(0.023)	(0.023)
	IVV	0.8675	0.0734	0.4705	0.4187
		(0.023)	(0.023)	(0.023)	(0.023)
BATS	SPY	0.9867	0.5736	0.7802	0.9093
		(0.023)	(0.023)	(0.023)	(0.023)
	IVV	0.4264	0.0133	0.2198	0.0907
		(0.023)	(0.022)	(0.022)	(0.023)
Arca	SPY	0.9236	0.1427	0.5331	0.5865
		(0.01)	(0.01)	(0.01)	(0.01)
	IVV	0.8573	0.0764	0.4669	0.4134
		(0.01)	(0.01)	(0.01)	(0.01)
Chicago Board Option Exchange (CBOE)	SPY	0.9675	0.5213	0.7444	0.8506
		(0.021)	(0.02)	(0.02)	(0.02)
	IVV	0.4787	0.0325	0.2556	0.1494
		(0.02)	(0.02)	(0.02)	(0.02)
National Stock Exchange (NSX)	SPY	0.999	0.996	0.998	0.999
		(0.001)	(0.001)	(0.001)	(0.001)
	IVV	0.004	0.001	0.002	0.001
		(0.001)	(0.001)	(0.001)	(0.001)
Boston Stock Exchange	SPY	0.7810	0.1088	0.4449	0.3975
		(0.021)	(0.021)	(0.021)	(0.021)
	IVV	0.8912	0.2190	0.5551	0.6025
		(0.021)	(0.021)	(0.021)	(0.021)
Philadelphia Stock Exchange	SPY	0.8720	0.1982	0.5351	0.5647
		(0.022)	(0.023)	(0.022)	(0.022)
	IVV	0.8018	0.1280	0.4649	0.4353
		(0.023)	(0.023)	(0.023)	(0.023)
EDGE A Stock Exchange	SPY	0.8676	0.4180	0.6428	0.6845
		(0.02)	(0.02)	(0.02)	(0.02)
	IVV	0.5820	0.1324	0.3572	0.3155
		(0.021)	(0.021)	(0.021)	(0.021)

Table 2.9: Price discovery measures (IS and PDS) in log of bid-ask mid-quotes of SPY in BATS and NASDAQ (from high frequency to low frequency data). Note: bootstrap s.e. in parenthesis.

Time-lag between each observation	Exchanges	Hasbrouck Information Share -Upper bound	Hasbrouck Information Share -Lower bound	Information Share- Mean	PDS
1 second interval	BATS	0.9215 (0.02)	0.0227 (0.02)	0.4721 (0.02)	0.3428 (0.02)
	NASDAQ	0.9773 (0.02)	.0785 (0.02)	0.5279 (0.02)	0.6572 (0.02)
5 seconds interval	BATS	0.9656 (0.05)	0.01 (0.04)	0.4828 (0.05)	0.0162 (0.05)
	NASDAQ	0.99 (0.05)	0.0344 (0.05)	0.5172 (0.05)	0.9838 (0.05)
10 seconds interval	BATS	0.98 (0.05)	0.01 (0.05)	0.4922 (0.05)	0.1028 (0.05)
	NASDAQ	0.99 (0.05)	0.02 (0.05)	0.5078 (0.05)	0.8972 (0.05)
20 seconds interval	BATS	0.99 (0.09)	0.001 (0.09)	0.4963 (0.09)	0.1821 (0.09)
	NASDAQ	0.999 (0.09)	0.01 (0.09)	0.5037 (0.09)	0.8179 (0.09)
30 Seconds interval	BATS	0.995 (0.10)	0.001 (0.10)	0.4978 (0.10)	0.2019 (0.10)
	NASDAQ	0.999 (0.10)	0.005 (0.10)	0.5022 (0.10)	0.7981 (0.10)
40 seconds interval	BATS	0.995 (0.11)	0.001 (0.11)	0.4978 (0.11)	0.1353 (0.11)
	NASDAQ	0.999 (0.11)	0.005 (0.11)	0.5022 (0.11)	0.8647 (0.11)
50 seconds interval	BATS	0.998 (0.11)	0.002 (0.11)	0.4999 (0.11)	0.4726 (0.11)
	NASDAQ	0.998 (0.11)	0.002 (0.11)	0.5001 (0.11)	0.5274 (0.11)
1 minute interval	BATS	0.998 (0.12)	0.001 (0.12)	0.4993 (0.12)	0.3276 (0.12)
	NASDAQ	0.999 (0.12)	0.002 (0.12)	0.5007 (0.12)	0.6724 (0.12)
1.5 minutes interval	BATS	0.996 (0.12)	0.001 (0.12)	0.4983 (0.12)	0.0283 (0.12)
	NASDAQ	0.999 (0.12)	0.004 (0.12)	0.5017 (0.12)	0.9717 (0.12)
2 minutes interval	BATS	0.999 (0.15)	0.001 (0.15)	0.50 (0.15)	0.1168 (0.15)
	NASDAQ	0.999 (0.15)	0.001 (0.15)	0.50 (01.15)	0.8832 (0.15)

Chapter 3

A Study on Bond Exchange-Traded Funds (ETFs) and Corporate Bond Liquidity

3.1 Introduction

Since the credit crunch during the financial crisis of 2008, the issue of corporate bond market liquidity has caught the attention of academics and policy-makers alike. In addition to this, the rising popularity of different basket securities for bonds is also raising new concerns regarding their possible effect on the constituent bonds market. In this paper, we study and analyze the effect of different basket securities of corporate bonds (e.g., bond Exchange-traded Funds and bond mutual funds) on the liquidity of their underlying corporate bonds. Given the unique features of bond Exchange-traded Funds (bond ETFs), we also give special attention to these new financial instruments.

We use different well-known measures for bond liquidity and study how the exposure of corporate bonds to ETFs and mutual funds affects these measures. We conduct this study in four different ways. First we use full set of sample which includes all kinds of corporate bonds. Second, we use a sample containing only ETF-bonds. Lastly, we use two subsamples – one containing only investment grade bonds and the other only high yield bonds.

Even a decade ago, corporate bonds could only be bought and sold in the over-the-counter (OTC) market with the help of dealers. Typically, the corporate bond market tends to be very illiquid, that is, it is often very difficult to trade corporate bonds. This is because a common practice in this market is to hold these bonds until maturity and this leads to a higher number of zero trading days and high bid-ask spreads in this market. With the advent of new financial vehicles like bond ETFs, this situation has been rapidly changing. Buyers and investors of corporate bonds now have ready access to these highly illiquid assets through bond ETFs. This has allowed them to trade in bonds indirectly throughout the day in any stock exchange. In addition to this, the low-expense ratio, low management fees, and high tax-efficiency make bond ETFs very attractive. A natural research agenda that arises from this trend is to understand the nature of the effect of bond ETFs on their underlying corporate bond liquidity.

Figure 3.1 shows the cross-sectional average over time for the major investors of corporate bonds: insurance companies, mutual funds, pension funds, and ETFs. It is clearly evident from Figure 3.1 that the bond ETFs' ownership of corporate bonds compared to other investors is still very low (around 1%). This has motivated us to take a closer look into bond mutual funds as well. In this paper we look into the effects of both ETFs and mutual funds together (as a proxy for the basket security ownership of bonds) and we also analyze their effects separately.

Bond mutual funds are similar to bond ETFs in the sense that they give investors the opportunity to invest with a diverse exposure to various sectors of the fixed income market. ETFs are usually considered as a subset of the mutual fund asset class (close-ended mutual funds). Although on average a bond mutual fund entails a comparatively higher expense ratio and management fees, it enjoys more flexibility compared to its ETF counterparts in terms of investing and repositioning its holdings when required.

Two contrasting views are primarily found in the previous literature regarding the perceived effect of basket securities on their underlying assets' liquidity. The theoretical studies done by Kyle (1985), Gammill and Perold (1989), Subrahmanyam (1991) and Gorton and Pennacchi (1993) support the "negative impact" view. A common hypothesis that all these studies postulate is that the introduction of basket securities encourages less informed traders to migrate from underlying markets to basket security markets. The informational disadvantage for a less informed trader dissipates in the basket security market, because the importance of better knowledge about a specific asset has less value in a diversified basket security market. The exit of less informed traders increases the adverse selection cost in underlying markets and in turn results in lower liquidity of underlying assets. Subrahmanyam (1991) and Gorton and Pennacchi (1993) further suggest that underlying assets that have lower weights in the newly introduced basket securities are more vulnerable to the resulting negative liquidity shock.

All the above models, however, rely on a strong assumption that different markets are perfectly integrated as far as information is concerned (Hegde and McDermott, 2004). The "positive impact" view is primarily supported by theoretical models that assume imperfect integration of information among markets and segmented financial markets. Fremault (1991), Kumar and Seppi (1994) and Holden (1995) utilize such models to demonstrate the positive impact of introducing a basket security into the market. The common focus in these models is the role of arbitrage opportunity between the two markets – the market for basket security and the market for constituent securities. These models find that the arbitrage opportunity reduces the information asymmetry across markets, reduces the temporary order imbalance across markets, increases the competition among the informed cross-market arbitrageurs, and reduces the arbitrage cost. All of these outcomes affect the liquidity in both markets in a positive way.

Support for the positive impact view of basket securities can also be found in a different theoretical setting of Merton (1986). This is also known as the recognition hypothesis. The hypothesis suggests that when an asset is included in a newly formed basket security, it gets recognized by the investors in a positive way and results in a broader investor base for that particular underlying asset. This ultimately results in higher liquidity for the underlying asset.

The empirical research has evolved into two schools of thought. Hamm (2010) finds modest evidence in support of an increase in the adverse selection cost of underlying stocks after they are held by newly introduced diversified ETFs. Dannhauser (2014) finds that ETF activity has a significant negative impact on investment grade bond liquidity and a statistically insignificant effect on high yield bond liquidity. Alternatively, Hegde and McDermott (2003) find that inclusion of new underlying securities into the S&P 500 index enhances the liquidity of those new entrants. Hegde and McDermott (2004) and Richie and Madura (2007) find that the inception of ETFs like Diamond and QQQ has a positive impact on their underlying securities in terms of liquidity. Using data from the French Stock Exchange, Winne, Gresse, and Platten (2011) find that index-stock spreads are lower for underlying stocks than those of non-index stocks after the introduction of the first ETF which was created to track CAC 40 index.

Apart from Dannhauser (2014), the empirical studies have mostly focused on equity ETFs. Our paper contributes to the existing literature by conducting an exhaustive investigation on bond ETFs and bond funds and their effects on underlying corporate bonds. Dannhauser (2014) does an excellent job in motivating the research question about bond ETFs and proposing a systematic approach to evaluate their impact on the constituents in terms of yield spread and liquidity. Nevertheless, the short range of her sample period (2009Q1 to 2013Q4) cannot tell us the implication of bond ETFs during the financial crisis of 2008. Furthermore, if exposure to bond ETFs makes the underlying bonds less liquid, then high yield bonds are more likely to be affected by this negative liquidity shock than are investment grade bonds. In this light, Dannhauser (2014)'s finding that investment grade bonds are more adversely affected in terms of liquidity than high yield bonds due to an increase in ETF ownership of these bonds is very interesting and requires further investigation. Our approach is very similar to that of Dannhauser (2014). However, our paper considers a larger sample period that includes the financial crisis, and we also conduct our investigation using samples with more diverse classification.

We utilize two dataset sources in this paper: the eMaxx database and the TRACE dataset. We access a wide range of investors' holding information of different corporate bonds from the eMaxx database, and from TRACE database we construct different measures of liquidity using various bond price information. We apply both fixed-effect panel regression and 2-way fixed-effect panel regression to study the effect of ETF ownership and mutual fund ownership of corporate bonds on our constructed liquidity measures.

Our first analysis relies on a comprehensive sample that include investment grade and high yield bonds, and ETF- and non-ETF bonds, and which covers the period from the 2002:Q3 to 2014:Q3. We find strong evidence that mutual fund ownership has a strong positive impact in improving corporate bond liquidity. This strong result is robust to the use of different liquidity measures and estimation techniques (fixed or two-way). We find either a positive effect or no effect on bond liquidity due to variation in ETF ownership of bonds. The result differs subject to the liquidity measure and estimation techniques used.

Next we construct a subsample from the complete database that contains only information regarding bonds that are bought and held by different ETFs. In this analysis, we find strong evidence of a positive impact on bond liquidity for both ETFs and mutual funds.

Our final empirical study considers the sample of investment grade bonds and the sample of high yield bonds in isolation. The idea is to assess the impact on corporate bond liquidity due to variation in ETF and mutual fund ownership within the group of bonds with similar credit rating. We find evidence that higher ownership of ETFs and mutual funds lead to an improvement in the liquidity of underlying high yield corporate bonds. However, the statistical significance of this impact varies for ETFs depending on the estimation technique and measure of liquidity used. In contrast, for investment grade bonds we do not find any statistically significant impact for ETFs. The mutual funds are found to have positive impact on bond liquidity only in the case of particular liquidity measure.

The rest of paper is organized as follows: Section 2 contains descriptions of the data and variables; section 3 examines the methodology; section 4 reports the results; and section 5 summarizes our findings and conclusions.

3.2 Description of Data and Variables

This section first describes the dataset used in this paper, and then explains in detail the two important variables used in this paper: measures for bond liquidity and corporate bond ownership of ETFs and mutual funds.

3.2.1 *Data Description*

The primary dataset used in this paper is the Thomson Reuters eMAxx Database. This dataset provides us with quarterly holding information of individual corporate bonds. The holding data are categorized into different kinds of investors, including mutual funds, ETFs, pension funds, and insurance companies. The sample obtained from eMaxx database provides information for 14,827 corporate bonds over the period 1998Q2 to 2014Q3, although information for most bonds does not cover the whole sample period. That is, our panel data is unbalanced in nature. In the eMaxx data, every corporate bond is identified with its unique 8-digit CUSIP and for each bond we have the holding information of 35 categories of investors[17].

ETFs are defined as a mutual funds in this database and, therefore, are not readily identified. To identify the ETFs, first of all, we filter all the mutual funds' issuers' names and CUSIPs. First we identify those ETFs which are cross-listed with the list of bond ETFs (287 bond ETFs) that are available in ETF database[18]. We also manually identify the ETFs by searching for key words in the names of issuers. A point to note here is that there are some bond ETFs which were listed as bond mutual funds during their inception and then later are listed as ETFs. Therefore, we also hand collect the inception date of all of the different bond ETFs that we have identified from the list. We then cross check the inception date with the observable time period in our panel data. The idea is to correctly identify an investor's category (whether it is a bond fund or bond ETF) over time.

---

[17] 1. 13F Filer, 2. 401K, 3. Annuity/Variable Annuity, 4. Annuity/VA - Money Market, 5. Bank-Portfolio, 6. Bank-Savings/Bldg Society, 7. Bank-Trust, 8. Church/Religious Org, 9. Corporation, 10. Credit Union, 11. Finance Company, 12. Foundation/Endowment, 13. Government, 14. Health Care Systems, 15. Hedge Fund, 16. Hospital, 17. Insurance Co-Diversified, 18. Insurance Co-Life/Health, 19. Insurance Co-Prop & Cas, 20. Investment Manager, 21. Mutual Fund – Balanced, 22. Mutual Fund - Money Mkt, 23. MutFd-OE/UnitTr/SICAV/FCP, 24. MutFd-CE/Inv Tr/FCP, 25. Mutual Fund-Equity, 26. Mutual Fund-Fund of Funds, 27. Nuclear De-Comm Trust, 28. Other, 29. Pension Fund-Corporate, 30. Pension Fund-Government, 31. Pension Fund-Union, 32. Reinsurance Company, 33. Small Business Invst Co., 34. Spezial Fund, 35. Unit Investment Trust

[18] ETF database (www.etfdb.com) is a web-based source of news and data on ETFs.

Besides the holding information, the eMaxx dataset also contains quantitative information like the effective date, date of maturity, coupon rate, S&P credit rating, Moody's Credit rating, total bond outstanding, and total amount of bonds offered also in addition to qualitative information like whether a particular bond is callable or not.

We assign different numeric values to S&P's and Moody's credit ratings. For the S&P rating, corporate bonds with aAAA rating get a numeric credit rating value of 1 and a C credit rating gets a value of 21. For Moody's rating, 'AAA' gets a value of 1 and a C credit rating gets a value of 24. We also create two subsamples based on credit rating: investment grade bonds and high yield bonds. For investment grade bonds, we make a subset of the sample based on whether S&P credit rating is less than or equal to 10 (credit rating: 'BBB-') and the rest are sampled as high-yield bonds.

We use the observed date and the date of maturity and their difference to calculate the date to maturity. We express the date to maturity in terms of months. From the eMaxx database, we removethose observations for which we get a negative date of maturity, which implies that the bond information was recorded after the maturity date and is therefore meaningless.

For the liquidity measure, we collect data from Financial Industry Regulatory Authority's (FINRA) TRACE. TRACE contains corporate bonds' trade-related data that is collected by FINRA from the secondary market of corporate bonds. We merge the eMaxx database of holding information with the TRACE database of liquidity by matching the CUSIPs and the observed year-quarter. After merging, we are left with 5594 bonds and 58,308 bond-quarter observations. Table 3.1 summarizes the descriptive statistics of the important variables in these two databases. In our regression analysis, we use the merged dataset.

3.2.2 Variable Description- ETF ownership and mutual fund ownership of corporate bonds

Using the eMaxx database, we also calculate the ownership of different major investors. We categorize them into four major groups: mutual funds (bond funds), ETFs (bond ETFs), insurance companies, and pension funds. To calculate the ETF ownership, we divide the ETF holding of each category of investors by the total amount of bond outstanding. We follow the same method to calculate bond mutual fund ownership. Due to some incorrect entries in the database, there are observations whose total bond outstanding is much less than the total holdings of all categories of investors[19]. Therefore, if we always use total bond outstanding as the normalizing factor, the percentage contributions often add up to more than 100%. To resolve this issue, we choose the normalizing factor from whichever is the greatest from the total bond outstanding, amount of bonds offered, or the sum of all holdings (or total holdings).

3.2.3 Variable Description – Measures of Bond Liquidity

For the liquidity analysis, we consider several measures of liquidity that have been proposed in different recent studies. As far as liquidity is concerned there is still a consensus to be reached among financial economists. A common practice in the previous studies on bond liquidity is to utilize multiple measures of liquidity to provide a complete picture. Our measures for liquidity are the price impact measure of illiquidity proposed by Amihud (2002); Roll (1984)'s measure of illiquidity; the percentage of zero trading days and zero return days (denoted by 'zeros') proposed by Lesmond, Ogden, and Trzinka (1999) and Chen et al. (2007); the median of imputed round-trip cost (IRC) by Feld$\ddot{u}$tter (2012); and bond's turnover. Next follows a brief description of these five estimators.

i. Amihud (2002)'s measure:

Amihud (2002)'s measure of illiquidity has been heavily used in the recent literature. Let $A_{i,q}$ be the Amihud (2002) illiquidity measure of corporate bond, i at quarter, , $return_{i,t}$ be the return on bond i on the t-th day of quarter q and $volume_{i,t}$ be the trading volume of bond i on the same day. If $N_{i,q}$ is the number of days in quarter q when transaction has been observed or return to volume ratio is available, then

$$A_{i,q} = \frac{1}{N_{i,q}} \sum_{t=1}^{N_{i,q}} \frac{|return_{i,t}|}{Volume_{i,t}} \quad (1)$$

A higher value of $A_{i,q}$would mean lower liquidity of the corporate bond i at quarter q. Figure 3.2 demonstrate cross sectional average of Amihud (2002) measure of illiquidity in each quarter. We show three time plots: for ETF bonds, for non-ETF bonds, and for the total bond market. The comparison between ETF bonds and non-ETF bonds is particularly interesting. Until 2008, the cross sectional averages of Amihud (2002)'s measure for both ETF and non-ETF bonds seem to

[19] We get the total holdings at any time, t by adding the holdings of all the 35 categories of investors at time t.

very close to each other. However, after 2008 there is a dramatic rise in the illiquidity of non-ETF bonds compared to ETF bonds following the collapse of Lehman Brothers.

ii. Roll (1984)'s measure:

Roll (1984)'s measure of illiquidity in terms of realized spread is another widely utilized measured to quantify a bond's liquidity. If any observed transaction price of a bond $'i'$ at time $'t'$, is $P_{i,t}$ then,

$$roll_{i,t} = 2 * \sqrt{-cov(\Delta P_{i,t}, \Delta P_{i,t-1})}$$

One particular drawback of this measure is that the adverse selection component is missing in its construction. The degree of this measure reflects the trading and order processing costs, which many believe to have a transient effect on the prices.

Figure 3 shows the plot for Roll (1984)'s measure of bid-ask spread. The plots look very similar to those of Amihud (2002)'s measure. We also notice the sharp increase in illiquidity during 2008 Q3. Again we observe that ETF bonds have, on average, higher liquidity than the non-ETF bonds. This is particularly true after 2008.

iii. Percentage of Zero Return and Zero Trade days (Zeros)

This is a measure of illiquidity first suggested by Lesmond, Ogden, and Trzinka (1999). Bonds whose prices stay stagnant over long periods or bonds that do not trade for long periods are likely to be less liquid. It is computed as:

$$Zeros_{i,t} = \frac{(Zero\ Return\ Days + Zero\ Trade\ Days)_{i,t}}{Trading\ Days_t} * 100$$

This measure is not without its limitations. Using actual transaction prices during the sub-prime mortgage crisis, Dick-Nielsen et al. (2012) show that there is no consistent relationship between the number of zero trading days and spreads. In fact, when traders holding less liquid bonds are frantically try to get rid of the risky bonds from their portfolio, the zero measure actually decreases, suggesting higher liquidity of bonds during crisis periods.

Figure 3.6 shows the time plot for zeros, or the percentage of zero trading days and zero return days. This measure of illiquidity seem to be always higher for non-ETF bonds but lately, this measure has also been converging towards that of ETF bonds. A big problem with this measure is that no dramatic increase was seen during 2008 Q3suggesting that it missed the liquidity crunch of that period.

iv. Median of Imputed Round-trip Cost (IRC):

This measure was developed by Feldütter (2012). Feldütter (2012) observes that there are usually two or three trades in the same bond on the same day with some volume. For each consecutive trade, the imputed round-trip cost (IRC) is calculated as follows:

$$IRC = \frac{P_{max} - P_{min}}{P_{min}} * 100$$

where $P_{max}$is the highest price and P_{min} is the lowest price within an IRT. The daily estimate of roundtrip cost is the average IRC for all IRTs in day and the monthly IRC is the median daily observation. A higher IRC is indicative of higher transaction costs and thus lower liquidity.

Figure 3.5 shows the cross sectional median plot for median IRC at each quarter. Here, we do not find any significant difference between ETF- and non-ETF bonds.

v. Turnover

Turnover is a measure of liquidity based solely on trading volume. More specifically, it is the ratio of trade volume to bonds outstanding expressed as percentage. Turnover of a bond i at quarter, q is defined as

$$Turnover_{i,q} = \frac{Trade\ Volume_{i,q}}{Amount\ of\ Bonds\ Outstanding_{i.q}} * 100$$

A high turnover of a bond will imply that out of the total bond issued for that particular bond, a higher proportion is being traded that quarter. This also implies that a corporate bond with higher turnover is considered to be more liquid.

Figure 3.6 shows the time plot of cross sectional average of 'turnover'. Before 2008, the trading activity, as measured by turnover, was always significantly higher for ETF bonds, suggesting they have higher liquidity compared to non-ETF bonds. In contrast, the post-2008 turnover of ETF bonds has been converging towards that of non-ETF bonds.

Table 3.2 reports the correlation among these five liquidity measures. Amihud (2002)'s and Roll (1984)'s measures are found to have highest pair-wise correlation (0.67).. Among the others, only turnover and zeros have pair-wise correlation which is worth mentioning (-0.33).

The failure of zeros and turnover to capture the 2008 credit crunch makes them a weak and less reliable measure of bond liquidity. Also, the low pair-wise cross correlation of the other three measures with either Amihud (2002)'s or Roll (1984)'s measures clearly suggests they are not measuring the same aspects of liquidity as Amihud or Roll. Median IRC captures the high illiquidity of corporate bonds during 2008, but is very weakly correlated with both Amihud (2002) and Roll (1984). In this paper, we extensively report the results derived from using Roll (1984) as the bond liquidity measure. We use the other measures of liquidity to check the robustness of our findings in Roll (1984)'s case.

3.3 Methodology

This section describes the empirical methodology used in this paper to investigate the effect of basket security ownership on corporate bond liquidity. We propose to estimate the effect using two different econometric specifications or models. Since we have panel data of corporate bond liquidity and their holding information, all of our models are predominantly panel regression models.

3.3.1 *Endogeneity concerns*

The issue of endogeneity is a common concern for these types of models; it can arise in several ways. First, at any time t, ETFs and mutual funds may choose to hold bonds based on their liquidity. Second, there may be an unobserved heterogeneous component of liquidity which varies across individual bonds. Third, bond liquidity may be affected systematically during certain times depending on the time-specific events. There are different ways to resolve these endogeneity concerns. For example, a more straightforward argument comes from Ben-David, Franzoni, and Moussahwi (2014), who suggest that the variation in ETF ownership is exogenous to the variable of interest (in this case, bond liquidity). Dannhauser (2014) explains this exogeneity from the fact that the main objective of ETFs is usually minimizing tracking error and replication. For mutual funds, this endogeneity concern still remains as their portfolio decisions are made by a fund manager whose objective is to maximize the absolute performance of the fund. A more systematic approach to tackling this issue is to take the lagged value of the ownership as the regressor in the model (Ben-David, Franzoni, and Moussawi, 2014; Dannhauser, 2014; and Hamm, 2012). The second and third endogeneity concerns can be easily tackled by considering individual fixed effects (in our case, bond fixed effects) and time fixed effects (quarter fixed effects in our case) in the model. Dannhauser (2014) also argues that the time specific fixed effect will control for the common trend in the corporate bond market.

3.3.2 Comparison of liquidity for different classes of corporate bonds

From Figures 3.2-3.6, it is quite evident that bonds that are held by bond ETFs are more liquid than those which are not held by any ETFs. Figures 3.7 and 3.8 demonstrate the individual bond ETFs' ownership and bond mutual funds' ownership, respectively, for investment grade bonds and high yield bonds. Interestingly, in recent times both ETF and bond mutual funds seem to be investing comparatively more in high yield bonds than in investment grade bonds.

Figure 3.7 is particularly interesting. Since their inception in 2002Q4, bond ETFs have exclusively invested in investment grade bonds. The first time bond ETFs started investing in high yield corporate bonds was in 2007Q1. After the crisis of 2008, higher yield bonds were bought by bond ETFs more and consequently, the bond ETF ownership of high yield bonds is larger than that of investment grade bonds since 2008.

Figures 3.7 and 3.8 provide us with evidence that the high liquidity of ETF bonds may not be a mechanical outcome due to ETFs holding only liquid bonds. Investment grade bonds naturally enjoy higher liquidity than high yield bonds due to fewer risk factors involved in trading them. Figure 3.9 demonstrates the liquidity difference between investment grade bonds and high yield bonds using Roll (1984)'s measure of bond illiquidity. We find that on average high yield bonds are less liquid, and during the crisis period this illiquidity is more prominent in high yield bonds.

3.3.3 Model Specification

Model 1 investigates the liquidity effect from the point of view of basket securities' ownership, which computes the ownership of bond ETF and bond funds together. We also control for non-ETF bonds (bonds that have ETF holding equal to zero) to capture whether there is any unexplained difference between the liquidity of ETF-bonds and of non-ETF bonds. We further control for average credit rating and date to maturity in the model. The average credit rating is calculated as the average of two ratings: the S&P rating and Moody's rating.

Model 1:

$$Liquidity\ Measure_{i,t} = \alpha_i + \alpha_t + \beta_1 Total\ Ownership\ of\ Basket\ Security_{i,t-1} + \beta_2 NonETF\ Dummy_{i,t} + \beta_3 Avg.Credit\ Rating_{i,t} + \beta_4 Date\ to\ Maturity_{i,t} + e_{i,t}$$

In Model 2, we isolate the effect of bond ETF and bond funds' ownership. The implication of the slope coefficient remains the same as before.

Model 2:

$$Liquidity\ Measure_{i,t} = \alpha_i + \alpha_t + \beta_1\ Bond\ ETF\ Ownership_{i,t-1} + \beta_2 Bond\ Fund\ Ownership_{i,t-1} + \beta_3 NonETF\ Dummy_{i,t} + \beta_4 Avg.Credit\ Rating_{i,t} + \beta_5 Date\ to\ Maturity_{i,t} + e_{i,t}$$

In Model 3, we control for other investors' ownership of corporate bonds. We introduce lagged pension fund and lagged insurance company ownership as additional regressors.

Model 3:

$$Liquidity\ Measure_{i,t} = \alpha_i + \alpha_t + \beta_1\ Bond\ ETF\ Ownership_{i,t-1} + \beta_2 Bond\ Fund\ Ownership_{i,t-1} + \beta_3 Insurance\ Ownership_{i,t-1} + \beta_4 Pension\ Fund\ Ownership_{i,t-1} + \beta_5 NonETF\ Dummy_{i,t} + \beta_6 Avg.Credit\ Rating_{i,t} + \beta_7 Date\ to\ Maturity_{i,t} + e_{i,t}$$

Model 4 does not include a time fixed effect, but it has an interaction term with thedummy variable for the year 2008 and a lagged ETF ownership variable. The idea is to study the effect of ETF ownership on bond liquidity during the time of crisis.

Model 4

$$Liquidity\ Measure_{i,t} = \alpha_i + \beta_1\ Bond\ ETF\ Ownership_{i,t-1} + \beta_2 Bond\ Fund\ Ownership_{i,t-1} + \beta_3 Insurance\ Ownership_{i,t-1} + \beta_4 Pension\ Fund\ Ownership_{i,t-1} + \beta_5 NonETF\ Dummy_{i,t} + \beta_6 Avg.Credit\ Rating_{i,t} + \beta_7 Date\ to\ Maturity_{i,t} + \beta_8 \left(Bond\ ETF\ ownership_{i,t-1} * 2008\ Year\ Dummy\right) + e_{i,t}$$

In Models 2-4, if the coefficients of lagged ETF ownership (β_1) and lagged mutual fund ownership (β_2) turn out to be negative and statistically significant, then it implies that higher ownership of corporate bonds by ETFs and mutual funds improves the liquidity of underlying or constituent corporate bonds. In Model 4, the slope coefficient of interaction term (β_8) will indicate the effect of lagged ETF ownership on bond liquidity during the financial crisis of 2008. If β_8 is positive and statistically significant, then this implies that ETF ownership of corporate bonds actually decreased the liquidity of underlying bonds.

3.4 Results

This section describes the regression results and their implications. We estimate our previously mentioned two models with four different samples. Our first results are based on the full sample which we derive by merging the eMaxx dataset for holding information with the TRACE database for liquidity measures. Next, we construct a sample which contains only ETF bonds and present our results. Lastly, we construct two samples, one of which contains only investment grade bonds and the other contains high yield bonds.

In Tables 3-4, as well as two-way fixed effect estimators, we also report the bond-fixed effect estimators to provide a comparison between the two results. In Tables 5-6, we only use two-way fixed effect estimators to report our results.

Table 3.3 reports different estimation results for all four models that that we have noted in the previous section. We use a full set of samples including both ETF bonds and non-ETF bonds and report results only using Roll (1984)'s measure of bond liquidity. The idea is to demonstrate how the sign and value of coefficients and their statistical significance change, if they do, across different models. The estimation of Model 1 finds that, regardless of the estimation technique used (fixed effect or two-way effect), the total lagged ownership of ETFs and mutual funds improves liquidity. Note that a higher value of Roll (1984)'s measure implies higher illiquidity of the underlying bonds. Therefore, negative coefficients imply a positive impact or improvement in liquidity due to higher ownership. In Models 2 and 3, individual (or bond) fixed effect estimators are similar to two-way fixed effect estimators in values and in sign. However, in terms of statistical significance these two estimators differ from each other as the two-way fixed effect estimators do not find any statistically significant impact for ETFs in either of the two models. In Model 4, we find that the overall impact of ETF ownership on bond liquidity is positive (i.e., negative and significant slope coefficients) but during the crisis of 2008, this impact is estimated to be negative and significant (the positive slope coefficient of the interaction term).

In the following analysis, we focus only on Model 2. The rationale behind this is that the adjusted R^2 does not improve in the estimation of Models 4 or 3 compared to model 2. In Table 3.4 we use only Model 2 and also report results from using two estimation techniques (bond fixed effect and two-way fixed effect) and Roll (1984)'s liquidity measure. The results are described in four pairs of columns. The first pair of columns reports the result using the full sample (this is the same result we derived from Table 3.1), the second pair of columns reports the result using a sample consisting only those bonds for which we find non-zero ETF ownership in the data, the third pair of columns reports results using sample of investment grade bonds only, and the last pair of columns reports results using sample of high yield bonds.

The statistical significances of the coefficients are again different depending on the estimation technique used. Most of the statistically significant results by bond fixed effect estimators are found to be insignificant by two-way fixed effect estimators. The R^2 values are also found to be higher in two-way fixed effect estimators. From the two-way fixed effect estimators, we find that within the group of ETF bonds (i.e., the bonds which have positive ETF ownership), lagged ETF ownership plays a positive role in improving underlying bond liquidity. More precisely, a 1% increase in lagged ETF ownership decreases the illiquidity of corporate bonds as measured by Roll (1984) by 0.1 basis points. We do not find any statistically significant impact of ETFs on bond liquidity measured for either investment grade bonds or high yield bonds. Lagged bond mutual fund ownership is found to have no effect on ETF bonds or on investment grade bonds. However, for high yield bonds, lagged mutual fund ownership plays a positive role as it reduces the illiquidity of underlying corporate bonds by 0.1 basis points.

Table 3.5 reports the estimation from Model 2 and the results derived by two-way fixed effect estimators. The first four columns report results using Amihud's (2002) price impact measure of bond illiquidity as a dependent variable. The last four columns report results using median Imputed Round-trip Cost (IRC) as a dependent variable. The main goal is to check the robustness of the results found in Table 3.4.

We find that in both cases, using Amihud (2002)'s measure and median IRC, lagged ETF ownership has a positive and significant impact on underlying bond liquidity only in the sample of ETF bonds. In the other sample, this impact is found to be statistically insignificant. Mutual fund ownership is also found to have a significant positive impact in improving bond liquidity only for high yield bonds. Both these results are similar in nature to those of Table 3.4.

Table 3.6 reports the estimation results using two other liquidity measures – using zeros and turnover as dependent variables. The first four columns report results for zeros and the last four columns report results for turnover. We find that, in these cases, lagged ETF ownership does not have a statistically significant impact on ETF bonds, unlike Roll (1984), Amihud (2002) and median IRC. However, an interesting result for both zeros and turnover is that lagged ETF ownership is found to have a statistically significant positive impact on high yield corporate bond liquidity.

3.5 Conclusion

We find a mix of evidence of ETF ownership's effect on bond liquidity depending on the liquidity measure used in the analysis. In the case of Amihud (2002), Roll (1984) and the median IRC measure of bond liquidity, we find that for the corporate bonds which are bought and held by ETFs, higher ETF ownership improves the liquidity of these bonds. Unlike Dannhuaser (2014), our study does not find statistically significant results using these three liquidity measures for either investment grade bonds or high yield bonds.

The other two measures of liquidity, zeros and turnovers, give us different results in terms of statistical significance. We find that lagged ETF ownership improves the liquidity of high yield corporate bonds. Bond mutual funds, on the other hand, are found to have a statistically significant effect on high yield bonds when considering the measures Roll (1984), Amihud (2002), Median IRC, and zeros. We find that the variable turnover has positive and statistically significant impact on bond liquidity in the ETF-bond sample.

The objective of this paper is to study the effect of ETF ownership on corporate bond liquidity. Our study finds a positive impact on ETF ownership for bonds that are bought by ETFs (or ETF bonds) and high yield bonds. The problem of endogeneity is a major concern in this study. We utilize the arguments and model specifications used in previous studies (Ben-David, Franzoni, and Moussawi, 2014 and Dannhauser, 2014) to overcome this issue. We acknowledge the main research question posed in this paper is very similar to that of Dannhauser (2014). However, our sample coverage and empirical approach is distinctive enough to shed new light on this important research agenda.

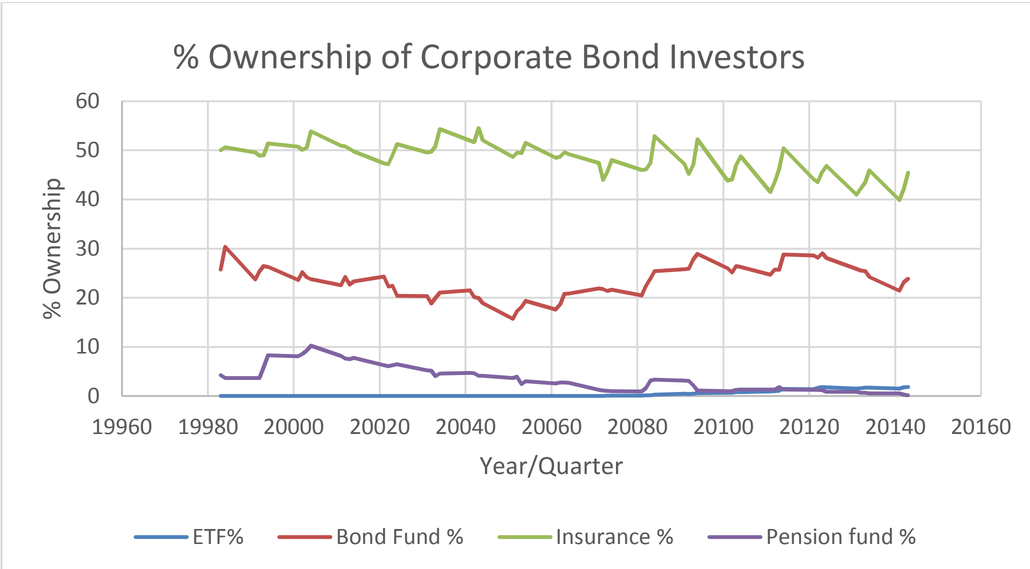

Figure 3.1: Percentage ownership or holdings of different Corporate Bond Investors

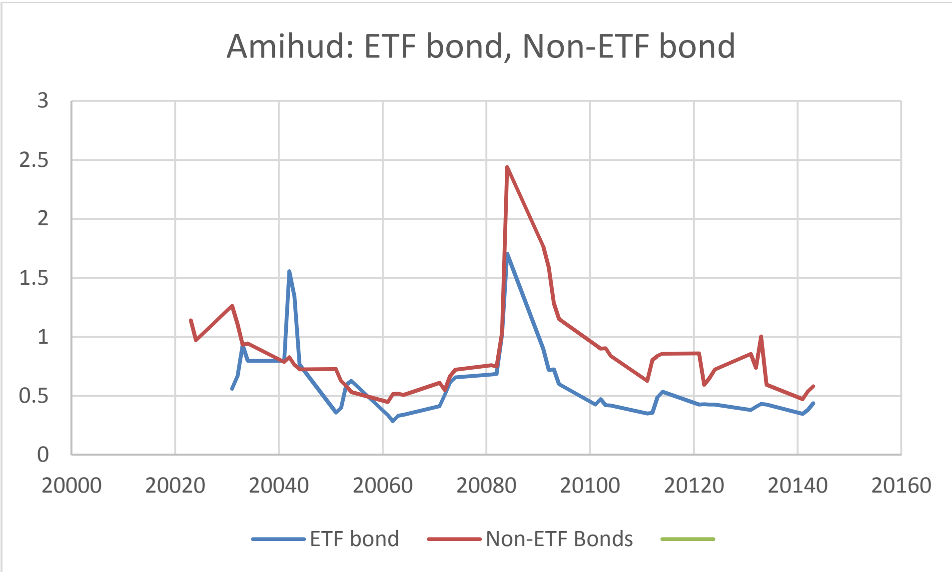

Figure 3.2: Amihud (2002) measure of Illiquidity for ETF bonds and Non-ETF Bonds.

Note: Higher Amihud (2002) values imply higher illiquidity of corporate bonds. We use the cross-sectional Average of liquidity at each time period.

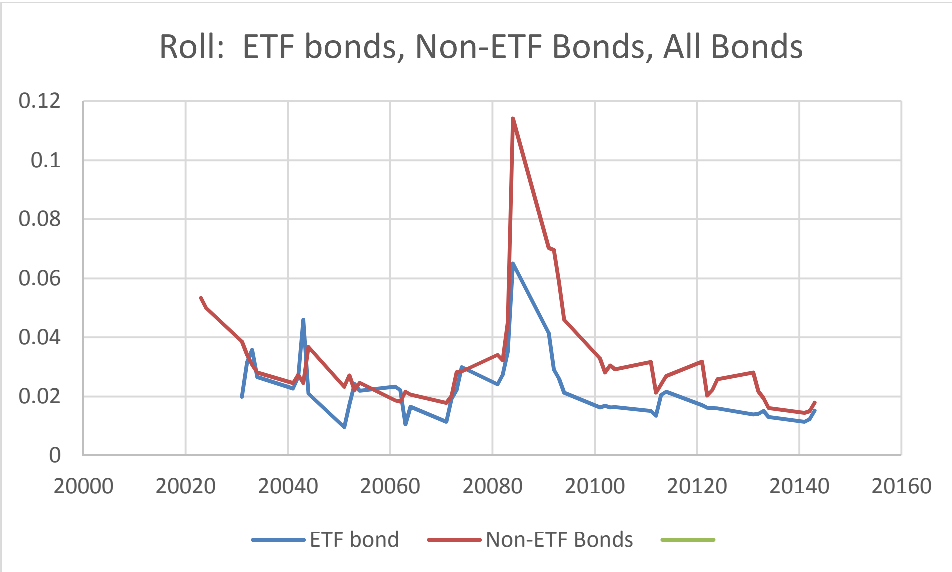

Figure 3.3: Roll (1984) measure of Illiquidity for ETF bonds and Non-ETF Bonds.

Note: Higher Roll (1984) values imply higher illiquidity of corporate bonds. We use the cross-sectional Average of liquidity at each time period.

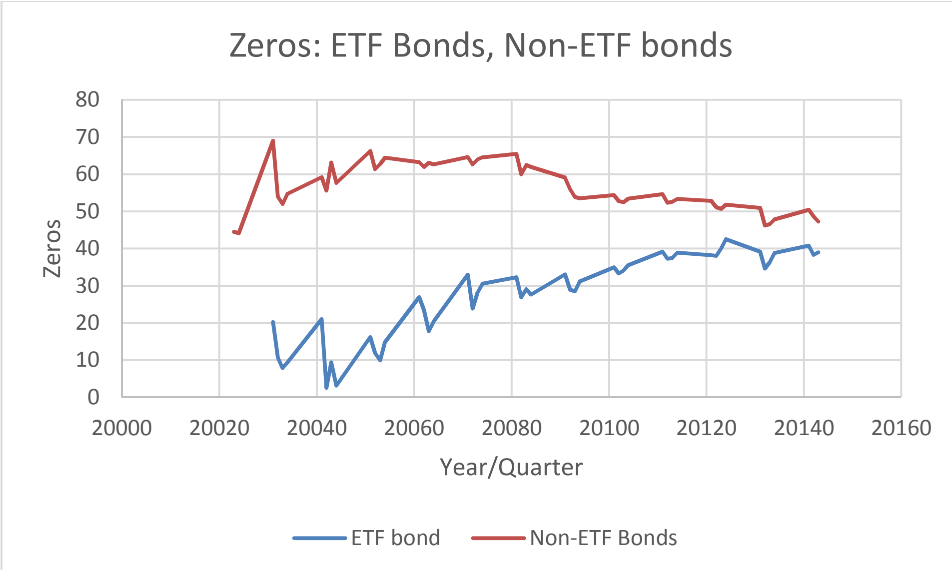

Figure 3.4: ‘Zeros’ measure of Illiquidity for ETF bonds and Non-ETF Bonds.

Note: Higher ‘zeros’ values imply higher illiquidity of corporate bonds. We use the cross-sectional Average of liquidity at each time period.

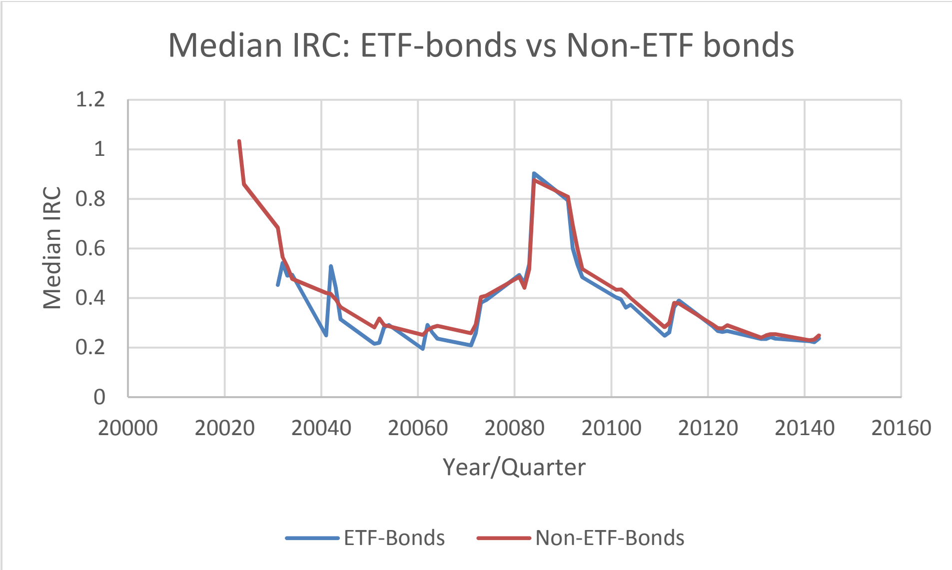

Figure 3.5: Median IRC measure of Illiquidity for ETF bonds and Non-ETF Bonds.

Note: Higher Median IRC values imply higher illiquidity of corporate bonds. We use the cross-sectional median of the liquidity at each time period.

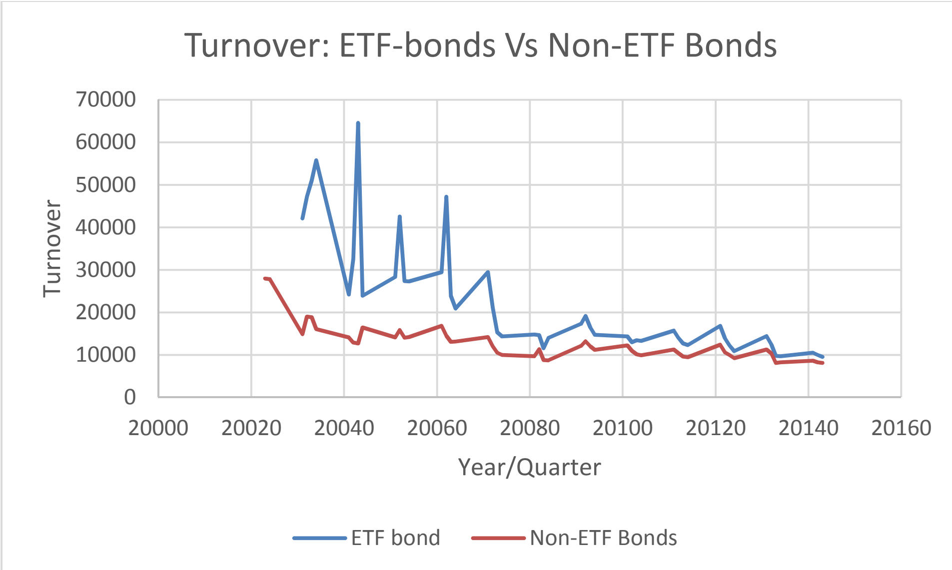

Figure 3.6: Turnover of different corporate bonds- ETF bonds and Non-ETF Bonds.

Note: Higher turnover values implies lower illiquidity of corporate bonds. We use the cross-sectional Average of liquidity at each time period.

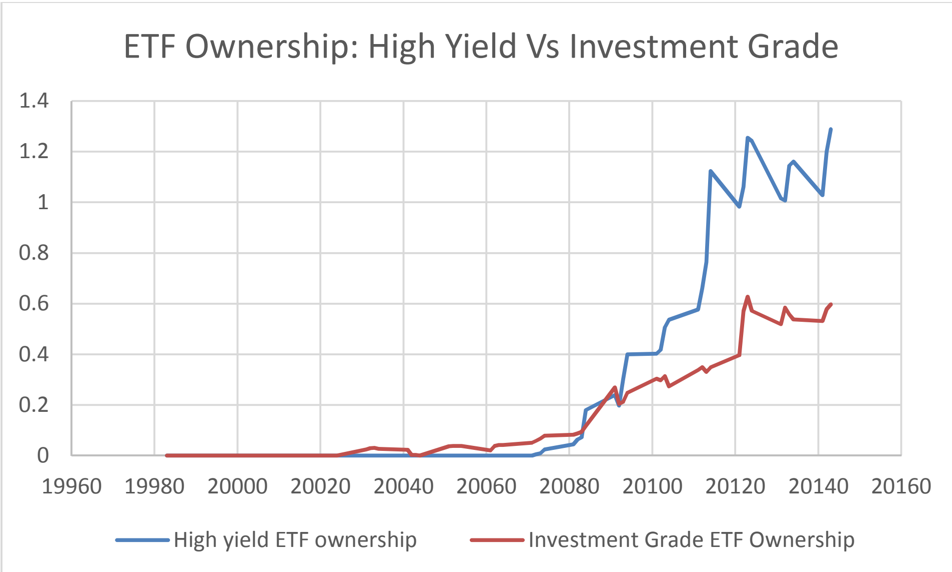

Figure 3.7: Comparison between Investment Grade Bonds and High Yield Bonds in terms of ETF Ownership.

Note: We use the cross-sectional average at each period.

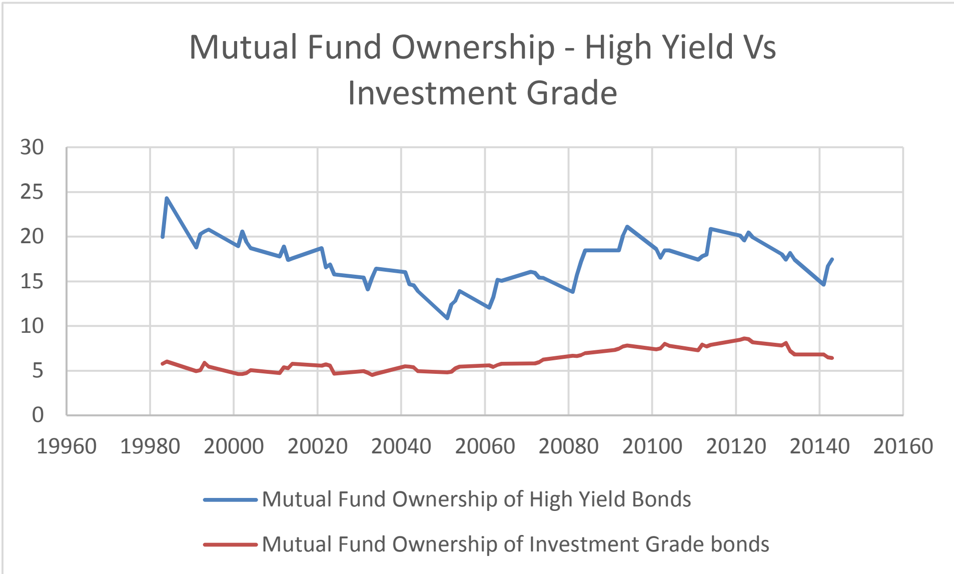

Figure 3.8: Comparison between Investment Grade Bonds and High Yield Bonds in terms of Bond Mutual fund Ownership.

Note: We use the cross-sectional average at each period.

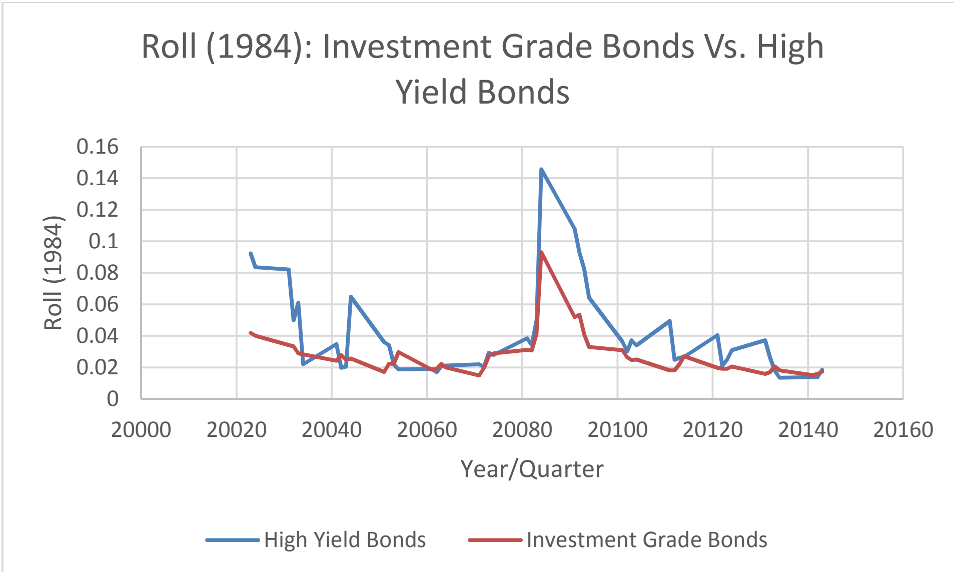

Figure 3.9: Illiquidity of bonds as measured by Roll (1984) for Investment grade bonds Vs High Yield bonds.

Note: We use cross-sectional average at each time period.

Table 3.1: Summary Statistics of eMaxx Database and Merged Database

	Mean	Median	Max	Min	Std.Dev
eMaxx database					
Coupon Rate	7.91	7.62	19	0.25	1.99
Amount Offered	372333	250000	15000000	125	386669.5
Average Credit Rating	11.03	10.5	22.5	1	4.27
Date to Maturity	87.95	60.03	1182.03	0.03	106.79
Dataset merged with TRACE					
Coupon Rate	7.23	7	18.5	0.65	1.83
Amount Offered	455762	325000	15000000	6500	430778
Avarage Credit Rating	10.42	10	22.5	1	4.18
Date to Maturity	85.15	53.03	1169.76	0.03	110.16

Table 3.2 : Correlation among Different Liquidity Measure

	Roll	Zeros	Amihud	Median IRC	Turnover
Roll	1	0.02	0.67	0.13	0.04
Zeros	0.02	1	0.02	0.01	-0.33
Amihud	0.67	0.02	1	0.19	-0.02
Median IRC	0.13	0.01	0.19	1	-0.00
Turnover	0.04	-0.33	-0.02	-0.00	1

Table 3.3: Different Regression Results using the full sample (all bonds) and Roll (1984) measure as liquidity measure. Note: High Roll(1984) implies lower bond liquidity.

Regressors	model#1 (fixed effect)	model#1 (2-way)	model#2 (fixed effect)	model#2 (2-way)	model#3 (fixed effect)	model#3 (2-way)	model#4(fixed effect)
ETF + bond-fund %(lag)	-0.001 (2.37)*	-0.001 (2.83)**					
ETF %(lag)			-0.002 (2.34)*	-0.001 (0.77)	-0.002 (2.32)*	-0.001 (0.77)	-0.002 (2.28)*
Mutual Fund % (lag)			-0.001 (2.11)*	-0.001 (2.66)**	-0.001 (2.20)*	-0.001 (2.67)**	-0.001 (2.23)*
Insurance Comp. % (lag)					-0.000 (3.44)**	-0.000 (0.41)	-0.000 (3.39)**
Pension Fund (lag)					0.001 (3.55)**	-0.000 (0.04)	0.001 (3.55)**
2008 Dummy * ETF % (lag)							0.006 (5.05)**
Non- ETF dummy	0.014 (4.95)**	0.010 (3.15)**	0.014 (4.58)**	0.011 (3.12)**	0.014 (4.44)**	0.011 (3.10)**	0.014 (4.52)**
Avg rating	0.001 (0.83)	-0.000 (0.55)	0.001 (0.82)	-0.000 (0.55)	0.000 (0.27)	-0.000 (0.63)	0.000 (0.29)
Date-to-maturity	-0.000 (3.08)**	0.000 (0.44)	-0.000 (3.14)**	0.000 (0.45)	-0.000 (3.07)**	0.000 (0.47)	-0.000 (3.09)**
Constant	0.048 (3.68)**	0.042 (3.61)**	0.048 (3.79)**	0.042 (3.68)**	0.058 (4.49)**	0.043 (3.77)**	0.043 (3.77)**
R^2	0.00	0.02	0.00	0.02	0.00	0.02	0.02
N	24,959	24,959	24,959	24,959	24,959	24,959	24,959

* $p<0.05$; ** $p<0.01$

Table 3.4: Estimation Results of Model 2 with Roll (1984) measure using different samples.

Regressors	model#2 (fixed)	model#2 (2-way)	model#1 (fixed)	model#1 (2-way)	model#1 (fixed)	model#2 (2-way)	model#1 (fixed)	model#2 (2-way)
	All Bonds	All bonds	ETF-bonds	ETF-Bonds	Investment Grade Bonds	Investment Grade Bonds	High Yield Bonds	High Yield Bonds
ETF %(lagged)	-0.002 (2.34)*	-0.001 (0.77)	-0.002 (2.53)*	-0.001 (2.08)*	-0.001 (3.70)**	0.000 (0.98)	-0.003 (2.57)*	-0.002 (1.47)
Mutual Fund % (lagged)	-0.001 (2.11)*	-0.001 (2.66)**	0.000 (1.85)	0.000 (1.47)	-0.000 (0.23)	-0.000 (1.08)	-0.000 (1.29)	-0.001 (1.97)*
Non- ETF dummy	0.014 (4.58)**	0.011 (3.12)**			0.008 (3.32)**	0.005 (2.30)*	0.012 (2.70)**	0.006 (1.54)
Avg rating	0.001 (0.82)	-0.000 (0.55)	-0.000 (0.52)	-0.001 (0.69)	0.002 (1.94)	0.001 (0.91)	0.001 (0.83)	0.000 (0.15)
Date-to-maturity	-0.000 (3.14)**	0.000 (0.45)	0.000 (3.44)**	0.000 (2.13)*	-0.000 (0.97)	0.000 (2.39)*	-0.001 (2.46)*	0.000 (0.27)
constant	0.048 (3.79)**	0.042 (3.68)**	0.012 (1.23)	0.016 (1.64)	0.012 (1.35)	0.009 (1.06)	0.053 (1.97)*	0.055 (2.05)*
R^2	0.00	0.02	0.01	0.06	0.00	0.06	0.01	0.02
N	24,959	24,959	7,642	7,642	12,258	12,258	9,278	9,278

* $p<0.05$; ** $p<0.01$

Table 3.5: Estimation of Model 2 for Amihud (2002) and Median IRC using only bond fixed effect estimators. Note: high Amihud (2002) and high median IRC implies lower liquidity.

Regressors	model#2 (All Bonds)	model#2 (ETF-Bonds)	model#2 (Investment Grade bonds)	model#2 (High Yield Bonds)	model#2 (All Bonds)	model#2 (ETF-Bonds)	model#2 (Investment Grade bonds)	model#2 (High Yield Bonds)
	Amihud (2002)				Median IRC			
ETF % (lagged)	0.004 (0.34)	-0.017 (2.45)*	-0.002 (0.12)	-0.021 (1.12)	-0.025 (0.82)	-0.011 (2.44)*	-0.008 (0.61)	-0.004 (0.39)
Bond Fund % (lagged)	-0.016 (3.91)**	0.003 (0.92)	-0.006 (1.47)	-0.016 (2.78)**	-0.021 (1.78)	0.002 (1.44)	-0.008 (1.16)	-0.003 (2.37)*
Non- ETF dummy	0.218 (2.63)**		0.017 (0.41)	0.257 (2.05)*	0.850 (1.27)		0.006 (0.18)	0.028 (1.05)
Avg rating	0.001 (0.06)	-0.094 (1.87)	-0.008 (0.20)	0.004 (0.10)	-0.067 (0.91)	-0.027 (2.24)*	-0.000 (0.01)	0.009 (0.89)
Date-to-maturity	0.002 (1.21)	0.000 (0.13)	0.003 (2.96)**	-0.015 (2.93)**	-0.004 (0.39)	0.002 (5.25)**	0.005 (4.65)**	0.013 (1.31)
constant	0.915 (3.75)**	1.350 (2.64)**	0.655 (2.28)*	1.427 (2.03)*	2.058 (1.66)	0.521 (4.17)**	0.349 (2.66)**	0.437 (2.90)**
R^2	0.02	0.10	0.09	0.02	0.00	0.17	0.04	0.06
N	29,098	8,400	14,643	11,884	32,191	8,525	16,572	13,073

* p<0.05; ** p<0.01

Table 3.6: Estimation of Model 2 for Zeros and Turnover using only bond fixed effect estimator. Note: high zeros implies lower liquidity and high turnover implies higher liquidity.

Regressors	model#2 (All Bonds)	model#2 (ETF-Bonds)	model#2 (Investment Grade bonds)	model#2 (High Yield Bonds)	model#2 (All Bonds)	model#2 (ETF-Bonds)	model#2 (Investment Grade bonds)	model#2 (High Yield Bonds)
	Zeros				Turnover			
ETF % (lagged)	-0.659 (3.40)**	-0.352 (1.69)	0.067 (0.22)	-1.098 (5.41)**	264.718 (2.16)*	125.03 (0.87)	-160.46 (0.63)	524.617 (3.61)**
Bond Fund % (lagged)	-0.332 (12.17)**	-0.270 (4.89)**	-0.328 (5.93)**	-0.272 (8.80)**	5.704 (0.21)	130.54 (3.08)**	-141.05 (1.83)	42.671 (1.72)
Non- ETF dummy	6.403 (10.37)**		3.566 (4.24)**	9.520 (11.77)**	-1,405.48 (3.56)**		-745.22 (1.26)	-2,198.6 (4.67)**
Avg rating	-0.079 (0.45)	-0.751 (1.24)	-0.784 (1.40)	0.097 (0.32)	-294.28 (2.47)*	-311.43 (1.61)	141.07 (0.54)	-609.92 (3.13)**
Date-to-maturity	0.084 (3.15)**	-0.053 (2.44)*	0.007 (0.27)	-0.170 (2.69)**	295.60 (10.30)**	343.25 (5.80)**	302.09 (8.99)**	258.371 (5.26)**
constant	57.47 (25.56)**	52.579 (8.11)**	63.060 (13.84)**	53.906 (11.62)**	1,323.19 (0.87)	-9,037.9 (2.53)*	-10,410.15 (4.15)**	17,061.47 (5.80)**
R^2	0.10	0.07	0.12	0.12	0.10	0.15	0.12	0.07
N	34,113	8,636	18,263	13,753	32,316	8,636	18,263	13,753

* p<0.05; ** p<0.01

BIBLIOGRAPHY

[1] Amihud, Y., 2002. Illiquidity and stock returns: cross-section and time-series effects. Journal of Financial Markets 5, 31–56.

[2] Bae K-H, Kang J-K, and Wan J, 2012, The Costs of ETF Membership: Valuation Effect of ETFs on Underlying Firms, Working Paper

[3] Baillie, R. T., Booth, G. G., Tse, Y., Zabotina, T., 2002, Price discovery and common factor models, Journal of Financial Markets 5, 309-321

[4] Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A., Shephard, N., 2008, Realised Kernels in Practice: Trades and Quotes, Econometrics Journal, Vol 4, 1-33

[5] Ben-David, I., Franzoni, F., Moussawi, R., 2014. Do ETFs Increase Volatility? National Bureau of Economic Research.

[6] Boehmer, B., Boehmer, E., 2003, Trading your neighbor's ETFs: Competition or fragmentation?, Journal of Banking & Finance 27, 1667-1703

[7] Borkovec, M, Domowitz, I., Serbin, V., Yegerman, H., 2010, Liquidity and Price Discovery in Exchange-Traded Funds, Investment Technology Group (ITG), Inc.

[8] Boudt, K., Cornelissen, J., Payseur, S., Highfrequency: Toolkit for the analysis of highfrequency financial data in R

[9] Bruder, B., Roncalli, T., 2012, Managing Risk Exposure Using the Risk Budgeting Approach, Working Paper, Lyxor Asset Management

[10] Caporale, G. M., Girardi, A., 2013, Price discovery and trade fragmentation in a multi-market environment: Evidence from the MTS system, Journal of Banking & Finance 37, 227-240

[11] Chakravarty, S., Gulen, H., Mayhew, S., 2004, Informed trading in stock and option markets, Journal of Finance, Volume 59, No. 3, 1235-1258

[12] Chen, L., Lesmond, D.A., Wei, J., 2007. Corporate yield spreads and bond liquidity. The Journal of Finance 62, 119–149.

[13] Dick-Nielsen, J., Feldhütter, P., Lando, D., 2012. Corporate bond liquidity before and after the onset of the subprime crisis. Journal of Financial Economics 103, 471–492.

[14] Dannhauser, C. D., 2014, The equitization of the corporate bond market: the impact of ETFs on Bond Yields and Liquidity, Working Paper

[15] Fang, Y., Sanger, G. C., 2012, Index Price Discovery in the Cash Market, Working paper,

[16] Feldhütter, P., 2012. The same bond at different prices: identifying search frictions and selling pressures. Review of Financial Studies 25, 1155–1206.

[17] Fremault, A., 1991. Stock index futures and index arbitrage in a rational expectations model. Journal of Business 64, 523–547.

[18] Gammill, J.F., Perold, A.F., 1989. The changing character of stock market liquidity. Journal of Portfolio Management 15, 13–18.

[19] Gonzalo, J., Granger, C. W. J., 1995, Estimation of common long-memory components in cointegrated systems, Journal of Business and Economic Statistics 13, 27-35

[20] Gorton, G., Pennacchi, G., 1993. Security baskets and index-linked securities. Journal of Business 66, 1– 27

[21] Grammig, J., Peter, F. J., 2013, Tell-Tale Tails: A data driven approach to estimate unique market information shares, Journal of Financial and Quantitative Analysis, Vol. 48, pp. 459-488.

[22] Hegde SP, McDermott JB (2004) The market liquidity of DIAMONDS, Q's, and their underlying stocks. Journal of Banking and Finance 28: 1043-1067.

[23] Kumar, P., Seppi, D., 1994. Information and index arbitrage. Journal of Business 67, 481–509

[24] Kyle, A.S., 1985. Continuous auctions and insider trading. Econometrica 53, 1315–1335.

[25] Hamm, Sophia J.W., 2010. The effect of ETFs on stock liquidity. Working paper. University of Pennsylvania.

[26] Harris, F. H. deB., McInish, T. H., Wood, R. A., 2002, Security price adjustment across exchanges: an investigation of common factor components in Dow Stocks, Journal of Financial Markets 5, 277-308

[27] Hasbrouck, J., 1995, One Security, Many Markets: Determining the Contribution to Price Discovery, Journal of Finance, Vol- 50, No. 4, 1175-1199

[28] Hasbrouck, J., 2002, Stalking the "efficient price" in market microstructure specifications: an overview, Journal of Financial Markets 5, 329-339

[29] Hasbrouck, J., 2003, Intraday Price Formation in U.S. Equity Index Markets, Journal of Finance, Vol LVIII, No. 6, 2375-2399

[30] Hegde Shantaram P., McDermott John B., (2003) The liquidity effects of revisions to the S&P 500 index: An empirical analysis. Journal of Financial Markets 6: 413–459.

[31] Hedge, Shantaram P., McDermott, John B., 2004, The market Liquidity of DIAMONDS, Q's and their underlying stocks, Journal of Banking and Finance 28, 1043-1067

[32] Lehman, B. N., 2002, Some desiderata for the measurement of price discovery across markets, Journal of Financial Markets 5, 259-276

[33] Lesmond, D., J. Ogden, and C. Trzcinka, 1999, A new estimate of transaction costs, Review of FinancialStudies 12, 1113–1141.

[34] Lien, D., Shrestha, K., 2009, A New Information Share Measure, The Journal of Futures Markets, Vol 29, No. 4, 377-395

[35] Marshall, B., R., Nguyen, N. H., Visaltanachoti, N., 2013, ETF Arbitrage: Intraday evidence, Journal of Banking & Finance 37, 3486-3498

[36] Mizarch, B., Neely C. J., 2008, Information Shares in the U.S. Treasury Market, Journal of Banking & Finance 32, 1221-1233

[37] Putninš, T. J., 2013, What do price discovery metrics really measure? , Journal of Empirical Finance 23, 68-83

[38] Roll, Richard, 1984, A simple implicit measure of the effective bid-ask spread in an efficient market, Journal of Finance 39, 1127–1139.

[39] Subrahmanyam A (1991) A theory of trading in stock index futures. Review of Financial Studies 4: 17-51.

[40] Theissen, Eric, 2002. Price discovery in floor and screen trading systems. Journal of Empirical Finance, 9, 455-474.

[41] Tse, Y., Martinez, V., 2007, Price discovery and informational efficiency of international iShares funds, Global Finance Journal 18, 1-15

[42] Winne Rudy D., Gresse, C., Platten, I., 2014, Liquidity and Risk Sharing Benefits from Opening an ETF market with Liquidity Providers: Evidence from CAC 40 Index, International Review of Financial Analysis, Forthcoming

[43] Yan, B., Zivot, E., 2007, The Dynamics of Price Discovery, University of Washington Mimeo,

[44] Yan, B., Zivot, E., 2010, A structural analysis of price discovery measure, Journal of Financial Markets 13, 1-19

[45] Yu, L., Basket securities, price formation, and informational efficiency, 2005, Working paper, University of Notre Dame

VITA

Syed Galib Sultan is a doctoral student of economics at University of Washington Seattle. His field of research is financial economics and applied time series econometrics. After finishing his PhD, Galib is expected to join the Model Validation Group at State Street as Quantitative Analyst / Assistant Vice President. Besides research, Galib likes to play cricket and watch cricket games.

In order to contact Galib, please email at galibsultan@gmail.com.