The value of ETF liquidity ☆

Marta Khomyn [a*] and Tālis J. Putniņš [a,b]

[a] *University of Technology Sydney, PO Box 123 Broadway, NSW 2007, Australia*
[b] *Stockholm School of Economics in Riga, Strelnieku Street 4a, Riga, LV 1010, Latvia*

December 02, 2018

Abstract

We investigate the apparent paradox of persistent fee differentials for exchange traded funds (ETFs) that track the same index, where counterintuitively more expensive ETFs often attract *more* investment. We show that this apparent paradox arises due to liquidity clienteles—investors with short holding horizons are attracted to the most liquid ETFs, thereby making them more liquid, and allowing the ETF issuers to charge a higher fee in equilibrium. Long horizon investors are more sensitive to the fee and therefore hold low-fee ETFs, which in turn are less liquid due to lower investor turnover. Liquidity clienteles also explain key features of ETFs competition, including the first-mover advantage and the ability for incumbent ETFs to maintain higher fees. We exploit the unique laboratory created by competing ETFs to measure the value of market liquidity to investors.

JEL classification: G11, G12, G14

Keywords: ETFs, liquidity, competition, trading volume

☆ We thank Shane Miller, Maureen O'Hara, Tom Smith, Damien Sherman, Elvira Sojli, Avanidhar Subrahmanyam, Damien Wallace, seminar participants at 8th Behavioral Finance and Capital Markets Conference, and seminar participants at 8th FIRN Annual Conference for helpful comments. Marta Khomyn gratefully acknowledges funding from the Capital Markets CRC Limited and is a visiting researcher at Chi-X Australia during the time spent on this study. Tālis J. Putniņš gratefully acknowledges funding from the Australian Research Council (ARC DE150101889).

[*] Corresponding author. Tel.: +61 481314349.

Email addresses: marta.khomyn@student.uts.edu.au (M. Khomyn) and talis.putnins@uts.edu.au (T. Putniņš).

1. Introduction

Can two identical baskets of securities trade at different prices? The law of one price says they should not, and yet there are many cases of exchange traded funds (ETFs) that replicate the same index, but charge different fees (management expense ratios, MERs). Moreover, the fee differentials are persistent, do not decrease through time or with competition, and in fact the higher fee ETF is often the one with *more* assets under management. So, what is it that investors are paying for when they choose higher cost ETFs that track the same basket of securities as a competitor? This paper shows that the answer is liquidity. What is more, we also show that the liquidity clienteles that give rise to the apparent violation of the law of one price are also instrumental to understanding the process by which ETFs compete and the equilibrium that is observed in this rapidly expanding market.

To illustrate the central points of this paper, consider the MERs of the three ETFs that track the S&P 500 index: State Street's SPY charges 9.4 basis points (bps) per annum, while Black Rock's IVV and Vanguard's VOO charge only 4 bps. Despite its fee being more than twice that of its competitors, the SPY is the largest of these ETFs by assets under management and there are no signs of the SPY's dominance declining in favor of its cheaper alternatives. We observe a similar situation in most same-index ETFs, where highly liquid first movers charge higher management fees compared to their cheaper competitors. What differentiates the relatively expensive ETFs like SPY from their competitors is the sheer amount of readily accessible liquidity.[1] SPY not only has more assets under management, but also higher turnover, leading to much greater daily traded volume.

If faced with a choice of multiple ETFs on the same index, which of these ETFs would an investor choose? Intuitively, and as we show formally, it depends on investment horizon. For short horizon investors (generally referred to in this paper as "high-turnover investors"), it's optimal to choose a more liquid ETF, even if it means paying a higher MER. Because high-turnover investors trade in and out of positions frequently, trading costs rather than MER costs constitute a higher proportion of their ETF holding fees, on an annualized basis. Therefore, high-turnover investors — typically institutions using ETFs for short-term tactical portfolio allocations — constitute a natural clientele of the highly liquid high-fee ETFs like SPY, which gives those ETFs greater secondary market turnover and reinforces their higher level of liquidity. At the other end of the spectrum, long term investors —for example, retail investors using ETFs as buy-and-hold vehicles — will not be

[1] Interestingly, despite higher MERs, SPY does not offer better performance in terms of tracking error: in fact, the opposite is the case. In case of SPY, that adds to the performance drag, resulting in IVV beating SPY by 0.48% over 10 years.

too concerned about liquidity because they will rarely trade, but will pay great attention to the fee differential as this is an important source of performance drag.

In effect, the "liquidity begetting liquidity" effect generated by the large and relatively short-horizon investor base allows SPY's issuers to extract a rent from the liquidity externalities. Liquidity externalities create a strong first-mover advantage and lead to less than perfectly competitive fee setting. The value of market liquidity to an investor is integral to the nature of competition among ETFs and the equilibrium in this market.

We develop a simple theoretical model to formalize the intuition in the above example and characterize the interplay between ETF fees, their liquidity, and investor clienteles. We also provide an empirical analysis of these relations, showing the above example is by no means an isolated case and liquidity and clienteles play an important role in how same-index ETFs compete. Our main contribution is in characterizing the important role played by liquidity in the ETF market: showing how it shapes competition between ETFs, drives fee setting in Nash equilibrium, and explains persistent fee differentials. From an investor's perspective, the Nash equilibrium has similarities with a prisoner's dilemma: while it would be optimal for all investors as a group to switch to a cheaper ETF, an individual liquidity-sensitive investor is worse off by switching, as he incurs the costs of illiquidity. Thus, our model helps resolve the apparent paradox of same-index ETFs charging vastly different MERs.

We test the model empirically and find that consistent with the model, higher MERs in ETFs tracking a given index tend to be associated with more liquidity: higher dollar volume and narrower relative bid-ask spreads. We also find evidence of the clientele effect. Higher-MER ETFs tend to have higher turnover (traded dollar volume divided by market capitalization), suggesting a clientele skewed towards shorter-horizon investors. This finding is in line with the industry view that that high ETF turnover is mainly due to institutions trading large short-term positions, for example, for short-term tactical allocation, hedging, or rebalancing. The common feature of these institutional traders is that they require substantial liquidity and trade in and out of their ETF positions frequently. Retail traders, on the other hand, use ETFs as an investment vehicle and therefore have longer holding horizons (Balchunas, 2016).

Our model also explains why new ETF launches, even if based on the same underlying index, do not necessarily prompt incumbents to lower their fees. For example, Box, Davis, and Fuller (2017) study the effects of new ETF introductions on the incumbent ETF's liquidity and MERs. They find that MERs of the incumbent do not decrease as a result of greater competition. We offer the liquidity explanation for this phenomenon and show why liquidity externalities prevent investors from switching to lower-MER competitors, thus allowing the incumbent ETF issuers to keep their MERs high.

Another contribution is that the model characterizes the conditions under which multiple ETFs per index are likely to emerge: namely, high proportion of high-turnover investors and significant differences in the holding horizons between the high-turnover and low-turnover clienteles (suggesting the high-MER issuer finds it more profitable to keep MERs high and serve solely the high-turnover clientele instead of lowering MERs to capture both the high-turnover and the low-turnover investors), as well as substantial combined AUM allocated to a given index, and relatively low fixed costs of issuers (suggesting issuers operate with significant economies of scale, which allows the competing low-liquidity ETF to charge .low enough MERs to attract the low-turnover clientele).

Our findings help explain the striking concentration of liquidity in a handful of major funds: 50% of ETF dollar volume is concentrated in the first 15 ETFs (out of the total of almost 2000 ETFs listed in the US). This oligopoly-like concentration of dollar volumes persists despite no shortage of newcomers: a new ETF is being launched on average every trading day?[2] By zooming in on indices with multiple ETFs, we capture almost half of all equity ETF dollar volumes, and just as our model suggests, those are the ETFs that attract significant institutional trading (i.e., high proportion of high-turnover investors), track major benchmark indices like S&P500 or Russell 2000, have substantial combined AUM allocated to them, and are held by highly heterogenous investors: on one side of the spectrum — the short-term traders (e.g., for hedging purposes or tactical positions), and on the other — the long-term buy-and-hold investors (e.g., for gaining broad market exposure at low cost).

Finally, we exploit unique features of the ETF market to provide novel measures of how investors value market liquidity. The standard approach in the asset pricing literature is very indirect and involves trying to infer the premium associated with illiquidity by measuring average asset returns for securities with different liquidity. In contrast, for competing same-index ETFs we can directly observe the fee differentials and relate them to underlying liquidity to measure how many basis points of return investors at the margin are willing to forgo for a given amount of additional liquidity. We find that investors are willing to pay 1.15% higher MERs in exchange for 1% higher dollar volumes (relative to a competing fund tracking the same index). In terms of trading costs, the average ETF investor pays 0.51 bps higher MER for each 1 bps of narrower spreads. This generates welfare transfers worth $780 million annually, which can be interpreted as

[2] According to ETF.com, there were 118 ETF launches in the first half of 2018 (January 1 – June 26), 271 launches in the year 2017, 247 in 2016, 284 in 2015, and 202 in 2014.

investors' payments for accessing superior liquidity provided by high-MER ETFs. The magnitude of this payment for liquidity represents 0.096% of the $815 billion of assets under management invested in indices with multiple ETFs.

To put our findings in a broader context, it's worth noting that ETFs have become an increasingly popular investment vehicle: in terms of money invested globally, they are set to cross a $5 trillion mark in 2018. In the year 2017 alone, ETFs saw $460 billion of new inflows, which amounts to $1.8 billion inflows on an average working day.[3] While ETFs are still a relatively small portion of all US assets by value, comprising just over 10% of $30tn US equity market capitalization, they account for over 30% of dollar volume traded. Seven out of ten most actively traded US securities in 2017 were ETFs rather than stocks (Financial Times, 2017). SPY alone is responsible for around one third of ETF traded dollar volume. It is also the most frequently traded security in the world, trading over 20 times a second. It is therefore important to understand the drivers of investors' and issuers' behavior in this rapidly growing market.

This paper contributes to several strands of literature. The first is the growing body of studies on ETFs (Madhavan, 2016), which are part of broader fund management literature. The second is the liquidity clientele literature pioneered by Amihud and Mendelson (1986) and followed by liquidity-adjusted asset pricing papers. The third is the market fragmentation literature studying network externalities when multiple trading venues compete for order flow, and traders behave strategically according to the Nash equilibrium dynamics.

The ETF literature has mostly focused on various effects of ETFs. For example, a number of studies investigate how ETFs affect market fragility by propagating market-wide demand shocks (Malamud, 2015; Ben-David, Franzoni, Moussawi, 2014; Krause, Ehsani, and Lien, 2014; Chinco and Fos, 2017; Bhatcharya and O'Hara, 2017). Multiple studies also explore the effects of ETFs on individual securities: Dannhauser (2016), Madhavan (2016), Lettau and Madhavan (2018), Madhavan and Sobczyk (2016), Glosten, Nallareddy, and Zou (2016), Wermers and Xue (2015), Marshall, Nguyen and Visaltanachoi (2013), and Li and Zhu (2016) argue that ETFs improve price discovery in the underlying securities. On the other hand, Hamm (2014), Da and Shive (2018), Israeli, Lee and Sridharan (2017) and Agarwal, Hanuona, Moussawi, and Stahel (2018) find that ETFs may harm informational efficiency in underlying securities through incorporating market-wide news rather than idiosyncratic news, and by causing their constituents to experience higher trading costs.

[3] *Financial Times*, 2018.

Little attention has been paid to how ETFs compete and why there is a considerable heterogeneity in the ETF investor base, which is what our paper brings to this literature. The existing ETF studies suggest that the high level of market liquidity offered by ETFs is one of their most attractive features compared to unlisted funds (Madhavan, 2016). Furthermore, industry practitioners point out that "Many institutional investors won't touch an ETF with volume less than $100 million a day or that isn't the most liquid ETF in the category" (Balchunas, 2016). Our paper recognizes that ETF investors value their liquidity and takes this notion one step further to show that liquidity also affects fee setting and competition between ETFs. In highlighting the dominance of institutional traders in the most liquid ETFs, our paper is consistent with Huang, O'Hara and Zhong (2018), Li and Zhu (2016), Easley, Michayluk, O'Hara, and Putnins (2018), and Xu, Yin, and Zhao (2018). However, unlike the above papers, we do not zoom in on specific institutional uses of ETFs (e.g., hedging industry-specific risk, circumventing short sale constraints or employing market timing strategies), but rather highlight how investor heterogeneity (i.e., the difference in holding horizons between institutional and retail investors) gives rise to liquidity clienteles and allows the most liquid ETFs to charge higher fees in equilibrium.

In the broader context of mutual funds literature, our paper is related to studies on how mutual funds compete and set fees. The seminal paper of Berk and Green (2004) that has shaped the thinking on this issue shows that in equilibrium mutual funds set fees such that the after-cost returns faced by investors are equalized across funds. Variation in fund manager ability to generate alpha drives variation in the fees across funds, with highly skilled managers being able to charge higher fees and capture more of the alpha generated. We show that competition between ETFs is quite different in two important regards. First, there is no attempt to generate alpha among standard index tracking ETFs, so the key driver of fee differences in mutual funds does not apply to ETFs. At the same time, secondary market liquidity, which does not play a role in competition between unlisted mutual funds is instrumental in the equilibrium for ETFs. Therefore, although both listed and unlisted funds are mutual investment vehicles, they operate and compete in rather different ways.

The competition strategies among the passive funds are particularly interesting in the light of recent developments in the competitive landscape. In July and August 2018, two big fund managers announced slashing their fees to zero: Fidelity did so for two of their index funds' MERs, and Vanguard for their brokerage platform fees, allowing investors to trade ETFs for free[4]. Just as

[4] See the news releases for Vanguard here https://www.cnbc.com/2018/07/02/vanguard-slashing-costs-on-nearly-all-etfs-even-rival-schwab.html , and for Fidelity here https://www.businessinsider.com/ap-fidelity-slashes-fees-as-funds-battle-for-investors-2018-8/?r=AU&IR=T

our model predicts, passive fund issuers can choose to compete at either extreme: by setting fees close to zero while being relatively illiquid (as is the case for Fidelity's index funds), or by charging above-zero MERs while being highly liquid (as is the case for Vanguard's ETFs that you can trade at zero cost).

We also contribute to the liquidity clientele literature that originates from the seminal study by Amihud and Mendelson (1986) and relates market liquidity to asset pricing. Amihud and Mendelson (1986) show that long-horizon investors can earn higher returns by holding less liquid stocks, while short-horizon investors are willing to sacrifice some returns in more liquid stocks for the benefit of immediate execution. A similar mechanism is at play in our model and drives the clientele effect whereby short-horizon investors hold highly liquid high fee ETFs and long-horizon investors hold less liquid low fee ETFs. However, our model differs in several important regards. Liquidity in our model is endogenous and a function of investor choices, whereas in Amihud and Mendelson the bid-ask spread is exogenously given. Importantly in our model, ETF issuers are aware of the liquidity/fee tradeoffs that investors face and set fees accordingly to capture particular clienteles. Thus, our model includes strategic clientele capture.

A number of empirical studies document the inverse cross-sectional relation between asset returns liquidity, after controlling for risk (Brennan and Subrahmanyam,1996; Amihud, 2002), while subsequent literature shows that investors demand compensation for liquidity risk (Hasbrouck and Seppi, 2001; Chordia, Roll, and Subrahmanyam, 2000; Huberman and Halka, 2001). We differ from these studies in taking a more direct approach of inferring liquidity premia from MERs of ETFs following identical underlying portfolios. In that respect, our approach is more similar to the bond market papers comparing yield differentials between different instruments of similar coupon and maturity (Amihud and Mendelson, 1991; Krishnamurthy, 2002; Longstaff, Mithal, and Neis, 2005; Subrahmanyam, Jankowitsch, and Friewald, 2012).[5] However, in case of OTC-traded instruments like bonds, CDS, swaps etc., liquidity is scarce and search costs are high, hence investors sacrifice yield to avoid extreme illiquidity (Duffie, Garleanu and Pedersen, 2005), while in case of ETFs investors are on the opposite side of the liquidity spectrum: they accept higher MER to access extremely liquid securities.

The common thread between these studies and ours is that liquidity differentials can in fact explain the apparent law of one price violations. While many papers point to limits to arbitrage as one of the major explanations for why such violations can persist (e.g., Shleifer and Vishny 1997), our results imply another driver can be differences in market liquidity. Differences in liquidity can

---

[5] See Holden, Jacobsen, and Subrahmanyam (2014) for a more detailed literature review on liquidity premia.

lead to apparent violations of LOOP not simply because illiquidity creates limits to arbitrage, but rather because investors value liquidity. Therefore, two assets with identical cash flows can have persistently different prices if they have different liquidity. This tendency is what drives the fee differentials among competing ETFs. Consequently, the LOOP principle could be expanded to a broader concept of liquidity-adjusted LOOP (LALOOP).

Third, at a broad level, our model is based on the classic microeconomic models of non-cooperating agents, with the resulting Nash equilibrium (Nash, 1951). What differentiates our model from many other settings that have been modelled is that the investors ("consumers") care not only about fees ("prices") but also a second dimension, that being liquidity. What makes our setting interesting is that liquidity differs from static product features in that it is endogenous. Fee setting by ETF issuers affects investor choices, which affects liquidity, which then effects fee setting, and so forth. Our model is therefore closest in spirit to microeconomic models of competition for goods that have network externalities (e.g., Katz and Shapiro, 1985; Economides,1996), which in our case are liquidity externalities. However, our model differs from standard network externalities models by having clienteles form due to heterogeneity in investment horizons, which is an important and realistic feature.

Finally, our paper is indirectly related to the vast literature on fragmentation of trading across competing trading venues. There are certain parallels between the fragmentation of investors across different ETFs and fragmentation of trading volumes across multiple exchanges.[6] For example, trading venues compete on speed, fees, and fee structures, which leads to clientele effects in fragmented markets (Faucault, Kadan and Kandel, 2005; Yao and Ye, 2018). Similar to Faucault et al. (2005), we model the high-turnover (impatient) ETF traders behaving strategically in choosing the cost-minimizing ETF (trading venue). However, the trading venue liquidity, unlike ETF liquidity, is not driven by investors' holding horizons. Hence, the nature of first mover advantage in the ETF market is different: while exchanges accumulate liquidity simply by the virtue of being the first mover, ETFs accumulate liquidity by attracting the high-turnover investors. In other words, sufficient investor heterogeneity is a necessary condition for separating equilibria among ETFs based on the same index, which is not the case for competition among trading venues.

Another parallel is with the models of fragmented trading, in which multiple equilibria are possible, depending on transaction costs and the traders' conjectures about trading volumes in different venues (Pagano, 1989). In a similar manner, we show that ETFs compete on liquidity and fees to attract investors with different holding horizons, but we take the model a step further by

[6] See Gomber, Sagade, Theissem, Weber, and Westheide (2017) for a more detailed review of consolidation vs competition literature.

introducing holding horizon heterogeneity. The fragmentation literature also investigates the welfare effects of competition. For example, Mendelson (1987) models the trade-off between market thinness in case of fragmentation and high order communication costs in case of consolidation, while Baldauf and Mollner (2016) model the trade-off between benefits from lower bid-ask spreads and the costs from cross-venue arbitrage, Colliard and Faucault (2012) suggest that competition is beneficial due to lower trading fees, and Pagnotta and Philippon (2011) argue that product differentiation benefits heterogenous investors, which is in line with our proposition about the clientele effect in the ETF market.

2. Theory model of ETF competition

2.1. Baseline model structure

Consider a simple oligopolistic market, with two competing ETFs based on the same index (as in Stackelberg, 1934). Profit-maximizing ETF issuers A and B choose their management fees to maximize profits, taking into account the strategy of the competing ETF issuer. Issuer A gets to choose his clientele first and charges the management fee (as a percentage of assets under management, AUM) f_A while B charges the management fee f_B.[7] Both issuers face the same annual costs f_{fix}, which cover payments to the index provider and the administrative costs of running the fund. The total investments (AUM) in the given ETF index is X, with ETF A having x_A share of the AUM (in %), and ETF B having x_B share (in %).

ETF issuers A and B face the following profit maximization problems, which involve choosing the fee that maximizes their profits (π_A and π_B, respectively):

$$\max_{f_A}(\pi_A) = max\{Xx_Af_A - f_{fix}\} \tag{1}$$

$$\max_{f_B}(\pi_B) = max\{Xx_Bf_B - f_{fix}\} \tag{2}$$

In this model, ETF issuers behave as Stackelberg (1934) duopolists: to arrive at equilibrium fees, they follow a sequential game and set their prices (fees) recognizing competitor's optimal response. In our model, A is a market leader in that he is the one with a more liquid ETF. However, we depart from the classic Stackelberg model in that investors care not only about fees (MERs), but also about liquidity of an ETF. Both dimensions — fees and liquidity — are endogenous:

[7] We show in Appendix 2 that $f_A > f_B$, as ETF A is more liquid and as a first mover gets to choose his clientele (high-turnover investors, who have lower sensitivity to MER). In reality, index issuers can sign exclusivity agreements with ETF issuers, essentially guaranteeing the embargo period during which the index won't be licensed to competing issuers. So monopolistic fee setting by the first issuer is a realistic feature of the model, and allows the first mover to cement the liquidity advantage before the competitor arrives.

issuers' fees affect investors' choice of ETFs, which in turn affect liquidity of ETFs, which again affects fee setting and so on.

There are two types of investors in this ETF index — high-turnover investors (with holding horizon h_F, measured in years on fractions of a year) and low-turnover ones (with holding horizon $h_S > h_F$). The proportion of high-turnover investors is w_F, the proportion of low-turnover is w_S, and each investor can choose one ETF only: either A or B. The average trading frequency of high-turnover investors is $\lambda_F = \frac{1}{h_F}$ — the number of times per year that high-turnover investors exit their ETF positions[8], and the average trading frequency of low-turnover investors is $\lambda_S = \frac{1}{h_S}$. A marginal high-turnover (low-turnover) investor's time until transaction (inter-arrival time) is therefore $H_F = \frac{1}{w_F \lambda_F} = \frac{h_F}{w_F}$ ($H_S = \frac{1}{w_S \lambda_S} = \frac{h_S}{w_S}$ for low-turnover).[9]

ETF investors trade though market makers, who set the bid-ask spread to recover the costs of liquidity provision. According to classic market microstructure models, the three main costs that market makers recover through the bid-ask spread are fixed order processing costs (e.g., admin costs), adverse selection costs and inventory costs (De Jong and Rindi, 2009). Adverse selection costs approach zero in the ETF market, given that the prices of constituent securities are readily available for market makers, which minimizes the risk of trading at stale quotes. Therefore, we argue that fixed costs and inventory costs are the most relevant costs of liquidity provision in the ETF market. We model these costs as a function of the inter-arrival time of investors: $C_{LIQ} = cH$, where c is a constant order processing cost, and H is the inter-arrival time. The rationale behind behind this functional form is that (i) fixed costs per trade are reduced when trades occur more frequently, and (ii) inventory holding risks are reduced when trade arrivals are more frequent. The proposed functional form is in line with Garman (1976), Ho and Stall (1980, 1981, 1983) and Amihud and Mendelson (1986), as well as Foucault, Kadan, and Kandel (2005) and Rosu (2009). The liquidity cost decreasing in trading activity is also consistent with the empirical evidence in Amihud and Mendelson (1986) and McInish and Wood (1992).

An individual investor i that considers a single investment in an ETF j involving buying the ETF, holding the ETF, then selling the ETF faces two types of costs:

[8] Investors' arrival intensity is modelled as a Poisson process with parameter λ_F for high-turnover investors, and λ_S for the low-turnover.

[9] Accounting for the fact that there are two sides to the transaction would scale the inter-arrival time by 2 (assuming balanced order flow), as an average high-turnover buyer would face the arrival intensity of high-turnover sellers equal to $0.5\lambda^F$. We do not model this explicitly, as in subsequent steps multiplying by a constant does not materially alter the model results.

(i) $C_{MER_{ij}} = f_j h_i$, the annual fee charged by the ETF j that he selects (the ETF's MER).[10]

(ii) $C_{LIQ_{ij}} = cH_{ij}$, a round-trip trading cost charged by the market maker, that is linear in the expected waiting time until a trade counterparty arrives, H_{ij}, is similar to Foucault, Kadan, and Kandel (2005) and Rosu (2009).

Investors arrive to the market knowing their holding horizon (h_i) and the round-trip transaction cost (cH_{ij}), hence they choose between two ETFs, A or B, to minimize their overall cost of the investing in the ETF market, $C_{TOTAL_{ij}} = C_{LIQ_{ij}} + C_{MER_{ij}} = cH_{ij} + f_j h_i$. ETF investors are competitive, and each is a small proportion of their respective group, hence individual investor's decision does not have a material effect on the proportion of high-turnover vs low-turnover investors.

2.2. Nash equilibrium for investors

Suppose high-turnover investors choose ETF A and low-turnover investors choose ETF B. Then, they incur the following total costs:

$$C_{TOTAL}^{F} = f_A h_F + \frac{ch_F}{w_F} \tag{3}$$

$$C_{TOTAL}^{S} = f_B h_S + \frac{ch_S}{w_S} \tag{4}$$

To identify the conditions under which this choice of ETFs is an equilibrium (a separating equilibrium as the two investor groups hold different ETFs), consider an investor's standpoint. If an individual high-turnover investor decided to deviate from this strategy and instead invest in ETF B, then he would incur total costs equal to $C_{TOTAL}^{F*} = f_B h_F + \frac{ch_S}{w_S}$. Note that the MER portion of his costs $C_{MER}^{F*} = f_B h_F$ is driven by the fees in ETF B and this investor's holding period is h_F. The liquidity portion of his costs $C_{LIQ}^{F*} = \frac{ch_s}{w_s}$ is driven by the time it takes for the trading counterparty to arrive. Because ETF B market is dominated by low-turnover investors, the inter-arrival time is $\frac{h_S}{w_S}$. An individual high-turnover investor will not have an incentive to deviate from the strategy of choosing ETF A, if his total cost of investing in A is lower than o choosing B:

$$f_A h_F + \frac{ch_F}{w_F} < f_B h_F + \frac{ch_S}{w_S} \tag{5}$$

Following the same logic, if an individual low-turnover investor chose to deviate from the strategy of investing in B, and invest in A instead, then his costs would be $C_{TOTAL}^{S*} = f_A h_S + \frac{ch_F}{w_F}$.

[10] Note that j can take two values: A or B, and i can take two values: high-turnover or low-turnover.

A low-turnover investor will choose not to deviate from the original strategy of investing in B, if his total cost of investing in B is lower than that of investing in A:

$$f_B h_S + \frac{ch_S}{w_S} < f_A h_S + \frac{ch_F}{w_F} \tag{6}$$

The inequalities (5) and (6) jointly determine the condition under which we obtain a separating Nash equilibrium (with the high-turnover investing in A, and the low-turnover in B), in which no investor wants to deviate from the chosen strategy:

$$\frac{c}{1-w_F} - \frac{ch_F}{w_F h_S} < f_A - f_B < \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F} \tag{7}$$

Condition (7) suggests that the separating equilibrium is stable (i.e., no investor has an incentive to deviate from the chosen strategy), if the fee differential between ETFs A and B is bounded from both sides. If the fee differential is too small (i.e., below $\Delta_{min} = \frac{c}{1-w_F} - \frac{ch_F}{w_F h_S}$), the low-turnover investors switch to the more expensive ETF (i.e., the MER discount in the low-liquidity ETF B becomes small enough to make MER saving too low for the low-turnover investor and incentivize him to switch to the high-MER ETF A, which offers lower liquidity costs). If the fee differential is too large (i.e., above $\Delta_{max} = \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F}$), the high-turnover investors switch to the less liquid ETF (i.e., the MER premium is too high for them to justify the liquidity advantage). Condition (7) also guarantees that ETF A held by the high-turnover is more liquid than ETF B held by the low-turnover. In other words, the cost of liquidity in A is lower than that in B. The full derivation is provided in Appendix 1.

2.3. Nash equilibrium for ETF issuers

As in classic duopoly models (Stackelberg, 1934), issuers A and B act independently and sequentially, and optimize their strategy given the competitor's optimal strategy. The issuers are aware of investors' optimum choice (i.e., that the high-turnover invest in A, and the low-turnover in B, if A is more liquid than B and fees are within the acceptable range), and take into account investors' cost sensitivity when setting their fees (i.e., setting the fee so that the fee differential satisfies inequality (3) derived above: $\Delta_{min} < f_A - f_B < \Delta_{max}$). The fee setting is summarized in Figure 1.

<Fig. 1 here>

As shown in Appendix 2, a separating equilibrium emerges when issuer B charges the fee just at the margin of $f_0 + \Delta_{max} - \Delta_{min}$ (where f_0 is B's breakeven fee), and A charges the fee just at the margin of $f_0 + \Delta_{max}$. For this equilibrium to be stable, it must hold that A's profit from the high-turnover clientele is higher than the potential profits from both high-turnover and low-turnover clienteles at a lower fee. Formally, the following inequality must hold for A to find it suboptimal to deviate from the chosen strategy:

$$f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max}) \tag{8}$$

Consider what happens if either issuer chooses to deviate from the conjectured strategy. Suppose A lowers his fee below $f_A = f_0 + \Delta_{max}$ to get both clienteles. Then, B responds by lowering his fee to maintain the Δ_{min} difference between his fee and A's, and keep the low-turnover clientele. This would result in the race to the bottom dynamics, whereas A can keep lowering his fee down to $f_A = f_0 + \Delta_{min}$, at which point he achieves a market share of 100% and competes B out of business. For A not to be interested in this scenario, A's profit from the high-turnover clientele at a higher fee must be greater than the profit from both clienteles at a lower fee: $f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max})$, which is outlined in condition (8) above.

Suppose A increases his fee, charging up to $f_A = f_0 + 2\Delta_{max} - \Delta_{min}$ without losing the high-turnover clientele, and earning a higher profit. However, at this fee level of A, B can reduce his fee by a marginally small amount and attract both the high-turnover and the low-turnover clienteles, which would prompt A to lower his fee, resulting in sequential lowering of fees by both issuers until they reach an equilibrium : either at $f_A = f_0 + \Delta_{max}$ and $f_B = f_0 + \Delta_{max} - \Delta_{min}$, or at $f_A = f_0 + \Delta_{min}$ and $f_B = f_0$ (with the second case essentially competing B out of business and A achieving 100% market share).

Hence, it is not optimal for A to either increase or decrease his fee relative to the conjectured strategy. Now, consider whether B has an incentive to change his fee.

Suppose B lowers his fee below $f_B = f_0 + \Delta_{max} - \Delta_{min}$. That would result in lower profits for B, which is not optimal. If B increases his fee, that decreases the fee differential below Δ_{min}, so B loses the low-turnover clientele to A. Hence, it is not optimal for B to either increase or decrease his fee relative to the conjectured strategy. The full derivation is provided in Appendix 2.

2.4. Conditions for the separating equilibrium and testable hypotheses

The separating equilibrium occurs when both investors' and issuers' conditions are satisfied. Recall from inequality (7) and Appendix 1 that the high-turnover investors will choose ETF A, and the low-turnover ETF B, if A is more liquid than B, and the fee differential between the two is neither too large (not to give the high-turnover an incentive to switch to the lower-MER ETF B), nor too small (not to give the low-turnover an incentive to switch to the more liquid ETF A). ETF issuers recognize this and set their fees so that to cater to their respective clienteles (i.e., A caters to the high-turnover investors, and B to the low-turnover). However, issuer A only allows this to happen, if his profits from catering to the high-turnover investors are greater than his potential profits from catering to both the high-turnover and the low-turnover at lower MERs. Combining investors' and issuers' conditions that must hold in separating equilibrium[11]:

$$\begin{cases} LIQ_A > LIQ_B \\ f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max}) \end{cases} \tag{9}$$

The second condition can be expressed as a function of liquidity differential $\Delta LIQ = LIQ_A - LIQ_B$:

$$\frac{f_{fix}}{X} + \Delta LIQ\left(\frac{1}{h_S} - \frac{w_F}{h_F}\right) < 0 \tag{10}$$

Recognizing that $\Delta LIQ > 0$, inequality (10) implies that the proportion of high-turnover investors (w_F) must be greater than the homogeneity parameter (h_{FS})[12], for issuer A to find it profitable to cater to only the high-turnover clientele instead of starting the price war trying to capture both the high-turnover and the low-turnover[13]:

$$w_F > h_{FS} \tag{11}$$

Solving the system of inequalities in (9) provides a set of restrictions on model parameters that guarantee the separating equilibrium. Given that there are factors beyond the model that might affect whether or not the separating equilibrium is achieved, characterizing the strength of forces driving the system into the separating equilibrium is worthwhile. Therefore, we proceed to outline the conditions for separating equilibrium as $y > 0$, where y is the propensity for separating equilibrium, re-expressed from inequality (10):

$$y = w_F^2\left(f_{fix}h_{FS} + cXh_{FS} + cX\right) - w_F h_{FS}\left(f_{fix} + 2cX + cXh_{FS}\right) + ch_{FS}^2 X \tag{12}$$

The separating equilibrium is achieved for the model parameters that allow for $y > 0$, and adhere to the restriction $w_F > h_{FS}$. In appendix 4, we show mathematically how this function behaves when parameter values change. Below, we provide the intuition and resulting hypotheses:

> ***Hypothesis 1****: The separating equilibrium is more likely when the proportion of high-turnover ETF investors in a given index is higher.*

[11] Denoting ETF A's liquidity as $LIQ_A = -C_{LIQ}^A$ and ETF B's liquidity $LIQ_B = -C_{LIQ}^B$.

[12] Note that $h_{FS} = \frac{h_F}{h_S}$.

[13] Note that inequality (11) is more restrictive than investors' condition on liquidities ($LIQ_A > LIQ_B \rightarrow w_F > \frac{h_{FS}}{1+h_{FS}}$). Hence, $w_F > h_{FS}$ restriction is driven by the competitive dynamics of the issuers that requires sufficiently high proportion of high-turnover investors with sufficiently short holding horizons (relative to the low-turnover) to enable A not to lower his fees.

Hypothesis 2*: The separating equilibrium is more likely when ETF investors' holding horizons in a given index are less homogenous.*

Intuitively, the separating equilibrium is more likely when the proportion of high-turnover investors is large (hence making it more attractive for issuer A to settle for the high-turnover clientele instead of starting the price war trying to capture both the high-turnover and the low-turnover). The separating equilibrium is also more likely when investors are highly dissimilar in their holding horizons (i.e., high heterogeneity, or low homogeneity parameter h_{FS}), because in that case, high-turnover investors are even more likely to accept high MERs charged by issuer A, while the low-turnover are even less likely, hence making it more attractive for issuer A to keep only the high-turnover clientele and charge them high MERs. We provide the full derivation in Appendix 3.

Hypotheses 3 and 4 characterize the separating equilibrium with respect to the fixed costs (f_{fix}) and AUM (X) parameters.

Hypothesis 3*: The separating equilibrium is more likely when the ETF issuers' fixed costs in a given index are lower.*

Hypothesis 4*: The separating equilibrium is more likely when the combined AUM of all ETFs in a given index is higher.*

Separating equilibrium is more likely when issuers' fixed costs are low. The fixed costs essentially cover payments to the index provider and overheads, hence unsubstantial costs make the barriers to entry lower, which in our model translates into the breakeven fee that is low enough for the low-MER ETF to be able compete with the incumbent.

A similar mechanism is at play when the combined assets under management of the two ETFs are high, as this allows for economies of scale that make the breakeven MER sufficiently low for ETF A to be able to compete with B. To provide further intuition for how parameter values affect separating equilibrium, we provide plots with respect to various model parameters in Figure 2.

<Fig. 2 here>

Figure 2 plots the system of inequalities (9) as a plain that characterizes the forces for separating equilibrium in a three-dimensional space. The y dimension can be interpreted as how strong the forces for separating equilibrium are for each of the plotted parameter combinations on x and z axes, keeping the other two parameters constant.[14] The plots provide useful intuition to complement the formal model. For example, Panel A shows that the force for separating equilibrium is stronger for higher values of w_F parameter and lower values of h_{FS}parameter. This suggests that indices that have a high proportion of high-turnover investors with highly dissimilar holding horizons are more likely to end up in separating equilibrium, keeping other factors constant.[15] Note that we fix the other two parameters — combined AUM and fixed costs — at levels actually observed for an average index in separating equilibrium, hence values of y are above zero for most of parameter combinations of h_{FS} and w_F, except the most disadvantageous — low w_F and high h_{FS}. So the way to interpret the y dimension on Panel A plot is by asking the following question: given the other conditions for separating equilibrium are satisfied, which combinations of h_{FS} and w_F parameters are more likely to result in separating equilibrium for a given index?

Panels B through E of Figure 2 provide a similar intuition with respect to other parameters. Panel B suggests that higher values of combined AUM per index (X) and higher proportion of high-turnover investors (w_F) are more likely to satisfy the conditions for separating equilibrium, keeping other parameters constant. Panel C suggests that higher values of combined AUM per index (X) and lower values of similarity of holding horizons (h_{FS}) are more likely to lead to separating equilibrium, keeping other factors constant. Panel D suggests that separating equilibrium is more likely for low values of fixed costs (f_{fix}), combined with high proportion of high-turnover investors (w_F). Panel E suggests that separating equilibrium is more likely for low values of fixed costs (f_{fix}), combined with low homogeneity parameter (h_{FS}).

2.3. Liquidity-fee relation in separating equilibrium and testable hypotheses

As we have shown in Subsection 2.3 and in Appendix 2, in Nash equilibrium, ETF issuers charge the following fees in equilibrium:

$$f_A = f_0 + \Delta_{max} \tag{13}$$

$$f_B = f_0 + \Delta_{max} - \Delta_{min} \tag{14}$$

Hence, the fee differential between ETF A MERs and ETF B MERs in equilibrium is:

$$f_A - f_B = \Delta_{min} = \frac{c}{w_S} - \frac{ch_F}{w_F h_S} \tag{15}$$

[14] Recall we have four parameters in total: h_{FS}, w_F, X, f_{fix}. We vary two of them in each plot, and fix the other two at average level observed for our sample ETFs.
[15] Note that the c parameter is calibrated to the data to minimize the sum of squared deviations from the equilibrium relation between fees and liquidity.

At the same time, investors incur the following costs when investing in ETFs A and B in Nash equilibrium:

$$C_{LIQ}^{A} = \frac{ch_F}{w_F} \tag{16}$$

$$C_{LIQ}^{B} = \frac{ch_S}{w_S} \tag{17}$$

Expressing the liquidity cost differential as a function of fee differential:

$$C_{LIQ}^{A} - C_{LIQ}^{B} = c\left(\frac{h_F}{w_F} - \frac{h_S}{w_S}\right) = -h_S(f_A - f_B) \tag{18}$$

Recognizing that the liquidity cost differential is just the liquidity differential with a minus sign (denoting the ETF A's liquidity as LIQ_A and ETF B's liquidity as LIQ_B), we obtain the following liquidity-fee relation for ETFs in separating equilibrium:[16]

$$LIQ_A - LIQ_B = h_S(f_A - f_B) \tag{19}$$

Note that in the above relation, the liquidity differential between the high-fee and the low-fee ETF depends only on the fee differential and on the low-turnover investors' holding horizon. This implies that in equilibrium, the value of ETF liquidity in a given index is set by the low-turnover investors. This result is driven by the Nash equilibrium dynamics (see Appendix 2), which introduces an asymmetry in the optimum fee setting behavior by issuers. In simple words, low-turnover investors hold a "swing vote" in whether or not to choose ETF B, hence it's the low-turnover investors' minimum acceptable fee differential that constraints how high B can set his fee, and it's therefore low-turnover investor's fee sensitivity that determines the fee differential between A and B.

To formalize this intuition, recall that in equilibrium, the fee difference between A and B is just at the margin of MER differential that corresponds to the low-turnover investors' indifference condition (as in inequality (6)). This means that the equilibrium fee differential between A (the high-MER ETF) and B (the low-MER ETF) is the lowest possible fee differential at which B does not lose the low-turnover clientele to A. Hence, it's the low-turnover investors' fee sensitivity that drives the trade-off between liquidity cost and MER cost in Nash equilibrium.

Another way to interpret the equilibrium relation between liquidity and MER differentials in Eq. (19) is to see it as the law of one price condition. For a low-turnover investor to be indifferent between investing in ETF A or ETF B, it must be that the annual savings in liquidity cost that he generates by choosing A ($C_{LIQ}^{B} - C_{LIQ}^{A}$) are the same as the annual savings in MER that he generates by choosing B ($h_S(f_A - f_B)$.). Note that h_S can be seen as a coefficient of proportionality that ensures liquidity costs and MER costs are both measured on an annual basis.

[16] The full derivation provided in Appendix 4.

Overall, the separating Nash equilibrium is characterized by the high-turnover investors (i.e., high-turnover investors) choosing the high-MER ETF, and the low-turnover (i.e., the low-turnover investors) the low-MER ETF. Additionally, the model suggests that in equilibrium, we should observe a positive relation between liquidity and MER differentials. We formalize these model predictions in the following hypotheses:

Hypothesis 5: *For ETFs tracking the same index, liquidity is positively related to fees.*

Hypothesis 6*: For ETFs tracking the same index, turnover is positively related to fees.*

3. Empirical analysis

3.1. Data and descriptive statistics

We obtain daily data from ETF Global, which covers the full universe of US-domiciled ETFs, and use CRSP to add daily spread and price measures for each ETF. We restrict our sample to equity ETFs traded on US markets, excluding ETNs (exchange-traded notes), leveraged or inverse ETFs, and ETFs that are hedged versions of the original fund. The full sample includes 1035 equity ETFs traded in the US in 2017. Since the model predictions concern static equilibrium relations, we test them in the cross-section of ETFs, computing an annual average for each variable for each ETF. The underlying data for each ETF is at daily frequency, and covers the year 2017.

For the tests involving indices in separating equilibrium, we filter the data for indices tracked by multiple ETFs and arrive at the core sample of 60 ETFs based on 24 indices. Each ETF in our sample shares the benchmark with at least one other ETF. In general, we have two or three ETFs per benchmark. See Table 1 for the full list of ETFs used in our analysis.

<Table 1 here >

We cross-validate the identified list of same index ETFs by using three alternative approaches, which confirm that our sample correctly identifies ETFs that invest in the same portfolio of stocks. Firstly, we manually check each ETF on ETF.com, an online provider of ETF statistics that allows to search for ETFs by ticker and identify competitor ETFs with the same portfolio exposure. Secondly, we compare our list with that of same benchmark ETFs provided in the appendix of Box, Davis, and Fuller (2017) who use Morningstar data in their analysis. Thirdly, we compute the empirical measure of portfolio similarity by using ETF Global data on portfolio weights pertaining to each ETF constituent. Specifically, the similarity measure is computed as follows:

$$Sim_{AB} = 1 - Dist_{AB} = 1 - \frac{1}{2}\sum_{i=1}^{N}|w_{iA} - w_{iB}| \tag{20}$$

Where w_{iA} is the weight of stock i in ETF A, w_{iB} is the weight of stock i in ETF B, N is the number of stocks in both ETFs A and B, and $Dist_{AB}$ is the distance measure between ETFs A and B. ETF pairs with similarity measure close to 1 (distance measure close to 0) have practically identical portfolio exposure, hence validating that they indeed follow the same benchmark index.

The descriptive statistics on key variables is provided in Table 2.

<Table 2 here >

As shown in descriptive statistics, our sample covers 1035 ETFs with a combined AUM of \$2.26 trillion, and combined daily value traded of \$52.78 billion. The subsample with multiple ETFs per index accounts for 36% of the total AUM and 47% of total daily dollar volume of all equity ETFs in the sample. Given that indices in separating equilibrium (i.e., with multiple ETFs per index) account for only 5.8% of all ETFs by count, the high concentration of AUM and dollar volume in those indices is quite striking.

Since we test for separating equilibrium relations in a given index, the descriptive statistics are computed as index-level averages. The average MERs are 22.67 bps in separating equilibrium, and more than twice higher in non-separating equilibrium (50.13bps). The difference in relative spreads is even wider: investors pay 6.19 bps on average for a round-trip transaction in ETFs in separating equilibrium, compared to 29.02 bps if an index is tracked by one ETF only. Separating equilibrium ETFs also tend to be broader (854 constituents on average, compared to 276), have significantly larger AUM (\$34.32 billion, compared to 1.48 billion), and larger daily dollar volumes (\$1.03 billion vs \$0.03 billion).Separating equilibrium ETFs are also two times more likely to be issued by Vanguard, BlackRock or State Street (the top 3 issuers) or track a major index (branded by MSCI, S&P or Russell).

3.2. OLS regression results

We test the theory model predictions empirically by exploring the relation between same index ETFs' liquidity and their management expense ratios. First, we plot the demeaned versions of relative spread, dollar volume and turnover against demeaned MERs for each ETF.

<Fig.3 here >

As illustrated in Figure 3, ETFs with lower relative spreads (compared to their same index peers) tend to have higher MERs (compared to those same index peers), as predicted by the model. However, spreads are imperfect measures of liquidity, especially for highly liquid tick-constrained ETFs. That is why we use dollar volumes as a more natural proxy for liquidity. The analysis for dollar volumes is also consistent with the theory model predictions. ETFs with higher dollar volumes (compared to their same index peers) tend to have higher MERs (compared to those same index peers). This is in line with the model intuition stated in Hypothesis 5: that investors "pay" for higher ETF liquidity by accepting higher MERs.

Next, we perform a similar exploratory analysis for the turnover-MER relation. As shown in Figure 3, ETFs with higher turnovers (compared to their same index peers) tend to have higher MERs (compared to those same index peers), consistent with Hypothesis 6. The high-MER high-turnover ETFs also tend to be those that were launched first (i.e., first movers). This is in line with the clientele effect, which arises due to heterogenous holding horizons among investors. Recall from the theory model derivation (see Appendix 2) that this effect emerges because the first mover accumulates liquidity before the second ETF is launched, and the superior liquidity serves as a source of monopolistic rents, because the high-turnover investors (i.e., those with high turnovers) prefer the highly liquid high-MER ETF, even when there's a low-MER alternative.

To formally test the above effects, we run cross-sectional OLS regressions with index fixed effects. The baseline form of the regression model is as follows:

$$MER_i = \alpha + \beta_1 LIQ_i + \beta_2 TrackingError_i + \mu_{IND_i} + \varepsilon_i \tag{21}$$

Where LIQ_i can refer to relative spread, log-turnover or log-dollar volume of ETF i, depending on the model specification, MER_i is the net expense ratio of ETF i, $TrackingError_i$ is the tracking error of ETF i, and μ_{IND_i} is the index dummy for ETF i. There are 23 index dummies (the omitted dummy is S&P MidCap 400 Value Index). Since there are multiple ETFs following each index in our sample, index fixed effects allow us to capture within-index variation in MERs and liquidity measures.[17] In other words, regressions with index fixed effects ask the question "For ETFs tracking the same index, how much MER are investors sacrificing, on average, for an extra unit of liquidity?" Note that we control for both tracking error and MER, as those are the only two factors that should account for any differences between ETFs that track the same underlying index. We report the regression results in Table 3.

<Table 3 here >

Regression results suggest that for an average ETF, 1 bps increase in relative spread is associated with 0.51 bps lower MER (compared to other same index ETFs). This result is statistically significant at 1% level, and economically meaningful. The coefficient on tracking error is insignificant, so it does not materially affect spread differences between same index ETFs. This

[17] Fixed effects regression with index fixed effects is equivalent to demeaning both dependent and independent variables with respect to the average value of each variable per index. For example, estimating the regression model in Eq. (21) is equivalent to estimating the following model (demeaning all variables using within-index transformation):
$MER_i - \overline{MER} = \alpha + \beta_1(LIQ_i - \overline{LIQ}) + \beta_2(TrackingError_i - \overline{TrackingError}) + \varepsilon_i$,
Where $\overline{MER}$ is the average MER of all ETFs tracking the same index, $\overline{LIQ}$ is the average liquidity measure (dollar volume, turnover or relative spread) for all ETFs tracking the same index, and $\overline{TrackingError}$ is the average tracking error of all ETFs tracking the same index.

result corroborates Hypothesis 5: indeed, investors value ETF liquidity and accept higher MERs in ETFs with lower costs of round-trip transaction, compared to same index ETF peers.

For convenience of interpretation, we run regressions for turnovers and dollar volumes using log-transformed variables. The results suggest that for an average ETF, 1% increase in turnover is associated with 0.24% higher MER, while 1% increase in dollar volume commands 1.15% higher MER. These results are in line with Hypotheses 5 and 6: indeed, high-turnover investors in a given index are more likely to transact in high-MER ETFs, and that allows issuers to charge them 1.15% higher MERs for 1% increase in the dollar volume.

We rule out alternative explanations for MER differences among same index ETFs by controlling for tracking error and introducing index fixed effects. Higher tracking error does decrease MERs slightly (compared to same index ETFs), but its effect is much weaker compared to the liquidity measures. For example, 1% increase in tracking error reduces MERs by 0.09%, while 1% increase in dollar volumes gives MERs a boost of 1.15%.

Note that the mathematical form of our regression equations flows directly from the theory model Eq. (19), which describes the equilibrium relation between MER and liquidity for ETFs tracking the same index. According to the theory model, ETF liquidity and MERs are jointly determined, given certain values of exogenous model parameters (heterogeneity of investors' holding horizons, combined AUM per index, issuers' fixed costs, and proportion of fast investors) that characterize the specific index which these ETFs track. Hence, our intention in regression analysis is not to capture the causal link between MERs and liquidity (and the theory model does not suggest there is such a link), but rather to estimate the average MER premium investors are paying for an extra unit of liquidity in same-index ETFs. The implicit assumption is that we observe ETFs in equilibrium. This assumption is plausible, given that we use one year averages to construct the cross section of MERs, liquidity and tracking error, and given that our sample ETFs had existed for a least three years before the period for which the measures were calculated, which is arguably a sufficient period to arrive at equilibrium.[18]

The regression results discussed above allow us to draw inferences with respect to the average ETF. However, it might be of interest to consider the MER-liquidity trade-offs with respect to the average dollar invested, especially since ETFs are rather dissimilar in terms of dollars invested (AUMs), so what holds for the average ETF might not hold for the average dollar invested. That is why we also report AUM-weighted least squares regression results to complement the baseline analysis. As shown in Table 3, the effects for dollar volumes and turnovers are stronger in

[18] The most recently launched ETFs in our sample are from year 2014, while we calculate the average variable values for the year 2017.

weighted OLS, compared to simple OLS. That is not surprising, as high-AUM ETFs tend to extract the highest liquidity rents in terms of higher MERs. As for relative spreads, the effect is present but not statistically significant, as the high-AUM ETFs tend to be tick-constrained and hence their liquidity advantages are manifested through dollar volumes rather than relative spreads.

3.4. Probit regression results

We also test the model predictions related to the driving forces of separating equilibrium. Recall that hypotheses 1–4 suggest that separating equilibrium is more likely to occur in indices with greater proportion of high-turnover investors, greater heterogeneity of investors' holding horizons, lower issuers' costs, and greater combined AUM. We test these predictions empirically by modelling the probability of separating equilibrium in all indices tracked by US-domiciled equity ETFs.

We do not observe the holding horizon differentials in single-ETF indices, nor do we have data on issuers' costs. The only readily available variable is the combined AUM of all ETFs per index. For the remaining variables, we use liquidity measures such as spreads and dollar volumes to proxy for the proportion of high-turnover investors in a given index, we use dummy variables for major index benchmarks to proxy for the extent of heterogeneity, and we use dummies for top 3 ETF issuers to proxy for the level of issuers' fixed costs. Results of probit regressions are reported in Table 4.[19]

<Table 4 here >

In general, regression results corroborate the model predictions. Specifically, we find that indices tracked by more liquid ETFs (both in terms of higher daily value traded and in terms of lower spreads) are more likely to be in separating equilibrium. Since greater ETF liquidity is the preferred habitat of the high-turnover investors, this is in line with Hypothesis 1. Further, the separating equilibrium is more likely to occur in major benchmark indices, such as those offered by S&P, Russell and MSCI. Because institutional investors are more likely to make short-term tactical allocations to those indices, the investor base in these indices is likely to be heterogenous. Hence, this result corroborates Hypothesis 2. We also find that indices with greater proportion of ETFs by top 3 issuers are more likely to be in separating equilibrium, in line with Hypothesis 3. The presence of Black Rock, Vanguard and State Street — the top 3 ETF issuers — is typically

[19] Given that indices in non-separating equilibrium significantly outnumber those in separating equilibrium, we perform the robustness check by randomly selecting 150 indices with single ETF per index, and combining them with the 24 indices in non-separating equilibrium, and run the same probit regression. Results are reported in Panel B of Table 4.

indicative of low-cost and low-fee environment. Finally, high combined AUM is a strong predictor of separating equilibrium, as suggested in Hypothesis 4.

4. Welfare analysis

4.1. Theoretical underpinnings

Recall that in separating equilibrium (i.e., when there are multiple ETFs on a given index), ETF A is necessarily more liquid than ETF B. The extra liquidity in ETF A is valuable from the high-turnover investors' perspective, as it allows them to turn over their positions at a lower cost. This enables issuer A to charge MERs over and above what would be a competitive level, hence generating oligopolistic rents in equilibrium. In this scenario, issuers A and B earn the following profits:

$$\pi_A = (f_0 + \Delta_{max})Xw_F - f_{fix} \tag{22}$$

$$\pi_B = (f_0 + \Delta_{max} - \Delta_{min})X(1 - w_F) - f_{fix} \tag{23}$$

The costs incurred by the high-turnover investors in separating equilibrium constitute issuer A's revenues:

$$C_F = (f_0 + \Delta_{max})Xw_F \tag{24}$$

The costs incurred by the low-turnover investors in separating equilibrium constitute issuer B's revenues:

$$C_S = (f_0 + \Delta_{max} - \Delta_{min})X(1 - w_F) \tag{25}$$

From the societal welfare perspective, the cost involved in delivering ETF liquidity in a given index in separating equilibrium is simply the sum of costs incurred by both issuers:

$$C_{SEP} = 2f_{fix} \tag{26}$$

All non-zero profits generated in separating equilibrium constitute a welfare transfer from investors to issuers. The value of this transfer can be interpreted as the value of liquidity to ETF investors:

$$V_{SEP_{Transfer}} = (f_0 + \Delta_{max})Xw_F + (f_0 + \Delta_{max} - \Delta_{min})X(1 - w_F) - 2f_{fix} \tag{27}$$

Now, let us consider the counterfactual scenario with homogenous liquidity across ETFs tracking a given index. If there are no differences in liquidity, no liquidity clienteles emerge and there is room for only one ETF issuer in a given index. This issuer would earn the following competitive profits: [20]

$$\pi^* = Xf^* - f_{fix} = 0 \tag{28}$$

---

[20] $f^* = \frac{f_{fix}}{X}$ is the breakeven fee of the sole ETF issuer. See Appendix 5 for the full derivation of equilibrium fees in the absence of liquidity clienteles.

The costs incurred by investors in this scenario are:

$$C^* = Xf^* \tag{29}$$

The costs of delivering ETF liquidity in non-separating equilibrium without liquidity clienteles are:

$$C_{NON-SEP} = f_{fix} \tag{30}$$

Under homogenous liquidity conditions, no welfare transfers occur ($V_{NON-SEP_{Transfer}} = 0$), as MERs are set at a competitive level and the issuer generates zero profit. Note that the difference in costs of delivering ETF liquidity in separating vs non-separating equilibrium gives the value of deadweight loss, which arises due to the liquidity externality (i.e., "liquidity begetting liquidity" effect). The deadweight loss emerges due to two ETFs delivering the value of liquidity rather than one:

$$V_{DWL} = C_{SEP} - C_{NON-SEP} = f_{fix} \tag{31}$$

Hence, the welfare analysis suggests the following key insights about the effects of ETF liquidity in separating equilibrium:

1) In separating equilibrium, the value of ETF liquidity to investors can be calculated as the sum of economic profits earned by issuers.
2) From the societal welfare perspective, the deadweight loss due to liquidity externalities in separating equilibrium can be calculated as one times fixed costs of an ETF issuer.

4.2. Quantitative analysis

We estimate the welfare effects in the sample of 24 indices tracked by 60 ETFs. Our estimates suggest that all US-traded ETFs in separating equilibrium generate the total value of $780 million annually to ETF investors. These are the excess profits of high-MER ETF issuers, which the high-turnover investors pay to access the liquidity pool in those instruments. The magnitude of this liquidity-enabled welfare transfer is comparable with the average AUM-weighted MER differential, representing 0.096% of the $815 billion of assets under management invested in the 60 ETFs in our sample. As a point of reference, the equal-weighted average MER differential for same index ETFs in our sample is 7.7 bps (or 0.077%). Figure 4 illustrates the MER differentials and liquidity value estimates per index.

<Fig. 4 here>

We also estimate the cost to society of having multiple ETF issuers tracking the same index. The annual value of $70 million is lost due to the liquidity externality effect, whereas the highly liquid ETF charges MERs at monopolistic levels. Effectively, it's the value of inefficiency in resource allocation in separating equilibrium compared to the counterfactual scenario in which all liquidity in a given index is concentrated in a single ETF. This inefficiency arises due to the high-MER ETF taking advantage of MER-insensitive high-turnover investors and leaving enough market share for the low-MER ETF to break-even by serving the low-turnover clientele. In dollar value terms, this cost to society is over ten times smaller compared to the value of accessing liquidity by the high-turnover investors.

To arrive at these estimates, we make the following assumptions:

1) The low-MER ETF[21] earns zero profits.
2) Fixed costs are homogenous across all ETFs tracking a given index.

These assumptions imply that the value of ETF liquidity to investors can be proxied by excess profits earned by the high-MER ETF over the low-MER ETF. Because we assume the low-MER ETF just breaks even, its revenues are equal to fixed costs:

$$\pi_B = f_B X x_B - f_{fix} = 0 \rightarrow \ f_{fix} = f_B X x_B \tag{32}$$

Where f_B is the MER of the low-MER ETF, x_B is the market share of the low-MER ETF based on AUM, X is the combined AUM of all ETFs in a given index. Hence, the value of liquidity from the perspective of high-turnover investors is simply the excess profit earned by the high-MER ETF over and above the competitive level.

$$V_{LIQ} = V_{SEP_{Transfer}} = \pi_A = f_A X x_A - f_{fix} \tag{33}$$

Note that these assumptions are conservative, as they are likely to over-estimate issuers' fixed costs and underestimate their profits. This suggests that our proxies for the value of ETF liquidity are likely to be the lower-bound estimates.

5. Discussion and conclusions

Our paper makes several contributions. Firstly, it advances the market microstructure literature by proposing a novel approach to measuring how much investors value liquidity. As such, the ETF market offers a unique laboratory for inferring liquidity premia from MER differentials. Our study makes a clear case that liquidity premia can explain the supposed law of one price violations in financial markets. Secondly, we shed light on the microeconomics of competition in the ETF market and show that liquidity externalities can lead to oligopolistic price-setting behavior by first-movers. Thirdly, we argue that ETF liquidity can be a double-edged sword in affecting

[21] In case of three ETFs per index, we assume that only the highest-MER ETF earns non-zero profits. This assumption effectively makes our estimates of the value of ETF liquidity more conservative. Hence, it is fair to interpret these estimates as representing the lower bound of the value of ETF liquidity.

investor welfare, as liquidity-sensitive investors are likely to pool in suboptimal equilibrium with higher fees. Table 5 summarizes our empirical findings as they relate to the testable hypotheses.

<Table 5 here >

5.1. The value of liquidity

Similar to Amihud and Mendelson (1986) model for stocks, we model the liquidity clienteles arising in ETFs due to investors' heterogenous holding horizons. However, ETFs, unlike stocks, offer a unique laboratory for inferring the liquidity premia from the readily observed MERs, quoted by issuers in basis points. We find that investors require an average MER discount of 0.51 bps to buy an ETF with 1 bps higher spread, compared to a peer ETF tracking the same index. Investors also accept 1.15% higher MERs in exchange for 1% higher volume in an average ETF, relative to same index peers. This effect arises due to high-turnover investors pooling in highly liquid ETFs, while low-turnover investors in less liquid low-fee ETFs.

We estimate that ETF liquidity is worth at least $780 million annually to ETF investors. This is how much high-turnover investors pay to ETF issuers for accessing liquidity in major US indices with multiple ETFs tracking them. Hence, we contribute to the market microstructure literature by relating the welfare analysis to standard microstructure models.

4.2. Competition dynamics in the ETF market

We propose a coherent explanation for the apparent puzzle of persistent MER differences between same index ETFs. We also show why first mover ETFs tend to keep their MERs high even after competing ETF launches. The model implication is that liquidity externalities prevent investors from switching to low-MER competitors, thus allowing the incumbent ETF issuers to keep their MERs high. We find empirical support for the model predictions in the sample of same-index US-domiciled equity ETFs.

4.3. Welfare implications of liquidity

We estimate that the annual value of $70 million is lost due to the liquidity externality effect, whereas the highly liquid ETF charges MERs at monopolistic levels. Effectively, this is the double-expense incurred by index trackers when liquidity in a given index is delivered by multiple ETFs rather than one. It arises due to the classical "coordination problem" among ETF investors, whereas it's suboptimal for a single investor too switch to the low-liquidity ETF, although it would be beneficial to do so as a group.

Our model suggests that liquidity can be a "double-edged sword" in affecting investor welfare. On one hand, highly liquid markets are beneficial to allocating resources efficiently. On the other hand, in fragmented markets the incumbent can extract monopolistic rents from liquidity-sensitive traders, as liquidity externalities can lead to suboptimal pooling equilibria.

Appendix 1

This appendix derives the conditions for Nash equilibrium in which high-turnover investors choose ETF A and low-turnover investors ETF B. An individual high-turnover (low-turnover) investor will not have an incentive to deviate from the strategy of choosing ETF A (B), if his costs in case of holding A (B) are lower than in case of holding B (A). Hence the following inequalities should hold for high-turnover and low-turnover investors respectively:

$$\begin{cases} f_A h_F + \frac{ch_F}{w_F} < f_B h_F + \frac{ch_S}{w_S} \\ f_B h_S + \frac{ch_S}{w_S} < f_A h_S + \frac{ch_F}{w_F} \end{cases} \tag{A1.1}$$

$$\begin{cases} f_A - f_B < \frac{ch_S}{w_S h_F} - \frac{c}{w_F} \\ f_B - f_A < \frac{ch_F}{w_F h_S} - \frac{c}{w_S} \end{cases} \tag{A1.2}$$

$$\begin{cases} f_A - f_B < \frac{ch_S}{w_S h_F} - \frac{c}{w_F} \\ f_A - f_B > \frac{c}{w_S} - \frac{ch_F}{w_F h_S} \end{cases} \tag{A1.3}$$

Hence, we have the following condition on investor behavior, under which the separating equilibrium emerges:

$$\frac{c}{1-w_F} - \frac{ch_F}{w_F h_S} < f_A - f_B < \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F} \tag{A1.4}$$

Intuitively, this condition suggests that for the separating equilibrium to emerge, it should be that the fee differential is bounded from both sides. If the fee differential is too small, the low-turnover investors switch to the more expensive ETF (the MER differential is not that high to accept the lack of liquidity in ETF B). If the fee differential is too large, the high-turnover investors switch to the less liquid ETF (the MER differential is too high for them to justify the liquidity advantage).

Simplifying the outside part of the compound inequality (A1.4) gives us the condition on liquidities:

$$\frac{h_S}{(1-w_F)h_F} + \frac{h_F}{w_F h_S} - \frac{1}{w_F} - \frac{1}{(1-w_F)} > 0 \tag{A1.5}$$

$$\frac{w_F h_S h_S + (1-w_F)h_F h_F - (1-w_F)h_F h_S - w_F h_F h_S}{w_F(1-w_F)h_F h_S} > 0 \tag{A1.6}$$

$$w_F > (1-w_F)\frac{(h_F h_S - h_F h_F)}{(h_S h_S - h_F h_S)} \tag{A1.7}$$

$$w_F > w_S \frac{h_F}{h_S} \tag{A1.8}$$

Recall that the cost of liquidity in A is: $C_{LIQ}^A = \frac{ch_F}{w_F}$, and in B: $C_{LIQ}^B = \frac{ch_S}{w_S}$. Expressing w_F and w_S and plugging into the inequality above:

$$\frac{ch_F}{C_{LIQ}^A} > \frac{ch_S}{C_{LIQ}^B} * \frac{h_F}{h_S} \tag{A1.9}$$

Recognizing that $c > 0, h_F > 0, h_S > 0$ the inequality simplifies to:

$$C_{LIQ}^A < C_{LIQ}^B \tag{A1.10}$$

Expressing this inequality in terms of liquidity rather than the cost of liquidity (denoting ETF A's liquidity as $LIQ_A = -C_{LIQ}^A$ and ETF B's liquidity $LIQ_B = -C_{LIQ}^B$):

$$LIQ_A > LIQ_B \tag{A1.11}$$

This condition guarantees that the ETF A held by the high-turnover is more liquid than the ETF B held by the low-turnover. In other words, the cost of liquidity in A is lower than that in B.

Appendix 2

This appendix provides the proof that there are two possible Nash equilibria resulting from the issuers' competitive dynamics:

(i) An equilibrium with two ETFs: A charging $f_0 + \Delta_{max}$ and B charging $f_0 + \Delta_{max} - \Delta_{min}$.

(ii) An equilibrium with A charging $f_0 + \Delta_{min}$ and being the only ETF in the market,

where $f_0 = \frac{f_{fix}}{Xw_S}$ is the breakeven fee for B, Δ_{min} is the minimum acceptable fee differential between A and B, under which B can keep the low-turnover clientele, Δ_{max} is the maximum acceptable fee differential, under which A can keep the high-turnover clientele.

Recall that we model issuers in line with the Stackelberg (1934) duopoly model, where one issuer (A) is the market leader, while another (B) is the follower. Issuers set their MERs taking into account investors' response, as well as competitor's response. Based on the condition investors' derived in Appendix 1, the follower always sets his fee (f_B) low enough relative to the leader's fee (f^A) to satisfy the following condition: $\frac{c}{1-w_F} - \frac{ch_F}{w_F h_s} < f_A - f_B < \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F}$. Let's call this fee differential Δ. With this notation, we can rewrite the inequality as $\Delta_{min} < \Delta < \Delta_{max}$. In line with the inequality above, the minimum acceptable fee differential under which B can keep his clientele of low-turnover investors is as follows:

$$\Delta_{min} = \frac{c}{1-w_F} - \frac{ch_F}{w_F h_S} \qquad \text{(A2.1)}$$

The maximum acceptable fee differential under which A can keep his clientele of high-turnover investors is as follows:

$$\Delta_{max} = \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F} \qquad \text{(A2.2)}$$

Suppose B sets his fee at f_0 and A sets his fee at $f_0 + \Delta_{max}$. Let's check whether this is a stable Nash equilibrium from issuers' perspective. With these fees, A will attract the high-turnover clientele, and B the low-turnover clientele. Hence, A's profits will be:

$$\pi_A = Xw_F(f_0 + \Delta_{max}) - f_{fix} \qquad \text{(A2.3)}$$

And B's profits:

$$\pi_B = Xw_S f_0 - f_{fix} = 0 \qquad \text{(A2.4)}$$

Let's check if either issuer has an incentive to deviate, and if so, under which conditions:

1) Keeping B's strategy as given ($f_B = f_0$):

a. If A increases his fee above $f_A = f_0 + \Delta_{max}$, then he loses his high-turnover clientele to B, hence ending up with negative profits ($\pi_{A*} = -f_{fix}$). This is not an optimal strategy.

b. If A decreases his fee below $f_A = f_0 + \Delta_{max}$, he can gain the low-turnover clientele and have the market share of 100%, if he charges $f_A = f_0 + \Delta_{min}$. For this strategy to be inferior for A, compared to the current strategy, it must be that his additional profits from the low-turnover clientele are lower than the profits he gives up by lowering his fee. Formally, the following inequality should hold for this strategy to be inferior to the conjectured strategy: $f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max})$

2) Keeping A's strategy as given ($f_A = f_0 + \Delta_{max}$):

a. If B decreases his fee below $f_B = f_0$, he will be earning negative profits, which is not an optimal strategy.

b. If B increases his fee above $f_B = f_0$, he can actually earn higher profits without losing his clientele up until the fee $f_B = f_0 + \Delta_{max} - \Delta_{min}$. The profits under this strategy are higher ($\pi_B = Xw_S(f_0 + \Delta_{max} - \Delta_{min}) - f_{fix} > 0$), hence this strategy will be preferred by B.

3) If B charges ($f_B = f_0 + \Delta_{max} - \Delta_{min}$) and A charges ($f_A = f_0 + \Delta_{max}$), neither has an incentive to deviate, as they recognize the following:

a. A can lower his fee marginally and get both clienteles, to which B will respond by lowering his fee to maintain the Δ_{min} difference between his fee and A's, and keep the low-turnover clientele. This would result in the race to the bottom dynamics, whereas A can keep lowering his fee down to $f_A = f_0 + \Delta_{min}$, at which point he achieves a market share of 100% and competes B out of business. For A not to be interested in this scenario, A's profit from the high-turnover clientele at higher fee must be higher than the profit from both clienteles at a lower fee: $f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max})$.

b. A can increase his fee, charging up to ($f_A = f_0 + 2\Delta_{max} - \Delta_{min}$) without losing the high-turnover clientele, and earning a higher profit. However, at this fee level of A, B can reduce his fee by a marginally small amount and attract both the high-turnover and the low-turnover clienteles, which would prompt A to lower his fee, resulting in sequential lowering of fees by both issuers until they reach an equilibrium: either at $f_A = f_0 + \Delta_{max}$ and $f_B = f_0 + \Delta_{max} - \Delta_{min}$, or at $f_A = f_0 + \Delta_{min}$ and $f_B = f_0$ (with the second case essentially competing B out of business and A achieving 100% market share).

Therefore, there are two possible Nash equilibria from the issuers' perspective:

1) An equilibrium with two ETFs: A charging $f_0 + \Delta_{max}$ and B charging $f_0 + \Delta_{max} - \Delta_{min}$. This equilibrium is stable under the following condition:

$$f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max}) \tag{A2.5}$$

2) An equilibrium with A charging $f_0 + \Delta_{min}$ and being the only ETF in the market. This equilibrium is stable under the following condition:

$$f_0 + \Delta_{min} > w_F(f_0 + \Delta_{max}) \tag{A2.6}$$

Appendix 3

This appendix shows how the strength of forces driving the system into the separating equilibrium varies with the model parameters. For separating equilibrium to occur, both investors' and issuers' conditions should be satisfied. Writing down investors' conditions (inequality A1.11) and issuers' conditions (inequality A2.5) as a system of inequalities:

$$\begin{cases} LIQ_A > LIQ_B \\ f_0 + \Delta_{min} < w_F(f_0 + \Delta_{max}) \end{cases} \tag{A3.1}$$

Recognizing that $LIQ_A = -\frac{ch_F}{w_F}$, $LIQ_B = -\frac{ch_S}{(1-w_F)}$ and $h_{FS} = \frac{h_F}{h_S}$ and re-expressing investors' condition:

$$\frac{LIQ_A}{LIQ_B} > 1 \tag{A3.2}$$

$$\frac{w_F}{1-w_F} * \frac{1}{h_{FS}} > 1 \tag{A3.3}$$

Recognizing that $1 - w_F > 0$ and $h_{FS} > 0$, the inequality simplifies to:

$$w_F > (1 - w_F)h_{FS} \tag{A3.4}$$

$$w_F > \frac{h_{FS}}{1+h_{FS}} \tag{A3.5}$$

Recognizing (from Appendix 2) that $f^0 = \frac{f_{fix}}{X(1-w_F)}$, $\Delta_{min} = \frac{c}{1-w_F} - \frac{ch_{FS}}{w_F} = \frac{LIQ_A}{h_S} - \frac{LIQ_B}{h_S}$ and $\Delta_{max} = \frac{LIQ_A}{h_F} - \frac{LIQ_B}{h_F}$ and denoting $\Delta LIQ = LIQ_A - LIQ_B$, we can rewrite issuers' condition as follows:

$$\frac{f_{fix}}{X(1-w_F)}\ (1 - w_F) + \Delta LIQ\left(\frac{1}{h_S} - \frac{w_F}{h_F}\right) < 0 \tag{A3.6}$$

$$\frac{f_{fix}}{X}\ + \Delta LIQ\left(\frac{1}{h_S} - \frac{w_F}{h_F}\right) < 0 \tag{A3.7}$$

Recognizing that $\frac{f_{fix}}{X} > 0$ and $\Delta LIQ > 0$, condition A3.7 implies:

$$\frac{1}{h_S} - \frac{w_F}{h_F} < 0 \tag{A3.8}$$

$$\frac{1}{h_S} < \frac{w_F}{h_F} \tag{A3.9}$$

$$w_F > \frac{h_F}{h_s} \tag{A3.10}$$

$$w_F > h_{FS} \tag{A3.11}$$

Restriction A4.11 is a necessary condition for the separating equilibrium to occur. Note that it is more restrictive than investors' condition derived above (inequality A3.5). Hence, the system of inequalities becomes:

$$\begin{cases} \frac{f_{fix}}{X(1-w_F)} + \frac{c}{1-w_F} - \frac{ch_F}{w_F h_S} < w_F \left(\frac{f_{fix}}{X(1-w_F)} + \frac{ch_S}{(1-w_F)h_F} - \frac{c}{w_F} \right) \\ w_F > h_{FS} \end{cases} \tag{A3.12}$$

Recognizing that $X > 0$, $w_F > 0$, $(1 - w_F) > 0$, $h_{FS} > 0$, the system of inequalities simplifies to:

$$\begin{cases} w_F^2 \left(f_{fix} h_{FS} + cXh_{FS} + cX \right) - w_F h_{FS} \left(f_{fix} + 2cX + cXh_{FS} \right) + ch_{FS}^2 X > 0 \\ w_F > h_{FS} \end{cases} \tag{A3.13}$$

Let us denote the function on the left-hand side of the first inequality as follows:

$$y = w_F^2 \left(f_{fix} h_{FS} + cXh_{FS} + cX \right) - w_F h_{FS} \left(f_{fix} + 2cX + cXh_{FS} \right) + ch_{FS}^2 X \tag{A3.14}$$

On the interval that satisfies the second condition ($w_F > h_{FS}$), y is the propensity for separating equilibrium, expressed as a function of model parameters from Eq. (3.14).

1. Characterizing the propensity for separating equilibrium with respect to the proportion of high-turnover investors (w_F)

Since $f_{fix} h_{FS} + cXh_{FS} + cX > 0$, y is an upward-facing parabola w.r.t. w_F parameter. $D \geq 0$ suggests that there exist w_F roots for $y = 0$ case:

$$D = h_{FS}^2 \left(f_{fix} + 2cX + cXh_{FS} \right)^2 - 4ch_{FS}^2 X \left(f_{fix} h_{FS} + cXh_{FS} + cX \right) \tag{A3.15}$$

$$D = h_{FS}^2 \left(f_{fix} - cXh_{FS} \right)^2 + 4cXf_{fix} h_{FS}^2 \geq 0 \tag{A3.16}$$

The y-minimizing value of w_F is computed by taking a derivative and setting it to zero:

$$\frac{dy}{dw_F} = 2w_F \left(f_{fix} h_{FS} + cXh_{FS} + cX \right) - h_{FS} \left(f_{fix} + 2cX + cXh_{FS} \right) \tag{A3.17}$$

$$w_F^* = \frac{h_{FS} \left(cXh_{FS} + f_{fix} + 2cX \right)}{2 \left(cXh_{FS} + f_{fix} h_{FS} + cX \right)} \tag{A3.18}$$

$y = 0$ gives the following two solutions:

$$w_{F_1} = \frac{h_{FS} \left(f_{fix} + 2cX + cXh_{FS} \right) - h_{FS} \sqrt{\left(f_{fix} - cXh_{FS} \right)^2 + 4cXf_{fix}}}{2 \left(f_{fix} h_{FS} + cXh_{FS} + cX \right)} \tag{A3.19}$$

$$w_{F_2} = \frac{h_{FS} \left(f_{fix} + 2cX + cXh_{FS} \right) + h_{FS} \sqrt{\left(f_{fix} - cXh_{FS} \right)^2 + 4cXf_{fix}}}{2 \left(f_{fix} h_{FS} + cXh_{FS} + cX \right)} \tag{A3.20}$$

Solving for $y > 0$ gives the following two intervals:

$$w_F \in (0; w_{F1}) \cup (w_{F2}; 1) \tag{A3.21}$$

Corollary 1

The propensity for separating equilibrium increases as the proportion of high-turnover investors (w_F) in a particular index increases, given other parameter restrictions for separating equilibrium are satisfied.

Proof

Combining inequalities:

$$\begin{cases} w_F \in (0; w_{F1}) \cup (w_{F2}; 1) \\ w_F > h_{FS} \end{cases} \tag{A3.22}$$

Consider two mutually exclusive and collectively exhaustive intervals:

a) $h_{FS} \in (0; w_{F1})$
b) $h_{FS} \in (w_{F1}; 1)$

On the interval $h_{FS} \in (w_{F1}; 1)$, taking into account the restriction $w_F > h_{FS}$, the following should hold:

$$h_{FS} - w_{F1} > 0 \tag{A3.23}$$

$$h_{FS} - \frac{h_{FS}(f_{fix}+2cX+cXh_{FS}) - h_{FS}\sqrt{(f_{fix}-cXh_{FS})^2 + 4cXf_{fix}}}{2(f_{fix}h_{FS}+cXh_{FS}+cX)} > 0 \tag{A3.24}$$

As $h_{FS} > 0$, the inequality simplifies to:

$$1 - \frac{(f_{fix}+2cX+cXh_{FS}) - \sqrt{(f_{fix}-cXh_{FS})^2 + 4cXf_{fix}}}{2(f_{fix}h_{FS}+cXh_{FS}+cX)} > 0 \tag{A3.25}$$

As $2(f_{fix}h_{FS} + cXh_{FS} + cX) > 0$, the inequality simplifies to:

$$2(f_{fix}h_{FS} + cXh_{FS} + cX) - (f_{fix} + 2cX + cXh_{FS}) + \sqrt{(f_{fix} - cXh_{FS})^2 + 4cXf_{fix}} > 0 \tag{A3.26}$$

$$2f_{fix}h_{FS} + cXh_{FS} - f_{fix} + \sqrt{(f_{fix} - cXh_{FS})^2 + 4cXf_{fix}} > 0 \tag{A3.27}$$

Inequality A4.27holds for all model parameters.[22] Therefore, we have proven that $h_{FS} - w_{F1} > 0$ holds, and $h_{FS} - w_{F1} < 0$ is not possible (i.e., the interval $h_{FS} \in (0; w_{F1})$ is not possible). A3.22 system of inequalities becomes:

$$\begin{cases} w_F \in (0; w_{F1}) \cup (w_{F2}; 1) \\ w_F > h_{FS} \\ h_{FS} \in (w_{F1}; 1) \end{cases} \tag{A3.28}$$

Selecting the most restrictive conditions suggests:

$$\begin{cases} w_F \in (w_{F2}; 1) \\ w_F > h_{FS} \\ h_{FS} \in (w_{F1}; 1) \end{cases} \tag{A3.29}$$

[22] No analytical solution is presented, as the inequality was checked numerically in Mathematica.

Because y is an upward-facing parabola w.r.t. w_F parameter, and $w_{F2} > w_F^*$, y is increasing in w_F on the interval $w_F \in (w_{F2}; 1)$. Because this interval is the only one compatible with $w_F > h_{FS}$ restriction, the y function is always increasing with the proportion of high-turnover investors.

2. Characterizing the propensity for separating equilibrium with respect to homogeneity parameter (h_{FS})

Rewriting the y function in A3.14 w.r.t. h_{FS}:

$$y = h_{FS}^2 cX(1 - w_F) + h_{FS} w_F \left(w_F \left(f_{fix} + cX \right) - \left(f_{fix} + 2cX \right) \right) + w_F^2 cX \tag{A3.30}$$

Since $cX(1 - w_F) > 0$, y is an upward-facing parabola w.r.t. h_{FS}. The discriminant can take both positive and negative values, suggesting $y = 0$ might or might not have solutions w.r.t. h_{FS}:

$$D = w_F^2 \left(w_F^2 \left(f_{fix} + cX \right)^2 - w_F \left(2f_{fix} + 6f_{fix} cX \right) + f_{fix}^2 + 4f_{fix} cX \right) \tag{A3.31}$$

$$D = 4f_{fix} \left(f_{fix} + 6f_{fix} cX - f_{fix}^3 - 6f_{fix}^2 cX - 4c^3 X^3 \right) \tag{A3.32}$$

The y-minimizing value of h_{FS} is computed by taking a derivative and setting it to zero:

$$\frac{dy}{dh_{FS}} = 2h_{FS}\, cX(1 - w_F) + w_F \left(w_F \left(f_{fix} + cX \right) - \left(f_{fix} + 2cX \right) \right) = 0 \tag{A3.33}$$

$$h_{FS}^* = w_F * \frac{f_{fix}(1 - w_F) + cX(1 - w_F) + cX}{2cX(1 - w_F)} \tag{A3.34}$$

Corollary 2

The propensity for separating equilibrium decreases as the homogeneity of investors' holding horizons (h_{FS}) in a particular index increases, given other parameter restrictions for separating equilibrium are satisfied.

Proof

Consider two mutually exclusive and collectively exhaustive intervals:

a) $h_{FS} \in (0; h_{FS}^*)$

b) $h_{FS} \in (h_{FS}^*; 1)$

Given that $w_F > h_{FS}$ and $h_{FS} \in (0; h_{FS}^*)$, the following condition is sufficient to rule out interval b):

$$w_F < h_{FS}^* \tag{A3.35}$$

$$h_{FS}^* - w_F > 0 \tag{A3.36}$$

$$w_F \left(\frac{f_{fix}(1 - w_F) + cX(1 - w_F) + cX}{2cX(1 - w_F)} - 1 \right) > 0 \tag{A3.37}$$

Given that $2cX(1 - w_F) > 0$ for all parameters, inequality A3.37 simplifies to:

$$\left(f_{fix} + cX \right)(1 - w_F) + cX - 2cX(1 - w_F) > 0 \tag{A3.38}$$

$$w_F < \frac{f_{fix}}{f_{fix} - cX} \quad \text{(A3.39)}$$

Inequality A3.39 can be considered on two intervals:

a) $f_{fix} < cX \rightarrow$ inequality A3.38 doesn't have a solution, as this interval is not compatible with $w_F > 0$ condition.

b) $f_{fix} > cX \rightarrow$ inequality A4.38 holds for all parameter values, as $\frac{f_{fix}}{f_{fix} - cX} > 1$ and $w_F < 1$.

Hence, we have proven that $w_F < h_{FS}^*$ holds, and it is a sufficient condition for $h_{FS} \in (0; h_{FS}^*)$ to hold. This suggests that the y function is decreasing with the homogeneity of investors' holding horizons.

3. Characterizing the propensity for separating equilibrium with respect to the fixed costs parameter (f_{fix})

Rewriting the y function in A3.14 w.r.t. f_{fix}:

$$y = f_{fix} h_{FS} w_F^2 - w_F h_{FS} f_{fix} + w_F^2 cX h_{FS} + cX w_F^2 - 2 w_F h_{FS} cX - w_F h_{FS}^2 cX + cX h_{FS}^2 \quad \text{(A3.40)}$$

Taking a partial derivative of y (Eq. (A4.40)) w.r.t. f_{fix}:

$$\frac{dy}{df_{fix}} = w_F^2 h_{FS} - w_F h_{FS} = w_F h_{FS}(w_F - 1) \quad \text{(A3.41)}$$

Corollary 3

The propensity for separating equilibrium decreases as the fixed cost of running a fund (f_{fix}) in a particular index increase.

Proof

The expression in A3.40 is always non-positive, as $w^F \leq 1 \rightarrow \; w_F - 1 \leq 0$, and $w_F h_{FS} > 0$ hence $\frac{dy}{df^{fix}} \leq 0$, suggesting the separating equilibrium is less likely for higher values of issuers' fixed costs.

4. Characterizing the propensity for separating equilibrium with respect to combined assets under management of all ETFs tracking a particular index (X)

Rewriting the y function in A3.14 w.r.t. X:

$$y = f_{fix} h_{FS} w_F^2 - w_F h_{FS} f_{fix} + w_F^2 cX h_{FS} + cX w_F^2 - 2 w_F h_{FS} cX - w_F h_{FS}^2 cX + cX h_{FS}^2 \quad \text{(A3.42)}$$

Taking a partial derivative of y (Eq. (A3.42)) w.r.t. X:

$$\frac{dy}{dX} = c(w_F^2(1 + h_{FS}) - 2 w_F h_{FS} + h_{FS}^2) \quad \text{(A3.43)}$$

Corollary 4

The propensity for separating equilibrium increases as the combined AUM (X) in a particular index increase.

Proof

The expression in A3.42 is always positive, as it is an upward-facing parabola with discriminant that is less than zero: $c > 0, (1 + h_{FS}) > 0$, and $D = 4h_{FS}^2 - 4h_{FS}^2(1 + h^{FS}) < 0$, suggesting $\frac{dy}{dX} > 0$ for all parameter values.

Summary table

<table>
<tr><td>Restrictions</td><td colspan="3">$$\begin{cases} f_{fix}h_{FS}w_F^2 - w_F h_{FS} f_{fix} + w_F^2 cXh_{FS} + cXw_F^2 - 2w_F h_{FS} cX - w_F h_{FS}^2 cX + cXh_{FS}^2 > 0 \\ w_F > h_{FS} \\ w_F \in \left(\frac{h_{FS}\left(f_{fix} + 2cX + cXh_{FS}\right) + h_{FS}\sqrt{\left(f_{fix} - cXh_{FS}\right)^2 + 4cXf_{fix}}}{2\left(f_{fix}h_{FS} + cXh_{FS} + cX\right)}; 1\right) \\ h_{FS} \in \left(0;\ w_F * \frac{f_{fix}(1 - w_F) + cX(1 - w_F) + cX}{2cX(1 - w_F)}\right) \end{cases}$$</td></tr>
<tr><td rowspan="5">Parameter directions</td><td>Parameter</td><td>Derivative of the y function w.r.t. parameter[23]</td><td>Implication</td></tr>
<tr><td>w_F</td><td>$\frac{dy}{dw_F} > 0$</td><td>Separating equilibrium is more likely when the proportion of high-turnover investors in a given index is high.</td></tr>
<tr><td>h_{FS}</td><td>$\frac{dy}{dh_{FS}} < 0$</td><td>Separating equilibrium is more likely when the homogeneity of holding horizons of investors in a given index is low.</td></tr>
<tr><td>f_{fix}</td><td>$\frac{dy}{df_{fix}} < 0$</td><td>Separating equilibrium is more likely when ETF issuers’ fixed costs in a given index are low.</td></tr>
<tr><td>X</td><td>$\frac{dy}{dX} > 0$</td><td>Separating equilibrium is more likely when the combined AUM of all ETFs in a given index is high.</td></tr>
</table>

[23] Given the restrictions for all other parameters are satisfied.

Appendix 4

This appendix derives the relation between the ETF fees and liquidity in separating equilibrium. As shown in Appendix 2, in equilibrium, issuers A and B charge the following fees:

$$f_A = f_0 + \Delta_{max} \tag{A4.1}$$

$$f_B = f_0 + \Delta_{max} - \Delta_{min} \tag{A4.2}$$

Hence, the equilibrium fee differential is:

$$f_A - f_B = f_0 + \Delta_{max} - f_0 - \Delta_{max} + \Delta_{min} = \Delta_{min} \tag{A4.3}$$

As shown in Appendix 2, this fee differential can be expressed as:

$$f_A - f_B = \Delta_{min} = \frac{c}{w_S} - \frac{ch_F}{w_F h_S} \tag{A4.4}$$

Recall that investors' liquidity cost in ETFs A and B are as follows:

$$C_{LIQ}^{A} = \frac{ch_F}{w_F} \tag{A4.5}$$

$$C_{LIQ}^{B} = \frac{ch_S}{w_S} \tag{A4.6}$$

Expressing the liquidity cost differential as a function of the fee differential, we have:

$$C_{LIQ}^{A} - C_{LIQ}^{B} = c\left(\frac{h_F}{w_F} - \frac{h_S}{w_S}\right) = -h_S(f_A - f_B) \tag{A4.7}$$

Recognizing that the liquidity cost differential is just the liquidity differential with a minus sign (denoting the ETF A's liquidity as LIQ_A and ETF B's liquidity as LIQ_B), we obtain the following liquidity-fee relation for ETFs in separating equilibrium:

$$LIQ_A - LIQ_B = h_S(f_A - f_B) \tag{A4.8}$$

Appendix 5

This appendix presents the theoretical underpinnings behind the welfare analysis.

1. Welfare effects in non-separating equilibrium without liquidity clienteles

Recall from Appendix 2 that in equilibrium, the costs to high-turnover investors (of investing in ETF j)[24] are:

$$C_{TOTAL}^{F} = f_j h_F + \frac{ch_F}{w_F} \tag{A5.1}$$

And the costs to low-turnover investors are:

$$C_{TOTAL}^{S} = f_j h_S + \frac{ch_S}{w_S} \tag{A5.2}$$

Recall also that in this model, the clientele effect arises due to one ETF having the liquidity advantage over another, which allows the former to charge monopolistic rents in terms of higher MERs. To eliminate these monopolistic rents, the liquidity costs would have to be zero for both ETFs. Hence, $c = 0$ in the counterfactual scenario without the liquidity begetting liquidity effect. In that scenario, the costs to high-turnover and low-turnover investors respectively are:

$$C_{TOTAL}^{F} = f_j h_F \tag{A5.3}$$

$$C_{TOTAL}^{S} = f_j h_S \tag{A5.4}$$

Although the low-turnover investors are still more fee-sensitive than the high-turnover (as $h_S > h_F$), in the absence of liquidity advantage, both investor types prefer an ETF with a lower fee. Hence, investor i's cost minimization[25] imposes the following condition on fees:

$$f_i = \min\{f_A, f_B\} \tag{A5.5}$$

What would be the issuers' optimum fee setting behavior in this case? We can show that in the absence of liquidity externalities, the only stable Nash equilibrium is one in which there's only one ETF issuer setting the fee at breakeven level of $f^* = \frac{f_{fix}}{X}$.

Suppose there are two ETFs in the market: A and B. If A sets his fee at $f_A > f_B$, he will lose all investors to B. Hence, both ETFs have an incentive to keep lowering their fees until the point of zero profits. Recall that the profit function for ETF A is:

$$\pi_A = X x_A f_A - f_{fix} \tag{A5.6}$$

Hence, ETF A's breakeven fee is:

$$f_A = \frac{f_{fix}}{X x_A} \tag{A5.7}$$

[24] Note that j can take two values: A or B.
[25] Note that i can take two values: high-turnover or low-turnover.

The lowest possible breakeven fee is achieved when $x_A = 1$ (A has 100% market share). In that scenario, A competes B out of the market. Decreasing the fee below $f^* = \frac{f_{fix}}{X}$ would result in negative profits for either issuer, hence is not optimal. Increasing the fee above $f^* = \frac{f_{fix}}{X}$ would result in losing 100% market share to the competitor, which is not optimal. Therefore, the only stable equilibrium emerges when there is only one ETF in the market, and he charges the breakeven fee of f^*.[26] The profit of this single ETF issuer is:

$$\pi^* = Xf^* - f_{fix} = 0 \tag{A5.8}$$

In the non-separating equilibrium without liquidity clienteles, total fees paid by investors are as follows:

$$C^* = \frac{f_{fix}}{X} * X = f_{fix} \tag{A5.9}$$

Hence, the following welfare effects arise in non-separating equilibrium without liquidity clienteles:

1) No welfare transfers occur, as MERs are set at a competitive level and the issuer generates zero profit:

$$V_{NON-SEP_{Transfer}} = 0 \tag{A5.10}$$

2) The cost of providing ETF liquidity in a given index is equal to the cost incurred by a single ETF issuer:

$$C_{NON-SEP} = f_{fix} \tag{A5.11}$$

2. Welfare effects in separating equilibrium with liquidity clienteles

In the separating equilibrium, issuers' profits are as follows: [27]

$$\pi_A = (f_0 + \Delta_{max} - \Delta_{min})Xw_F - f_{fix} \tag{A5.12}$$

$$\pi_B = (f_0 + \Delta_{min})X(1 - w_F) - f_{fix} \tag{A5.13}$$

The costs to high-turnover and low-turnover investors respectively are as follows:

$$C_F = (f_0 + \Delta_{max})Xw_F \tag{A5.14}$$

$$C_S = (f_0 + \Delta_{max} - \Delta_{min})X(1 - w_F) \tag{A5.15}$$

Hence, the following welfare effects arise in the separating equilibrium with liquidity clienteles:

[26] One can argue that there is a possible equilibrium with two ETFs charging identical fees, earning zero profits, and having equal market shares: $x_A = x_B = 0.5$. However, this equilibrium is not stable in the sense that each issuer has an incentive to undercut on fees. Because even marginally lower fee results in capturing 100% market share, it would result in the race to the bottom dynamics, until the breakeven fee of f^* is reached.

[27] See Appendix 2 for the Nash equilibrium derivation.

1) The welfare transfer from investors to issuers that is also the payment for accessing ETF liquidity:

$$V_{SEP_{Transfer}} = (f_0 + \Delta_{max})Xw_F + (f_0 + \Delta_{max} - \Delta_{min})X(1 - w_F) - 2f_{fix} \qquad \text{(A5.16)}$$

2) The cost of providing ETF liquidity in a given index is equal to the cost incurred by two ETF issuers:

$$C_{SEP} = 2f_{fix} \qquad \text{(A5.17)}$$

3. Welfare effects in separating equilibrium with liquidity clienteles compared to non-separating equilibrium without liquidity clienteles

The value of welfare transfers from investors to issuers does not affect the overall welfare from the perspective of society as a whole. The remaining stream of costs and benefits is simply the cost incurred by issuers. To arrive at the net effect on societal welfare, we compare the costs of delivering ETF liquidity in separating equilibrium with liquidity clienteles to that in non-separating equilibrium without liquidity clienteles. The deadweight loss emerges due to two ETFs delivering the value of liquidity rather than one:

$$V_{DWL} = C_{SEP} - C_{NON-SEP} = f_{fix} \qquad \text{(A5.18)}$$

References

Agarwal, V., Hanouna, P., Moussawi, R., and Stahel, C. W. (2018). Do ETFs increase the commonality in liquidity of underlying stocks? *Social Science Research Network Working Paper Series.*

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of Financial Economics*, 17(2), 223–249.

Amihud, Y., and Mendelson, H. (1991). Liquidity, maturity, and the yields on US Treasury securities. *The Journal of Finance,* 46(4), 1411-1425.

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

Balchunas, E. (2016). The Institutional ETF Toolbox: How Institutions Can Understand and Utilize the Fast-Growing World of ETFs. John Wiley & Sons.

Baldauf, M., and Mollner, J. (2016). Fast traders make a quick buck: The role of speed in liquidity provision. *Social Science Research Network Working Paper Series.*

Ben-David, I., Franzoni, F., and Moussawi, R. (2014). *Do ETFs increase volatility?* National Bureau of Economic Research.

Berk, J. B., and Green, R. C. (2004). Mutual fund flows and performance in rational markets. *Journal of Political Economy*, 112(6), 1269-1295.

Bhattacharya, A., and O'Hara, M. (2017). Can ETFs increase market fragility? Effect of information linkages in ETF markets. *Social Science Research Network Working Paper Series.*

Box, T., Davis, R., and Fuller, K. (2017). ETF Competition and Market Quality. *Social Science Research Network Working Paper Series.*

Brennan, M. J., and Subrahmanyam, A. (1996). Market microstructure and asset pricing: On the compensation for illiquidity in stock returns. *Journal of financial economics*, 41(3), 441-464.

Chinco, A., and Fos, V. (2017). The sound of many funds rebalancing. *Social Science Research Network Working Paper Series.*

Chordia, T., Roll, R., and Subrahmanyam, A. (2000). Commonality in liquidity. *Journal of Financial Economics*, 56(1), 3-28.

Colliard, J. E., and Foucault, T. (2012). Trading fees and efficiency in limit order markets. *The Review of Financial Studies*, 25(11), 3389-3421.

Da, Z., and Shive, S. (2018). Exchange traded funds and asset return correlations. *European Financial Management*, 24(1), 136-168.

Dannhauser, C. (2016). The impact of innovation: Evidence from corporate bond ETFs. *Journal of*

Financial Economics, Forthcoming.

De Jong, F., and Rindi, B. (2009). The microstructure of financial markets. Cambridge University Press.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, 73(6), 1815-1847.

Easley, D., Michayluk, D., O'Hara, M., and Putniņš, T. J. (2018). The Active World of Passive Investing. *Social Science Research Network Working Paper Series.*

Economides, N. (1996). The economics of networks. *International Journal of Industrial Organization*, 14(6), 673-699.

Foucault, T., Kadan, O., and Kandel, E. (2005). Limit order book as a market for liquidity. *The Review of Financial Studies*, 18(4), 1171-1217.

Garman, M. B. (1976). Market microstructure. *Journal of Financial Economics*, 3(3), 257-275.

Glosten, L., Nallareddy, S., and Zou, Y. (2016). ETF trading and informational efficiency of underlying securities. *Social Science Research Network Working Paper Series.*

Gomber, P., Sagade, S., Theissen, E., Weber, M. C., and Westheide, C. (2017). Competition between equity markets: A review of the consolidation versus fragmentation debate. *Journal of Economic Surveys*, 31(3), 792-814.

Hasbrouck, J., and Seppi, D. J. (2001). Common factors in prices, order flows, and liquidity. *Journal of Financial Economics*, 59(3), 383-411.

Ho, T., and Stoll, H. R. (1980). On dealer markets under competition. *The Journal of Finance*, 35(2), 259-267.

Ho, T., and Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial Economics*, 9(1), 47-73.

Ho, T. S., and Stoll, H. R. (1983). The dynamics of dealer markets under competition. *The Journal of Finance*, 38(4), 1053-1074.

Holden, C. W., Jacobsen, S., and Subrahmanyam, A. (2014). The empirical analysis of liquidity. *Foundations and Trends in Finance*, 8(4), 263-365.

Huang, S., O'Hara, M., and Zhong, Z. (2018). Innovation and informed trading: Evidence from industry ETFs. *Social Science Research Network Working Paper Series.*

Huberman, G., and Halka, D. (2001). Systematic liquidity. *Journal of Financial Research*, 24(2), 161-178.

Israeli, D., Lee, C. M., and Sridharan, S. A. (2017). Is there a dark side to exchange traded funds? An information perspective. *Review of Accounting Studies*, 22(3), 1048–1083.

Katz, M.L., and Shapiro, C. (1985). Network externalities, competition, and compatibility. *The*

American economic review, 75(3), 424–440.

Krause, T., Ehsani, S., and Lien, D. (2014). Exchange-traded funds, liquidity and volatility. *Applied Financial Economics*, 24(24), 1617–1630.

Krishnamurthy, A. (2002). The bond/old-bond spread. *Journal of Financial Economics*, 66(2-3), 463–506.

Lettau, M., and Madhavan, A. (2018). Exchange-Traded Funds 101 for Economists. *Journal of Economic Perspectives*, 32(1), pp.135–54.

Li, F. W., and Zhu, Q. (2016). Synthetic shorting with ETFs. *Social Science Research Network Working Paper Series.*

Longstaff, F. A., Mithal, S., and Neis, E. (2005). Corporate yield spreads: Default risk or liquidity? New evidence from the credit default swap market. *The Journal of Finance*, 60(5), 2213-2253.

Madhavan, A., and Sobczyk, A. (2015). Price dynamics and liquidity of exchange-traded funds. *Social Science Research Network Working Paper Series*.

Madhavan, A. N. (2016). Exchange-Traded Funds and the New Dynamics of Investing. Oxford University Press.

Malamud, S. (2016). A dynamic equilibrium model of ETFs. *Social Science Research Network Working Paper Series.*

Marshall, B. R., Nguyen, N. H., and Visaltanachoti, N. (2013). ETF arbitrage: Intraday evidence. *Journal of Banking and Finance*, 37(9), 3486–3498.

Mendelson, H. (1987). Consolidation, fragmentation, and market performance. *Journal of Financial and Quantitative Analysis*, 22(2), 189-207.

Nash, J. (1951). Non-cooperative games. *Annals of Mathematics*, 286-295.

Pagano, M. (1989). Trading volume and asset liquidity. *The Quarterly Journal of Economics*, 104(2), 255–274.

Pagnotta, E. S., and Philippon, T., (2011). The Welfare Effects of Financial Innovation: High-Frequency Trading in Equity Markets. In *2011 Meeting Papers* (No. 1246). Society for Economic Dynamics.

Roşu, I. (2009). A dynamic model of the limit order book. *The Review of Financial Studies*, 22(11), 4601-4641.

Shleifer, A., and Vishny, R. W. (1997). The limits of arbitrage. *The Journal of Finance*, 52(1), 35-55.

Subrahmanyam, M. G., Jankowitsch, R., and Friewald, N. (2012). Illiquidity or credit deterioration: A study of liquidity in the US corporate bond market.

Von Stackelberg, H. (1934). Marktform und gleichgewicht. J. Springer.

Wermers, R., and Xue, J. (2015). Intraday ETF trading and the volatility of the underlying. *Social Science Research Network Working Paper Series*.

Wigglesworth, R. (2018). ETF growth is "in danger of devouring capitalism". *The Financial Times,* February 5 Special Report.

Yao, C., and Ye, M. (2018). Why trading speed matters: A tale of queue rationing under price controls. *The Review of Financial Studies*, 31(6), 2157-2183.

Xu, L., Yin, X., and Zhao, J. (2016). Differently Motivated ETF Trading Activities and the Volatility of the Underlying Index. *Social Science Research Network Working Paper Series.*

Table 1
Equity ETFs with identical index exposure

This table provides the list of ETFs that share the same index benchmark with at least one other ETF. The sample contains 60 same-index US-domiciled ETFs based on 24 unique indices. The ETF characteristics are daily averages for the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is the annualized percentage ratio of daily dollar volume divided by assets under management (AUM).

Index benchmark	ETF issuer	Ticker	Inception date	MER, bps	Relative spread, bps	Dollar volume, $ mln	AUM, $ mln	Turnover, %
FTSE Emerging NR USD	Charles Schwab	SCHE	2010/01/14	13.00	4.11	21.91	3404.86	161.72
FTSE Emerging NR USD	Vanguard	VWO	2005/03/04	14.22	2.40	456.54	55862.66	205.71
JPX-Nikkei 400 Net Total Return Index	Deutsche Bank	JPN	2015/06/24	35.67	17.83	0.09	12.83	184.94
JPX-Nikkei 400 Net Total Return Index	Blackrock	JPXN	2001/10/23	48.00	12.55	0.25	88.26	70.19
MSCI ACWI Ex USA NR USD	Blackrock	ACWX	2008/03/26	32.09	2.67	24.82	2393.42	261.13
MSCI ACWI Ex USA NR USD	State Street	CWI	2007/01/10	30.00	4.29	7.77	1295.34	150.90
MSCI ACWI Low Carbon Target Index	Blackrock	CRBN	2014/12/08	20.00	11.54	0.94	403.20	58.68
MSCI ACWI Low Carbon Target Index	State Street	LOWC	2014/11/25	20.00	17.98	0.11	140.78	19.61
MSCI EAFE 100% Hedged NR USD	Deutsche Bank	DBEF	2011/06/09	35.00	3.39	48.11	7919.25	153.04
MSCI EAFE 100% Hedged NR USD	Blackrock	HEFA	2014/01/31	35.94	4.27	33.49	4075.72	218.31
MSCI Japan 100% Hedged NR USD	Deutsche Bank	DBJP	2011/06/09	45.00	3.40	19.63	1896.35	260.98
MSCI Japan 100% Hedged NR USD	Blackrock	HEWJ	2014/01/31	49.00	3.44	32.11	1098.26	734.91
NASDAQ 100 Equal Weighted TR USD	First Trust	QQEW	2006/04/19	60.00	4.82	2.78	506.10	138.18
NASDAQ 100 Equal Weighted TR USD	Direxion	QQQE	2012/03/21	35.00	15.57	1.36	124.43	275.71
Russell 1000 Growth TR USD	Blackrock	IWF	2000/05/22	20.00	0.93	181.09	35900.41	127.05
Russell 1000 Growth TR USD	Vanguard	VONG	2010/09/20	12.00	3.72	5.26	1162.83	113.69
Russell 1000 TR USD	Blackrock	IWB	2000/05/15	15.00	1.00	118.77	18490.70	161.76
Russell 1000 TR USD	State Street	ONEK	2005/11/08	10.00	18.39	0.56	137.28	102.48
Russell 1000 TR USD	Vanguard	VONE	2010/09/20	12.00	4.60	2.51	717.25	88.37
Russell 1000 Value TR USD	Blackrock	IWD	2000/05/22	20.00	0.92	230.89	36874.94	157.76
Russell 1000 Value TR USD	Vanguard	VONV	2010/09/20	12.00	3.60	4.35	1067.79	102.48
Russell 2000 Growth TR USD	Blackrock	IWO	2000/07/24	24.58	1.81	95.36	8078.46	297.28
Russell 2000 Growth TR USD	Vanguard	VTWG	2010/09/20	20.00	6.18	0.92	193.21	120.04
Russell 2000 TR USD	Blackrock	IWM	2000/05/22	20.00	0.71	3651.68	38912.53	2365.30
Russell 2000 TR USD	State Street	TWOK	2013/07/08	10.00	13.04	1.51	204.92	185.24
Russell 2000 TR USD	Vanguard	VTWO	2010/09/20	15.00	5.17	8.53	1050.01	204.53
Russell 2000 Value TR USD	Blackrock	IWN	2000/07/24	24.58	1.41	111.87	8766.73	321.59

Russell 2000 Value TR USD	Vanguard	VTWV	2010/09/20	20.00	7.54	1.06	173.43	154.20
Russell 3000 TR USD	Blackrock	IWV	2000/05/22	20.00	1.15	28.34	7686.42	92.86
Russell 3000 TR USD	State Street	THRK	2000/10/04	10.00	15.21	1.49	416.72	89.92
Russell 3000 TR USD	Vanguard	VTHR	2010/09/20	15.00	4.82	1.29	343.30	94.48
S&P 500 Growth TR	Blackrock	IVW	2000/05/22	18.00	1.14	91.79	17775.55	130.00
S&P 500 Growth TR	State Street	SPYG	2000/09/25	12.72	5.86	3.91	769.41	128.47
S&P 500 Growth TR	Vanguard	VOOG	2010/09/07	15.00	3.98	6.02	1544.69	98.02
S&P 500 TR USD	Blackrock	IVV	2000/05/15	4.00	0.51	834.05	112581.34	186.46
S&P 500 TR USD	State Street	SPY	1993/01/22	9.40	0.40	17012.97	240662.22	1780.40
S&P 500 TR USD	Vanguard	VOO	2010/09/07	4.32	0.61	434.58	68728.50	159.18
S&P 500 Value TR USD	Blackrock	IVE	2000/05/22	18.00	1.16	83.42	13687.09	153.51
S&P 500 Value TR USD	State Street	SPYV	2000/09/25	12.72	8.60	2.10	330.83	160.68
S&P 500 Value TR USD	Vanguard	VOOV	2010/09/07	15.00	5.75	2.70	726.10	93.46
S&P Global Infrastructure TR USD	State Street	GII	2007/01/25	40.00	38.18	0.85	142.82	149.67
S&P Global Infrastructure TR USD	Blackrock	IGF	2007/12/10	47.42	5.94	9.75	1617.15	151.72
S&P MidCap 400 Growth TR	Blackrock	IJK	2000/07/24	25.00	2.25	23.19	6607.17	88.41
S&P MidCap 400 Growth TR	Vanguard	IVOG	2010/09/07	20.00	3.47	2.59	685.03	95.17
S&P MidCap 400 Growth TR	State Street	MDYG	2005/11/08	15.00	8.06	3.70	665.06	139.74
S&P MidCap 400 TR	Blackrock	IJH	2000/05/22	7.00	0.80	214.32	39471.90	136.73
S&P MidCap 400 TR	Vanguard	IVOO	2010/09/07	15.00	2.79	3.49	721.24	121.88
S&P MidCap 400 TR	State Street	MDY	1995/05/04	25.00	0.69	359.18	19412.60	466.23
S&P MidCap 400 Value TR USD	Blackrock	IJJ	2000/07/24	25.00	2.50	21.94	5821.33	94.95
S&P MidCap 400 Value TR USD	Vanguard	IVOV	2010/09/07	20.00	6.23	2.15	651.70	82.98
S&P MidCap 400 Value TR USD	State Street	MDYV	2005/11/08	15.00	8.44	3.26	457.36	179.29
S&P SmallCap 600 Growth TR	Blackrock	IJT	2000/07/24	25.00	4.08	19.30	4391.47	110.71
S&P SmallCap 600 Growth TR	State Street	SLYG	2000/09/25	15.00	8.12	5.84	1209.82	121.60
S&P SmallCap 600 Growth TR	Vanguard	VIOG	2010/09/07	20.00	5.01	1.25	249.79	125.54
S&P SmallCap 600 TR USD	Blackrock	IJR	2000/05/22	7.00	1.64	198.20	29655.62	168.49
S&P SmallCap 600 TR USD	State Street	SLY	2005/11/08	15.00	10.44	3.31	765.07	109.02
S&P SmallCap 600 TR USD	Vanguard	VIOO	2010/09/07	15.00	3.49	4.55	635.85	179.86
S&P SmallCap 600 Value TR	Blackrock	IJS	2000/07/24	25.00	4.01	21.62	4779.00	114.00
S&P SmallCap 600 Value TR	State Street	SLYV	2000/09/25	15.00	8.33	6.12	977.80	157.62
S&P SmallCap 600 Value TR	Vanguard	VIOV	2010/09/07	20.00	4.86	1.26	215.83	146.98

Table 2
Descriptive statistics

This table reports descriptive statistics for the variables used in regression analysis. Panel A contains 60 same-index US-domiciled ETFs, based on 24 unique indices with combined assets under management of $823.58 billion and combined daily dollar volume of $24.79 billion. Panel B contains 975 indices, each tracking one ETF, with combined assets under management of $1,445.42 billion and combined daily dollar volume of $27.98 billion. All variables are calculated from the daily frequency data and averaged per index over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM).

	Mean	Standard Deviation	25th percentile	50th percentile	75th percentile
Panel A. Indices with multiple ETFs per index (in separating equilibrium)					
MER, bps	22.67	11.80	15.11	20.00	26.66
Relative Spread, bps	6.19	4.92	3.45	4.88	6.68
Turnover, %	331.21	283.01	191.47	244.11	328.47
Number of Constituents	854.52	716.24	338.46	590.67	1176.98
AUM, $ bn	34.32	85.85	4.79	8.70	34.34
Daily Dollar Volume, $ mln	1033.00	3801.70	318.00	895.00	200.10
Panel B. Indices with one ETF per index (in non-separating equilibrium)					
MER, bps	50.13	40.25	35.00	48.00	62.20
Relative Spread, bps	29.02	66.83	6.13	14.37	31.81
Turnover, %	534.35	1458.76	171.21	289.54	495.36
Number of Constituents	276.50	532.47	44.16	100.49	276.48
AUM, $ bn	1.48	5.44	0.02	0.12	0.69
Daily Dollar Volume, $ mln	28.80	177.90	0.20	0.90	4.60

Table 3
OLS regressions

This table reports cross-sectional regression results for six different models with index fixed effects, where MER is the dependent variable, and independent variables are reported in the first column. Panel A reports the results for simple OLS regressions. Panel B reports the results for AUM-weighted least squares regressions. The sample contains 60 same index US-domiciled ETFs based on 24 unique indices. All variables are calculated from the daily frequency data and averaged per ETF over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM), tracking error is the standard deviation of the difference in daily returns between an ETF and its benchmark index. T-statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	MER (1)	*MER (2)*	*Log MER (3)*	*Log MER (4)*	*Log MER (5)*	*Log MER (6)*
Panel A. Ordinary least squares regressions						
Intercept	23.12***	22.97***	2.06***	1.83***	1.29	2.28
	(8.37)	(8.62)	(4.98)	(4.82)	(0.19)	(0.36)
Relative Spread	-0.55***	-0.51***				
	(-4.01)	(-3.83)				
Log Turnover			0.19**	0.24***		
			(2.42)	(3.21)		
Log Dollar Volume					1.21***	1.15***
					(3.12)	(3.11)
Tracking Error		-0.08*				
		(-1.92)				
Log Tracking Error				-0.12***		-0.09**
				(-3.00)		(-2.13)
Adjusted R^2	0.85	0.86	0.73	0.78	0.83	0.84
Fixed Effects	Index	Index	Index	Index	Index	Index
Panel B. AUM-weighted least squares regressions						
Intercept	26.38***	26.44***	1.46***	1.44***	-2.07	-1.76
	(4.37)	(4.33)	(5.58)	(5.54)	(-0.31)	(-0.26)
Relative Spread	-0.69	-0.66				
	(-1.11)	(-1.03)				
Log Turnover			0.37***	0.37***		
			(14.66)	(14.70)		
Log Dollar Volume					1.57***	1.56***
					(5.01)	(4.95)
Tracking Error		-0.25				
		(-0.49)				
Log Tracking Error				-0.15		-0.26
				(-1.29)		(-0.68)
Adjusted R^2	0.74	0.73	0.93	0.93	0.84	0.84
Fixed Effects	Index	Index	Index	Index	Index	Index

Table 4
Probit regressions

This table reports results for probit regressions modelling the probability of an index being in separating equilibrium (i.e., having multiple ETFs tracking it). Independent variables are reported in the first column. Panel A sample contains 24 indices with multiple ETFs per index and 975 indices with one ETF per index, Panel B sample contains 24 indices with multiple ETFs per index and 150 randomly selected indices with one ETF per index. All variables are calculated from the daily frequency data and averaged per index over the year 2017. Relative spread (in basis points) is absolute bid-ask spread divided by midpoint, major index is one issued by MSCI, S&P or Russel, top 3 ETF issuers are Vanguard, BlackRock and State Street. Dollar volume and AUM are in $ billion, number of constituents is in hundreds. Chi-squared statistics are reported in parentheses. ***, **, and * indicate statistical significance at 1%, 5%, and 10% levels, respectively.

	Prob (1)	*Prob (2)*	*Prob (3)*	*Prob (4)*
Panel A. All equity ETFs				
Intercept	-2.87***	-2.87***	-2.37***	-2.32
	(127.11)	(126.87)	(68.87)	(73.69)
Dollar Volume		-0.07	0.44**	0.42**
		(0.06)	(4.21)	(3.83)
Relative Spread			-0.03**	-0.03**
			(5.32)	(4.10)
Major Index Dummy	0.56**	0.56**	0.56**	
	(3.60)	(3.62)	(4.32)	
AUM	0.04***	0.04***		
	(23.76)	(17.22)		
Top3 Issuer Dummy	0.51*	0.51***	0.56**	0.82***
	(3.12)	(3.07)	(4.09)	(10.24)
Number of Constituents				0.02**
				(4.94)
Panel B. Randomly selected ETFs in non-separating equilibrium				
Intercept	-2.18***	-2.18***	-1.61***	-1.48**
	(46.38)	(46.19)	(15.83)	(14.46)
Dollar Volume		0.72	2.26*	2.31**
		(0.28)	(3.17)	(3.46)
Relative Spread				-0.03
				(2.40)
Major Index Dummy	0.55	0.54	0.54	
	(2.11)	(2.04)	(2.19)	
AUM	0.04***	0.03**		
	(9.43)	(4.42)		
Top3 Issuer Dummy	0.75**	0.74**	0.59	0.87***
	(3.87)	(3.67)	(2.44)	(6.03)
Number of Constituents				0.01
				(0.17)

Table 5
Summary of empirical results

Hypotheses	Empirical support
Hypothesis 1: *The separating equilibrium is more likely when the proportion of high-turnover ETF investors in a given index is higher.*	Yes
Hypothesis 2: *The separating equilibrium is more likely when ETF investors' holding horizons in a given index are less homogenous.*	Yes
Hypothesis 3: *The separating equilibrium is more likely when the ETF issuers' fixed costs in a given index are lower.*	Yes
Hypothesis 4: *The separating equilibrium is more likely when the combined AUM of all ETFs in a given index is higher.*	Yes
Hypothesis 5: *For ETFs tracking the same index, liquidity is positively related to fees.*	Yes
Hypothesis 6: *For ETFs tracking the same index, turnover is positively related to fees.*	Yes

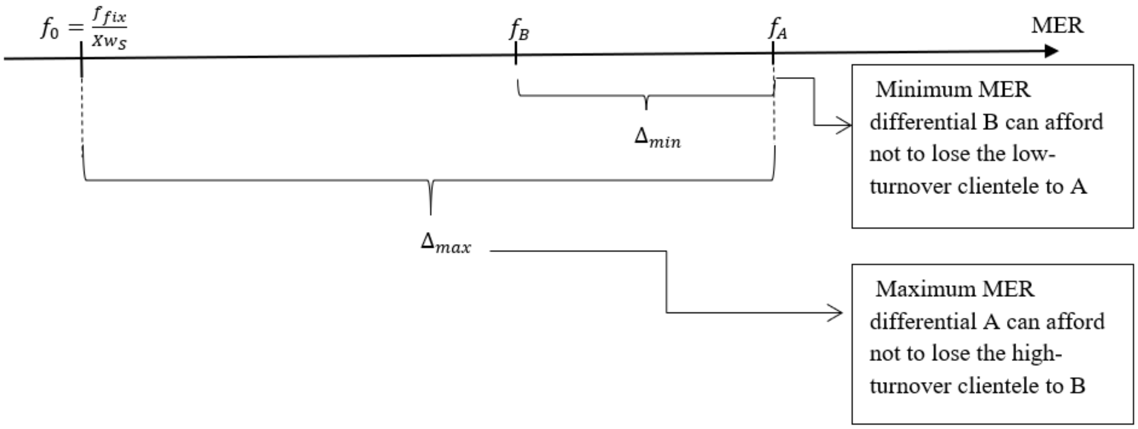
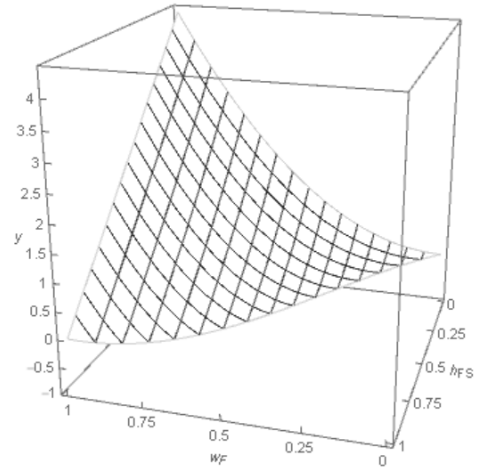
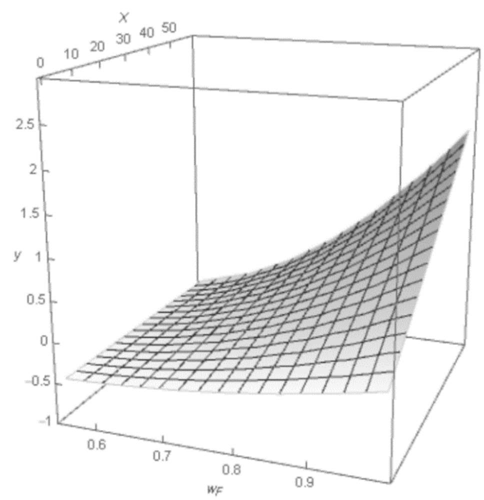


Fig. 1. Fee setting behavior in Nash equilibrium.
This figure summarizes the fee setting behavior in Nash equilibrium. f_A is issuer A's optimum fee, f_B is issuer B's optimum fee, f_0 is issuer B's breakeven fee, f_{fix} is the fixed cost faced by ETF issuers, X is the total AUM of all ETFs on a given index, w_S is the proportion of low-turnover investors in the ETF market. In separating Nash equilibrium (i.e., when two ETFs track the same index), neither issuer has an incentive to deviate from the chosen strategy, if issuer A (the high-MER issuer of the more liquid ETF A) charges the fee $f_A = f_0 + \Delta_{max}$, and B charges the fee $f_B = f_0 + \Delta_{max} - \Delta_{min}$. If A charges MER higher than $f_A = f_0 + \Delta_{max}$, high-turnover investors incur too high MER fees and have an incentive to switch to ETF B, despite the higher liquidity costs in B. If B charges MER higher than $f_B = f_0 + \Delta_{max} - \Delta_{min}$, low-turnover investors do not save enough on MER fees to justify their higher liquidity costs in B, and hence have an incentive to switch to ETF A. The full derivation of the Nash equilibrium in this model is presented in Appendix 2.

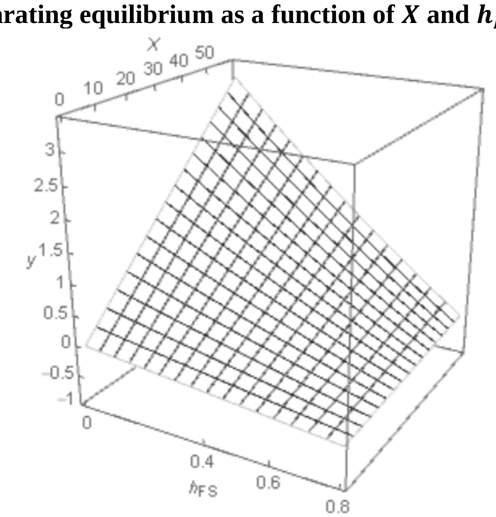
Panel A: Propensity for separating equilibrium as a function of w_F and h_{FS} parameters



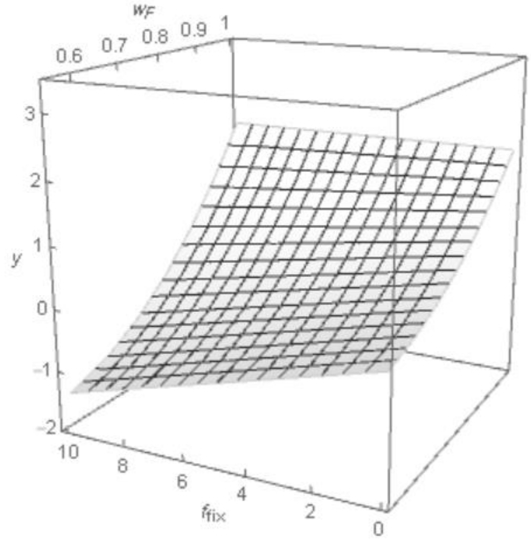
Panel B: Propensity for separating equilibrium as a function of X and w_F parameters



Panel C: Propensity for separating equilibrium as a function of X and h_{FS} parameters



Panel D: Propensity for separating equilibrium as a function of f_{fix} and w_F parameters



Panel E: Propensity for separating equilibrium as a function of f_{fix} and h_{FS} parameters

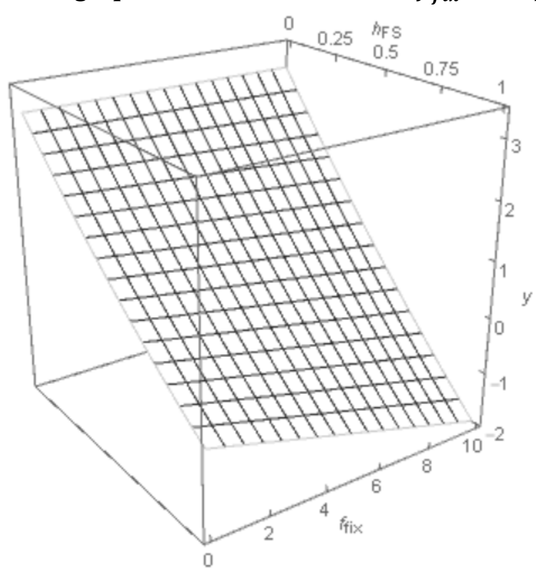
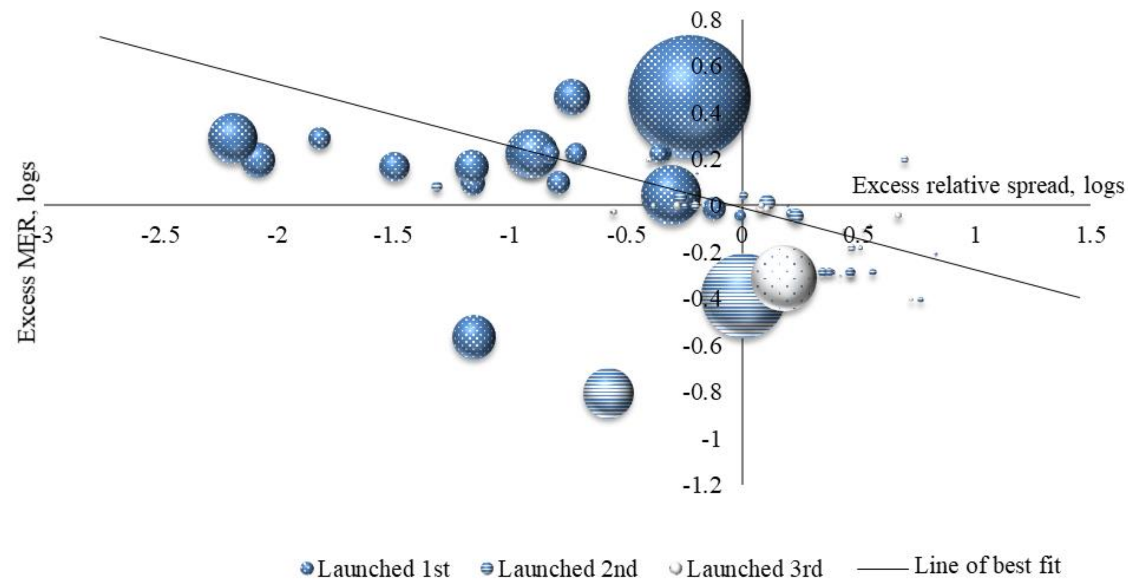


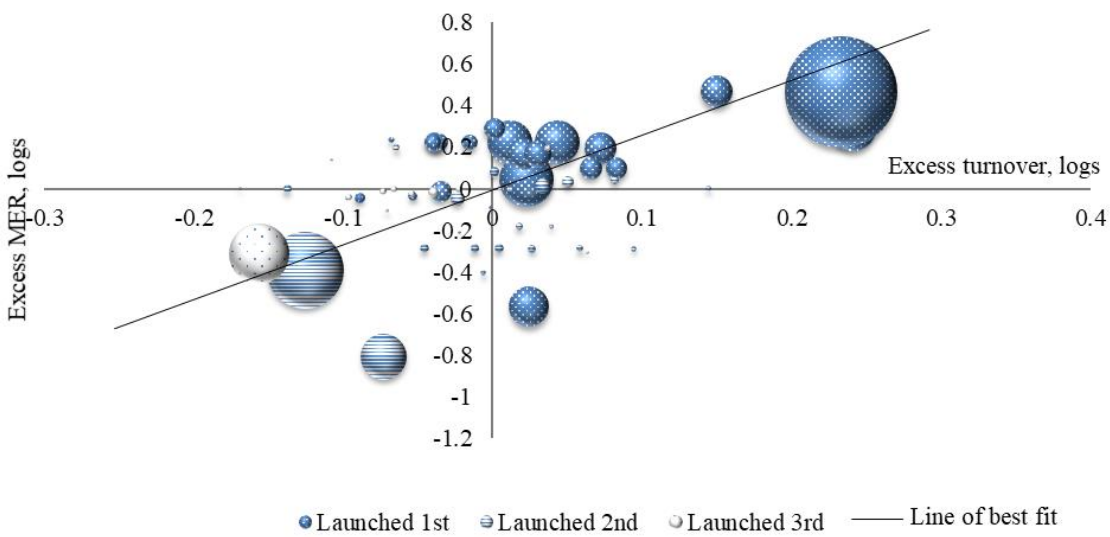
Fig. 2. Propensity for separating equilibrium as a function of model parameters.

This figure plots the propensity for separating equilibrium (y) against the model parameters. w_F is the share of AUM in an ETF with the highest MER per index, h_{FS} is the ratio of holding horizons of the high-MER and low-MER ETFs per index, f_{fix} is the fixed cost faced by ETF issuers in a given index (in \$ mln), X is the combined AUM of all ETFs in a given index (in \$ bn). Separating equilibrium occurs when $y > 0$ and $w_F > h_{FS}$. The non-varying model parameters are fixed as follows: $w_F = 0.83$, $f_{fix} = 3.28$, $h_{FS} = 0.55$, $c = 0.078$, $X = 56.34$. The parameter values are estimated by computing index-level average from the sample of 24 indices. For the derivation of separating equilibrium conditions, see Appendix 3.

Panel A: MER vs bid-ask spread for same index ETFs



Panel B: MER vs turnover for same index ETFs



Panel C: MER vs dollar volume for same index ETFs

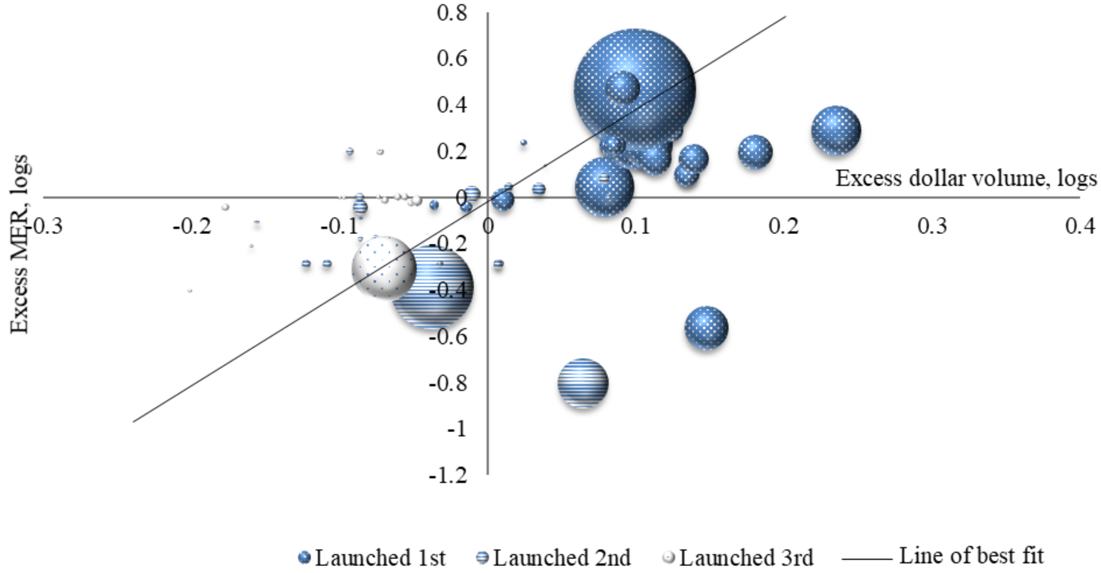


Fig. 3. MERs vs liquidity measures for ETFs tracking the same index.

This figure plots excess MERs against excess liquidity measures on a log-scale. An excess MER in an ETF is the percentage difference between this ETF's MER and the average MER across all ETFs tracking the same index as this ETF. The log transformation of MERs and liquidity measures is done by taking a natural logarithm of *(1+% Excess MER)* and *(1+% Excess Liquidity Measure)* respectively. The bubble size is proportional to assets under management (AUM) of a given ETF. The sample contains 60 same index US-domiciled ETFs based on 24 unique indices. All variables are calculated from the daily frequency data and averaged per ETF over the year 2017. MER is net expense ratio, relative spread is absolute bid-ask spread divided by midpoint, turnover is annualized percentage ratio of daily dollar volume divided by assets under management (AUM).

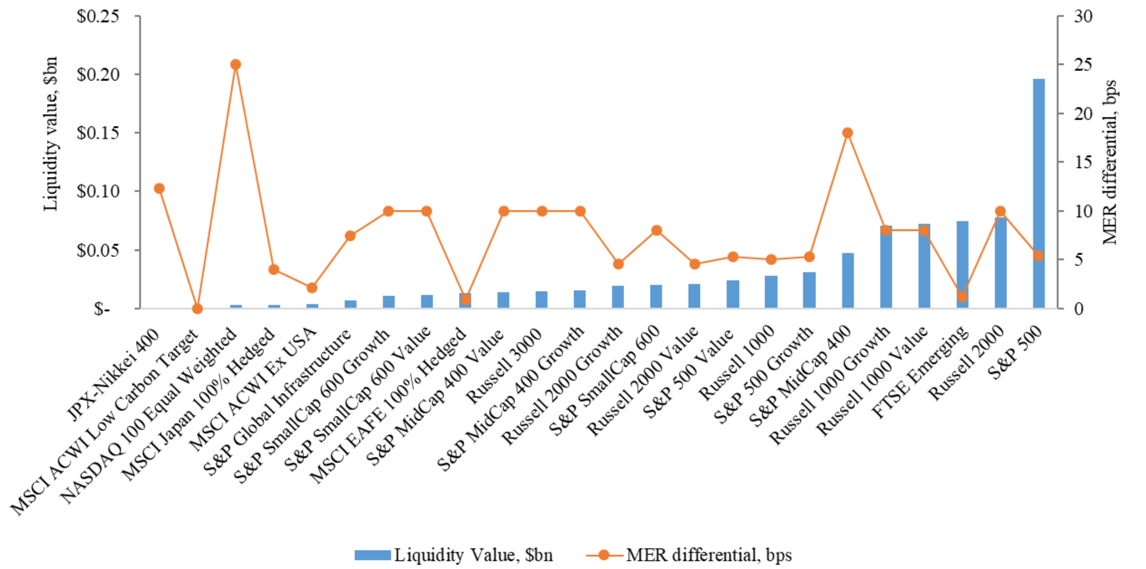


Fig. 4. MER differentials and the value of ETF liquidity.
This figure plots the MER differentials between same index ETFs, as well as the value of liquidity from in each index tracked by multiple ETFs. The MER differential is the difference between management expense ratios (MERs) the high-fee ETF and the low-fee ETF tracking the same index. The value of ETF liquidity is estimated as the profit of the high-fee ETF issuer less the profit of the low-fee ETF issuer. For details on welfare analysis, see Appendix 5. The sample contains 60 same-index US-domiciled ETFs based on 24 unique indices. All variables are calculated from the daily frequency data and averaged per ETF over the year 2017.