The life cycle of investment management when "today's alpha is tomorrow's beta"

Georgios Magkotsios*

Marshall School of Business, University of Southern California

May 1, 2017

Abstract

I discuss the effects of competition among managers within a class of investment funds. I assume that the fund class has a finite set of profitable investment opportunities. I show a connection among economies of aggregate scale, the curvature of the flow-performance relation, and the distribution of surplus among investors and managers. Initially, the fund class operates under increasing returns to aggregate scale. The flow-performance relation is concave, and the investor surplus gradually dominates the manager surplus. As the investment opportunities within the fund class diminish, the returns to aggregate scale reverse to decreasing, the flow-performance relation becomes convex, and the total surplus declines to zero. The aggregate risk is reduced through "closet indexing". The average returns from active investing are not persistent, and the fund class transforms eventually to a scalable pool of passively invested capital.

Keywords: investment management; alpha; liquidity; returns to scale; flow-performance relation; network externality.

JEL Classification Numbers: G11, G12, G23

*Georgios.Magkotsios.2017@marshall.usc.edu. I am grateful to my supervisors Wayne Ferson, Kevin Murphy, Kenneth Ahern, and Arthur Korteweg, as well as John Matsusaka, Oguzhan Ozbas, Gerard Hoberg, and the participants of the 2016 Trans-Atlantic Doctoral Conference at London Business School for their valuable comments.

I. Introduction

A fundamental question about asset management is whether the investor benefits more from active than passive strategies. In a pioneering paper, Berk and Green (2004) show that a single fund manager can extract all the surplus from investment by optimally increasing his fee. I focus on the competition among managers for investor capital and superior performance. The implications for fees, returns, risk, size, and the distribution of surplus among investors and managers are related to the investment opportunities within a class of funds. I show that competition diminishes these opportunities as the total number of funds increases over time.

The model provides an empirically testable connection between the effect of size on fund returns and subsequent investor flows. The declining availability of profitable investment opportunities implies a life cycle for the fund class. Managers in incipient fund classes create a new market and provide liquidity to investors by reducing the cost of active investing. The aggregate demand increases as more talented cohorts of managers replace previous incumbents. As a result, the fund class experiences increasing returns to aggregate scale. The investor flows are more sensitive to bad performing and less sensitive to good performing managers. This implies a concave flow-performance relation. The opposite is true for mature fund classes with a large number of funds and limited profitable opportunities. The fund class operates under diseconomies of aggregate scale, and the flow-performance relation is convex.

Two features define a fund class in the model. The first is its benchmark, namely a unique risk and return relation that reflects the underlying securities and the strategies used by the incumbent managers. The second feature is the finite supply of profitable opportunities that are available on aggregate. Each fund class is subject to a regulatory regime that limits the set of permissible active strategies. The regulatory constraints limit the opportunities for abnormal returns within the class. For instance, equity mutual fund managers are not permitted to short stocks. As a result, they must abstain from shorting strategies.

The managers enter the fund class with an exogenous level of talent that affects their returns

from active investing. The key variables in my model are the cross-sectional mean and variance of managerial talent. The mean provides a standard of comparison among the incumbents and defines the benchmark for the fund class. A fund's abnormal return (alpha) is its risk-adjusted return in excess of the benchmark. The cross-sectional variance of talent is a proxy for the availability of profitable opportunities within the fund class. A large dispersion allows the most talented managers to outperform their benchmark by a wide margin and yield large abnormal returns. On the other hand, the managers are alike in talent when the variance is low. The homogeneity in talent suggests that the managers crowd into similar trading strategies and investment opportunities, resulting in low alphas overall. The opportunities for abnormal returns within the fund class eventually deplete as the cross-sectional variance is reduced to zero.

During the outset of the life cycle, a small population of managers create a market and supply liquidity to investors by injecting capital into unexploited profitable opportunities. These opportunities would be unavailable or too costly for the investors to capture on their own. Similarly to Berk and Green (2004), the managers can extract parts of the total surplus from investment by indexing a fraction of their assets under management. As the number of funds rises over time, the competition resembles Schumpeter's "creative destruction" and forces to exit the managers who underperform with respect to the benchmark. The aggregate demand and investor surplus from alpha surge, because the competition increases the expected abnormal returns and decreases the fund fees. As a result, the fund class operates under increasing returns to aggregate scale during its early growth stage.

As the fund class matures, the escalating competition diminishes the opportunities for alpha at the aggregate level. Although the incumbent managers are on average more talented over time, they also become more homogeneous in their strategies. An increasing number of funds crowd into a diminishing opportunity set over time. The managers invest in the same direction altogether, and the increasing trading costs result in more elusive alphas. Eventually, the returns to aggregate scale reverse from increasing to decreasing, and the total surplus from active investing declines. The managers gradually index larger portions of their assets as their fund size grows and their alpha

depletes. Throughout its life cycle, a fund class transforms from a risky investment vehicle that is rich in opportunities for alpha to a set of funds whose performance and risk are similar to the class benchmark. This is the concept of "today's alpha is tomorrow's beta".[1] It implies that the opportunities for alpha at the aggregate level and active trading wane, as the entry of more talented managers over time increases the level of sophistication for the class benchmark.

The main feature in Berk and Green (2004) is diminishing returns to scale at the fund level. The underlying mechanism in my setup is based on class-wide variables such as the number of funds, aggregate size, and the cross-sectional variance in talent as a proxy for the remaining profitable opportunities. My model predicts that every fund class operates initially with increasing returns, followed by decreasing returns to aggregate scale as it matures over time. As a result, the competition among managers provides a potential explanation for the diminishing returns to aggregate scale that is observed in equity mutual funds (Pástor et al., 2015). Further empirical work on other fund classes such as fixed income or alternative investments could test this model for increasing returns to scale. In addition, the model predicts that the fees and returns of mutual funds during their early growth should be larger than what we observe today, while the fees and returns in asset classes such as hedge funds will decrease in the future.

The flow-performance relation is monotonically increasing, implying larger allocations to the most talented managers. This relation is concave during the early stages of the life cycle, when the fund class is abundant in profitable opportunities. Concave flows imply that investors are more sensitive to bad performance and less sensitive to good performance. The insensitivity of capital reallocation among talented managers follows from the investor's preference for variety, aiming to exploit a broader set of opportunities for alpha and diversify his risk. The least talented managers underperform the benchmark by a wide margin when the dispersion of talent is large, triggering thus large capital outflows from them.

During the late stages of the life cycle, the curvature of the flow-performance relation reverses to convex. Convexity indicates that flows are more sensitive to good performance and less sensitive to

[1] This is a popular expression among market practitioners that was coined by Andrew Lo. See also Cho (2017) for a costly arbitrage-based argument where alphas turn to betas.

bad performance. The returns of most managers are little different than the benchmark return when the opportunities for alpha are scarce across the fund class. This implies less sensitive outflows for those who underperform the benchmark, because their losses are relatively small. However, the reward for managers who yield positive alpha is disproportionately large, because it is more strenuous to yield positive alpha when all managers are similar in talent.

Following Ippolito (1992) and Sirri and Tufano (1998), an extensive empirical literature shows a convex flow-performance relation for equity mutual funds. My model implies that the convexity of investor flows is related to a scarcity in profitable investment opportunities within this fund class. On the other hand, Goldstein et al. (2017), Kaplan and Schoar (2005), and Getmansky (2012) show concave flows for investors in corporate bond mutual funds, private equity, and hedge funds respectively. My model implies that these fund classes should also operate under increasing returns to aggregate scale, both effects related to a substantial availability of opportunities for alpha.

I discuss the life cycle of investment management from an industrial organization point of view, and with symmetric information across the market. Garleanu and Pedersen (2016) model the competition among managers based on information inefficiency. The diminishing returns to aggregate scale in their model arise because the market becomes more efficient as the fund class grows in size. The poor abnormal returns within large classes such as mutual funds stem from a relatively large number of noise allocators, who invest mostly in uninformed managers. As a result, their contract is not first-best. The investors in my model reward positive innovations in performance with a finite elasticity of substitution, and the contract is first-best.

Garleanu and Pedersen (2016) also predict that fees decline over time, and the population of managers reduces to a small number of large funds that dominate the class. In the asymptotic limit, the investor search costs diminish, along with the fees and the number of funds. As a result, the fund class dies out. Contrary to their prediction, the number of funds increases in my model and it stabilizes to a finite value by the end of the life cycle. Mature fund classes are still valuable to investors, because they offer a low-cost diversification over investment strategies that are particular to that class.

II. The model

The model describes the competition among managers within a single class of funds. I assume that there is no managerial mobility across different fund classes. The manager's only alternative to active investing within the class is indexing his assets under management. The model is neo-classical and there is no moral hazard or adverse selection. The markets make rational expectations and investors allocate resources efficiently. As a result, the implied contract among investors and managers is first-best.

A. Individual fund and class-wide features

A fund class in this paper consists of all funds whose returns are subject to the same set of risk factors, and it may involve multiple investing styles. For instance, the class of equity mutual funds includes growth, income, and blend styles. The risk factors define a unique relation between a fund's returns and systematic risk within the class. This relation provides a class-wide benchmark to compare with each fund. The managers are compensated for their abnormal returns, namely their returns from active investing that are in excess of the class benchmark. The term "alpha" refers to the abnormal returns of a fund.

Each potential entrant to the fund class is endowed with an exogenous talent in active investing τ_i whose value cannot be modified after entry. The distribution of talent for the potential entrants is $H(\tau_i)$, and it is common for all entry cohorts. A manager's talent reflects his ability in exploiting profitable opportunities and yielding abnormal returns that do not stem from luck. The level of talent is unobservable and imperfectly known to the market, including the manager himself. All market participants observe the realized fund returns over time, and update their estimates about every manager's talent. I assume that the uncertainty about a manager's talent is resolved within one period after entry, implying that a single innovation is sufficient for the market to estimate the true value of his talent.

The class benchmark is itself a value of talent $\tau_B(t)$ that provides a standard of comparison

among the incumbent managers. Some incumbents outperform and others underperform with respect to the benchmark. In principle, the benchmark talent could have any value within the distribution of talent among the incumbent managers. I specify this value to be the cross-sectional weighted average of talent estimates. Thus, I define the benchmark $\tau_{_B}(t)$ as

$$\tau_{_B}(t) \equiv \begin{cases} \dfrac{1}{N_t} \displaystyle\sum_{i=1}^{N_t} q_{it} \hat{\tau}_{it} & \text{for } t > 0 \\ 0 & \text{for } t = 0 \end{cases} \tag{1}$$

where N_t is the number of incumbent managers only, and it does not include those who have exited at time t or before, while q_{it} is the size of fund i at time t. The term $\hat{\tau}_{it}$ is the market's estimate for the talent of manager i conditional on information until time t. Since the uncertainty about managerial talent is resolved within a period, the estimate is equal to the true value of talent for managers who were incumbents before time t. However, $\hat{\tau}_{it}$ for managers who enter the fund class at time t is an estimate from the distribution $H(\tau_i)$.

The gross abnormal return for fund i that is realized at time $t+1$ is

$$R_{it+1} = 1 + \tau_i - \tau_{_B}(t) + \varepsilon_{it+1} \ , \quad i = 1, \ldots, N_t \ , \quad t \geqslant 0 \tag{2}$$

As a result, the market uses the innovations at $t+1$ to update each manager's talent level individually, while comparing those estimates with the benchmark at time t. The term $\tau_i - \tau_{_B}(t)$ expresses the manager's alpha, and it can be positive or negative. The noise terms ε_{it+1} are jointly distributed over time and across managers with zero mean. These shocks express the component of luck in the realized return R_{it+1}.

I define a proxy for the remaining profitable opportunities within the fund class. The proxy is the market estimate of the cross-sectional variance in talent within the fund class, conditional on

information until time t

$$\hat{\sigma}_\tau(t) \equiv \frac{1}{N_t - 1} \sum_{i=1}^{N_t} (\hat{\tau}_{it} - \tau_{_B}(t))^2 \tag{3}$$

Since the uncertainty about talent is resolved during a period, $\hat{\sigma}_\tau(t)$ is equal to the true variance of talent in the population of incumbent managers if there is no entry and exit of funds. The cross-sectional variance $\hat{\sigma}_\tau(t)$ is a key variable to my model, because it provides a measure for the remaining profitable opportunities through active investing within the fund class. When $\hat{\sigma}_\tau(t)$ is large, the heterogeneity in talent across the incumbent managers implies a strong potential for abnormal returns by those on the right tail of the talent distribution, but also the potential for large realized losses by managers on the left tail of the distribution.

On the other hand, a large number of managers with a small variation in talent is related to a shortage of opportunities. When the active strategies of the incumbents span the full opportunity set for the fund class, then the managers are forced to crowd into similar strategies. As more funds chase the same opportunities, the rising trading costs make these opportunities more elusive. Even for the case where every manager is very talented in identifying profitable opportunities, the resulting alphas are small because all rival managers trade in the same direction. As a result, the combination of large number of funds N_t and small talent dispersion $\hat{\sigma}_\tau(t)$ implies a scarcity of profitable trades.

B. The investor's problem

The investors in Berk and Green (2004) supply flows to a single manager with perfect elasticity, based on innovations over time about the manager's talent. They reward positive abnormal returns from active investing with fund inflows, and punish negative abnormal returns with outflows. In my model, the fund class at time t includes N_t managers with varying levels of talent in active investing τ_i, where $i = 1, \ldots, N_t$. Similarly to Berk and Green, the supply of flows to each manager is based on performance.

The investor has mean-variance preferences for the allocations of capital within his portfolio of funds. His utility is quadratic, implying that the funds are imperfect substitutes. The finite elasticity of substitution between managers expresses the investor's tradeoff between superior talent and diversification. His goal is to exploit the profitable opportunities at the aggregate level. For this purpose, he employs the most talented managers. However, he also diversifies his wealth across multiple funds to mitigate his risk and benefit from labor pooling effects, because different managers exploit in principle different profitable opportunities.

The portfolio of funds at time t involves N_t active managers and one passive index. I introduce the passive fund to have the aggregate size of actively invested assets as an endogenous variable in the model. The investor's problem during a single period is the following:

$$\max_{\boldsymbol{q_t}} U = \boldsymbol{E_t}[\boldsymbol{q_t'}(\boldsymbol{r_{t+1}} - \boldsymbol{f_t})] - \frac{a}{2}\, \boldsymbol{q_t'} \boldsymbol{V_t} \boldsymbol{q_t} \quad s.t. \quad \boldsymbol{q_t'} \mathbf{1} \leqslant W_t \tag{4}$$

where a is a parameter related to the investor's risk aversion and elasticity of substitution, and W_t is the exogenous total wealth invested among the active funds and the passive index at time t. All vectors have $N_t + 1$ elements, where $\boldsymbol{q_t}$, $\boldsymbol{f_t}$, and $\boldsymbol{r_{t+1}}$ are the vectors for fund sizes, fees, and nominal returns realized at $t+1$. In addition, $\mathbf{1}$ is a vector of ones, and $\boldsymbol{V_t}$ is the $(N_t+1)\times(N_t+1)$ covariance matrix for the returns of the passive index and the incumbent active managers within the fund class. The fund fee represents the price that the investor pays for active management.[2] Lemma 1 shows the solution to the investor's problem.

LEMMA 1. *The equilibrium size for fund i at time t is given by*

$$q_{it}^{*} = \frac{b_0}{N_t + 1} + b_1 E_t[R_{it+1} - f_{it}] - \frac{b_2}{N_t + 1} \sum_{j=0}^{N_t} E_t[R_{jt+1} - f_{jt}] \tag{5}$$

where b_0 is a scalar that encapsulates the investor's total wealth W_t, $b_2 < b_1$, and $E_t[R_{jt+1} - f_{jt}]$ is the expected net-of-fee alpha for fund j.

[2] I do not make a distinction between management and performance fees.

The equilibrium demand for a fund i includes the net-of-fee alpha that the investor receives from that fund.[3] This is similar to the equilibrium solution in Berk and Green (2004). The single manager in their setup can act as a discriminating monopolist and raise his fee until he absorbs all surplus from investment. The new feature in this model is the dependence of the investor demand on rival net-of-fee alphas within the fund class. Equation (5) shows that any arbitrary attempt from a manager to raise his own fees may trigger outflows that are redistributed to rival funds. This equation also shows that the size of any individual fund is affected by the total number N_t of incumbent managers. This variable is a proxy for the intensity of the competition among the managers within a fund class.

C. The competition among managers

The managers within a fund class compete for investor capital and profitable investment opportunities. The manager of fund i sets his fee f_{it} at time t, and the investors determine the fund size q_{it} through flows that satisfy their demand in equation (5). The number of incumbents N_t is a measure of the competition's intensity among the managers, and the cross-sectional variance of talent is a proxy for the remaining profitable opportunities within the fund class. These are the underlying variables that determine the equilibrium fees and fund sizes. The abnormal returns R_{it+1} reflect a manager's talent in active investing. A manager's talent helps in generating returns that can potentially outperform the benchmark, but he has no control or influence over the performance of rival funds and the benchmark returns.

The cost of active investing has two components. The first stems from bid-ask spreads that every manager pays to implement his strategies. This cost is related to the fund's own size, but it is also affected by the availability of profitable opportunities. The second component is an opportunity cost to the manager for maintaining a particular strategy. The total cost function for active investing

[3]The gross alpha for the index fund is zero by definition, implying that $E_t[R_{0t+1}] \equiv 0$ always.

is

$$C(q_{it}, \hat{\sigma}_\tau(t)) = \frac{cq_{it}^2}{\hat{\sigma}_\tau(t)} + hq_{it} \tag{6}$$

where c and h are constants. The quadratic term captures the diminishing returns to scale at the fund level. The cross-sectional variance of talent describes the scarcity of opportunities for abnormal returns at the aggregate level. When $\hat{\sigma}_\tau(t)$ is large, the fund class is abundant in investment opportunities and the trading costs for every manager are relatively low. On the other hand, a small $\hat{\sigma}_\tau(t)$ implies that the profitable opportunities are scarce, and that managers must trade in the same direction. As a result, the trading costs rise for all managers and impede them from outperforming the class benchmark.

The linear term hq_{it} in equation (6) expresses an additional opportunity cost. This is the cost that the manager incurs when he forgoes investment opportunities from restrictions on his portfolio, imposed either by investors or the manager himself.[4] As a result, the managers are differentiated both by talent level and the opportunities for alpha that they chase. The total cost of active investing is unobservable to the investor. It affects the manager's profits from fees, and it also affects the investor's surplus indirectly by decreasing the total value added to the initial investment.

The fund's revenue is based on fees, and it is a fraction of the value that the manager generates for the investor. Each manager i chooses his fee at time t to maximize his expected profits from active investing. The manager's problem is

$$\max_{f_{it}} E_t[\Pi_{it+1}] = f_{it} q_{it} E_t[R_{it+1}] - \frac{cq_{it}^2}{\hat{\sigma}_\tau(t)} - hq_{it} \tag{7}$$

where q_{it} is the investor demand in equation (5) and it clears the market. The Nash equilibrium for

[4] For instance, it is quite common for institutional investors to allocate capital to managers under the condition that they do not trade certain securities.

equations (5) and (7) yields the fund's response function

$$f_{it}^{*} = \Bigg[2b_0 b_1 c + (N_t+1) b_1 h \hat{\sigma}_\tau(t) - 2b_1 b_2 c \sum_{j=0}^{N_t} E_t\left[R_{jt+1} - f_{jt}^{*}\right] + E_t[R_{it+1}] \Bigg(2b_1^2 c(N_t+1) + b_0 \hat{\sigma}_\tau(t) + b_1(N_t+1)\hat{\sigma}_\tau(t) - b_2 \hat{\sigma}_\tau(t) \sum_{j=0}^{N_t} E_t\left[R_{jt+1} - f_{jt}^{*}\right] \Bigg) \Bigg] \cdot \Big[2b_1(N_t+1)\Big(b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t)\Big) \Big]^{-1} \quad (8)$$

given the approximation $\partial q_{it}/\partial f_{it} = -b_1 + b_2/(N_t+1) \approx -b_1$, and also $R_{it+1}^2 \approx R_{it+1}$ in equation (8). The fund's response function depends on the sum of expected returns and the sum of equilibrium fees for all incumbent managers. Adding up equations (8) for all funds allows to solve for the sum of fees, and then retrieve each equilibrium fee. Substituting equation (8) in equation (5) gives the equilibrium fund size. The equilibrium aggregate size Q_t^* is the sum of the fund sizes for the N_t incumbent active managers.

An approximation that is useful for the comparative statics analysis hereafter, is that the sum of returns $SR \equiv \sum_j E_t[R_{jt+1}]$ is not affected by perturbations to the expected return of a single fund. This implies that any innovations about a manager's talent cannot affect the average estimated talent for the fund class. The following Lemma discusses the monotonicity of the relation between fund flows and performance, and the comparative statics for the number of incumbent funds N_t.

LEMMA 2. *An increasing number of managers is detrimental to fund fees and size, but it correlates positively with the aggregate investor demand. The flow-performance relation for each fund is monotonically increasing, but it declines to zero as the profitable opportunities within the fund class diminish. Specifically,*

$$\frac{\partial f_{it}^{*}}{\partial N_t} < 0\ , \quad \frac{\partial q_{it}^{*}}{\partial N_t} < 0\ , \quad \frac{\partial Q_t^{*}}{\partial N_t} > 0\ , \ \textit{and} \quad \frac{\partial q_{it}^{*}}{\partial E_t[R_{it+1}]} \gtrless 0 \quad \forall i,t \quad (9)$$

The competition among managers affects adversely the funds and benefits the investors. The

investor has preference for variety in active management, and diversifies his capital across multiple managers to exploit the profitable opportunities within the fund class. As a result, the aggregate size increases when new managers enter the fund class, while the assets and fees of each fund decline.

The distribution of surplus among investors and managers depends on the availability of profitable opportunities within the fund class. I assume that fees are not reinvested in the next period, making them a dead-weight cost. As a result, the total investment surplus from a single fund i at time t after fees and trading costs is equal to

$$TS_{it} = q^*_{it}\left(E_t[R_{it+1}] - f^*_{it}\right) - \frac{c\left(q^*_{it}\right)^2}{\hat{\sigma}_\tau(t)} - hq^*_{it} \tag{10}$$

The manager surplus is the fund profit from fees, and the investor surplus is what remains from the total surplus after the manager extracts his surplus through fees.

The manager may extract a larger portion of the total surplus if he invests a fraction of his assets under management into a passive index. This is termed "closet indexing", because the investor cannot monitor whether all capital is actively traded or not. Since indexing is unobservable, I assume that the investor's demand is the same as in equation (5), and the equilibrium fees are still determined by equation (7). This implies that indexing does not affect the population of incumbent managers and the allocation of capital. The manager trades actively $x_{it}q_{it}$ of his assets under management, and invests the rest of his capital $(1 - x_{it})q_{it}$ in passive strategies. The optimal fraction $x^*_{it} < 1$ is found by maximizing the following objective:

$$\max_{x_{it}} \left\{ f^*_{it}q^*_{it}(x_{it}E_t[R_{it+1}] + 1 - x_{it}) - \frac{c\left(x_{it}q^*_{it}\right)^2}{\hat{\sigma}_\tau(t)} - hq^*_{it} \right\} \tag{11}$$

Equation (11) implies that passively invested capital does not earn abnormal returns. The asterisks denote equilibrium values for capital and fees that are determined by equations (5) and (7).

LEMMA 3. *Each manager i may choose to trade actively at time t only a fraction of his assets*

under management

$$x_{it}^{*} = \frac{f_{it}^{*}(E_t[R_{it+1}] - 1)\hat{\sigma}_{\tau}(t)}{2cq_{it}^{*}} = \frac{f_{it}^{*}(\tau_i - \tau_{_B}(t))\hat{\sigma}_{\tau}(t)}{2cq_{it}^{*}} \tag{12}$$

to appropriate some of the investment surplus without affecting the population of rival managers, fund fees, and the investor's demand.

Lemma 3 shows that the managers with the largest alphas $\tau_i - \tau_{_B}(t)$ are more active, and tend to increase their fraction in active management when fees are large. However, the managers tend to become more passive as their total fund size increases, or when the profitable opportunities diminish through a declining cross-sectional variance of talent.

III. The life cycle of investment management

The model in Section II describes a snapshot of the competition among managers. In this section, I discuss the time-series effects of this competition. Potential entrant managers will enter the fund class if their expected profits are positive, while incumbents with negative expected profits will exit. In equilibrium, the marginal entrant at time t is the manager who has zero expected profits. As a result, the number of incumbent managers N_t at time t is specified by

$$E_t[\Pi_{kt+1}] = f_{kt}^{*}q_{kt}^{*}E_t[R_{kt+1}] - \frac{c\left(q_{kt}^{*}\right)^2}{\hat{\sigma}_{\tau}(t)} - hq_{kt}^{*} = 0 \tag{13}$$

where k is the marginal entrant fund, and the asterisks denote equilibrium values that are specified by the solution to equations (4) and (7). The solution of equation (13) yields the number of incumbents $N_t\Big(\hat{\sigma}_{\tau}(t)\Big)$ as a function of the cross-sectional dispersion of talent within the fund class. Unfortunately, this equation has no closed-form solution. Alternatively, I model an exogenous entry of managers.

The fundamental assumption about entry during the life cycle of the fund class is that the competition among managers intensifies over time. Historical data on the mutual fund industry seem

to support this assumption. Figure 1 illustrates a growth in the number of funds and aggregate demand, especially during the 1990s. The number of funds seems to reach a plateau after the dot-com crisis, while the aggregate demand has maintained its ascending track. I make the following assumption about fund entry over time.

Entry assumption: *The rate of net entry is positive and proportional to the aggregate profits of the incumbent managers. Specifically,*

$$\frac{dN_t}{dt} \propto E_t \left[\sum_{j=1}^{N_t} \Pi_{jt+1} \right] > 0 \ , \ \forall t \geqslant 0 \tag{14}$$

The expected profits for funds at any time are affected by the availability of profitable investment opportunities. The assumption above guarantees a positive net entry throughout the life cycle. However, the net entry is lessened over time as the opportunities within the fund class diminish from competition. Entry ceases at the end of the life cycle, when all opportunities are depleted and the cross-sectional dispersion of talent declines to zero, implying zero profits from active investing for all incumbent managers.

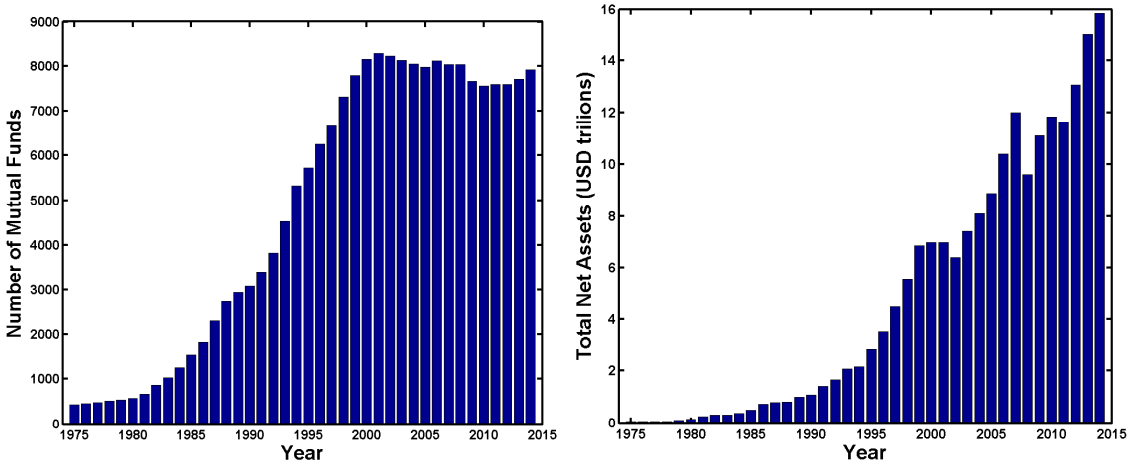

Figure 1: Time series of the total number of mutual funds in US (left), and total net assets in US trillions (right). Data are from the Investment Company Institute.

A. Investor flows and managerial competition

The investor has mean-variance preferences and chooses to diversify his wealth across multiple funds. The matching between capital and managerial talent is assortative, implying that more talented managers administer larger portions of assets. The investors learn about managerial talent through a stochastic process that utilizes the innovations from realized returns. The uncertainty about an incumbent's talent is resolved within a single period in this model. Lemma 2 shows a monotonically increasing relation between investor flows and fund performance. This implies that the investor withdraws capital from managers who underperform with respect to the benchmark, and reinvests it to managers who outperform the benchmark. He supplies inflows to the funds with positive innovations in anticipation of larger returns and investment surplus in the future. The following Lemmas show the impact of flows on the class benchmark, and compare the flows across different entry cohorts over time.

LEMMA 4. *The performance-based flows by the investor impede managers progressively from outperforming the class benchmark, i.e.* $\tau_{_B}(t) > \tau_{_B}(t-1), \forall t$.

LEMMA 5. *Funds in earlier entry cohorts receive comparatively larger flows than funds in subsequent entry cohorts. Specifically, between two managers of identical talent and realized returns who entered at different cohorts, the manager who entered first will receive larger flows over time.*

The competition among managers and the investor's response through capital flows characterize the evolution of the class benchmark $\tau_{_B}(t)$. Managers with $\tau_i < \tau_{_B}(t)$ will have negative expected abnormal returns from time t and beyond, and they will underperform the benchmark. These managers will experience outflows from investors until they liquidate the fund and exit. The rising $\tau_{_B}(t)$ implies that many incumbents who yield positive abnormal returns at a certain time will eventually be surpassed by younger and more talented managers in the future. This feature of the competition is intuitively similar to a *"creative destruction"* (Schumpeter, 1942). The less

talented managers are forced to exit, and they are replaced by younger cohorts of more talented managers.[5]

More importantly, the increasing $\tau_B(t)$ and the exogenous prior distribution $H(\tau_i)$ for the talent of potential entrants imply that the cross-sectional variance $\hat{\sigma}_\tau(t)$ for incumbent managers must decrease over time. The effects of managerial turnover during the life cycle of the fund class are summarized below.

PROPOSITION 1. *The competition among managers decreases their fees and increases the average talent over time. However, it depletes the opportunities for alpha within the class by decreasing the cross-sectional variance in talent, i.e.*

$$\frac{d\hat{\sigma}_\tau(t)}{dt} < 0 \tag{15}$$

Proposition 1 has significant implications to the evolution of the life cycle. The sensitivity of fund fees to performance depends on the dispersion of talent. For every fund i at time t, the limiting values for the relation between fees and performance are

$$\lim_{\hat{\sigma}_\tau(t)\to\infty} \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = -\frac{\Big[2(N_t+1)b_1 - N_t b_2\Big]h}{2(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^2} < 0 \tag{16}$$

$$\lim_{\hat{\sigma}_\tau(t)\to 0} \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = \frac{(N_t+1)b_1 - N_t b_2}{(N_t+1)(b_1-b_2)} > 0 \tag{17}$$

When the fund class is bountiful in profitable opportunities, the managers decrease their fees in expectation of good performance in the future. This novel result implies that the competition among managers provides a means to counteracting increases in fund fees, by augmenting investor flows to rival funds. The investors have preferences for variety, and diversify their wealth across multiple funds with allocations proportional to managerial talent. If a manager attempts to increase his fees in expectation of a positive innovation, then the investors can react by redirecting their flows to

[5] Pástor et al. (2015) show empirically a rising talent over time in equity mutual funds (see Figure 2 in their paper). They explain this trend in terms of changes to the population of managers.

rival managers of similar talent. On the contrary, the manager decreases his fee to boost his future capital inflows.

Proposition 1 suggests that the remaining profitable opportunities become gradually more elusive, as it becomes more strenuous to outperform the class benchmark over time. Below a certain level of $\hat{\sigma}_\tau(t)$, it is impractical for the investor to diversify capital over similarly talented managers with low alphas, and managers who can still yield large abnormal returns are more valuable. This situation is similar to the setup of Berk and Green (2004), where the manager can increase his fee in expectation of good performance. The following Lemma shows that the threshold level of talent dispersion where the manager can increase his fee is unique, and its value depends on the level of intensity for the competition among his peers.

LEMMA 6. *There exists a unique positive root σ_{thr} that is common to the following equations*

$$\frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = 0 \quad , \quad \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = 0 \quad , \; and \quad \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} = 0 \tag{18}$$

and its value is increasing in the number of incumbent managers N_t.

$$\begin{aligned}\sigma_{thr}(N_t) = b_1 c \Bigg\{ & (N_t+1)b_1 - b_0 + b_2 \left(\sum_{j=0}^{N_t} E_t[R_{jt+1}] - (N_t+1) \right) + \\ & \Bigg[4h(N_t+1)^2(2b_1-b_2)(b_1-b_2) + \left(b_2 \left(\sum_{j=0}^{N_t} E_t[R_{jt+1}] - (N_t+1) \right) + \right. \\ & \left. b_1(N_t+1) - b_0 \right)^2 \Bigg]^{1/2} \Bigg\} \cdot \Big[h(N_t+1)(2b_1-b_2) \Big]^{-1}\end{aligned} \tag{19}$$

Lemma 6 suggests that the life cycle of a fund class may be separated in two stages, depending on the availability of profitable investment opportunities. The evolution of the life cycle has consequences for the economies of aggregate scale, the curvature of the flow-performance relation, and the distribution of investment surplus among the managers and investors. I discuss these consequences below for each stage of the life cycle.

B. The early stages of the life cycle

Incipient fund classes have few managers and little aggregate capital. The threshold variance σ_{thr} in equation (19) is nearly zero. The fund class is abundant in profitable opportunities, as implied by $\hat{\sigma}_\tau(t) > \sigma_{thr}$. The managers compete for superior performance and investor flows. The fees are negatively correlated with expected performance, because the managers try to maintain their share of investor flows while new funds enter. The number of funds grows over time and the competition reduces $\hat{\sigma}_\tau(t)$ as opportunities are exploited. The aggregate capital within the fund class increases over time too, as the following Lemma shows.

LEMMA 7. *The equilibrium aggregate demand Q_t^* increases during the early stages of the life cycle, i.e. $Q_{t+1}^* > Q_t^*$.*

The equilibrium demand for a single fund is heterogeneous in the manager's own performance, but homogeneous in the cross-sectional dispersion of talent. This feature along with preferences for variety introduce a positive network externality among investors. An investor that diversifies his capital across multiple funds induces the managers to decrease their fees. The capital allocation by a single investor benefits the other investors too, because it decreases their fee costs. As a result, the concurrent growth in the number of funds and aggregate demand is advantageous to investors. This is a network externality similar to Katz and Shapiro (1985), where "consumers are assumed to be heterogeneous in their basic willingness to pay for the product, but homogeneous in their valuation of the network externality".

The network externality, also known as demand-side economies of scale, and the competition among managers trigger an increase in expected fund performance. The declining fees (Proposition 1) are associated with larger expected returns (see equation 16). Since the aggregate demand and expected fund performance correlate positively, the fund class operates under increasing returns to aggregate scale during the early stage of its life cycle. Figure 2 shows that the correlation between aggregate demand and expected fund returns is positive for $\hat{\sigma}_\tau(t) > \sigma_{thr}$.

The flow-performance relation for nascent fund classes is concave. Figure 2 shows that the

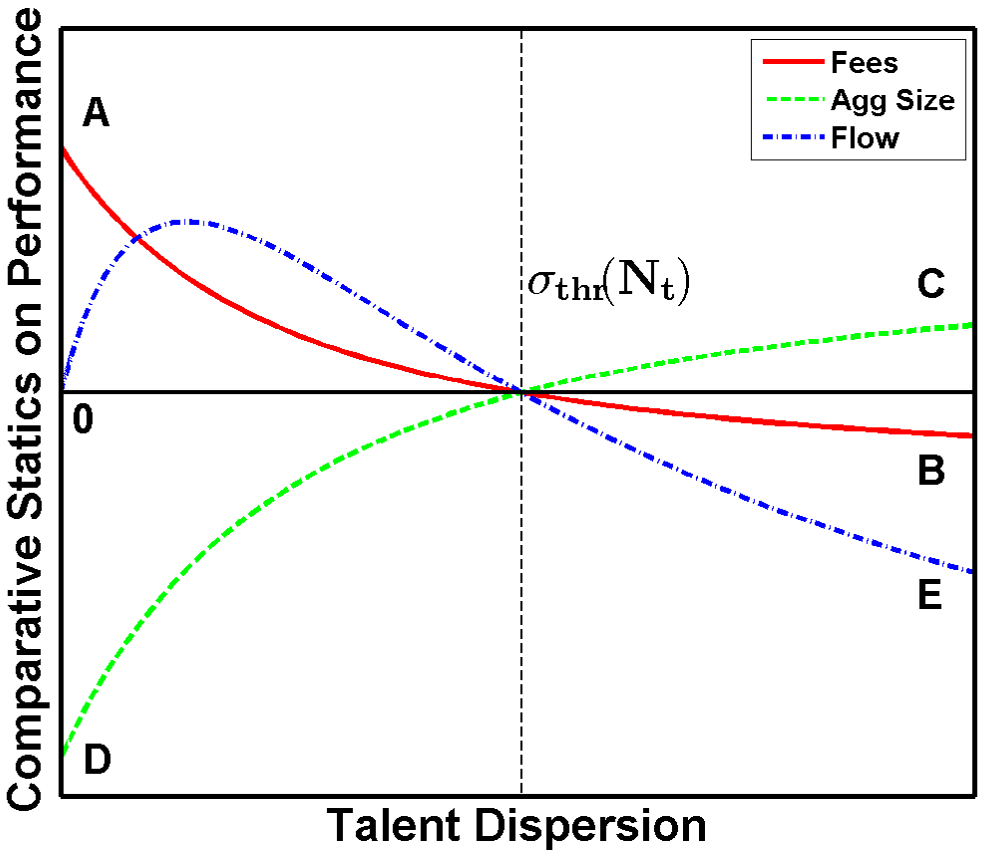

Figure 2: Comparative statics of fees on performance (red solid line), returns to aggregate scale (green dashed line), and flow-performance relation (blue dot-dashed line) as functions of the cross-sectional dispersion of managerial talent. The vertical dashed line marks the threshold variance σ_{thr} where the returns to aggregate scale reverse from increasing to decreasing, and the curvature of the flow-performance relation switches from concave to convex. The graph also shows limiting values for all curves when $\hat{\sigma}_\tau(t) \to 0$ and $\hat{\sigma}_\tau(t) \to \infty$. These are

$$\text{(A)} \quad \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = \frac{(N_t+1)b_1 - N_t b_2}{(N_t+1)(b_1-b_2)}$$

$$\text{(B)} \quad \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = -\frac{(2(N_t+1)b_1 - N_t b_2)h}{2(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^2}$$

$$\text{(C)} \quad \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = \frac{b_1((N_t+1)b_1 - N_t b_2)h}{(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^2}$$

$$\text{(D)} \quad \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = -\frac{b_1((N_t+1)b_1 - N_t b_2)}{(N_t+1)(b_1-b_2)}$$

$$\text{(E)} \quad \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} = -\frac{b_1(2(N_t+1)b_1 - (N_t+2)b_2)h}{2(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^3}$$

second derivative of fund size with expected returns is negative. Concavity implies that investor flows are more sensitive to bad performance, and less sensitive to good performance. The concavity is related to a large dispersion of talent within the fund class. The investor faces a tradeoff between seeking the most talented managers and diversifying his wealth to exploit potentially new investment opportunities. When there are plenty of these opportunities available (large $\hat{\sigma}_\tau(t)$), the investor favors diversification across multiple funds. Investors who are the earliest in allocating capital to young funds are those who supply the largest flows over time (see Lemma 5) and benefit the most from future performance, until the uncertainty about each manager's talent is resolved. As a result, the relatively small sensitivity to good performance stems from the investor's attempt to claim promptly a broad set of profitable opportunities. On the other hand, a large dispersion of talent implies the possibility for large realized losses too. Managers who underperform the benchmark during this stage of the life cycle are very harmful to the investor's portfolio. As a result, the investor flows are most sensitive to bad performance.

The talent dispersion also affects the distribution of total surplus among investors and managers. The total surplus from a single fund is given by equation (10), and it plateaus to a nearly constant value for large talent dispersion. The manager can extract a part of the surplus that he generates if he indexes a fraction of his assets under management (Lemma 3). During the early stages of the life cycle, the manager surplus is approximately linear in talent dispersion. The manager surplus may be either larger or smaller than the investor surplus at the onset of the life cycle, depending on the values of the model parameters. However, as the profitable opportunities are exploited over time and $\hat{\sigma}_\tau(t)$ decreases, the investor surplus increases and eventually dominates the manager surplus.

PROPOSITION 2. *The returns to aggregate scale are increasing during the early stages of the life cycle, stemming from demand-side economies of scale and the competition among managers. The flow-performance relation is concave. The total surplus from investment of a single fund is positive and split between the investor and the manager, with the investor enjoying the larger portion.*

C. *The late stages of the life cycle*

Despite the improvement in average talent within the class over time, the decreasing $\hat{\sigma}_\tau(t)$ prognosticates the eventual depletion of the opportunities for abnormal returns. The scarcity of profitable opportunities gradually attenuates the benefit of the network externality by raising the cost of active investing (see equation (6)). As the competition intensifies, an increasing number of managers chase a diminishing set of investment opportunities at higher cost. When the dispersion of talent reduces below the critical level where $\hat{\sigma}_\tau(t) \leqslant \sigma_{thr}$, the life cycle of the fund class transitions to its late stages. Managers who outperform the benchmark by a wide margin are rare, because of the low dispersion of talent.

The correlation between fund fees and performance changes to positive (ceteris paribus), implying that the effect of the network externality deteriorates. Since the fund fees decrease over time (Proposition 1), the expected returns decline. However, the aggregate demand may increase or decrease during the late stages of the life cycle. Following the deterioration of the network externality and the low prospects for future abnormal returns (low $\hat{\sigma}_\tau(t)$), the returns to aggregate scale reverse from increasing to decreasing. The transition is rooted in the rise of trading costs for all incumbents through the reduction of profitable opportunities within the fund class. Figure 2 shows that the correlation between the aggregate demand and expected fund returns is negative for $\hat{\sigma}_\tau(t) < \sigma_{thr}$ values of talent dispersion.

The flow-performance relation during the late stages of the life cycle is convex. Figure 2 shows that the second derivative of fund size with expected returns is positive for $\hat{\sigma}_\tau(t) < \sigma_{thr}$, and σ_{thr} is the inflection point for the flow-performance curve. Convexity implies that investor flows are more sensitive to good performance, and less sensitive to bad performance, and it is related to the dispersion of talent. The managers crowd around the few remaining profitable opportunities within the fund class when $\hat{\sigma}_\tau(t)$ is small. They must trade in the same direction to exploit these opportunities, raising thus the cost of active trading and making the opportunities more elusive. As a result, it becomes more strenuous to outperform the benchmark. The convexity of the flow-

performance relation implies that the investors value significantly those managers who still yield positive alpha within this environment.[6] In addition, the investor outflows from managers who underperform with respect to the benchmark are moderate, because all incumbents are alike in talent and realized loses are smaller compared to earlier stages of the life cycle.

The diminishing dispersion of talent during the late stages of the life cycle is critical to the investment surplus and aggregate risk within the fund class. Figure 3 shows the decline of total investment surplus when the cross-sectional dispersion of talent is small. The deterioration of the network externality allows the managers to increase their fees in expectation of good performance. Depending on the values for the model parameters, the manager can potentially capture a larger portion of the total surplus against the investor. However, the total surplus deteriorates to zero along with the dispersion of talent. This implies that active investing becomes less valuable to both investors and managers when there are not many profitable opportunities within the class.

On the other hand, every manager indexes the majority of his assets when $\hat{\sigma}_\tau(t)$ is very small (Lemma 3). Indexing is less risky than active investing, suggesting that the aggregate risk within the fund class is reduced during the late stages of the life cycle. The number of funds is still increasing, until all the opportunities for alpha are depleted and every incumbent fully indexes his assets. At that time, all entries cease and the number of funds N_t plateaus, because the potential entrants are indifferent between indexing and not entering. Throughout its life cycle, the fund class transforms from an investment vehicle that yields surplus from active trading to a large pool of capital where risk and performance are similar to indexing at the margin. This is the intuition of "today's alpha" becoming inevitably "tomorrow's beta", as the profitable investment opportunities within the fund class are competed away.

PROPOSITION 3 (*Today's alpha is tomorrow's beta*). *During the late stages of the life cycle, the returns to aggregate scale reverse from increasing to decreasing and the flow-performance relation is convex. The total surplus from active investing declines asymptotically to zero. The aggregate risk is smaller than earlier stages of the life cycle, because managers index larger portions of their*

[6] Berk and Green (2004) also derive a convex flow-performance relation when managerial talent is scarce.

assets under management.

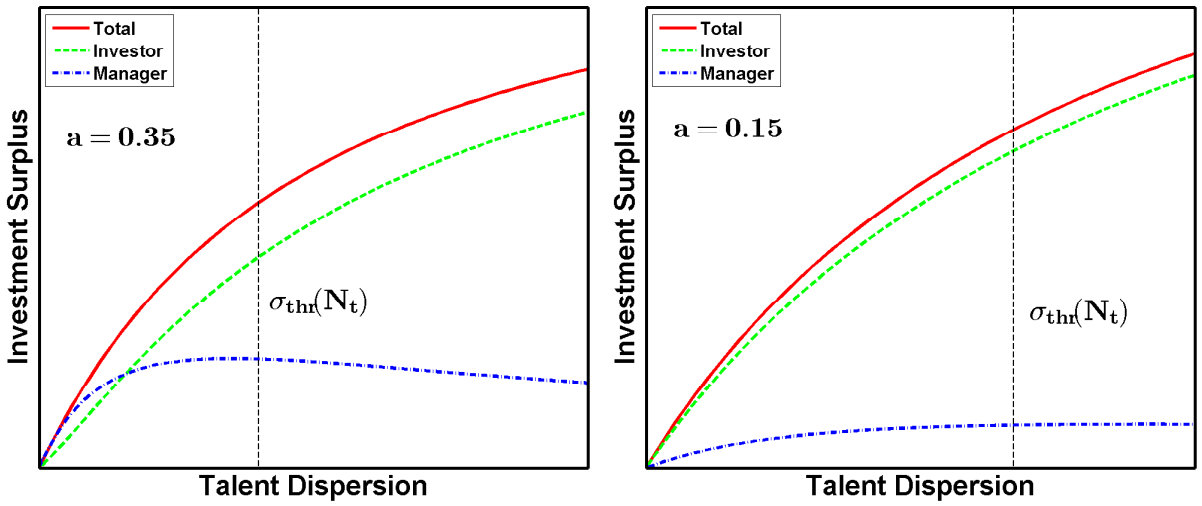

Figure 3: Investment surplus as a function of the cross-sectional dispersion of talent. The parameter values are $c = 0.01$, $h = 0.1$, $q_0 = 1$, $N_t = 3$, and $b_0 = 1.5$ for both plot. The manager surplus exceeds the investor surplus as $\hat{\sigma}_\tau(t) \to 0$ for $a = 0.35$ in the left plot ($b_1 = 2.857$, $b_2 = 1.43$), but it is always smaller for $a = 0.15$ in the right plot ($b_1 = 6.667$, $b_2 = 3.333$).

D. The life cycle and aggregate liquidity

The model shows how the availability of profitable investment opportunities within the fund class is critical to the evolution of its life cycle. A potential interpretation of the dynamics discussed above is based on aggregate liquidity. The term liquidity describes broadly the ability to trade a specific asset at low spreads and without affecting adversely the asset's price. Berk and Green (2004) assume diminishing returns to scale through a convex cost function for active management. The single manager in their setup is liquidity-constrained, because he expends resources to pay for bid-ask spreads that rise dramatically with his fund size and affect his returns. The managers in this model still have to pay for bid-ask spreads, but the scarce resource is the set of profitable opportunities that are available within a fund class.[7]

Within nascent fund classes there are plenty of opportunities for alpha, and the managers can exploit them at relatively low cost. By creating a market for active investing they supply liquidity

[7]Glode and Green (2011) use a similar concept to explain the persistence in performance for hedge funds.

to investors. The investor diversifies his wealth across multiple funds to mitigate risk, but also benefits from the network effect that arises from the diversification over multiple active trading strategies. The cost of this liquidity is an indirect reduction to each fund's net-of-fee alpha through the manager's opportunity cost of following a particular strategy (see equation (6)). As a result, hQ_t is the total liquidity premium that the investor pays for the network effect at time t. This premium determines the magnitude for the economies of scale at the aggregate level, as the asymptotic limit for $\partial Q_t / \partial E_t[R_{it+1}]$ in Figure 2 shows.

However, the liquidity at the aggregate level gradually diminishes as the profitable opportunities are exploited over time. Eventually, a large number of managers compete with each other and trade in the same direction to capture the few remaining opportunities. This increases the trading cost and affects the returns of all managers simultaneously. Therefore, the availability of profitable opportunities within the fund class is inherently connected to aggregate liquidity. The managers are liquidity-constrained at the aggregate level, rather than the fund level.

IV. Empirical implications of the model

Recent empirical evidence by Pástor et al. (2015) show that diminishing returns to aggregate scale for equity mutual funds dominate the returns to scale at the fund level. My model provides a potential theoretical explanation for the origin of the diminishing returns to aggregate scale, which is the depletion of the opportunities for alpha within the fund class (Propositions 1 and 3). The equilibrium solution involves efficient allocations, and all outcomes are first-best. Garleanu and Pedersen (2016) provide an alternative explanation for diseconomies of scale, based on the competition among managers and informational inefficiency in investment management. Consequently, the competition among managers is the key extension to Berk and Green (2004) that accounts for the existence of diminishing returns to aggregate scale. In addition, Proposition 2 suggests that the equity mutual fund industry must have operated under increasing returns to scale during the early growth phase of its life cycle. The same result should be true for other fund classes, such as bond

mutual funds, hedge funds, and other alternative investment asset classes.

Proposition 3 demonstrates the challenges in measuring alpha and investment surplus within mature classes like mutual funds. The empirical literature shows insignificant net abnormal returns and low gross abnormal returns on average. These results have been interpreted through rent-seeking managers that successfully absorb all surplus from investment (e.g. Berk and Green (2004)). I argue that mature classes like mutual funds have little alpha to offer, which makes it hard to estimate. This is consistent with the empirical results of Fama and French (2010), who show that most equity mutual funds yield zero alpha and managers at the tails have small positive and negative alphas.

My model shows that managerial performance is not persistent at the fund and aggregate levels. Diseconomies of scale at the fund level and investor flows after positive innovations undermine the future performance of a specific manager. This is also the result of Berk and Green (2004). At the aggregate level, the model reveals a non-monotonic evolution over time for the average fund return. It increases initially, but eventually it declines. Depending on the choice of sample period, average performance may seem persistent due to the effect of the network externality. However, the model predicts that performance must eventually erode, along with the profitable investment opportunities at the aggregate level.

The model predicts that the aggregate size of equity mutual funds will continue to grow in the long run even with poor abnormal returns, as they have evolved into a relatively safe investment vehicle that can manage large pools of capital at low cost. This result is similar to Glode (2011), who justifies the negative expected performance as an insurance premium that investors pay to protect themselves against bad states of the economy. It is also in agreement with Pástor and Stambaugh (2012), who assume that investor preferences account for diminishing returns to aggregate scale to explain the demand increase with mediocre performance. The small aggregate risk in my model for mature fund classes stems from a large fraction of indexed assets. This is consistent with Cremers and Petajisto (2009), who show an increase by 30% for the fraction of closet indexers among equity mutual fund managers since 1980.

The competition among managers and the availability of profitable opportunities relate the returns to aggregate scale with the curvature of the flow-performance relation. Fund classes with limited investment opportunities should have decreasing returns to aggregate scale and convex flows. Equity mutual funds are an asset class that fits this description. The literature on the convexity of flows within mutual funds is extensive.[8] In addition, Christoffersen (2001) documents convex flows in money market funds, and Kacperczyk and Schnabl (2013) show that investor flows are highly responsive to money market fund yields. These fund classes should have diminishing returns to aggregate scale, and Pástor et al. (2015) verify this for equity mutual funds. However, fund classes that have sufficient investment opportunities should have increasing returns to aggregate scale and concave flows. Goldstein et al. (2017), Kaplan and Schoar (2005), and Getmansky (2012) show concave flows among corporate bond funds, private equity, and hedge funds respectively. My model predicts that these fund classes should have increasing returns to aggregate scale.

The mechanism for the life cycle of a fund class allows the creation of "mega funds", namely funds with significantly larger assets under management than rivals. The most successful managers throughout the life cycle will be those who entered the fund class early with a talent level that is deep in the right tail of the prior distribution $H(\tau_i)$. These managers receive the largest inflows over time (Lemma 5), and their long tenure allows them to amass large amounts of capital. The existence of mega funds does not affect the final number of funds for the class. This is another indication that the depletion of alpha within the fund class is the root of the diminishing returns to aggregate scale, rather than illiquidity of large funds combined with their own diminishing returns.

My model has different implications than Garleanu and Pedersen (2016) on the final outcome of a fund class and the asset allocation of large investors. They predict that the informational inefficiencies are eliminated in the asymptotic limit, and the fund class will cease to exist. My model shows instead that a mature fund class is still valuable to investors, even in the absence of opportunities for alpha. The number of funds is stabilized to a finite value by the end of the life cycle. Large investors in Garleanu and Pedersen (2016) have a comparative advantage in active

[8] Ippolito (1992), Chevalier and Ellison (1997), Sirri and Tufano (1998), Lynch and Musto (2003), Huang et al. (2007), and Brown and Wu (2015) among others. See also the literature survey by Christoffersen et al. (2014).

investing, because their searching costs are a smaller fraction of their total investment compared to smaller investors. My model suggests that large investors allocate predominantly in mature classes to mitigate the risk of active investing.[9]

V. Potential extensions to the model

My model aims to explain the long-term regularities for the life cycle of a fund class. I assume an increasing level of competition, and show that the cross-sectional dispersion of talent declines over time. Throughout the analysis, I ignore performance fees and I assume an incubation constraint for entry. Certain fund classes like hedge funds have non-linear fee structures due to performance fees. However, the performance fee by itself does not fully explain the fund's performance (Agarwal et al., 2009). The high-water mark and hurdle rates mitigate potential risk-shifting behavior from performance fees (Goetzmann et al., 2003). In addition, indirect incentives from investor flows are at least 1.4 times larger than direct incentives from performance fees (Lim et al., 2016). As a result, considering only a simple fee structure in the model is not a significant restriction. However, the inclusion of performance fees could extend the interplay between the investor and managers to cases of moral hazard. For instance, models of decentralized investment management that consider hierarchical structures for the institutional investors show a loss in diversification (Van Binsbergen et al., 2008).

More importantly, the model lacks a mechanism to increase $\hat{\sigma}_\tau(t)$ by means of creating endogenously new investment opportunities. A potential extension to my model could include business cycle disturbances that may arise during the life cycle. These disturbances could be spawned either by technological innovation or investors who create opportunities for profit by injecting large amounts of capital to new firms in the real economy. The impact on the life cycle for a fund class could involve an increase in the cross-sectional dispersion of talent, reflecting a surge in profitable

[9]For instance, large investors such as Calpers and GE pension funds report historically an asset allocation that is roughly 88% on traditional fund classes (equity, debt, and real assets). On the other hand, smaller institutional investors such as university endowments tend to have a more balanced asset allocation, with roughly equal fractions of total assets among traditional and alternative fund classes according to annual reports from NACUBO.

opportunities. Eventually, these opportunities should decline in the long term, as active managers exploit them over time.

Another exogenous assumption is the ever-growing population of incumbent managers. Empirical trends for multiple fund classes seem to support this assumption. The only exception are hedge funds. This class started to grow after the dot-com crisis, when institutional investors began reallocating capital from equity mutual funds (Fung and Hsieh, 2012). However, many hedge funds have liquidated in the aftermath of the financial crisis of 2008, and the total number of hedge funds has been decreasing during the period 2012-2016. A potential extension to my model could derive the number of incumbent managers endogenously and study its interplay with the availability of investment opportunities.

VI. Conclusion

This paper discusses the life cycle of investment management. Fund managers compete for investor flows and profitable opportunities, and may either outperform or underperform relatively to the class benchmark. The benchmark is determined endogenously by a threshold level of talent that is required for entry. The markets learn about managerial talent through the innovations from the fund's track record. The equilibrium allocations are efficient, and the implicit contract among the investors and managers is first-best.

The driving forces for the life cycle are the competition among funds and the dispersion of managerial talent in active investing. The evolution of the life cycle resembles the concept of "today's alpha is tomorrow's beta". Alpha is abundant during the early stages of the life cycle, but it is eventually competed away by the managers. As a result, performance is not persistent. A fund class that initially provides opportunities for alpha, eventually becomes a platform for relatively safe investments of large pools of capital with returns similar to a passive index.

During the early stages of the life cycle, the competition among managers triggers a network externality for the investors and the aggregate demand increases. The fund class grows and oper-

ates under increasing returns to aggregate scale. The flow-performance relation is concave. The investor surplus from alpha increases during this stage. As the competition among managers becomes more intense, the opportunities for abnormal returns are curtailed and the fund class reverses to diminishing returns to scale. The flow-performance relation becomes convex, and the total surplus from active investing is depleted by the end of the life cycle.

The model provides an explanation for the coexistence of diminishing returns to aggregate scale with convex flows among equity mutual funds. It also predicts that early data on mutual funds or other data on fund classes with plenty of profitable opportunities should exhibit increasing returns to scale and concave flows. In addition, the model shows that the empirical measurement of alpha in mature classes such as equity mutual funds is challenging, because alpha is practically diminished during the latest stages of the life cycle. This result holds irrespectively of the quality of empirical risk factors and benchmark returns that are used to estimate alphas empirically.

Appendix

A. Derivations and proofs

The following auxiliary variables

$$A_t = \sum_{j=0}^{N_t} \frac{1}{b_1 c + E_t[R_{jt+1}]\hat{\sigma}_\tau(t)} \tag{A.1}$$

$$B_t = \sum_{j=0}^{N_t} \frac{E_t[R_{jt+1}]}{b_1 c + E_t[R_{jt+1}]\hat{\sigma}_\tau(t)} \tag{A.2}$$

$$D_t = 2b_1(N_t + 1) - 2b_1 b_2 c A_t - b_2 B_t \hat{\sigma}_\tau(t) > 0 \tag{A.3}$$

are used extensively in the proofs below. Equation (A.3) is positive for all values of $\hat{\sigma}_\tau(t) \geqslant 0$, since it has no positive roots and its asymptotic values are

$$\lim_{\hat{\sigma}_\tau(t)\to 0} D_t = 2(N_t + 1)(b_1 - b_2) > 0 \tag{A.4}$$

$$\lim_{\hat{\sigma}_\tau(t)\to\infty} D_t = N_t + 1 > 0 \tag{A.5}$$

Another useful inequality is

$$b_0 > b_2 \sum_{j=0}^{N_t} E_t[R_{jt+1}] \tag{A.6}$$

which stems from the restriction that the sum of equilibrium fees for large values of $\hat{\sigma}_\tau(t)$ must be a positive number.

Proof of LEMMA 1. To simplify the notation for this proof, I omit the time subscript t from vectors. The Lagrangian for the investor's problem (see equation (4)) is

$$\mathcal{L} = \boldsymbol{E_t}[\boldsymbol{q'}(\boldsymbol{r} - \boldsymbol{f})] - \frac{a}{2}\, \boldsymbol{q'V q} - \lambda(\boldsymbol{q'1} - W_t) \tag{A.7}$$

where $\boldsymbol{q} = (q_{0t}, q_{1t}, \ldots, q_{Nt})'$ is the fund size vector, $\boldsymbol{r} = (r_{0t+1}, r_{1t+1}, \ldots, r_{Nt+1})'$ is the fund nominal return vector, $\boldsymbol{f} = (f_{0t}, f_{1t}, \ldots, f_{Nt})'$ is the fund fee vector, $\mathbf{1}$ is a $(N+1) \times 1$ vector of ones, and $\boldsymbol{V}$ is the $(N+1) \times (N+1)$ covariance matrix for the returns of the passive index and the incumbent active managers within the fund class. The first-order condition for $\boldsymbol{q'}$ is

$$\boldsymbol{E_t}[\boldsymbol{r} - \boldsymbol{f}] - a\boldsymbol{V q} = \lambda \mathbf{1} \tag{A.8}$$

and the optimal fund sizes are given by

$$\boldsymbol{q^*} = \frac{1}{a}\, \boldsymbol{V^{-1}}\big[\boldsymbol{E_t}[\boldsymbol{r} - \boldsymbol{f}] - \lambda \mathbf{1}\big] = \frac{1}{a}\, \big[\boldsymbol{E_t}[\boldsymbol{R} - \boldsymbol{f}] - \lambda \boldsymbol{V^{-1}}\mathbf{1}\big] \tag{A.9}$$

with $\boldsymbol{R} = (0, R_{1t+1}, \ldots, R_{Nt+1})$ the vector of risk-adjusted fund returns in excess of the benchmark, i.e. the fund gross alphas. Notice that the passive index has zero alpha by definition. Multi-

plying equation (A.9) by $\mathbf{1'}$ allows to reconstruct the budget constraint and solve for λ

$$\mathbf{1'q^*} = \frac{1}{a}\left[\mathbf{1'E_t[R-f]} - \lambda\left(\mathbf{1'V^{-1}1}\right)\right] = W_t \Rightarrow \tag{A.10}$$

$$\lambda = \frac{\mathbf{1'E_t[r-f]}}{(\mathbf{1'V^{-1}1})} - \frac{aW_t}{(\mathbf{1'V^{-1}1})} \tag{A.11}$$

Substituting for λ in equation (A.9) yields the equilibrium demand function

$$\boldsymbol{q^*} = \frac{W_t}{(\mathbf{1'V^{-1}1})}\mathbf{V^{-1}1} + \frac{1}{a}\,\boldsymbol{E_t[R-f]} - \frac{\mathbf{1'E_t[R-f]}}{a\,(\mathbf{1'V^{-1}1})}\mathbf{V^{-1}1} \tag{A.12}$$

As a result, the demand function for fund i at time t is given by

$$q^*_{it} = \frac{W_t \sum\limits_{j=0}^{N_t}\omega_{ij}}{\sum\limits_{k=0}^{N_t}\sum\limits_{j\geqslant k}^{N_t}\omega_{kj}} + \frac{1}{a}E_t[R_{it+1}-f_{it}] - \frac{\sum\limits_{j=0}^{N_t}\omega_{ij}}{a\sum\limits_{k=0}^{N_t}\sum\limits_{j\geqslant k}^{N_t}\omega_{kj}}\sum_{j=0}^{N_t}E_t[R_{jt+1}-f_{jt}] \tag{A.13}$$

where ω_{ij} the matrix element of $\mathbf{V^{-1}}$ at row i and column j. The following ratio has order of magnitude

$$\frac{\sum\limits_{j=0}^{N_t}\omega_{ij}}{\sum\limits_{k=0}^{N_t}\sum\limits_{j\geqslant k}^{N_t}\omega_{kj}} = O\left(\frac{1}{N_t+1}\right) \tag{A.14}$$

implying that the optimal fund size may be written as

$$q^*_{it} = \frac{b_0}{N_t+1} + b_1E_t[R_{it+1}-f_{it}] - \frac{b_2}{N_t+1}\sum_{j=0}^{N_t}E_t[R_{jt+1}-f_{jt}] \tag{A.15}$$

with $b_2 < b_1$.

□

Proof of LEMMA 2. The correlations of the number of incumbents with fund fees, fund size, and aggregate size are

$$\frac{\partial f_{it}^*}{\partial N_t} = -\frac{b_1 X(2b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t))}{D_t^2(b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t))} < 0 \tag{A.16}$$

$$\frac{\partial q_{it}^*}{\partial N_t} = -\frac{b_1^2 E_t[R_{it+1}] X\hat{\sigma}_\tau(t)}{D_t^2(b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t))} \leqslant 0 \tag{A.17}$$

$$\frac{\partial Q_t^*}{\partial N_t} = \frac{b_1 X}{D_t^2}\Big(2b_1 + 2b_1(b_1 - b_2)cA_t + (b_1 - b_2)B_t\hat{\sigma}_\tau(t)\Big) > 0 \tag{A.18}$$

where A_t, B_t, and D_t are defined in equations (A.1) to (A.3) and

$$X \equiv 2b_0 + b_2\Big(- 2\sum_{j=0}^{N_t} E_t[R_{jt+1}] + hA_t\hat{\sigma}_\tau(t) + (2b_1 c + \hat{\sigma}_\tau(t))B_t\Big) > 0 \tag{A.19}$$

because $b_0 > b_2 \sum_{j=0}^{N_t} E_t[R_{jt+1}]$ from equation (A.6). As a result, equations (A.16) and (A.17) are negative and non-positive respectively. Equation (A.17) is equal to zero only when all the profitable opportunities are depleted for $\hat{\sigma}_\tau(t) = 0$. However, equation (A.18) is positive.

The monotonicity of the flow-performance relation is described by

$$\frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} \propto \frac{\hat{\sigma}_\tau(t)}{2D_t^2(N_t + 1)(b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t))^3} \tag{A.20}$$

where D_t is defined in equation (A.3). This implies that investor flows decline to zero as the investment opportunities within the fund class deplete. Equation (A.20) has no root for positive values of $\hat{\sigma}_\tau(t)$, and its limit for large values of talent dispersion is

$$\lim_{\hat{\sigma}_\tau(t)\to\infty}\left[\frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]}\right] = \frac{b_1\left(2E_t[R_{it+1}]^2 + h\right)}{2E_t[R_{it+1}]^2} > 0 \tag{A.21}$$

As a result,

$$\frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} > 0 \quad \text{for} \quad \hat{\sigma}_\tau(t) > 0 \tag{A.22}$$

which implies a monotonically increasing flow-performance relation in the presence of opportunities for alpha within the fund class.

□

Proof of LEMMA 3. The first-order condition from equation (11) for fund i at time t is

$$f_{it}^{*} q_{it}^{*} (E_t[R_{it+1}] - 1) - \frac{2c x_{it} (q_{it}^{*})^2}{\hat{\sigma}_\tau(t)} = 0 \tag{A.23}$$

which implies a fraction of assets under management

$$x_{it}^{*} = \frac{f_{it}^{*}(E_t[R_{it+1}] - 1)\hat{\sigma}_\tau(t)}{2c q_{it}^{*}} = \frac{f_{it}^{*}(\tau_i - \tau_{_B}(t))\hat{\sigma}_\tau(t)}{2c q_{it}^{*}} \tag{A.24}$$

that is actively invested in equilibrium.

□

Proof of LEMMA 4. Lemma 2 shows that managers with larger positive alphas are assigned more capital. The class benchmark is defined in equation (1) as the weighted average talent among the incumbent managers, with fund assets under management as weights. Managers with negative alphas lose capital, while managers with positive alphas gain capital. This implies that the investor flows gradually decrease the weights in equation (1) from the left tail of the talent distribution within the fund class. On the other hand, the flows increase the weights on the right tail of the distribution over time. As a result, the weighted average talent shifts upward and

$$\frac{d\tau_{_B}(t)}{dt} > 0 \tag{A.25}$$

□

Proof of LEMMA 5. Let $t_1 < t_2$. The variance of the prior for every manager is larger for the early entry cohort at t_1 than the subsequent cohort at t_2 (Lemma 4). This implies a larger uncertainty in the talent estimates $\hat{\tau}_{it}$ for the funds i in the early cohort, which translates to larger flows during

the learning process. For instance, the flow at time $t_1 + 1$ for a fund i in the early cohort is larger on average than the corresponding flow at time $t_2 + 1$ to a fund j of the later cohort, even if the managers have identical talent and realized returns.

□

Proof of PROPOSITION 1. In the long run, the incumbents with talent $\tau_i < \tau_{_B}(t)$ are forced to exit by liquidation. In addition, the prior distribution for potential entrants is truncated within the range $\tau_i \in [\tau_{_B}(t), \infty)$. Since the lower bound increases over time (Lemma 4), the cross-sectional dispersion $\hat{\sigma}_\tau(t)$ among the incumbent managers must decline, i.e.

$$\frac{d\hat{\sigma}_\tau(t)}{dt} < 0 \tag{A.26}$$

The decrease in $\hat{\sigma}_\tau(t)$ implies the diminishing of the opportunities for alpha within the class. In addition, all the incumbents have progressively similar levels of talent. The homogeneity in talent impedes managers from outperforming the benchmark, since the difference $\tau_i - \tau_{_B}(t)$ is near zero for every fund i. As a result, the opportunities for alpha within the class become more elusive.

The evolution of fund fees for all incumbents i over time is given by

$$\frac{df^*_{it}}{dt} = \frac{\partial f^*_{it}}{\partial N_t}\frac{dN_t}{dt} + \frac{\partial f^*_{it}}{\partial \hat{\sigma}_\tau(t)}\frac{d\hat{\sigma}_\tau(t)}{dt} \tag{A.27}$$

The first term is negative, because N_t increases over time and $\partial f^*_{kt}/\partial N_t$ is negative (Lemma 2). The second term is positive, because $\hat{\sigma}_\tau(t)$ declines over time from equation (A.26) and

$$\lim_{\hat{\sigma}_\tau(t)\to\infty}\left[\frac{\partial f^*_{it}}{\partial \hat{\sigma}_\tau(t)}\right] = 0 \tag{A.28}$$

$$\lim_{\hat{\sigma}_\tau(t)\to 0}\left[\frac{\partial f^*_{it}}{\partial \hat{\sigma}_\tau(t)}\right] \approx \frac{-b_0 E_t[R_{it+1}] - (N_t+1)(b_1-b_2)(E_t[R_{it+1}]-h)}{2c(N_t+1)(b_1-b_2)^2} < 0 \tag{A.29}$$

without a root for positive values of talent dispersion. As a result, the sign of equation (A.27) is indeterminate in general. However, as the average profitability for a fund decreases from the

shrinking opportunity set, the derivative $\partial f_{it}^{*}/\partial\hat{\sigma}_{\tau}(t)$ becomes positive above a threshold value for N_t. This threshold is given by

$$\begin{aligned}\ln N_{thr} = \frac{1}{b_1[b_1 E_t[R_{it+1}] + b_2(E_t[R_{it+1}]-1)]}\big(&-b_1b_2 - b_0b_1b_2 - b_1^3h - b_1^2b_2h \\ &+ b_0b_1^2E_t[R_{it+1}] - b_1b_2E_t[R_{it+1}] + b_0b_1b_2E_t[R_{it+1}] - b_2^2E_t[R_{it+1}] \\ &+ b_1^2E_t[R_{it+1}]^2 + 2b_1b_2E_t[R_{it+1}]^2 + b_2^2E_t[R_{it+1}]^2\big)\end{aligned} \tag{A.30}$$

Since N_t increases over time, the threshold N_{thr} will eventually be surpassed. As a result, both terms in equation (A.27) will be negative, implying that the fund fees decline for all incumbents in the course of time.

□

Proof of LEMMA 6. The roots of equation

$$\frac{\partial f_{it}^{*}}{\partial E_t[R_{it+1}]} = 0 \tag{A.31}$$

are

$$\begin{aligned}\sigma_{thr} = b_1c\Bigg\{&(N_t+1)b_1 - b_0 + b_2\left(\sum_{j=0}^{N_t}E_t[R_{jt+1}] - (N_t+1)\right) + \\ &\Bigg[4h(N_t+1)^2(2b_1-b_2)(b_1-b_2) + \left(b_2\left(\sum_{j=0}^{N_t}E_t[R_{jt+1}] - (N_t+1)\right) + \right. \\ &\left.b_1(N_t+1) - b_0\right)^2\Bigg]^{1/2}\Bigg\}\cdot\Big[h(N_t+1)(2b_1-b_2)\Big]^{-1}\end{aligned} \tag{A.32}$$

$$\begin{aligned}\sigma_1 = b_1c\Bigg\{&(N_t+1)b_1 - b_0 + b_2\left(\sum_{j=0}^{N_t}E_t[R_{jt+1}] - (N_t+1)\right) - \\ &\Bigg[4h(N_t+1)^2(2b_1-b_2)(b_1-b_2) + \left(b_2\left(\sum_{j=0}^{N_t}E_t[R_{jt+1}] - (N_t+1)\right) + \right.\end{aligned}$$

$$b_1(N_t+1)-b_0\Big)^2\Big]^{1/2}\Big\}\cdot\Big[h(N_t+1)(2b_1-b_2)\Big]^{-1} \tag{A.33}$$

$$\sigma_2=-\frac{b_1 c}{E_t[R_{it+1}]}\ ,\ \forall i \tag{A.34}$$

From these roots, σ_{thr} is the only positive one. From equation (A.6) the derivative of σ_{thr} is

$$\frac{d\sigma_{thr}}{dN_t}\propto b_1 c\left(b_0-b_2\sum_{j=0}^{N_t}E_t[R_{jt+1}]\right)\sigma_{thr}>0 \tag{A.35}$$

implying that the threshold cross-sectional variance in talent increases with N_t. In addition,

$$\frac{\partial Q_t^*}{\partial E_t[R_{it+1}]}=-\frac{2b_1((N_t+1)b_1-N_t b_2)(b_1 c+E_t[R_{it+1}]\hat{\sigma}_\tau(t))}{b_2(2b_1 c+E_t[R_{it+1}]\hat{\sigma}_\tau(t))+(b_1 c+E_t[R_{it+1}]\hat{\sigma}_\tau(t))D_t}\frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} \tag{A.36}$$

$$\begin{aligned}\frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2}=&-\hat{\sigma}_\tau(t)\Big(b_1 c(D_t-2(N_t+1)(1-b_1))+(D_t-2(N_t+1)(1-b_1)\\&-b_2)E_t[R_{it+1}]\hat{\sigma}_\tau(t)\Big)\cdot\Big(b_1 c(b2+D_t-2(N_t+1)(1-b_1))+(D_t\\&-2(N_t+1)(1-b_1))E_t[R_{it+1}]\hat{\sigma}_\tau(t)\Big)\cdot\Big[D_t(b_1+N_t(b_1-b_2))\\&(b_1 c+E_t[R_{it+1}]\hat{\sigma}_\tau(t))^3\Big]^{-1}\frac{\partial Q_t^*}{\partial E_t[R_{it+1}]}\\\approx&-\frac{\hat{\sigma}_\tau(t)\Big(D_t-2(N_t+1)(1-b_1)\Big)^2}{b_1 D_t(N_t+1)(b_1 c+E_t[R_{it+1}]\hat{\sigma}_\tau(t))}\cdot\frac{\partial Q_t^*}{\partial E_t[R_{it+1}]}\end{aligned} \tag{A.37}$$

where D_t is defined in equation (A.3). As a result, the economies of aggregate scale and the curvature of the flow-performance relation are zero when $\hat{\sigma}_\tau(t)=\sigma_{thr}$.

□

Proof of LEMMA 7. The evolution of aggregate demand is given by

$$\frac{dQ_t^*}{dt}=\frac{\partial Q_t^*}{\partial N_t}\frac{dN_t}{dt}+\frac{\partial Q_t^*}{\partial\hat{\sigma}_\tau(t)}\frac{d\hat{\sigma}_\tau(t)}{dt} \tag{A.38}$$

The first product is positive, because N_t increases over time and Q_t^* is positively correlated with N_t

(Lemma 2). In addition, $\hat{\sigma}_\tau(t)$ is decreasing (Proposition 1). The correlation between Q_t^* and $\hat{\sigma}_\tau(t)$ is positive for all $\hat{\sigma}_\tau(t) > 0$ values. As a result, the second term in equation (A.38) is negative, implying in general that the sign of that equation is indeterminate. However, the asymptotic limits for $\partial Q_t^*/\partial\hat{\sigma}_\tau(t)$ are

$$\lim_{\hat{\sigma}_\tau(t)\to\infty}\left[\frac{\partial Q_t^*}{\partial\hat{\sigma}_\tau(t)}\right] = 0 \tag{A.39}$$

$$\lim_{\hat{\sigma}_\tau(t)\to 0}\left[\frac{\partial Q_t^*}{\partial\hat{\sigma}_\tau(t)}\right] \approx \frac{b_1 + N_t(b_1 - b_2)}{2c(N_t+1)(b_1-b_2)^2} > 0 \tag{A.40}$$

As a result, only the first term in equation (A.38) describes the evolution of aggregate demand during the early stages of the life cycle, implying that the aggregate demand increases over time.

□

Proof of PROPOSITION 2. Lemma 6 shows that σ_{thr} is the unique root for three key comparative statics. These are the sensitivity of a fund's fee to its performance, the returns to aggregate scale, and the concavity of the flow-performance relation. During the early stage of the life cycle $\hat{\sigma}_\tau(t) > \sigma_{thr}$. The sensitivity of fees to performance is equal to

$$\begin{aligned}\frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = &- \frac{b_1 c(2b_2 + D_t) + (b_2 + D_t)E_t[R_{it+1}]\hat{\sigma}_\tau(t)}{2D_t^2\Big(b_1 c + E_t[R_{it+1}]\hat{\sigma}_\tau(t)\Big)^3} \cdot \\ &\Big[4b_1^3c^2(N_t+1) + 2b_1c(b_0 - b_2SR - D_t)\hat{\sigma}_\tau(t) + hD_t\hat{\sigma}_\tau(t)^2 + \\ &2b_1^2c\Big((N_t+1)\hat{\sigma}_\tau(t) - 2cD_t\Big)\Big]\end{aligned} \tag{A.41}$$

where SR is the sum of fund returns with $b_0 > b_2SR$ from equation (A.6), and D_t is defined in equation (A.3). When $\hat{\sigma}_\tau(t) > \sigma_{thr}$ the term $hD_t\hat{\sigma}_\tau(t)^2$ in equation (A.41) dominates all other terms, yielding a negative correlation between fees and performance. The asymptotic value for this correlation is

$$\lim_{\hat{\sigma}_\tau(t)\to\infty}\frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = -\frac{(2(N_t+1)b_1 - N_t b_2)h}{2(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^2} < 0 \tag{A.42}$$

Equation (A.36) in the proof of Lemma 6 shows that the economies of aggregate scale have opposite sign to equation (A.41). Equation (A.37) in the same proof shows that the curvature of the flow-performance relation has opposite sign to the economies of aggregate scale, i.e. the same sign as equation (A.41). As a result, the fund class operates under increasing returns to aggregate scale and the flow-performance relation is concave during the early stages of the life cycle. The signs for the two comparative statics of interest are also verified by the following asymptotic values

$$\lim_{\hat{\sigma}_\tau(t)\to\infty} \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = \frac{b_1((N_t+1)b_1 - N_t b_2)h}{(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^2} > 0 \tag{A.43}$$

$$\lim_{\hat{\sigma}_\tau(t)\to\infty} \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} = -\frac{b_1(2(N_t+1)b_1 - (N_t+2)b_2)h}{2(N_t+1)(2b_1-b_2)E_t[R_{it+1}]^3} < 0 \tag{A.44}$$

for all funds i. As a result, the returns to aggregate scale are increasing and the flow-performance is concave within the range $\hat{\sigma}_\tau(t) \in (\sigma_{thr}, \infty)$.

The total surplus from investment of fund i at time t is given by equation (10). Its limit for large values of $\hat{\sigma}_\tau(t)$ is

$$\begin{aligned}\lim_{\hat{\sigma}_\tau(t)\to\infty} TS_{it} = {} & \frac{q_0(E_t[R_{it+1}]-h)}{N_t \prod_{j=1}^{N_t} E_t[R_{jt+1}]} \\ & \Bigg\{ \Big[E_t[R_{it+1}](b_0 - \ln N_t + E_t[R_{it+1}]) - b_1 h\Big] N_t \prod_{j\neq i} E_t[R_{jt+1}] \\ & - b_2 h \left[\prod_{j\neq i} E_t[R_{jt+1}] + \sum_{j=1}^{N_t-2} \prod_{k=j}^{N_t-3} E_t[R_{kt+1}] \cdot \left(\prod_{m=k+2}^{N_t} E_t[R_{mt+1}] \right) \right. \\ & \left. + \prod_{j=1}^{N_t-2} E_t[R_{jt+1}] \cdot (E_t[R_{N_t-1,t+1}] + E_t[R_{N_t,t+1}]) \right] \Bigg\} > 0 \end{aligned} \tag{A.45}$$

because $E_t[R_{it+1}] > h$ for all funds i, otherwise the expected profits would be negative. Notice that the total surplus plateaus to a constant for large $\hat{\sigma}_\tau(t)$.

The surplus MS_{it} that manager i extracts at time t during the early stages of the life cycle is

approximately linear in $\hat{\sigma}_\tau(t)$, because

$$\lim_{\hat{\sigma}_\tau(t)\to\infty} MS_{it} = \frac{h^2(E_t[R_{it+1}]-1)^2\hat{\sigma}_\tau(t)}{4cE_t[R_{it+1}]^2} \tag{A.46}$$

The remaining surplus to the investor from fund i at time t is $IS_{it} = TS_{it} - MS_{it}$. As a result, the investor surplus increases as more opportunities are exploited over time and $\hat{\sigma}_\tau(t)$ decreases. The investor surplus may be larger or smaller than the manager surplus at the onset of the life cycle. However, IS_{it} increases over time, while MS_{it} decreases. As a result, the investor surplus will inevitably surpass the manager surplus.

□

Proof of PROPOSITION 3. Lemma 6 shows that σ_{thr} is a root both for the returns to aggregate scale and the concavity of the flow-performance relation. During the late stage of the life cycle $\hat{\sigma}_\tau(t) < \sigma_{thr}$. Proposition 2 shows that

$$\frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} > 0 \quad \text{and} \quad \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} < 0 \tag{A.47}$$

for all funds i and $\hat{\sigma}_\tau(t) > \sigma_{thr}$. As a result, each of these correlations has the opposite sign within the range $\hat{\sigma}_\tau(t) \in (0, \sigma_{thr})$. Therefore, the returns to aggregate scale are decreasing and the flow-performance is convex during the late stages of the life cycle.

The equilibrium demand for an active manager is zero when the profitable opportunities within the fund class are depleted. As a result, the total surplus from active investing in equation (10) also declines to zero. However, the bulk of the assets under management for every fund is invested passively, as shown in equation (12). In the limit where $\hat{\sigma}_\tau(t) \to 0$, the manager invests only in passive strategies. Since every manager performs closet indexing, the aggregate risk within the fund class is reduced as the profitable opportunities diminish.

□

References

Agarwal, Vikas, Naveen D. Daniel, and Narayan Y. Naik, 2009, Role of managerial incentives and discretion in hedge fund performance, *Journal of Finance* 64, 2221–2256.

Berk, Jonathan B., and Richard C. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295.

Brown, David P., and Youchang Wu, 2015, Mutual fund flows and cross-fund learning within families, *Journal of Finance* in press.

Chevalier, Judith, and Glenn Ellison, 1997, Risk taking by mutual funds as a response to incentives, *Journal of Political Economy* 105, 1167–1200.

Cho, Thummim, 2017, Turning alphas into betas: Arbitrage and endogenous risk, *Working paper* .

Christoffersen, Susan E. K., 2001, Why do money fund managers voluntarily waive their fees?, *Journal of Finance* 56, 1117–1140.

Christoffersen, Susan E. K., David K. Musto, and Russ Wermers, 2014, Investor flows to asset managers: Causes and consequences, *Annual Review of Financial Economics* 6, 289–310.

Cremers, K. J. Martijn, and Antti Petajisto, 2009, How active is your fund manager? a new measure that predicts performance, *Review of Financial Studies* 22, 3329–3365.

Fama, Eugene F., and Kenneth R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance* 65, 1915–1947.

Fung, William, and David A. Hsieh, 2012, Hedge funds, in George M. Constantinides, Milton Harris, and René M. Stulz, eds., *Handbook of the Economics of Finance*, volume 2B, chapter 16, 1063–1126 (North-Holland).

Garleanu, Nicolae, and Lasse H. Pedersen, 2016, Efficiently inefficient markets for assets and asset management, *NBER Working Paper No. w21563* .

Getmansky, M., 2012, The life cycle of hedge funds: Fund flows, size, competition, and performance, *Quarterly Journal of Finance* 02, 1250003–1–1250003–53.

Glode, Vincent, 2011, Why mutual funds "nderperform", *Journal of Financial Economics* 99, 546–559.

Glode, Vincent, and Richard C. Green, 2011, Information spillovers and performance persistence for hedge funds, *Journal of Financial Economics* 101, 1–17.

Goetzmann, William N., Jonathan E. Ingersoll, and Stephen A. Ross, 2003, High-water marks and hedge fund management contracts, *Journal of Finance* 58, 1685–1718.

Goldstein, Itay, Hao Jiang, and David T. Ng, 2017, Investor flows and fragility in corporate bond funds, *Journal of Financial Economics* in press.

Huang, Jennifer, Kelsey D. Wei, and Hong Yan, 2007, Participation costs and the sensitivity of fund flows to past performance, *Journal of Finance* 62, 1273–1311.

Ippolito, Richard A., 1992, Consumer reaction to measures of poor quality: Evidence from the mutual fund industry, *Journal of Law and Economics* 35, 45–70.

Kacperczyk, Marcin, and Philipp Schnabl, 2013, How safe are money market funds?, *Quarterly Journal of Economics* 128, 1073–1122.

Kaplan, S. N., and A. Schoar, 2005, Private equity performance: Returns, persistence, and capital flows, *Journal of Finance* 60, 1791–1823.

Katz, Michael L., and Carl Shapiro, 1985, Network externalities, competition, and compatibility, *American Economic Review* 75, 424–440.

Lim, Jongha, Berk A. Sensoy, and Michael S. Weisbach, 2016, Indirect incentives of hedge fund managers, *Journal of Finance* 71, 871–918.

Lynch, Anthony W., and David K. Musto, 2003, How investors interpret past fund returns, *Journal of Finance* 58, 2033–2058.

Pástor, Ľuboš, and Robert F. Stambaugh, 2012, On the size of the active management industry, *Journal of Political Economy* 120, 740–781.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23–45.

Schumpeter, Joseph, 1942, *Capitalism, Socialism and Democracy* (Harper and Brothers).

Sirri, Erik R., and Peter Tufano, 1998, Costly search and mutual fund flows, *Journal of Finance* 53, 1589–1622.

Van Binsbergen, Jules H., Michael W. Brandt, and Ralph S. J. Koijen, 2008, Optimal decentralized investment management, *Journal of Finance* 63, 1849–1895.