The life cycle of investment management when “today’s alpha is tomorrow’s beta”

Georgios Magkotsios*

Marshall School of Business, University of Southern California

September 12, 2017

Abstract

Fund entry and investor demand for asset management affect the performance and risk of the average active fund. When the funds operate under increasing (decreasing) returns to aggregate scale, the flow-performance relation is concave (convex). Active funds outperform passive benchmarks initially. As the profitable opportunities within a growing set of competing funds diminish, the total investment surplus declines to zero and most of the incumbent managers underperform relative to passive funds. The average returns from active investing are not persistent, and the aggregate risk is reduced through “closet indexing”. Eventually, all active funds transform together to a large and 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, Odilon Câmara, Arthur Korteweg, and other faculty at USC. I have received valuable feedback from Thierry Foucault, Raman Uppal, and the participants of the 2016 Trans-Atlantic Doctoral Conference in London, the 2017 FMA European meeting in Portugal, the 2017 EFMA meeting in Greece, and the 2017 FMA meeting in Boston. I am also very thankful to John Doukas for sponsoring the best PhD paper award at the EFMA meeting.
I. Introduction

A current debate in asset management is about the benefit of investors from active or passive strategies. The bulk of empirical research on this topic has focused on equity mutual funds because of their predominant scale. The average equity fund underperforms relative to passive indexes, but a minority of talented managers outperform their benchmark and receive excess new capital from investors.\(^1\) More recent evidence reveal a heterogeneity in the relations between scale, performance, and investor flows across various types of funds that is puzzling for existing theories.\(^2\) This paper provides a theoretical model of the competition among active managers that introduces a life cycle for asset management. The main results are a novel connection between economies of scale and the flow-performance relation across asset classes, and the conditions for successful active asset management in terms of benefit to investors.

The model’s key feature is that competition affects adversely the returns of every incumbent fund by depleting investment opportunities in the aggregate. When competition is moderate, the average active fund outperforms passive benchmarks. On the other hand, strong competition gradually erodes performance until the gross returns of the average active fund fall behind the returns of passive indexes. Active management in my model operates initially with increasing returns to aggregate scale, followed by decreasing returns as its life cycle matures. In addition, increasing returns to aggregate scale coincide with a concave flow-performance relation, while the flow-performance relation becomes convex during decreasing returns to aggregate scale.


\(^2\) The performance of the average equity fund declines with the fund’s own assets (Harvey and Liu, 2017) and the aggregate scale of all equity funds together (Pástor et al., 2015). Kaplan and Schoar (2005), Getmansky (2012), and Goldstein et al. (2017) show concave flow-performance relations for private equity, hedge funds, and corporate bond funds respectively. Magkotsios (2017) finds increasing returns to aggregate scale for these funds. The returns of index funds and ETFs are inert to variations in fund or aggregate scale.
Previous theoretical work has pondered on the relation between scale and performance. Berk and Green (2004) assume diminishing returns with fund size for a monopolist manager. The manager can extract all surplus from investment by optimally increasing his fee in equilibrium. Their model implies positive returns before fees and zero net-of-fee returns in excess of passive benchmarks. In reality, managers compete for investor capital and superior performance. With competition, it is not obvious that managers can extract all surplus. Pástor and Stambaugh (2012) and Feldman et al. (2016) assume that fund returns decline with aggregate scale to explain the persistence of poor track records for active managers. The competition among managers with asymmetric information in García and Vanden (2009) and Gârleanu and Pedersen (2016) generates diseconomies of aggregate scale in equilibrium. Diseconomies of aggregate scale arise in their models as the market becomes more efficient and attenuates the comparative advantage of informed managers. My model is different, because it also predicts increasing returns to aggregate scale for moderate competition. In addition, the unique link that I show between economies of scale and the sensitivity of investor flows to performance has not been documented before in the literature.

Competition in my model limits the opportunities for profitable investments in the aggregate. Every manager in my model enters with an exogenous level of talent in active investing that affects his abnormal returns (alpha). A manager’s alpha is his risk-adjusted return in excess of an exogenous benchmark. The main variables are the number of funds and the cross-sectional mean and dispersion of talent among the incumbents. These three variables together define a proxy for the availability of profitable investment opportunities to managers. For instance, opportunities are plentiful for a small group of managers with large dispersion of talent, because competition is moderate and the most talented managers can harvest large alphas. On the other hand, the opportunity set depletes faster across a large group of very talented managers with small dispersion of talent in the cross-section. Strong competition among similarly talented managers implies that they are gradually forced to crowd into similar trading strategies to exploit the remaining opportunities, resulting in poor alphas overall.
Initially, a small population of managers in the model creates a new market for active management and injects capital into unexploited investment opportunities. As the number of funds rises, the competition resembles Schumpeter’s “creative destruction” and forces the worst performing managers to exit. The aggregate demand and investor surplus from alpha surge, because the expected abnormal returns increase and fund fees decrease from competition. Therefore, the funds operate under increasing returns to aggregate scale during the early growth stage. The escalating competition gradually depletes the opportunities for alpha. Although the incumbent managers are on average more talented over time, they also become more homogeneous in their strategies and crowd into a diminishing opportunity set. Investing in the same direction altogether increases trading costs and results in more elusive alphas.

The rising trading costs hurt the returns of the average fund until it underperforms relative to passive funds. The total surplus from active investing declines and returns to aggregate scale change from increasing to decreasing, because the opportunities to earn alpha are limited. Fund managers are better off by indexing a fraction of their assets under management when competition in active investing is fierce. Over the life cycle of asset management, the funds transform from a risky investment vehicle that is rich in opportunities for alpha to a set of funds whose performance and risk are similar to those of passive funds. This is the concept of “today’s alpha is tomorrow’s beta”.

The flow-performance relation is monotonically increasing, assigning larger allocations to the most talented managers in equilibrium. This relation is concave early on the life cycle of asset management. 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 diversification across funds, aiming to exploit a broader set of opportunities for alpha while mitigating risk and transaction costs. The least talented managers underperform their peers by a wide margin when the dispersion of talent is large, triggering large capital outflows from them.

---

3This expression was coined by Andrew Lo. See also Cho (2017) for a costly arbitrage-based argument where alphas turn to betas.
During the late stages of the life cycle, the flow-performance relation becomes convex. Convexity indicates that flows are more sensitive to good performance and less sensitive to bad performance. The returns of most managers are little different than the benchmark return when the opportunities for alpha are scarce at the aggregate level. This implies less sensitive outflows for those who underperform the benchmark, because their losses are relatively small. However, the reward for managers who achieve positive alpha is disproportionately large, because it is more strenuous to outperform rivals when all managers are similar in talent.

II. The model

The model describes the competition among active managers for superior performance and investor flows of capital. A manager’s only alternatives to active investing are either indexing his assets under management or exiting. The markets make rational expectations, and there is no moral hazard or adverse selection.

A. Individual fund and aggregate features

Every potential entrant manager is endowed with an exogenous talent in active investing $\tau_i$ whose value cannot be modified after entry. Potential entrants draw their talent from a distribution $H(\tau_i)$ that is common to all entry cohorts over time. A manager’s talent reflects his ability in exploiting profitable opportunities that achieve abnormal returns in excess of an exogenous benchmark. The term “alpha” refers to the abnormal returns of a fund. 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 based on a stochastic learning process under symmetric information. 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.\(^4\)

Let \( q_{it} \) be the size of fund \( i \) at time \( t \), and \( Q_t \) the aggregate size of a total number \( N_t \) of competing funds. The fund’s gross risk-adjusted return from active investing that is realized at time \( t + 1 \) is

\[
R_{it+1} = 1 + \alpha_{it} + \epsilon_{it+1} \\
= 1 + \frac{\tau_i}{\bar{\tau}(t)} - R_b(t) + \epsilon_{it+1}, \quad i = 1, \ldots, N_t, \quad t \geq 0
\]

(1)

\[
\bar{\tau}(t) \equiv \frac{1}{Q_t} \sum_{i=1}^{N_t} q_{it} \tau_i ,
\]

(2)

where \( \bar{\tau}(t) \) is the size-weighted cross-sectional average of talent levels at time \( t \), and \( R_b(t) \) is the return of the benchmark portfolio. Examples for the benchmark portfolio include the underlying market or exogenous factor-mimicking portfolios. The noise terms \( \epsilon_{it+1} \) are jointly distributed over time and across managers with zero mean. These shocks reflect the component of luck in the realized return \( R_{it+1} \).

The term \( \alpha_{it} \equiv \frac{\tau_i}{\bar{\tau}(t)} - R_b(t) \) is the manager’s alpha, and it can be positive or negative. Equation (1) is the first example of the effect that competition and the cross-sectional distribution of managerial talent have on investment value. A very talented manager has the potential to identify and exploit profitable investment opportunities, and can attain large alphas when he competes against mediocre rivals. However, when the same manager competes against similarly talented rivals, then everyone has the potential to exploit the same opportunities for alpha and affects adversely the performance of others. Therefore, equation (1) implies that a manager’s return from active investing is affected by his talent relative to the environment that he competes, rather than his individual level of talent only.

The second example of the distribution’s effect on investment value is through the cost of active trading. A complimentary variable to \( \bar{\tau}(t) \) is \( \sigma_{\tau}(t) \), the size-weighted cross-sectional dispersion
of talent at time $t$ among the incumbents. The total cost function for manager $i$ is

$$C(q_{it}, \sigma_\tau(t)) = \frac{cN_t q_{it}^2}{2\sigma_\tau(t)} + h q_{it},$$  

(3)

where $c$ and $h$ are constants. The quadratic term describes the fraction of transaction costs that affect the manager’s profits. Transaction costs stem from bid-ask spreads that every manager pays to implement his strategies. The dispersion of talent $\sigma_\tau(t)$ and $N_t$ capture together the effect of correlated trading among a rising number of similarly talented competitors. A large dispersion $\sigma_\tau(t)$ among few managers reduces the transaction costs in total, because the heterogeneity of talent allows every manager to find and exploit his own investment ideas that do not necessarily correlate with strategies from rivals. On the other hand, a small dispersion $\sigma_\tau(t)$ combined with a large number of funds increases transaction costs. Homogeneity in talent by itself does not imply correlated trading, because highly talented managers can still identify profitable investments. However, correlated trading will emerge when a large number of similarly talented managers compete for a finite number of investment ideas. Even when all managers are very talented, they must use the same strategies to exploit those ideas.

The number of funds $N_t$, the weighted mean $\bar{\tau}(t)$, and the weighted dispersion $\sigma_\tau(t)$ define together a proxy for the remaining profitable investment opportunities among the competing managers. A small number of funds with large dispersion of talent implies moderate competition and a broad opportunity set. On the other hand, a large number of very talented and relatively homogeneous managers is related to a shortage of opportunities. When $N_t$ is large, then many managers are forced to crowd into similar strategies. As more funds pursue the same opportunities, the rising trading costs make these opportunities more elusive. Despite the fact that every manager is very talented in identifying opportunities, the resulting alphas are small because all rival funds trade in the same direction. This intuition is similar to that of Foster and Viswanathan (1996), who show that trading is less profitable when many agents chase the same information signal. Therefore, the combination of large $N_t$, large $\bar{\tau}(t)$, and small $\sigma_\tau(t)$ induces increased transaction costs from
correlated trading and reduces performance (see equations (3) and (2) respectively), implying a scarcity of profitable trades in total.

The linear term $hq_{it}$ in equation (3) expresses the fund’s operation costs. These costs may reflect unique features of the fund that are irrelevant to transaction costs, including the fund’s membership in a fund family, tax exposure, provision of financial advice services to investors, quality of account services, etc. For instance, Hortaçsu and Syverson (2004) attribute the existence of a large number of index funds partly to non-portfolio related features. In addition, the operation costs may involve the maintenance of a level of liquidity within the fund’s portfolio to insure against fire sales and liquidation of the fund. Therefore, this term is also an opportunity cost to the manager. In order to cover the costs of operation, he may have to forgo investment opportunities that could add value to the fund and increase its returns to investors. As a result, the managers are differentiated both by talent level and the opportunities for alpha that they pursue.

**B. The investor’s problem**

The investor has finite wealth and mean-variance preferences for the allocations of capital within his portfolio of funds. He supplies flows to each manager based on performance. At time $t$ the investor has resolved the uncertainty for the talent level of incumbents, but cannot observe the talent of managers entering this period. He rewards positive abnormal returns with fund inflows, and punishes negative abnormal returns with outflows. His utility is quadratic, implying that the funds are imperfect substitutes. The finite elasticity of substitution across managers expresses the investor’s tradeoff between superior talent and diversification. His goal is to exploit all profitable opportunities in the aggregate. 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 fund. 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:

$$\max_{q_t} U = E_t[q_t'(r_{t+1} - f_t)] - \frac{a}{2} q_t' V_t q_t - \frac{1}{2} q_t' \Lambda_t q_t \quad \text{s.t.} \quad q_t' 1 \leq W_t,$$

(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 passive index at time $t$. All vectors have $N_t + 1$ elements, where $q_t$, $f_t$, and $r_{t+1}$ are the vectors for fund sizes, fees, and nominal returns realized at $t + 1$. In addition, $1$ is a vector of ones, and $V_t$ is the $(N_t + 1) \times (N_t + 1)$ covariance matrix for the returns of the passive index and the incumbent active managers. The fund fee represents the price that the investor pays for active management. I do not make a distinction between management and performance fees.

The investor’s problem also includes a quadratic term for transaction costs

$$TC_t = \frac{1}{2} q_t' \Lambda_t q_t,$$

(5)

where $\Lambda_t$ is a symmetric positive-definite matrix. The diagonal elements of $\Lambda_t$ describe the fraction of transaction costs that affects the investor’s surplus. These elements and the quadratic terms in equation (3) for every fund capture the total transaction costs associated with active investing. The non-diagonal elements of $\Lambda_t$ reflect the loss in investment value from reallocating capital across funds. Since transaction costs increase with correlated trading, I assume that $\Lambda_t$ is proportional to the covariance matrix $V_t$.

Lemma 1 shows the solution to the investor’s problem.

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

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

(6)

---

5See Gärleanu et al. (2009) and Gärleanu and Pedersen (2013) for the same approximation in derivative pricing and portfolio optimization respectively.
where $E_t[R_{jt+1} - f_{jt}]$ is the expected net-of-fee alpha for fund $j$, $b_0$ is a scalar that encapsulates the investor’s total wealth $W_t$, $b_1$ and $b_{it} < b_1$ are related to the investor’s finite elasticity of substitution between two funds, and $b_{it}/(N_t+1)$ also encapsulates the return covariances between fund $i$ and rival managers.

The equilibrium demand for a fund $i$ depends on the fund’s own net-of-fee alpha, but also includes the average of net-of-fee alphas from every competing fund. Equation (6) 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. The finite elasticity of substitution across managers in coefficients $b_1$ and $b_{it}$ has a key role in the investor’s balance between the demand for superior talent and diversification. This tradeoff is also affected by risk, because $b_{it}$ includes the covariances for the returns of manager $i$ with his rivals. However, Lemma 1 implies that investor flows are always more sensitive to the fund’s own performance than the average performance among the competing funds.

C. The competition among managers

The fund managers participate in a monopolistic competition. They compete for investor capital and profitable investment opportunities. Each manager is differentiated from his rivals based on his individual talent in active investing, but cannot affect the aggregate indices of returns and fees. Monopolistic competition is a special case of differentiated Bertrand competition, and requires a relatively large number of funds. The next section includes a general Bertrand competition for two or more funds, and shows that the model’s predictions are consistent with monopolistic competition during the early growth stages of active management.

The manager of fund $i$ sets his fee $f_{it}$ at time $t$, and the investor determines the fund size $q_{it}$ through performance-based flows that satisfy his demand function. Monopolistic competition implies that the sum of returns $SR = \sum_j E_t[R_{jt+1}]$ and the sum of fees $\sum_j f_{jt}$ cannot be affected by perturbations to the expected return and fees respectively of a single fund. The innovations about
a single manager’s talent don’t affect the cross-sectional distribution of talent, including the average estimated value and its dispersion. Similarly, an arbitrary raise by a manager of his own fee has negligible effects on the average fee among the competitors. However, this action would impact investor flows for this manager, as Lemma 1 suggests.

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

\[
\max_{f_{it}} E_t[\Pi_{it+1}] = f_{it} q_{it} E_t[R_{it+1}] - \frac{c N_t q_{it}^2}{2 \sigma_r(t)} - h q_{it},
\]

(7)

where \( q_{it} \) is the investor demand that clears the market. The quadratic term in the manager’s cost function implies diminishing returns to scale at the fund level. The Nash equilibrium for equations (6) and (7) gives the fund’s response function

\[
f_{it}^* = \left[ b_0 b_1 c N_t + (N_t + 1) b_1 h \sigma_r(t) - b_1 b_1 c N_t \sum_{j=0}^{N_t} E_t \left[ R_{jt+1} - f_{jt}^* \right] + E_t[R_{it+1}] \left( b_1^2 c N_t (N_t + 1) + b_0 \sigma_r(t) + b_1 (N_t + 1) \sigma_r(t) - b_{it} \sigma_r(t) \sum_{j=0}^{N_t} E_t \left[ R_{jt+1} - f_{jt}^* \right] \right) \right] \cdot \left[ b_1 (N_t + 1) \left( b_1 c N_t + 2 E_t[R_{it+1}] \sigma_r(t) \right) \right]^{-1},
\]

(8)

given the approximation \( R_{it+1}^2 \approx R_{it+1} \) 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 gives the solution for the sum of fees. Then the fee for each fund can be retrieved by substitution to equation (8). The equilibrium fund size is given by equation (6).

The equilibrium aggregate size \( Q_t^* \) is the sum of the fund sizes for the \( N_t \) active managers.

The Herfindahl–Hirschman index \( HH_t \) is defined as

\[
HH_t \equiv \sum_{i=1}^{N_t} \left( \frac{q_{it}^*}{Q_t^*} \right)^2.
\]

(9)
This index is the average market share of a fund, weighted by the market share itself. It is an alternative measure for the intensity of competition among managers, with more weight on larger funds. The number of incumbent active managers $N_t$ is an equally-weighted measure of competition. The Herfindahl index ranges from zero to one, and it decreases as active investing becomes more competitive.

The following Lemma discusses the monotonicity of the relation between fund flows and performance, and the comparative statics for the number of funds.

**LEMMA 2.** An increasing number of managers is detrimental to fund fees and size, but it correlates positively with the aggregate investor demand. The Herfindahl index decreases with the number of funds. The flow-performance relation for each fund is monotonically increasing, but it declines to zero as $\sigma_\tau(t)$ diminishes. Specifically,

$$
\frac{\partial f_{it}^\ast}{\partial N_t} < 0, \quad \frac{\partial q_{it}^\ast}{\partial N_t} < 0, \quad \frac{\partial Q_t^\ast}{\partial N_t} > 0, \quad \frac{\partial H H_t}{\partial N_t} < 0, \quad \text{and} \quad \frac{\partial q_{it}^\ast}{\partial E_t[R_{it+1}]} \geq 0 \quad \forall i, t. \quad (10)
$$

Competition affects managers adversely, but it benefits the investors. The investor has preference for variety in active management, and diversifies his capital across multiple funds to mitigate risk and exploit a wide range of profitable opportunities. As a result, the aggregate size increases when new managers enter, while the assets and fees of each fund decline.

The distribution of surplus among investors and managers depends on the availability of investment opportunities. The total investment surplus from a single fund $i$ at time $t$ after trading costs is equal to

$$
T S_{it} = q_{it}^\ast E_t[R_{it+1}] - \frac{1}{2} (q_{it}^\ast)^2 \Lambda_{ii,t} - \left( c N_t (q_{it}^\ast)^2 \frac{1}{2\sigma_\tau(t)} - h q_{it}^\ast \right) 
= q_{it}^\ast E_t[R_{it+1}] - K_t \left( \frac{c N_t (q_{it}^\ast)^2}{2\sigma_\tau(t)} - h q_{it}^\ast \right). \quad (11)
$$

The term $K_t > 1$ at time $t$ reflects the assumption for the investor’s problem that the fraction of the cost for active asset management that is paid by the investor $q_{it}^2 \Lambda_{ii,t}/2$ (diagonal elements in
equation (5)) is proportional to the corresponding cost that is paid by the manager (equation (3)). Therefore, $K_t$ in equation (11) is used to capture the total cost of active investing in the estimation of total investment surplus. The manager surplus is the fund profit from fees, and the residual surplus is absorbed by the investor.

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 (6), 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 that have zero alpha by definition. The optimal fraction $x_{it}^* < 1$ is found by maximizing the following objective:

$$\max_{x_{it}} \left\{ f_{it}^* q_{it}^* (x_{it} E_t[R_{it+1}] + 1 - x_{it}) - \frac{cN_t (x_{it}q_{it}^*)^2}{2\sigma^*_t(t)} - hq_{it}^* \right\}. \tag{12}$$

The asterisks in equation (12) denote equilibrium values for capital and fees that are determined by equations (6) and (7).

**Lemma 3.** In equilibrium, each manager $i$ trades actively at time $t$ only a fraction of his assets under management given by

$$x_{it}^* = \frac{f_{it}^* (E_t[R_{it+1}] - 1)\sigma^*_t(t)}{cN_t q_{it}^*} = \frac{f_{it}^* (\tau_i / \tau(t) - R_b(t))\sigma^*_t(t)}{cN_t q_{it}^*}, \tag{13}$$

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 are more active, and tend to increase their actively managed assets when fees are large. However, managers tend to become more pas-
sive for increasing fund size, number of funds, average talent among competitors, or decreasing dispersion of talent.

III. The life cycle of investment management

In this section, I extend the single-period model to multiple time periods. I ignore dynamic hedging effects, implying that the equilibrium is a simple superposition of the single-period model from the previous section. Potential entrant managers enter if their expected profits are positive, while incumbents with negative expected profits sustain capital outflows until they liquidate and exit. In equilibrium, the marginal incumbent 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 the following condition

$$E_t[\Pi_{mt+1}] = f_{mt}^* q_{mt}^* E_t[R_{mt+1}] - \frac{cN_t (q_{mt}^*)^2}{2\sigma(t)} - h q_{mt}^* = 0,$$  \hspace{1cm} (14)

where $m$ is the marginal incumbent fund, and the asterisks denote equilibrium values that are specified by the solution to equations (4) and (7). The solution of equation (14) gives the number of incumbents $N_t(\tau(t), \sigma(t))$ as a function of the cross-sectional distribution of talent, although it has no closed-form solution. On the other hand, an endogenous evolution for $N_t$ is not critical to demonstrating the effects of competition on performance and investor flows. Therefore, I assume an exogenous entry of managers hereafter for simplicity.

The fundamental assumption about entry during the life cycle of active management is that competition 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.

---

6See Gârleanu and Pedersen (2013) for a closed-form solution of mean-variance portfolio optimization with transaction costs and dynamic hedging effects. However, Basak and Chabakauri (2010) show that mean-variance allocations in a dynamic setting can result in sub-optimal solutions in equilibrium.
**Entry assumption:** The rate of net entry is non-negative and proportional to the aggregate profits of the incumbent managers. Specifically,

\[
\frac{dN_t}{dt} \propto E_t \left[ \sum_{j=1}^{N_t} \Pi_{jt+1} \right] \geq 0, \quad \forall t \geq 0. \tag{15}
\]

The assumption above guarantees a positive net entry throughout the life cycle. However, net entry is lessened over time as profits decline from competition. Entry ceases at the end of the life cycle, when active investing earns zero profits for all incumbent managers and potential entrants are indifferent between active and passive investing.

![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.](image)

A. **Investor flows and managerial competition**

The assignment of capital to funds is assortative, with more talented managers administering larger portions of the assets. Lemma 2 shows a monotonically increasing relation between investor flows and fund performance. This implies that the investor withdraws capital from managers who have negative alpha, and reinvests it to managers who have positive alpha. He supplies inflows to funds with positive innovations in anticipation of larger returns and investment surplus in the future. The following Lemma shows the impact of flows on talent.
LEMMA 4. The performance-based flows by the investor increase the weighted average talent, i.e. $\tau(t) > \tau(t - 1), \forall t$.

Competition and the investor’s response to performance influence the cross-sectional distribution of talent. Managers with $\tau_i/\tau(t) - R_b(t) < 0$ will have negative alpha from time $t$ and beyond, and they will underperform relative to their peers and relative to the passive benchmark. These managers will experience outflows from investors until they liquidate the fund and exit. As a result, the less talented managers are forced to exit, and they are replaced by younger cohorts of more talented managers. The rising $\tau(t)$ implies that many incumbents who attain positive abnormal returns at a certain time will eventually be surpassed by more talented entrants in the future. This feature of competition is intuitively similar to a “creative destruction” (Schumpeter, 1942). Pástor et al. (2015) document empirically a rising average talent over time in equity mutual funds, and explain it in terms of changes to the population of managers.

More importantly, the increasing $\tau(t)$ combined with the exogenous prior distribution $H(\tau_i)$ for the talent of potential entrants imply that the cross-sectional dispersion $\sigma_\tau(t)$ for incumbent managers must decrease over time. The effects of managerial turnover during the life cycle of active management are summarized below.

PROPOSITION 1 (Depletion of opportunities). The competition among incumbent managers shifts the cross-sectional distribution of talent toward larger values, i.e.

$$\frac{d\tau(t)}{dt} > 0 \quad \text{and} \quad \frac{d\sigma_\tau(t)}{dt} < 0$$

as $N_t$ increases. Therefore, competition gradually depletes the investment opportunities for all managers.

The gradual diminishing of opportunities for alpha that are available to managers affects the performance-based flows of capital from the investor. The following Lemma compares the flows across different entry cohorts over time.
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 sensitivity of fund fees to performance depends on the distribution of talent. For every fund $i$ at time $t$, the limiting values for the relation between fees and performance are

$$
\lim_{\sigma_{\tau}(t) \to \infty} \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = -\frac{[2(N_t + 1)b_1 - \sum_{j \neq i} b_{jt}]h}{2(2(N_t + 1)b_1 - \sum_{j} b_{jt})E_t[R_{it+1}]^2} < 0 \quad (17)
$$

$$
\lim_{\sigma_{\tau}(t) \to 0} \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = \frac{(N_t + 1)b_1 - \sum_{j \neq i} b_{jt}}{(N_t + 1)b_1 - \sum_{j} b_{jt}} > 0. \quad (18)
$$

When profitable opportunities are bountiful, managers decrease their fees in expectation of good performance in the future. This novel result implies that competition provides a means to countering increases in fund fees, by augmenting investor flows to rival funds. The investor diversifies his 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 investor can react by redirecting flows to rival managers of similar talent. On the contrary, the manager decreases his fee to boost his future capital inflows. For instance, Christoffersen (2001) shows that money market funds relinquish fees to increase future investor flows.

Proposition 1 suggests that investment opportunities become gradually more elusive with competition, as it becomes more strenuous to outperform rivals over time. Below a certain value of $\sigma_{\tau}(t)$, it is impractical for the investor to diversify capital over similarly talented managers with small alphas, and managers who can still achieve large abnormal returns are more valuable. Those managers can increase their fees in expectation of good performance, similarly to the manager in the model of Berk and Green (2004). The Lemma below shows that this threshold value of talent dispersion is unique, and its value depends on the number of funds.

LEMMA 6. The following equations

$$
\frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = 0 \quad , \quad \frac{\partial Q_{it}^*}{\partial E_t[R_{it+1}]} = 0 \quad , \quad \text{and} \quad \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} = 0 \quad (19)
$$
have a common and unique positive root $\sigma_{thr}(N_t)$ that is increasing in $N_t$. The root is approximately equal to

$$
\sigma_{thr}(N_t) \approx b_1cN_t \left\{ (N_t + 1)b_1 - b_0 + \sum_{k=0}^{N_t} b_{kt} \left( \sum_{j=0}^{N_t} E_t[R_{jt+1}]/(N_t + 1) - 1 \right) + 4h \left( 2(N_t + 1)b_1 - \sum_{k=0}^{N_t} b_{kt} \right) \left( (N_t + 1)b_1 - \sum_{k=0}^{N_t} b_{kt} \right) + \left( \sum_{k=0}^{N_t} b_{kt} \sum_{j=0}^{N_t} E_t[R_{jt+1}]/(N_t + 1) \right) + b_1(N_t + 1) - b_0 \right)^2 \right\}^{1/2} + \left[ 2h \left( 2(N_t + 1)b_1 - \sum_{j=0}^{N_t} b_{jt} \right) \right]^{-1}.
$$

Proposition 1 and Lemma 6 suggest that the life cycle of active management may be separated in two stages. The first stage involves a relatively small number of funds with large dispersion of talent such that $\sigma_{\tau}(t) > \sigma_{thr}$. The second stage involves a large number of very talented managers on average (large $N_t$ and $\tau(t)$), but with small dispersion of talent such that $\sigma_{\tau}(t) < \sigma_{thr}$. Each stage has different implications for economies of aggregate scale, the curvature of the flow-performance relation, and the distribution of surplus among investors and managers. I discuss these effects below for each stage of the life cycle.

B. The early stages of the life cycle

Assuming a small number of funds and little aggregate capital initially, the threshold dispersion $\sigma_{thr}$ in equation (20) is nearly zero. Profitable opportunities are abundant, as implied by $\sigma_{\tau}(t) > \sigma_{thr}$ and the small $N_t$. The fund fees are negatively correlated with expected performance, because the incumbent managers try to maintain their share of investor flows while new funds enter. The following Lemma shows that the aggregate capital increases, and the Herfindahl index decreases as the number of funds grows.
LEMMA 7. *The equilibrium aggregate demand* $Q^*_t$ *increases during the early stages of the life cycle. The Herfindahl index decreases at the same time, i.e.* $Q^*_{t+1} > Q^*_t$ *and* $HH_{t+1} < HH_t$ *for every period.*

Lemma 7 suggests that the aggregate demand for active investing increases faster than the number of funds. All else equal, a rising aggregate demand decreases the market share of every manager. However, the Herfindahl index could increase in principle if there is a large number of new entrants or the aggregate capital is allocated only to a small fraction of funds. The declining Herfindahl index shows that net entry during every period is small enough, such that the sum of squared fund sizes in equation (9) rises slower than aggregate demand.

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 diversification introduce a positive network externality among investors. An investor that distributes his capital across multiple funds induces managers to decrease their fees. The capital allocation of a single investor benefits other investors too, because it decreases their cost in fees. 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 are associated with larger expected returns (see equation 17), and active funds outperform passive benchmarks on average. Since the aggregate demand and expected fund performance correlate positively, the funds operate under increasing returns to aggregate scale during the early stage of the life cycle. Figure 2 shows that for $\sigma_{T}(t) > \sigma_{thr}$, the correlation between aggregate demand and expected fund returns is positive.

The flow-performance relation during the early stages of the life cycle is concave. Figure 2 shows that the second derivative of fund size with expected returns is negative. Concavity implies
Figure 2: Comparative statics of fees on performance (red solid line), returns to aggregate scale (green dashed line), and curvature of the flow-performance relation (blue dot-dashed line) as functions of the cross-sectional dispersion of managerial talent. The vertical dashed line marks the threshold dispersion \( \sigma_{\text{thr}} \) where the returns to aggregate scale change 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 \( \sigma_\tau(t) \to 0 \) and \( \sigma_\tau(t) \to \infty \). These are

(Point A) \[ \lim_{\sigma_\tau(t) \to 0} \frac{\partial f^*_it}{\partial E_t[R_{it+1}]} = \frac{(N_t + 1)b_1 - \sum_{j \neq i} b_{jt}}{(N_t + 1)b_1 - \sum_j b_{jt}} \]

(Point B) \[ \lim_{\sigma_\tau(t) \to \infty} \frac{\partial f^*_it}{\partial E_t[R_{it+1}]} = -\frac{(2(N_t + 1)b_1 - \sum_{j \neq i} b_{jt})h}{2(2(N_t + 1)b_1 - \sum_j b_{jt})E_t[R_{it+1}]^2} \]

(Point C) \[ \lim_{\sigma_\tau(t) \to \infty} \frac{\partial Q^*_it}{\partial E_t[R_{it+1}]} = \frac{b_1((N_t + 1)b_1 - \sum_j b_{jt})h}{2(2(N_t + 1)b_1 - \sum_j b_{jt})E_t[R_{it+1}]^2} \]

(Point D) \[ \lim_{\sigma_\tau(t) \to 0} \frac{\partial Q^*_it}{\partial E_t[R_{it+1}]} = -b_1 \]

(Point E) \[ \lim_{\sigma_\tau(t) \to \infty} \frac{\partial^2 q^*_it}{\partial E_t[R_{it+1}]^2} = -\frac{b_1(2(N_t + 1)b_1 - \sum_j b_{jt} - b_{it})h}{(2(N_t + 1)b_1 - \sum_j b_{jt})E_t[R_{it+1}]^3} \]
that investor flows are more sensitive to bad performance and less sensitive to good performance. 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 (small $N_t$, large $\sigma_{\tau}(t)$), 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.

The relatively small sensitivity to a single fund’s good performance stems from the investor’s attempt to claim promptly a broad set of profitable opportunities. This also explains why good performing funds decrease their fees to attract investor flows. On the other hand, a small number of funds with large dispersion of talent implies the possibility for large realized losses from managers on the left tail. Managers with negative alpha 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, and least sensitive to good performance.

Competition and the availability of investment opportunities also affect the distribution of total surplus among investors and managers. The total surplus from a single fund is given by equation (11), 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 can be either larger or smaller than the investor’s surplus at the onset of the life cycle. However, as the profitable opportunities are exploited over time and $\sigma_{\tau}(t)$ decreases, the investor surplus increases and eventually dominates the manager surplus.

**PROPOSITION 2 (Network effect).** 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 aggregate gross alpha is positive, and the average active fund outperforms passive funds. The total surplus from a fund’s investment
is positive and split between the investor and the manager, with the investor’s portion increasing over time.

The results of Proposition 2 are based on monopolistic competition among managers, which requires a relatively large number of funds. It also implies that an individual fund cannot affect the aggregate indices of returns and fees. This explains why the average manager decreases his fees in expectation of good performance during the early stages of the life cycle, aiming to benefit from future investor inflows instead. Since the fee and return indices are unaffected by individual fund fees and returns respectively, a good performing manager can substantially increase his inflows by decreasing his fee below the fee index.

On the other hand, a manager’s performance and fees within an incipient group of only a few funds can significantly affect the corresponding indices. As a result, the manager benefits more from increasing his fee than attracting investor flows in expectation of good performance. However, the major empirical predictions of monopolistic competition during the early stages of the life cycle are the same as the predictions from a general differentiated Bertrand competition that is valid for two or more competing funds.

**COROLLARY 1.** The results of Proposition 2 for the early stages of the life cycle are identical to the predictions from a differentiated Bertrand competition among $N_t \geq 2$ funds. However, managers within a small group of funds increase their fees in expectation of good performance.

### C. The late stages of the life cycle

Despite the initial improvement in performance, the combination of rising $N_t$, rising $\bar{\tau}(t)$, and decreasing $\sigma_\tau(t)$ prognosticate the eventual depletion of investment opportunities. The shortage of profitable opportunities gradually attenuates the benefit of the network externality by raising excessively the cost of active investing. As the competition intensifies, an increasing number of managers crowd into a diminishing set of investment opportunities at higher cost. When the dispersion of talent reduces below the critical threshold $\sigma_{thr}$, the life cycle of active investing transitions
to its late stages. Managers who outperform their peers by a wide margin are rare during these stages, because most incumbents are alike in talent.

The correlation between fund fees and performance changes to positive, signaling the deterioration of the network externality. The fund fees and expected returns decrease over time. The returns to aggregate scale are decreasing in the absence of network effects. The transition is rooted in the rise of trading costs through the reduction of profitable opportunities. Figure 2 shows that the correlation between the aggregate demand and expected fund returns is negative for values of talent dispersion below the threshold $\sigma_{thr}$ (equation (20)).

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 $\sigma_\tau(t) < \sigma_{thr}$, and $\sigma_{thr}$ 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 when $N_t$ and $\bar{\pi}(t)$ are large, while $\sigma_\tau(t)$ 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. The convexity of the flow-performance relation implies that the investor values significantly those managers who still earn positive alphas within this environment.\textsuperscript{7} In addition, investor outflows from managers with negative alpha are moderate, because all incumbents are alike in talent and realized losses are smaller compared to earlier stages of the life cycle.

The lack of profitable opportunities is critical to investment surplus and aggregate risk. The total surplus deteriorates to zero as the dispersion of talent diminishes, implying that active investing becomes less valuable to both investors and managers. Since the total surplus from active investing is negligible, the distribution of surplus among investors and managers during the late stages of the life cycle is irrelevant. On the other hand, every manager indexes most of his assets when $\bar{\pi}(t)$ is large and $\sigma_\tau(t)$ is small (Lemma 3). Indexing is less risky than active investing, suggesting that aggregate risk is reduced. The number of funds is still increasing, until all the opportunities for

\textsuperscript{7}Berk and Green (2004) also derive a convex flow-performance relation when managerial talent is scarce.
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. Over its life cycle, the funds transform from an investment vehicle that offers positive 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 are competed away.\(^8\)

**PROPOSITION 3 (Today’s alpha is tomorrow’s beta).** *During the late stages of the life cycle, the returns to aggregate scale change 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 assets under management.*

In the limit where $\sigma_T(t) \to 0$, the Herfindahl index $HH_t$ and its rate of change over time $dHH_t/dt$ decline to zero. This implies that the index has low values during the late stages of the life cycle, and large groups of funds are very competitive. The slowdown in the decrease of the Herfindahl index is consistent with the decline in the net entry of new funds, as the profitable opportunities diminish over time. At the end of the life cycle, the net entry ceases and the Herfindahl index stabilizes to a small value.

Proposition 3 shows that aggregate risk decreases from indexing during the late stages of the life cycle. However, the cross-correlations of active returns among managers are expected to increase as their strategies crowd into a shrinking opportunity set. This broad increase in correlations has important implications for the performance of the average fund.

**COROLLARY 2.** *The aggregate gross alpha is negative during the late stages of the life cycle. The average active fund underperforms relative to passive funds both before and after fees.*

Corollary 2 suggests that very competitive groups of funds with a large $N_t$ will have only a few managers who achieve positive alpha. The lack of profitable opportunities and managerial

\(^8\)For instance, Stulz (2007) predicts that the performance gap between hedge funds and mutual funds will narrow, and hedge funds will become more regulated and less risky in the future.
competition impose negative alphas for most incumbents even before costs. This result is consistent with evidence for equity mutual funds in Kosowski et al. (2006) and Fama and French (2010), who find that only a small fraction of managers at the right tail of the distribution attain risk-adjusted returns in excess of their benchmark.

Pástor and Stambaugh (2012) explain the positive demand for active management even after poor average track records by assuming investor preferences that account for decreasing returns to aggregate scale. Feldman et al. (2016) extend the model of Pástor and Stambaugh and show that competition depletes the opportunities for alpha. However, they assume diminishing returns to aggregate scale. As a result, their setup does not include the benefits of competition to investors and cannot explain the emergence of increasing returns to aggregate scale.

My model suggests that most active managers underperform passive funds because of competition and the lack of opportunities for alpha. A large and very competitive group of active funds may be little different than passive funds in terms of the variety of securities offered to investors. However, the aggregate demand for active management remains large despite the poor track records, because active managers cede more control to the investor than passive funds over the choice of security portfolios and strategies. The higher fees relative to passive funds may be considered as the premium for this service to investors. Although a very competitive group of funds is deficient in alpha, it provide a platform to investors for large managed pools of capital that are well diversified and scalable. The scalability stems from indexing the bulk of managed assets, because the returns of indexed capital are unaffected by scale.

The equilibrium from competition among funds is very different than that from competition among investors in Berk and Green (2004). Their setup implies that gross alpha is positive and net alpha is zero both at the fund and aggregate levels. The investor flows in their model raise the cost of active investing disproportionately relative to the manager’s talent until the net alpha is zero for every fund. Propositions 2 and 3 imply a similar but weaker condition about persistence. The intuition is that managerial competition initially creates a market where the alpha of the average fund increases by exploiting new investment opportunities. However, as these opportunities
diminish over time, the alpha of the average fund decreases until it becomes negative eventually. Consequently, the performance of the average fund is not persistent throughout the life cycle of active management. This result holds at the aggregate level, and it does not restrict all managers from outperforming their benchmark over long periods of time. As a result, it is possible that a small minority of very talented managers at the right tail of the distribution earn consistently large positive net alphas.

D. The life cycle and aggregate liquidity

The model has shown how the availability of investment opportunities is critical to the life cycle of active management. A potential interpretation of the dynamics discussed above is related to 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. The single manager in Berk and Green (2004) is liquidity-constrained, because he expends resources to pay for bid-ask spreads that rise dramatically with his fund size and affect his returns. Managers in my model pay for bid-ask spreads too, but the scarce resource is the set of investment opportunities that are available.\(^9\)

Within nascent groups of funds there are plenty of opportunities for alpha, and managers can exploit them at relatively low cost. By creating a market for active investing, managers supply liquidity to investors. The investor diversifies his wealth across multiple funds to mitigate risk and transaction costs, but also benefits from the network externality that arises from the diversification over multiple active trading strategies. The manager’s cost for supplying liquidity to the investor is a reduction to his profits though the opportunity cost of following a particular strategy (linear term \(h_{q_{it}}\) in equation (3)). As a result, \(h_{Q_t}\) is the total liquidity premium for the network effect at time \(t\). This premium determines the magnitude of 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

---

\(^9\)Glode and Green (2011) use a similar concept to explain the persistence in the performance of hedge funds.
in the same direction to capture the few remaining opportunities. This increases the trading cost through $\sigma_T(t)$ and affects the returns of all managers simultaneously through $\tau(t)$. Therefore, the availability of profitable opportunities is inherently connected to aggregate liquidity, and managers are liquidity-constrained at the aggregate level.

IV. Empirical implications

Recent 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 (Propositions 1 and 3). Gârleanu and Pedersen (2016) provide an alternative explanation for diseconomies of scale, based on the competition among managers and informational inefficiency in investment management. 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 prediction applies for other fund classes too, such as bond mutual funds, hedge funds, and other alternative investment funds.

Proposition 3 demonstrates the challenges in measuring alpha and investment surplus within mature fund classes like equity 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 fund classes have little alpha to offer, which makes it hard to measure statistically. This is consistent with the empirical results of Kosowski et al. (2006) and Fama and French (2010), who show that most equity mutual funds have zero alpha and managers at the tails have small positive and negative alphas.

My model shows that fund 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 return of the average fund. It increases initially, but eventually 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. The reduced aggregate risk in my model during strong competition 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 investment opportunities link the returns to aggregate scale with the curvature of the flow-performance relation. Funds with limited opportunities 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 equity mutual funds is extensive.\(^{10}\) The model predicts that this fund class has diminishing returns to aggregate scale, and Pástor et al. (2015) verify this prediction. However, when managers have sufficient investment opportunities they should have increasing returns to aggregate scale and concave flow-performance relations. Goldstein et al. (2017), Kaplan and Schoar (2005), and Getmansky (2012) find concave flows among corporate bond funds, private equity, and hedge funds respectively. These fund classes are good candidates to test for increasing returns to aggregate scale.

The mechanism for the life cycle of active management 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 early with a talent level that is deep in

\(^{10}\)See also a survey for this literature by Christoffersen et al. (2014).
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. This is another indication that the depletion of alpha stems from diseconomies of aggregate scale, rather than illiquidity of large funds combined with their own diminishing returns.

V. Conclusion

This paper provides a model for the life cycle of investment management. Fund managers compete for investor flows and profitable opportunities. The markets learn about managerial talent through the innovations from the fund’s track record. The driving forces for the life cycle are the competition among funds and the distribution of managerial talent in active investing.

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 number of funds grows and managers operate 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 funds operate under diminishing returns to aggregate 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 subsets of early data for this asset class should exhibit increasing returns to aggregate scale and concave flow-performance relation. In addition, the model explains why the empirical measurement of alpha in mature asset classes such as equity mutual funds is challenging, because alpha is practically diminished during the latest stages of their life cycle.
Appendix

A. Model parameter restrictions

The following auxiliary variables

\[ A_t \equiv \sum_{j=0}^{N_t} b_1 c N_t + 2 E_t[R_{jt+1}] \sigma_\tau(t) \]  
(A.1)

\[ B_t \equiv \sum_{j=0}^{N_t} b_j t b_1 c N_t + 2 E_t[R_{jt+1}] \sigma_\tau(t) \]  
(A.2)

\[ D_t \equiv \sum_{j=0}^{N_t} E_t[R_{jt+1}] \]  
(A.3)

\[ F_t \equiv \sum_{j=0}^{N_t} b_j E_t[R_{jt+1}] b_1 c N_t + 2 E_t[R_{jt+1}] \sigma_\tau(t) \]  
(A.4)

are used extensively in the proofs below. Another useful relation is

\[ b_0 - b_{it} \sum_{j=0}^{N_t} E_t[R_{jt+1}] > 0 , \]  
(A.5)

which stems from the liquidation of a fund whose expected loss within the next period equals the total amount of assets managed. Specifically, the investor will liquidate the fund \( i \) at time \( t \) if the expected gross return \( E_t[R_{it+1}] \) is zero. As a result, the size of the liquidated fund is

\[ q_{it}^* = \frac{b_0}{N_t + 1} - \frac{b_{it}}{N_t + 1} \sum_{j=0}^{N_t} E_t[R_{jt+1}] - b_1 f_{it}^* + \frac{b_{it}}{N_t + 1} \sum_{j=0}^{N_t} f_{jt}^* = 0 , \]  
(A.6)

which implies equation (A.5), because \( b_1 > b_{it} \) and the investor is more sensitive to a fund’s own fee than rival fees.
B. Proofs

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

$$
\mathcal{L} = E_t[q'(r - f)] - \frac{a}{2} q'Vq - \frac{1}{2} q'\Lambda q - \lambda(q'1 - W_t) \quad \text{(B.1)}
$$

$$
= E_t[q'(r - f)] - \frac{d}{2} q'Vq - \lambda(q'1 - W_t) \quad , \quad \text{(B.2)}
$$

where $q = (q_0t, q_1t, \ldots, q_Nt)'$ is the fund size vector, $r = (r_{0t+1}, r_{1t+1}, \ldots, r_{Nt+1})'$ is the fund nominal return vector, $f = (f_0t, f_1t, \ldots, f_Nt)'$ is the fund fee vector, $1$ is a $(N + 1) \times 1$ vector of ones, $V$ is the $(N + 1) \times (N + 1)$ covariance matrix for the returns of the passive index and the incumbent active managers, $\Lambda$ the transaction costs matrix, while $a, d, W_t,$ and $\lambda$ are constants. The substitution from equation (B.1) to (B.2) implies that $\Lambda$ is proportional to $V$.

The first-order condition for $q'$ is

$$
E_t[r - f] - dVq = \lambda1 \quad \text{(B.3)}
$$

and the optimal fund sizes are given by

$$
q^* = \frac{1}{d} V^{-1}[E_t[r - f] - \lambda1] \equiv \frac{1}{d} \left[ E_t[R - f] - \lambda V^{-1}1 \right] \quad , \quad \text{(B.4)}
$$

where the net-of-fee alpha is defined as the net-of-fee Sharpe ratio

$$
E_t[R - f] \equiv V^{-1} E_t[r - f] \quad \text{(B.5)}
$$

with $R = (1, R_{1t+1}, \ldots, R_{Nt+1})$ the vector of risk-adjusted fund returns in excess of the benchmark, i.e. the fund gross alphas. The first element of $R$ corresponds to the passive index, and it has zero alpha by definition. Multiplying equation (B.4) by $1'$ allows to reconstruct the budget
constraint and solve for $\lambda$

\[
1'q^* = \frac{1}{d} \left[ 1'E_t[R - f] - \lambda \left( 1'V^{-1}1 \right) \right] = W_t \Rightarrow \quad (B.6)
\]

\[
\lambda = \frac{1'E_t[r - f]}{(1'V^{-1}1)} = \frac{W_t d}{(1'V^{-1}1)}. \quad (B.7)
\]

Substituting for $\lambda$ in equation (B.4) gives the equilibrium demand function

\[
q^* = \frac{W_t}{(1'V^{-1}1)}V^{-1}1 + \frac{1}{d} E_t[R - f] - \frac{1'E_t[R - f]}{d(1'V^{-1}1)}V^{-1}1. \quad (B.8)
\]

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

\[
q_{it}^* = \frac{W_t}{N_t} \sum_{j=0}^{N_t} \sum_{k=0}^{N_t} \frac{\omega_{ij}}{N_t} + \frac{1}{d} E_t[R_{it+1} - f_{it}] - \frac{\sum_{j=0}^{N_t} \omega_{ij}}{d} \sum_{j=0}^{N_t} \sum_{k=0}^{N_t} \frac{\omega_{kj}}{N_t} E_t[R_{jt+1} - f_{jt}] \quad (B.9)
\]

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

\[
\frac{\sum_{j=0}^{N_t} \omega_{ij}}{\sum_{k=0}^{N_t} \sum_{j=0}^{N_t} \omega_{kj}} = O \left( \frac{b_{it}}{N_t + 1} \right) \quad (B.10)
\]

with $0 < |b_{it}| < 1$, implying that the optimal fund size may be written as

\[
q_{it}^* = \frac{b_0}{N_t + 1} + b_1 E_t[R_{it+1} - f_{it}] - \frac{b_{it}}{N_t + 1} \sum_{j=0}^{N_t} E_t[R_{jt+1} - f_{jt}] \quad (B.11)
\]

with $b_1 > b_{it}$ and of course $b_1 > b_{it}/(N_t + 1)$ for all $N_t \geq 1$. 

\[\square\]
Proof of LEMMA 2. The correlation between the number of incumbents and fund fee is

\[
\frac{\partial f_{it}^*}{\partial N_t} = -\left\{ \left[ b_1^2 b_{it}(N_t + 1)^2 \left( b_1 c D_t - \sum_{j=0}^{N_t} E_t[R_{jt+1}] + (D_t + h A_t) \sigma(t) \right) + b_0 \left( F_t \sigma(t)^2 (F_t - b_{it} D_t) + b_1 b_{it} \sigma(t) (D_t (2(N_t + 1) - c B_t) - F_t (2c B_t + 2(N_t + 1) + b_{it} c A_t)) + b_1^2 \left( c^2 B_t^2 - c B_t (2(N_t + 1) + b_{it} c A_t) + (N_t + 1)(N_t + 1 + 2b_{it} c A_t)) \right) \right] \left( b_1 c + E_t[R_{it+1}] \sigma(t) \right)^{-1} \right\},
\]

(B.12)

where \( A_t, B_t, D_t, \) and \( F_t \) are defined in equations (A.1) to (A.4). Equation (B.12) as a function of \( \sigma(t) \) has no positive root when the parameter restrictions in equation (A.5) and \( b_1 > b_{it} / (N_t + 1) \) apply. The function’s limiting values are

\[
\lim_{\sigma(t) \to 0} \frac{\partial f_{it}^*}{\partial N_t} = \frac{b_0}{b_1 (N_t + 1)^2 (b_1 (N_t + 1) - Sb)} \cdot \left( b_1^2 (N_t + 1)^2 + (Sb - 2b_1 (N_t + 1))(Sb - (N_t + 1)b_{it}) \right) < 0 \quad \text{(B.13)}
\]

\[
\lim_{\sigma(t) \to \infty} \frac{\partial f_{it}^*}{\partial N_t} = -\frac{b_0 - \sum_{j=0}^{N_t} E_t[R_{jt+1}]}{2b_1 (N_t + 1)^2} < 0, \quad \text{(B.14)}
\]

where \( Sb \equiv \sum_{j=0}^{N_t} b_{jt} \) and in terms of order of magnitude, \( Sb - (N_t + 1)b_{it} \approx 0 \). As a result, equation (B.12) is negative and

\[
\frac{\partial q_{it}^*}{\partial N_t} = \frac{b_1 E_t[R_{it+1}] \sigma(t)}{b_1 c + E_t[R_{it+1}] \sigma(t)} \frac{\partial f_{it}^*}{\partial N_t} \leq 0 \quad \text{(B.15)}
\]
where equation (B.15) is equal to zero when $\sigma_\tau(t)$ is zero. Moreover, the correlation between the number of incumbents and aggregate investor demand is

$$\frac{\partial Q^*_t}{\partial N_t} = \left[ b_1^2(N_t + 1)^2(b_1 cB_t - Sb + F_t \sigma_\tau(t)) \left( b_1 cD_t - \sum_{j=0}^{N_t} E_t[R_{jt+1}] + (D_t + hA_t)\sigma_\tau(t) \right) + b_0 \left( b_1^2((N_t + 1 - cB_t)^2 + cA_t(b_1(N_t + 1)^2 - (2(N_t + 1) - cB_t)Sb) \right) + b_1\sigma_\tau(t) \left( b_1 D_t(N_t + 1)^2 - (2(N_t + 1) - cB_t)D_tSb - F_t(2(N_t + 1) - 2cB_t - cA_tSb) \right) + (F_t + D_tSb)F_t \sigma_\tau(t)^2 \right] - 1 \right) \cdot \left[ (N_t + 1)^2(b_1(N_t + 1 - cB_t) - F_t \sigma_\tau(t))^2 \right]^{-1}, \tag{B.16}$$

and it also lacks a positive root under the model’s parameter restrictions. Its limiting values are

$$\lim_{\sigma_\tau(t) \to 0} \frac{\partial Q^*_t}{\partial N_t} = \frac{b_0(N_t + 2)}{(N_t + 1)^2} > 0 \tag{B.17}$$

$$\lim_{\sigma_\tau(t) \to \infty} \frac{\partial Q^*_t}{\partial N_t} = \frac{b_0 + Sb \sum_{j=0}^{N_t} E_t[R_{jt+1}]}{(N_t + 1)^2} > 0 \tag{B.18}$$

implying that equation (B.16) is positive. The correlation of the Herfindahl index with $N_t$ is

$$\frac{\partial HH_t}{\partial N_t} = \frac{2q^*_{it}}{(Q^*_t)^3} \left( Q^*_t \frac{\partial q^*_{it}}{\partial N_t} - q^*_{it} \frac{\partial Q^*_t}{\partial N_t} \right) < 0 \tag{B.19}$$

from equations (B.15) and (B.16).

Similarly, the monotonicity of the flow-performance relation is described by

$$\frac{\partial q^*_{it}}{\partial E_t[R_{it+1}]} \geq 0, \tag{B.20}$$
because this derivative has no root for positive values of $\sigma(t)$, and its limiting values are

$$\lim_{\sigma(t) \to 0} \left[ \frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} \right] = 0$$

$$\lim_{\sigma(t) \to \infty} \left[ \frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} \right] = \frac{2b_1(N_t + 1) \left(2E_t[R_{it+1}]^2 + h\right)}{4(N_t + 1)E_t[R_{it+1}]^2} > 0. \tag{B.22}$$

As a result,

$$\frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} > 0 \quad \text{for} \quad \sigma(t) > 0,$$

which implies a monotonically increasing flow-performance.

Proof of LEMMA 3. The first-order condition from equation (12) for fund $i$ at time $t$ is

$$f_{it}^* q_{it}^* (E_t[R_{it+1}] - 1) - \frac{cN_t x_{it}(q_{it}^*)^2}{\sigma(t)} = 0 \tag{B.24}$$

which implies a fraction of assets under management

$$x_{it}^* = \frac{f_{it}^* (E_t[R_{it+1}] - 1)\sigma(t)}{cN_t q_{it}^*} = \frac{f_{it}^* (\tau_i/\tau(t) - R_b(t))\sigma(t)}{cN_t q_{it}^*} \tag{B.25}$$

that is actively invested in equilibrium.

Proof of LEMMA 4. Lemma 2 shows that managers with larger positive alphas are assigned more capital. Equation (2) defines $\bar{\tau}(t)$ 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 (2) from the left tail of the distribution for talent. 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

\[ \frac{d\bar{\tau}(t)}{dt} > 0. \]  

(B.26)

Proof of PROPOSITION 1. In the long run, the incumbents with talent \( \tau_i < \bar{\tau}(t) \) are forced to exit by liquidation, and the distribution of talent among the incumbents is truncated within the range \( \tau_i \in [\tau(t), \infty) \). Since the lower bound increases over time (Lemma 4), the cross-sectional dispersion \( \sigma_{\tau}(t) \) among the incumbent managers must decline, i.e.

\[ \frac{d\sigma_{\tau}(t)}{dt} < 0. \]  

(B.27)

The combination of rising \( N_t \), rising \( \bar{\tau}(t) \), and decreasing \( \sigma_{\tau}(t) \) implies the shrinkage of opportunities for alpha. As the group of funds becomes more competitive over time, the managers have progressively similar levels of talent. The homogeneity in talent impedes managers from outperforming the benchmark, and the opportunities for alpha become more elusive.

Proof of LEMMA 5. Let \( t_1 < t_2 \). Proposition 1 shows that \( \sigma_{\tau}(t_1) > \sigma_{\tau}(t_2) \). A larger dispersion of talent among all managers implies a larger uncertainty about the talent of potential entrants, which translates to larger flows after the resolution of uncertainty from 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 LEMMA 6. The equation

\[ \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = 0 \]  

(B.28)
is a polynomial of degree five. According to the Abel–Ruffini theorem, polynomials with abstract coefficients of degree five or higher may lack a closed-form solution. However, equation (B.28) has the following form

\[ A + B\sigma_\tau(t) + C\sigma_\tau(t)^2 + D\sigma_\tau(t)^3 + E\sigma_\tau(t)^4 - F\sigma_\tau(t)^5 = 0, \quad (B.29) \]

where the coefficients \(A\) to \(F\) are positive. Descartes’ rule of signs implies that this polynomial has a single positive root, and four negative or complex roots. The unique positive root for \(b_{it} = b_{jt}, \forall i, j\) is

\[
\sigma_{thr} = b_1 c N_t \left\{ (N_t + 1)b_1 - b_0 + \sum_{k=0}^{N_t} b_{kt} \left( \sum_{j=0}^{N_t} E_t[R_{jt+1}]/(N_t + 1) - 1 \right) + \left[ 4h \left( 2(N_t + 1)b_1 - \sum_{k=0}^{N_t} b_{kt} \right) \left( (N_t + 1)b_1 - \sum_{k=0}^{N_t} b_{kt} \right) + \left( \sum_{k=0}^{N_t} b_{kt} \left( \sum_{j=0}^{N_t} E_t[R_{jt+1}]/(N_t + 1) - 1 \right) + b_1(N_t + 1) - b_0 \right) \right]^{1/2} \right\}.
\]

\[
\left[ 2h \left( 2(N_t + 1)b_1 - \sum_{j=0}^{N_t} b_{jt} \right) \right]^{-1}, \quad (B.30)
\]

which is approximately equal to the positive root for the most general case where \(b_{it} \neq b_{jt}\), because the polynomial coefficients in equation (B.28) are mildly sensitive to variations in \(b_{it}\). The threshold variance \(\sigma_{thr}\) as a function of \(N_t\) has a limiting value

\[
\lim_{N_t \to \infty} \sigma_{thr}(N_t) = \frac{b_1 c N_t (1 + \sqrt{1 + 8h})}{4h} > 0, \quad (B.31)
\]

and the following root

\[
\sigma_{thr}(N_t^0) = 0 \Rightarrow N_t^0 = -\frac{b_1 - \sum_{j=0}^{N_t} b_{jt}}{b_1}. \quad (B.32)
\]
If \( b_1 - \sum_{j=0}^{N_t} b_{jt} > 0 \), then this root is negative, implying that \( \sigma_{thr} \) is positive for any \( N_t \). If \( b_1 - \sum_{j=0}^{N_t} b_{jt} < 0 \), then the root in equation (B.32) is \( 0 < N_t^0 < 1 \). However, this cannot be true, since it must be \( N_t \geq 1 \). This shows that \( \sigma_{thr} \) is positive for all \( N_t \). Moreover, its derivative relative to \( N_t \) is proportional to

\[
\frac{d\sigma_{thr}(N_t)}{N_t} \propto (2 + 16h)(N_t + 1)^2. \tag{B.33}
\]

As a result, the threshold cross-sectional dispersion of talent \( \sigma_{thr} \) increases with \( N_t \).

The correlations for returns to aggregate scale and the curvature of the flow performance relation are

\[
\frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = -\left[ b_1(b_1 c + 2E_t[R_{it+1}]\sigma_t(t)) \left( b_1(N_t + 1) - \sum_{j=1}^{N_t} b_{jt} \right) \right] \cdot \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]}
\]

\[
\frac{\partial^2 q_t^*}{\partial E_t[R_{it+1}]^2} = -2\sigma_t(t) \left[ 2b_1^2 c^2(N_t + 1 - cB_t)^2 + 4F_t\sigma_t(t)^3(2F_t\sigma_t(t) + b_{it})E_t[R_{it+1}]^2 + b_1^2 c(N_t + 1 - cB_t)(8(N_t + 1)E_t[R_{it+1}]\sigma_t(t) - c(4F_t\sigma_t(t) - b_{it} + 8B_tE_t[R_{it+1}]\sigma_t(t)) + b_1^2 \sigma_t(t)(8(N_t + 1)^2E_t[R_{it+1}]\sigma_t(t)^2 - 16c(N_t + 1)E_t[R_{it+1}]\sigma_t(t)(F_t + B_tE_t[R_{it+1}]) + c^2(2F_t^2\sigma_t(t) - F_t b_{it} + 16B_tF_tE_t[R_{it+1}]\sigma_t(t) + 8B_tE_t[R_{it+1}]\sigma_t(t)^2) \right] - b_1 E_t[R_{it+1}] \sigma_t(t) \cdot \left( 4(N_t + 1)\sigma_t(t)(4F_t\sigma_t(t) + b_{it})E_t[R_{it+1}] - c(8F_t^2\sigma_t(t)^2 + 16B_tF_t\sigma_t(t)^2E_t[R_{it+1}] - b_{it}(b_{it} - 4B_tE_t[R_{it+1}]\sigma_t(t))) \right) \cdot \left[ (b_1(N_t + 1 - B_t) + F_t\sigma_t(t))(b_1 c + 2E_t[R_{it+1}]\sigma_t(t))^3 \right]^{-1} \cdot \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]}, \tag{B.35}
\]
where $B_t$ and $F_t$ are defined in equations (A.2) and (A.4) respectively. The equations above show that these two correlations are proportional to the sensitivity of fund fees to performance. As a result, they have the same positive root $\sigma_{thr}$.

\begin{proof}
Proof of Lemma 7. The evolution of aggregate demand is given by

$$
\frac{dQ^*_t}{dt} = \frac{\partial Q^*_t}{\partial N_t} \frac{dN_t}{dt} + \frac{\partial Q^*_t}{\partial \sigma(t)} \frac{d\sigma(t)}{dt}.
$$

(B.36)

The first product is positive, because $N_t$ increases over time and $Q^*_t$ is positively correlated with $N_t$ (Lemma 2). In addition, $\sigma(t)$ is decreasing (Proposition 1). The correlation between $Q^*_t$ and $\sigma(t)$ is positive for all $\sigma(t) > 0$ values. As a result, the second term in equation (B.36) is negative, implying in general that the sign of that equation is indeterminate. However, the asymptotic limits for $\partial Q^*_t/\partial \sigma(t)$ are

$$
\lim_{\sigma(t) \to \infty} \left[ \frac{\partial Q^*_t}{\partial \sigma(t)} \right] = 0 \quad \text{and} \quad \lim_{\sigma(t) \to 0} \left[ \frac{\partial Q^*_t}{\partial \sigma(t)} \right] \approx \left[ (b_1(N_t + 1) - \sum_{j=0}^{N_t} b_{jt}) \right].
$$

Thus, the second term in equation (B.36) is much smaller than the first term during the early stages of the life cycle, implying that the aggregate demand increases over time.

Similarly, the asymptotic limit for the correlation between the Herfindahl index and the cross-sectional dispersion is

$$
\lim_{\sigma(t) \to \infty} \left[ \frac{\partial HH_t}{\partial \sigma(t)} \right] = 0.
$$

(B.39)

\end{proof}
which implies that the evolution of the index is described by

$$\frac{dHH_t}{dt} = \frac{\partial HH_t}{\partial N_t} dN_t < 0 \quad (B.40)$$

because the first term is negative (Lemma 2) and $N_t$ increases over time. As a result, the Herfindahl index declines during the early stages of the life cycle.

Proof of PROPOSITION 2. Lemma 6 shows that $\sigma_{thr}$ 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 $\sigma_\tau(t) > \sigma_{thr}$. The asymptotic value for these correlations are

$$\lim_{\sigma_\tau(t) \to \infty} \frac{\partial f^*_t}{\partial E_t[R_{it+1}]} = -\frac{(2(N_t + 1)b_1 - \sum_{j \neq i} b_{jt})h}{2(2(N_t + 1)b_1 - \sum_{j=1}^{N_t} b_{it})E_t[R_{it+1}]^2} < 0 \quad (B.41)$$

$$\lim_{\sigma_\tau(t) \to \infty} \frac{\partial Q^*_t}{\partial E_t[R_{it+1}]} = \frac{b_1((N_t + 1)b_1 - \sum_{j \neq i} b_{jt})h}{2(N_t + 1)b_1 - \sum_{j=1}^{N_t} b_{it})E_t[R_{it+1}]^2} > 0 \quad (B.42)$$

$$\lim_{\sigma_\tau(t) \to \infty} \frac{\partial^2 q^*_t}{\partial E_t[R_{it+1}]^2} = -\frac{b_1(2(N_t + 1)b_1 - 2b_{it} - \sum_{j \neq i} b_{jt})h}{(2(N_t + 1)b_1 - \sum_{j=1}^{N_t} b_{it})E_t[R_{it+1}]^3} < 0 \quad (B.43)$$

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

The expected return for a fund with zero alpha is $E_t[R_{it+1}] = 1$ (see equation (1)). This also implies that if all incumbent managers had zero alpha, then the sum of all gross alphas would be

$$SR \equiv \sum_{j=0}^{N_t} E_t[R_{jt+1}] = N_t + 1 \quad (B.44)$$

39
with zero alpha for the passive index, i.e. \( E_t[R_{0t+1}] \equiv 1 \) by definition. A fund’s fee must be non-negative, resulting in the following inequality for large values of \( \sigma_\tau(t) \)

\[
(b_0 - b_{it} \sigma) E_t[R_{it+1}] + b_1 (N_t + 1) (E_t[R_{it+1}] + h) \geq 0 \Rightarrow \\
\sigma \geq \frac{b_0 E_t[R_{it+1}] + b_1 (N_t + 1) (E_t[R_{it+1}] + h)}{b_{it} E_t[R_{it+1}]} > \frac{b_1 (N_t + 1)}{b_{it}} > N_t + 1 , \quad (B.45)
\]

implying that the aggregate gross alpha is positive during the early stages of the life cycle. In addition, equation (B.45) implies that the average active fund outperforms passive funds, since the average return is \( SR/(N_t + 1) > 1 \).

The total surplus from investment of fund \( i \) at time \( t \) is given by equation (11). Its limit for large values of \( \sigma_\tau(t) \) is

\[
\lim_{\sigma_\tau(t) \to \infty} TS_{it} = \left\{ 2(N_t + 1) b_1^2 (E_t[R_{it+1}] - h) \prod_{j \neq i} E_t[R_{jt+1}] + b_0 \left( N_t b_{it} - \sum_{j \neq i} b_{jt} \right) \cdot \right.
\]

\[
\prod_{j=1}^{N_t} E_t[R_{jt+1}] + (N_t + 1) b_1 \left( 2b_0 + (N_t - 2SR) b_{it} - \sum_{j \neq i} b_{jt} \right) \prod_{j=1}^{N_t} E_t[R_{jt+1}] + h \sum_{j \neq i} b_{jt} \cdot \prod_{j \neq i} E_t[R_{jt+1}] + b_{it} h \left( 2 \prod_{j \neq i} E_t[R_{jt+1}] + \sum_{j=1}^{N_t-2} \prod_{k=j}^{N_t-3} E_t[R_{kt+1}] \right)
\]

\[
\left( \prod_{m=k+2}^{N_t} E_t[R_{mt+1}] \right) + \left( \prod_{j=1}^{N_t-2} E_t[R_{jt+1}] \cdot \left( E_t[R_{N_t-1,t+1}] + E_t[R_{N_t,t+1}] \right) \right) \right\}.
\]

\[
\left\{ 2(N_t + 1) b_1^2 (E_t[R_{it+1}] - h - 2h E_t[R_{it+1}]) \prod_{j \neq i} E_t[R_{jt+1}] - b_0 \left( N_t b_{it} - \sum_{j \neq i} b_{jt} \right) \cdot \right.
\]

\[
\prod_{j=1}^{N_t} E_t[R_{jt+1}] + (N_t + 1) b_1 \left( -2b_0 + (N_t + 2 - 2SR) b_{it} + \sum_{j \neq i} b_{jt} \right) \prod_{j=1}^{N_t} E_t[R_{jt+1}] - h(1 + 2E_t[R_{it+1}]) \sum_{j \neq i} b_{jt} \cdot \prod_{j \neq i} E_t[R_{jt+1}] + b_{it} h \left( 2 \prod_{j \neq i} E_t[R_{jt+1}] + \sum_{j=1}^{N_t-2} \prod_{k=j}^{N_t-3} E_t[R_{kt+1}] \right)
\]

\[
\left( \prod_{m=k+2}^{N_t} E_t[R_{mt+1}] \right) + \left( \prod_{j=1}^{N_t-2} E_t[R_{jt+1}] \cdot \left( E_t[R_{N_t-1,t+1}] + E_t[R_{N_t,t+1}] \right) \right) \right\}.
\]
\[
\left[ b_1 \left( 2(N_t + 1) \left( 2(N_t + 1)b_1 - \sum_{j \neq i} b_{jt} \right) \right) ^2 \prod_{j=1}^{N_t} E_t[R_{jt+1}]^2 \right] > 0 ,
\]

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 \( \sigma_\tau(t) \).

The surplus \( MS_{it} \) that manager \( i \) extracts at time \( t \) during the early stages of the life cycle is approximately increasing and concave in \( \sigma_\tau(t) \), because

\[
MS_{it} \approx \sigma_\tau(t) \left[ 2b_1(N_t + 1)(b_0 - b_{it}SR)E_t[R_{it+1}](E_t[R_{it+1}] - h) + b_1^2(N_t + 1)^2 \left( h^2 - 2hE_t[R_{it+1}] + E_t[R_{it+1}]^2 - 2E_t[R_{it+1}] - 1 \right) + (b_0 - b_{it}SR)^2 E_t[R_{it+1}] \right] \cdot \frac{2b_1(N_t + 1)^2(b_1c + 2E_t[R_{it+1}]\sigma_\tau(t))^{-1}}{2} \tag{B.47}
\]

for large values of \( \sigma_\tau(t) \). 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 \( \sigma_\tau(t) \) 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.

\[
\square
\]

**Proof of COROLLARY 1.** The sensitivity of fund fees to performance for a differentiated Bertrand competition and large values of \( \sigma_\tau(t) \) is

\[
\lim_{\sigma_\tau(t) \to \infty} \frac{\partial f_{it}^*}{\partial E_t[R_{it+1}]} = \frac{((N_t + 1)b_1 - b_{it}) \left( E_t[R_{it+1}]^2 - h \right)}{(2(N_t + 1)b_1 - b_{it})E_t[R_{it+1}]^2} > 0. \tag{B.48}
\]

The first and second derivatives of fund size on performance are

\[
\lim_{\sigma_\tau(t) \to \infty} \frac{\partial q_{it}^*}{\partial E_t[R_{it+1}]} = \left[ ((N_t + 1)b_1 - b_{it}) \left( (N_t + 1)b_1 \left( E_t[R_{it+1}]^2 + h \right) - b_{it}E_t[R_{it+1}]^2 \right) \right] \cdot \left( (N_t + 1)(2(N_t + 1)b_1 - b_{it})E_t[R_{it+1}]^2 \right)^{-1} > 0 \tag{B.49}
\]
\[
\lim_{\sigma_{\tau}(t) \to \infty} \frac{\partial^2 q_{it}^*}{\partial E_t[R_{it+1}]^2} = - \frac{2b_1 h((N_t + 1)b_1 - b_{it})}{(2(N_t + 1)b_1 - b_{it})E_t[R_{it+1}]^2} < 0 ,
\]

which imply that the flow-performance relation is monotonically increasing and concave. The correlation between aggregate demand and fund performance is

\[
\lim_{\sigma_{\tau}(t) \to \infty} \frac{\partial Q_t^*}{\partial E_t[R_{it+1}]} = b_1 - \frac{1}{N_t + 1} \sum_{j=1}^{N_t} b_{jt} > 0 ,
\]

and the funds operate under increasing returns to aggregate scale. A fund’s fee must be non-negative, giving the following constraint for the index of returns

\[
SR \equiv \sum_{j=0}^{N_t} E_t[R_{jt+1}] \geq \frac{b_0 E_t[R_{it+1}] + b_1(N_t + 1)(E_t[R_{it+1}]^2 + h) - b_{it}h}{b_{it}E_t[R_{it+1}]}
\]

\[
> \frac{b_1(N_t + 1)}{b_{it}} > N_t + 1
\]

proving that the aggregate gross alpha is positive, and the average active fund outperforms passive funds. As a result, the results from a differentiated Bertrand competition are consistent with those from monopolistic competition during the early stages of the life cycle.

\[
\boxdot
\]

Proof of PROPOSITION 3. Lemma 6 shows that \(\sigma_{thr}\) 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 \(\sigma_{\tau}(t) < \sigma_{thr}\). Proposition 2 shows that

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

for all funds \(i\) and \(\sigma_{\tau}(t) > \sigma_{thr}\). As a result, each of these correlations has the opposite sign within the range \(\sigma_{\tau}(t) \in (0, \sigma_{thr})\). 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 are depleted. As a result, the total surplus from active investing in equation (11) also declines to zero. However, the bulk of the assets under management for every fund is invested passively, as shown in equation (13). In the limit where $\sigma_{\tau}(t) \to 0$, the manager invests only in passive strategies. Since every manager performs closet indexing, the aggregate risk is reduced as the profitable opportunities diminish.

Proof of COROLLARY 2. The expected return for a fund with zero alpha is $E_t[R_{it+1}] = 1$ (see equation (1)). This also implies that if all incumbent managers had zero alpha, then the sum of all gross alphas would be

$$SR \equiv \sum_{j=0}^{N_t} E_t[R_{jt+1}] = N_t + 1,$$  \hspace{1cm} (B.54)

with zero alpha for the passive index, i.e. $E_t[R_{0t+1}] \equiv 1$ by definition.

The average fee is part of the investor’s demand (Lemma 6). As a fee value itself, it is restricted between zero and one for every time period. These restrictions also hold in the limit where $\sigma_{\tau}(t) \to 0$, giving the following relation

$$\frac{SR}{N_t + 1} \leq 1 - \frac{b_0}{(N_t + 1)b_1 - \sum_{j=0}^{N_t} b_{jt}} < 1,$$  \hspace{1cm} (B.55)

where $b_1 > b_{it}$ from Lemma 1. The average active fund has a return $SR/(N_t + 1)$ that is smaller than the benchmark return during the late stages of the life cycle. In addition, the aggregate alpha before costs is negative, because $SR < (N_t + 1)$. As a result, the average active fund underperforms the benchmark and the aggregate alpha is negative before and after the cost of active investing is considered.

\end{proof}
References


Feldman, David, Konark Saxena, and Jingrui Xu, 2016, Is the active fund management industry concentrated enough?, *Working paper*.


Wermers, Russ, 2000, Mutual fund performance: An empirical decomposition into stock-picking