Hedge Fund Innovation

Arjen Siegmann*[1], Denitsa Stefanova†[1], and Marcin Zamojski ‡[1]

[1]*Department of Finance, VU University Amsterdam*

This version: December 13, 2013

ABSTRACT

We study first-mover advantages in the hedge fund industry by clustering hedge funds based on the type of assets and instruments they trade in, sector and investment focus, and fund details. We find that early entry in a cluster is associated with higher excess returns, longer survival, higher incentive fees and lower management fees compared to funds that arrive later. Moreover, the latest entrants have a high loading on the returns of the innovators, but with lower incentive fees, and higher management fees. Cross-sectional regressions show that the out-performance of innovating funds are declining with age. The results are robust to different parameters of clustering and backfill-bias, and are not driven by the possible existence of flagship and follow-on funds. Our results show that the reported characteristics of hedge funds can be used to infer strategy-related information and suggest that specific first-mover advantages exist in the hedge fund industry.

Keywords: hedge funds, first-mover advantage, innovation, clustering.

JEL classification codes: G15, G23.

*a.h.siegmann@vu.nl

†d.g.stefanova@vu.nl

‡Corresponding author: Marcin Zamojski (m.zamojski@vu.nl); VU University Amsterdam; Department of Finance; De Boelelaan 1105; NL-1081HV Amsterdam; The Netherlands.
Marcin Zamojski thanks the Dutch National Science Foundation (NWO) for financial support. We would like to thank Susan Christoffersen, Petri Jylha, André Lucas, Bas Peeters, Marco Rossi, Clemens Sialm, Patrick Verwijmeren, Mike Burkart, two hedge fund insiders, and seminar participants at the Tinbergen Institute, Mathematical Finance Days 2013 at HEC Montreal, European Finance Association 2013, Luxembourg Asset Management Summit 2013 for useful comments and suggestions.

1. Introduction

The large and increasing size of the hedge fund industry suggests that hedge funds are offering value to investors that is not available elsewhere. As of April 2012, the hedge fund industry has grown in size to approximately USD 2 trillion of assets under management (Hedge Fund Research, 2012). The number of funds is estimated to be around 10,000. The fee structure and light regulation give hedge funds the opportunity to follow investment strategies that are not directly available to mutual funds, for example. The excess returns and changing risk exposures as documented in the literature are a witness of their exceptional institutional structure, see (Fung and Hsieh, 1997, 2001; Agarwal and Naik, 2000, 2004; Ackermann, Mcenally, and Ravenscraft, 1999; Brunnermeier and Nagel, 2005; Agarwal, Daniel, and Naik, 2009; Patton and Ramadorai, 2013).

There is agreement in the literature on the fact that some funds show persistent out-performance (Jagannathan, Malakhov, and Novikov, 2010). Moreover, the out-performance seems to be related to changing risk exposures in reaction to (or in anticipation of) changing market conditions, see (Criton and Scaillet, 2011; Patton and Ramadorai, 2013).

In this paper, we analyse to what extent hedge fund performance is related to the inception date of a fund within a group of hedge funds with the same characteristics. We assume these hedge funds follow closely related strategies. Specifically, we test whether there is evidence that early-entrants perform better than similar funds that arrive later. Such an advantage has been shown to exist in investment banking, where innovation in financial products is visible through the ever-increasing number of products that are being offered (Herrera and Schroth, 2011). For the hedge fund industry, explicit information on—especially new—strategies or streams of income is not available. In this paper we introduce a novel approach to identify early-entrants, which might

be the carriers of new ideas, and followers, who appear later but have similar characteristics.

We group funds into clusters based on the characteristics that are supplied by the funds when registering in the Lipper TASS database. These characteristics cover the focus of the asset instruments (stocks, bonds, futures, etc.), sector and investment focus (emerging markets, US equities, etc.), and fund details (use of managed accounts, leverage, etc.). We call these characteristics the 'institutional design' of a hedge fund. The test we provide in this paper is whether the institutional design, i.e., a particular set of fund characteristics, provides information on the type of strategy followed by a hedge fund. If the initial characteristics convey little information on important return-generating aspects of the investment strategy, we should not find any performance-related effects if a hedge fund is established prior to others in a similar group of hedge funds. However, if the innovation that is necessary to set up a hedge fund affects both static characteristics and return patterns, our approach measures the benefits of early entrants in the hedge fund industry.

To identify early entrants, we sort hedge funds into clusters and measure the funds' moments of entry relative to the starting time and length of their cluster. Early entrants are the hedge funds that appear in the first quintile of their cluster. Likewise, latecomers can be identified relative to the starting date of the cluster and its length. The definition of a cluster is key, and we develop an algorithm specifically for the purpose of the paper. A custom-made algorithm, which we call Fast Binary Clustering, is necessary because we have 144 binary variables on which to cluster. Existing algorithms are either not suitable to binary data or exhibit problems with the high dimensionality. A related approach that uses clustering in finance is in Hoberg and Phillips (2010), who use the cosine distance in a text-based analysis to identify firms with related products.

In the literature, not all innovations are considered equal. Abernathy and Clark (1985) identify four types (architectural, niche, regular, and revolution-

ary) which differ in market impact. Furthermore, the definition of an innovator is flexible. It needs not be the first company in the new market. Instead, a notion of an early-entry is adopted where a group of firms is considered innovative (Christensen, Suárez, and Utterback, 1998; Utterback and Suárez, 1993; Makadok, 1998; Tufano, 1989). We first consider this definition of an innovator and group hedge funds according to the proximity of their arrival in a cluster to its inception date, which allows us to label funds as early-entrants. Conversely, Christensen, Suárez, and Utterback (1998) consider disadvantages of early-entry, and put forward a competing notion of a learning window around the time a final specification of a product (dominant design) is established. They argue that only firms which enter within this window can obtain a competitive advantage. We use this idea of 'dominant design' to test the alternative hypothesis that funds which enter a cluster during the highest growth phase have the best performance and profit the most from innovation. Thus, we consider funds who either appear in a cluster at the stage of its maximum growth, or at the moment when it decreases in size for the first time (late-stage entry).

The following are our findings. First, hedge funds that are first in a cluster earn a significantly higher excess return than funds that come later. Taken over all funds, the difference in excess performance between the first 20% and the last 20% of funds is 0.32% per month. We do not find evidence for a mechanism of a dominant design (a 'learning window') or benefits to late entry. The results are robust across hedge fund styles and to alternative specifications of risk factors, including the Pastor and Stambaugh (2003) and Sadka (2010) liquidity factors.

Second, we find evidence for pricing benefits in terms of higher incentive fees charged for the earliest quintile of funds in a cluster. Funds that arrive later in the life of the cluster have significantly lower incentive fees. The effect for management fees is the opposite: innovators charge a lower management fee than later entrants.

Third, we find that the portfolios sorted on entry time all load significantly on the first quintile portfolio (Q1), and their alphas decrease markedly. We take this as evidence that the returns of the first quintile portfolio, the 'innovators', contains non-systematic hedge funds' risk that is not in the standard risk factors. This is corroborated by the fact that including the fifth quintile portfolio (Q5) does not lead to a decrease in alpha. The fifth quintile of funds might contain non-systematic hedge fund risk, but it is not performance-related. Similar results hold when we regress hedge fund index-returns on the Q1 and Q5 portfolio returns.

Finally, panel and cross-sectional regressions show that the benefits of innovators are declining with the age of the fund and with net flows. This is consistent with a rational hedge fund market, where innovating hedge funds capture a large portion of investment flows and deliver alpha only to the earliest investors. A skilled hedge fund manager and the initial investors capture the excess performance. Later investors obtain only the marginal cost of capital, as in Berk and Green (2004). The returns on non-innovative hedge funds could still be attractive to investors, who might find it difficult to replicate systematic exposures themselves, either because of institutional or technological restrictions, or considerations of operational risk.

Our findings are related to the analysis of first-mover advantages in investment banking as well as the mutual fund and the pension fund industries, see Tufano (1989); Herrera and Schroth (2011); Lounsbury and Crumley (2007); Makadok (1998); Lopez and Roberts (2002). There, the findings are that first-movers obtain a higher share, but do not necessarily obtain a higher margin or higher fees. Our results suggest that an early-mover advantage also exists in the hedge fund industry, and is associated with higher returns, longer survival, and higher incentive fees. The higher incentive fees of early-movers, which is not found in the other industries, might reflect the decreasing returns to scale of hedge fund strategies, as witnessed by a negative size-return relationship of hedge funds, see Getmansky (2004).

Another contribution of our paper is to hedge fund classification. It is well known that self-reported styles are indicative of the exposures to risk factors, see Fung and Hsieh (1997), Agarwal and Naik (2004). Our results show that static characteristics other than style can be used to make groupings that have a bearing on performance and provide better peer candidates when evaluating hedge funds.

Our paper is related to studies on the factors that drive out-performance of hedge funds and early-stage investors, see Agarwal, Nanda, and Ray (2013), Sun, Wang, and Zheng (2012), Fung, Hsieh, Naik, and Teo (2013). We show how institutional design can be used to single out innovating hedge funds, and that early entrants in a cluster show out-performance that declines with age. It stresses the importance for investors of investing in an early stage, if they want to capture out-performance from hedge funds.

Finally, our results have some bearing on the issue of systemic risk in the hedge fund industry. If early movers gather a following of hedge funds that mimic the systematic risk exposures, systemic risk might be increased, following from the externality of the simultaneous unwinding of similar positions, see for example Khandani and Lo (2011), Aragon and Strahan (2011).

The remainder of the paper is structured as follows. Section 2 describes the data. Section 3 introduces the methodology for clustering and the construction of entry-time variables. Section 4 presents the results and Section 5 tests for the robustness of the results. Section 6 concludes.

2. Data

We use static and monthly data from January 1994 to January 2012 on both live and defunct funds from the Lipper TASS database. The TASS is a commercial database to which reporting is voluntary and it is currently commonly used in the literature on hedge funds. The sample consists of 16,051 hedge funds. We cluster these funds based on 144 binary variables that de-

scribe what are the used asset classes, investment focus and fund details of each fund, as given in the TASS database. See Appendix A for a list of the binary variables that we use. Table 1 shows summary statistics for hedge funds in the TASS database grouped by style classification.

Insert **Table 1** here

In Table 1 we see that the most hedge funds are Fund of Funds (36%) and the second largest group engage in Long/Short Equity hedging (21%). The remaining styles represent each less than 11% of the sample. We do not include Fund of Funds in our analysis.

3. Clustering and entry-time variables

The empirical approach is to make clusters of hedge funds based on their characteristics as listed in Appendix A. Given a cluster of hedge funds we define the degree of innovation of a hedge fund based on the relative time of entry in its cluster. This information allows us to test whether early-entry is optimal or if another window of opportunity might exist.

3.1. Clustering hedge funds by institutional design

Hedge funds are sorted into clusters based on similarities in their institutional design, which we define as the zeros and ones in the set of 144 binary variables listed in Appendix A. We thus infer structure of a network of knowledge based on the binary variables. To form clusters, we use a clustering algorithm specifically developed for this purpose, which we call Fast Binary Clustering (FBC). It builds on existing algorithms such as the k-means algorithm (Lloyd, 1982; Steinhaus, 1956; Ball and Hall, 1965; MacQueen, 1967) and density-based algorithms (Kailing, Kriegel, and Kröger, 2004; Ester, Kriegel, Sander, and Xu, 1996; Böhm, Kailing, Kriegel, and Kröger, 2004). In short, FBC is an agglomerative hybrid clustering algorithm that combines hierarchi-

cal, centroid, and density-connected algorithms. Each iteration of the clustering algorithm consists of a centroid and density step. In the centroid step, the algorithm assigns an archetype 'genome' to each cluster by averaging the characteristics of all observations in it. In the density step, previously identified clusters of hedge funds which are close enough (depending on a pre-defined distance metric ε) are merged. In the next iteration, the distance at which clusters are formed are increased by Δ_ε and the centroid and density steps are repeated. The distance is increased until the maximum allowed distance for joining two clusters is reached. For the distance between hedge funds and clusters we use the cosine distance measure, as in Hoberg and Phillips (2010); Watts and Strogatz (1998); Granovetter (1973).

The end result of the FBC-algorithm is a deterministic partition of the data given the distance between clusters ε and the size of its increments Δ_ε. Further details of the FBC algorithm are given in Appendix B. The logic underlying the algorithm is illustrated in Figure 1.

For the type of data we use, the clustering results are most affected by one input variable—the maximum distance between two clusters and/or funds. In the following, we work with clusters based on a maximum distance of 0.12, which leads to clusters with good properties from a clustering perspective[1]. We assess the sensitivity of our results to the distance parameter in Subsection 5.3.

Each cluster is assigned a starting month date, a duration (lifespan), and a size. The starting month of the cluster is the inception month of the first hedge fund[2] in the cluster. The duration of the cluster is defined as the time period between the inception dates of the last fund and the first fund in the cluster. We discard clusters with less than 5 funds.

[1]On average, the computational burden of clustering the hedge fund takes 2 hours on the Dutch National Computer Cluster (Lisa), which is comprised of a Dell Xeon InfiniBand cluster, 20 TFlop/sec.

[2]A hedge fund is considered to have been established in month t if its inception occurred after 15th of $t-1$ and before 16th of t.

From our sample of hedge funds from 1994 to 2012, a total of 2,771 hedge funds (26%) are in a cluster. 4,233 (40%) are not in any cluster, and 3,553 (34%) are not considered clustered because of a cluster size smaller than 5 funds, the minimum threshold.

Table 2 has the summary statistics of the resulting clusters of hedge funds, reported at clusters' inception years.

Insert **Table 2** here

From Table 2 we observe that a total of 172 clusters is identified by the clustering algorithm, with an average cluster size of 15.84 funds and duration of 54.47 months.

Our cluster methodology only uses the binary variables for the fund strategies. A first test to see whether clustering leads to distinct return properties per cluster of hedge funds is by comparing the mean returns within and between newly formed clusters. The test statistics of equal mean returns are in the last columns of Table Table 2. For most years, the F-statistics are high and significant, which gives some indication that our clustering methodology picks up differences in return distributions.

To prevent all hedge funds starting in 1994 to be deemed innovators, and to have equal representation of early and late arrivals in cluster, our subsequent analyses will use return data from the 2003-2010 period. The ultimate row of the table the averages for the 2003-2010 time period.

3.2. Entry-time Variables

In the literature on innovation, the optimal time of entry differs between industries and depends on, among others, industry structure and legal environment. On the one hand, an early-entrant has the advantage of limited competition (Tufano, 1989; Makadok, 1998; Lopez and Roberts, 2002). On the other hand, e.g., Utterback (1971); Utterback and Suárez (1993); Chris-

tensen, Suárez, and Utterback (1998) suggest that delayed-entry is optimal due to lower R&D costs which in the hedge fund industry could translate to lower search costs. Thus, to measure who benefits from innovation in the hedge fund industry, we need to be able to position the inception date of a hedge fund relative to other funds in the cluster in a systematic way. We construct three different cluster-specific variables (*FirstEntry*, *MaxGrowth* and *NegGrowth*) which should approximate optimal entry point in time. *FirstEntry* measures the time when first-entry occurred. *MaxGrowth* is the month in which the number of funds increases the most in absolute terms. *NegGrowth* is the month is in which the number of funds in the cluster decreases for the first time, i.e., when we observe (within a cluster) that more hedge funds stop reporting than there are new entrants. Both *MaxGrowth* and *NegGrowth* are determined based on the 6-month moving averages of the number of new entries and exits per month. Then, for each fund in a cluster, we compute distance in months vis-á-vis the three aforementioned entry-variables.

4. Results

We split hedge funds into quintiles based on the absolute distance between their inception date and the cluster variable, which is either *FirstEntry*, *MaxGrowth*, or *NegGrowth*. Dividing up in quintiles in this way is similar to Lopez and Roberts (2002); Utterback (1971); Utterback and Suárez (1993); Christensen, Suárez, and Utterback (1998).

For *FirstEntry* the first quintile (Q1) consists of hedge funds that belong to the first 20% of entrants in their cluster and the last quintile (Q5) has the 20% of funds which enter last.

For *MaxGrowth*, the first quintile has the 20% funds which enter the closest to the maximum-growth point. The last quintile comprises of the 20% of hedge funds which were opened furthest away from the maximum-growth point. Typically, both the first- and last-entrants are in the last quintile. Division into quintiles based on *NegGrowth* is done in a similar fashion.

4.1. Entry-time sorted quintiles

We first compute simple summary statistics, the average per-fund Fung and Hsieh (2004) 7-factor alpha, and average survival (in months) for quintiles of hedge funds based on any of the three anchor points. With new clusters being formed every year (see Table Table 2) each quintile contains hedge funds with inception dates spread out over several years. If there are benefits from innovation (better performance, survival, etc.) at a certain anchor point, we would expect to see a monotonic pattern in the population of hedge funds for the summary statistics, e.g., significantly lower returns of imitators. Thus, we report significance levels for differences in means from the first quintile of average returns, alpha and duration. The results are shown in Table 3.

Insert **Table 3** here

Panel A of Table 3 has the results for hedge funds sorted according to the *FirstEntry* anchor point. Mean return, median and alpha are monotonically decreasing from the first quintile to the last quintile. The difference in mean returns is 0.28 percentage points. For alpha it is 0.32 percentage points. Both are statistically significant. The average R-squared is slightly higher for Q5 funds. There is no monotonic pattern for durations, although the duration of Q1 funds is on the high end, with 38 months against 21 for Q5 funds. In all, this is suggestive of benefits from investing in hedge funds that are first in a cluster. Criton and Scaillet (2011); Patton, Ramadorai, and Streatfield (2012); Boyson (2010) associate a similar level of out-performance with evidence of skill.

As seen in Table 2 some 75% of all hedge funds were not clustered, either because the cluster size is too small (less than 5 funds per cluster), or because the necessary distance to include them in a valid cluster is larger than the threshold set for clustering. Given the high fraction of non-clustered hedge funds, this could be seen as evidence that distinctiveness is important, and

that benefits to imitation are low (or barriers are high). From the summary statistics for unclustered funds we observe excess returns and durations for unclustered funds which are significantly higher than Q5 funds, but lower than Q1-funds. This suggests that being unclustered is a proxy for distinctiveness, which comes with a better performance than being a late entrant in a cluster, i.e., Q5-funds.

In Panels B and C, we report results for the other anchor points (*MaxGrowth* and *NegGrowth*). We do not see any clear patterns in average returns, alphas or survival times. Therefore, we find no evidence for a 'window of opportunity' effect, around the time of maximum growth, nor an effect of higher efficiency for late entrants. In the following, we limit our attention to early-entry advantages, and thus the *FirstEntry* anchor point.

4.2. Portfolio results

We sort hedge funds into portfolios in the following way: first—to focus on the early stages of the hedge fund life cycle—we discard returns beyond 24 months for each fund[3]. This allows us to compare innovation and imitation occurring in similar periods of time. Then, at each month of the sample period all hedge funds with returns in that month are grouped into equally-weighted portfolios based on their quintile of entry. For each portfolio we compute Fung and Hsieh (2004) 7-factor alphas and present the results in Table 4.

Insert **Table 4** here

The portfolio results in Table 4 are similar to the statistics of the quintiles: there is a decreasing pattern for the mean return and alpha over the quintile portfolios. The portfolio with a long position in Q1 and short in Q5 has a mean return of 0.55 and an alpha of 0.46. The portfolio with Q1 hedge funds and a short position in not-clustered funds has a mean return and alpha which are

[3]Using a complete history of returns produces results which are qualitatively similar.

not significantly different from zero. This reinforces the idea that unclustered hedge funds could be regarded as innovative, just as hedge funds in Q1. In what follows, we keep these funds as a separate category, to see in what respects they are similar to Q1-funds. The portfolio of funds that come latest in the cluster (Q5) has no significant mean excess return or alpha. The difference between Q1 and Q5 portfolios is more pronounced than for the quintiles (where the complete return histories are used). This suggests that the innovation benefits that accrue to investors are located in the initial stages of the lifespan of clusters.

4.3. Characteristics of Innovation Quintiles

Table 5 presents the average characteristics per quintile of entry and the non-clustered (NC) hedge funds.

Insert **Table 5** here

The results in Table 5 lead to a number of interesting observations. First, there is an interesting pattern in the fees. In the quintiles of innovation, the average incentive fee of the earliest quintile (Q1) is 2.41% higher than that of Q5. The management fee is -0.24% lower. These patterns are consistent with the idea that early-arriving hedge funds are innovators, who obtain a high reward for their innovation only if it is successful, and an accordingly lower management fee. The pattern for the management fee is most pronounced, monotonically increasing from Q1 to Q5.

Second, Q1 funds significantly differ in characteristics from funds in other quintiles. Q1 funds have leverage more often than in Q5 (0.66 against 0.54), but with a lower average level (1.67% against 16.9%). A lower fraction of Q1 managers has personal capital invested, minimum investment is lower and use of a high-water mark is less frequent, compared to Q5. It remains to be seen whether the out-performance of Q1 funds can be attributed to their early-

entry in a cluster, or whether it is a result of their characteristics (or both). We test for this in a later subsection, using a cross-sectional Fama-MacBeth regression where the characteristics in Table 5 are taken into account.

Note that the results in Table 5 are not due to time-trends in characteristics, as hedge funds in the quintiles enter and exit at various times in the sample period.

4.4. Early and Late Entry as Factors

The observed excess performance of early-entry funds does not necessarily mean that they are imitated by hedge funds that come later in the cluster. It could be that hedge funds in later quintiles are entering similar markets as the innovators, but with different strategies and return characteristics. To test for this, we regress the returns of the other quintile portfolios on the returns of the first quintile portfolio and the 7 Fung and Hsieh (2004) risk factors. The results are in Table 6.

Insert **Table 6** here

Panel A of Table 6 shows the loadings of the quintile portfolios and unclustered funds on the returns of the Q1-portfolio of funds, noted as F_Q1. It also shows the alpha and R-squared. All portfolios seem to load significantly on the Q1-portfolio return. This shows that the Q1-portfolio captures systematic hedge fund risk that is not covered by the standard risk factors. The alphas are insignificant and decreasing in the quintile portfolios. Only the portfolio with unclustered funds has a significant alpha of 0.26.

Panel B shows the factor loadings and alphas when the Q5-portfolio is used as a risk factor. Here, both the loadings and the alphas are significant. Moreover, we observe that the alphas are are only slightly smaller than to the portfolio alphas in Table 4. This indicates that the Q5-portfolio contains only a

small part of non-systematic hedge fund risk. This is consistent with the insignificant alpha of the Q5-portfolio return in Table 4.

A second approach to analyzing the properties of the Q1 and Q5-portfolios is to test for their explanatory power in style regressions of hedge fund index-returns. This is reported in Table 7.

Insert **Table 7** here

Table 7 has the results of three different style regressions. The first model has the Fung and Hsieh (2004) 7-factor model for a portfolio of all funds, and the separate styles. We report only the alpha and R-squared of the regression. It shows that a large fraction of index-return variation can be explained by standard risk factors, which is a well known feature of hedge fund indices. The styles with the lowest R-squared are 'Options Strategy' , 'Managed Futures' and 'Global Macro'. The second model has the Q1-portfolio as an added risk factor. The loadings on Q1 are significant for all of the styles, except Fixed Income Arbitrage and Options Strategy. Across all styles, the alphas decrease and the R-squares increase. Thus the Q1-portfolio seems to capture a substantial part of non-systematic hedge fund risk.

Inclusion of the Q5-portfolio in the style regression, the third model, has the same effect on R-squares as with the Q1-portfolio. However, for all funds, the alpha with Q5 (0.52) is far higher than with Q1 (0.25), and a similar pattern is seen for all but a few hedge fund styles. The modest or absent decrease in alpha, compared to the first model, is consistent with Table 6 and again suggests that the Q5-portfolio has far less non-systematic hedge fund risk than the Q1-portfolio.

4.5. Panel Regressions

We know from Table 5 that the Q1-portfolio of hedge funds is associated with specific characteristics that differ between Q1 and Q5 funds. For example,

it might be that the higher incentive fees are co-determined with being innovative, so that performance is not driven by innovation alone, but also by the incentive structure. Additionally, we want to test whether out-performance due to innovation is decreasing with the age of the fund. If hedge fund returns are decreasing to scale, and the provision of capital is competitive, we should see a declining effect of being a Q1 fund over the fund's lifetime, as theorized by Berk and Green (2004). To control for characteristics and test for age and flow-effects, we estimate a panel regression.

To test for the impact of characteristics, age and flow-effects, we first perform a panel regression with the hedge fund alphas as dependent variable. The alphas are obtained from estimating the Fung and Hsieh (2004) 7-factor model for each fund with a rolling window of 24 months. To test for robustness, we also estimate Fama-MacBeth regressions, see for example Fung, Hsieh, Naik, and Teo (2013) and Sun, Wang, and Zheng (2012). This entails the estimation of cross-sectional regressions of alpha for each month and reporting time-series averages of the coefficients. Table 8 has the results, for three different specifications.

Insert **Table 8** here

The first thing to note from the results in Table 8 is that the outcomes for the panel regressions vis-à-vis the Fama-MacBeth estimations differ in size and significance for many controls, Q1, and for every specification. This suggest that either the panel regressions are misspecified, or that the coefficients on the explanatory variables are not stable over time. In the context of hedge funds, the latter explanation seems the most likely. Therefore, we focus on the Fama-MacBeth outcomes.

In all models, the significant characteristics are as expected from Table 5 and consistent with the existing literature on the sources of hedge fund out-performance. For example, age has positive sign, with implies that older funds

have a better performance. This is generally assumed to be a selection effect: good performing funds survive longer.

For model 1, the coefficient for Q1 is negative for the Fama-MacBeth regression. This suggests that the property of early-entry of the hedge funds in the Q1-portfolio might not be the sole reason for its out-performance, at least not for the complete lifetime. The performance results in Table 4 use the first 24 months of returns, while here the complete return histories are used. The intuition of innovation-driven out-performance in the early years of the fund is confirmed by the results in model 2. Here, we include interaction terms of Q1 with age. The coefficient for Q1 is 0.56 and significant, the coefficient for the interaction term of Q1 and age is -0.14 and significant. Model 3 adds lagged flows as a control, which turns out have a positive and significant impact on alphas. The coefficient on Q1 remains significant at 0.39. The decrease of innovation benefits with age can be compared with the impact of age on performance in the context of hedge funds entering emerging markets, see Aggarwal and Jorion (2010).

The results in Table 8 are consistent with Berk and Green (2004) in the context of hedge funds: hedge fund investors are sophisticated and invest in funds that are innovative. Over the lifetime of the fund, both performance and flows decrease. Managers obtain the rents from their skills through the fee structure and the increase in assets under management that decrease the potential out-performance but increases management fees.

5. Robustness

5.1. Correction for backfill bias

Backfill bias could influence results for hedge funds with an initial reporting date that is later than the inception date. Returns before the initial reporting date are called 'back-filled'. In our analysis we assumed innovation is especially beneficial to an innovator only shortly after it enters the market

due to increasing competition from imitators. As such we chose not to control for 'back-fill' bias before. However, our results are potentially affected by back-filled returns, which might not reflect actual investment returns and possibly overstate the benefits from being early.

To analyse the sensitivity to the backfill bias, we remove the first twelve months of returns of each fund and re-do our analysis[4]. Table 9 has the results for excess returns and loadings on the early entrants.

Insert **Table 9** here

The results in Table 9 are qualitatively similar to those in Table 4 and Table 6. Excess returns are significantly positive for the first quintile portfolio and monotonically decreasing over in the quintiles.

5.2. Additional Risk Factors

It might be that innovation is a proxy for an omitted risk factor in the Fung and Hsieh (2004) 7-factor model. One possible candidate is the return on an emerging markets index, on which hedge funds usually load significantly. Adding this factor to the model does not change the results (see Panel C of Table 10).

Insert **Table 10** here

An alternative explanation of our results is that hedge fund innovators are the first to find new markets that are initially less liquid. Then, the excess return for innovative funds might be a reflection of the liquidity premium in a new market or for new investment opportunities. Once other funds start following the same investment strategy, liquidity increases and the earliest funds earn an excess return. To correct for the effect of liquidity, or liquidity

[4]The clustering remains identical, as return information is not used for clustering.

timing, we include the Pastor and Stambaugh (2003) liquidity factor as well as the permanent-variable liquidity factor of Sadka (2010) (see Panel A and B in Table 10). Portfolio results remain unchanged.

5.3. Sensitivity to Clustering Parameters

Our results depend on how well the clustering algorithm is able to group funds with similar characteristics. In our analysis so far we set the maximum distance between two clusters and/or funds to 0.12. However, given that distance parameter, some funds are not assigned to a cluster or their cluster is too small and they are discarded, as denoted in Table 3, Table 4, and Table 5 with the label 'No Cluster'. Changing the maximum distance parameter on which clustering is based allows us to increase the sample size, but it may also affect the results. To assess the sensitivity of the results to a different maximum distance parameter, we redo our estimations for zero-distance clustering. This is equivalent to making clusters based on identical funds only. Based on the whole time sample used in clustering (1994–2012) we are able, in this case, to assign only 2,116 (20%) of funds to a quintile. 2,300 (22%) funds are found to be in clusters which do not satisfy our minimum requirement of 5 funds per cluster, while 6,141 (58%) funds are not clustered at all. Moreover, we identify 14 fewer clusters of innovation in the relevant time period of 2003–2010.

We construct quintile portfolios as before, and compute Fung and Hsieh (2004) 7-factor alphas. Table 11 reports the results.

Insert **Table 11** here

Table 11 confirms our findings on the excess returns of innovators vs. laggards. Overall, early entrants display higher mean returns and alphas than funds in higher quintiles. The portfolio with a long position in Q1-funds and short in Q5 has a significantly positive mean return and alpha (both higher than those obtained under the optimal clustering distance of 0.12). As well,

unclustered funds have similar characteristics as Q1-funds, in line with our previous findings.

Our results might also be sensitive to the number of clustering variables used for grouping hedge funds, and some variables might not be relevant to clustering. To test for the sensitivity of our results to the choice of clustering variables, we cluster using the binary variables in Appendix B, but without those from the category Fund Details. This leaves us with 129 (out of 144) variables. The portfolio results are in Table 12.

Insert **Table 12** here

Table 12 shows that the patterns for mean returns and alpha remain, with a decrease over the quintile portfolios from Q1 to Q5.

5.4. Fund Families

It might be that we are picking up the flagship funds of hedge fund management companies as innovators, as in Fung, Hsieh, Naik, and Teo (2013). And our Q5-funds could then be the follow-on funds. To test to what extent this is driving our results, we measure the degree in which funds from the same fund families are determining our clusters. To identify fund families, we use fuzzy matching of (partial) fund names with a hand-checked final test of similarity. Table 13 reports the degree of agreement between our clusters of hedge funds and those resulting from fund family identification.

Insert **Table 13** here

In order to quantify the degree of similarity between clusters and fund families, we perform two exercises. We first assume that true identification is obtained with the FBC algorithm. In 2003–2010 period there are 157 FBC clusters with 2,579 funds which come from 905 different fund families. This

already suggests that the overlap between the FBC clusters and fund families is not high. We also perform formal tests of cluster quality and report three entropy based measures: homogeneity, completeness, and their harmonic mean (V-measure). A homogeneous candidate cluster consists only of funds belonging to the same true cluster. Completeness is obtained if all funds form the same true cluster are grouped into the same candidate cluster. The V-measure in this case amounts to 0.66, where 1.00 corresponds to full agreement.

The results in the table show that our clusters, using the strategy descriptors are composed of different funds than the family-clusters. This suggests that fund families do not have the strategy components in common, but rather that they operate in different markets (low completeness). Moreover, fund families are more likely to expand operations in the markets they are already present in than to enter a new market (homogeneity is relatively higher).

The entropy measures tend to be inflated for higher number of clusters. To mitigate this bias, we also report two normalized measures: Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI). Both measures indicate very low overlap between FBC clusters and fund families.

Alternatively, we consider fund family membership to be true classification. The 905 fund families—in fact—consist of 6,141 funds about 60% of which are not clustered. The results confirm the findings of the previous scenario.

6. Conclusions

In this paper we cluster hedge funds by their use of assets instruments, sector and investment focus, and fund details. We find that funds that enter a cluster early have a higher excess return than funds that enter the cluster at a later date. The effect is found for the cross-section of clusters as well as for portfolios sorted on entry time in the cluster.

The results show that it is possible to define clusters of hedge funds based on descriptive characteristics, other than the investment style. It suggests

that the characteristics are actually related to the strategy followed by a hedge fund, and can be used to proxy for innovation taking place in the industry. In turn, early entrance in a cluster of similar hedge funds appears to be a signal of skill. The benefits to investors of the out-performance that is related to innovation, decrease with the age of the funds.

With respect to fees, we find that early entrants charge higher incentive fees and lower management fees than funds that enter later in the cluster. Together with the effect of age, we take this as further evidence that there is a competitive market for hedge fund assets, with decreasing returns to scale. Successful investors mirror the skills of hedge fund managers in that the timing of the investment is important. Later-stage unsophisticated investors can not be expected to receive a significant excess return. Nonetheless, this does not rule out demand for the alternative risk exposures and associated risk premiums that hedge funds can provide from investors who are otherwise limited in their investment strategies.

References

Abernathy, William J., and Kim B. Clark, 1985, Innovation: mapping the winds of creative destruction, *Research policy* 14, 3–22.

Ackermann, Carl, Richard Mcenally, and David Ravenscraft, 1999, The Performance of Hedge Funds: Risk, Return, and Incentives, *The Journal of Finance* 54, 833–874.

Agarwal, Vikas, Naveen D. Daniel, and Narayan Y. Naik, 2009, Role of Managerial Incentives and Discretion in Hedge Fund Performance, *The Journal of Finance* 64, 2221–2256.

Agarwal, Vikas, and Narayan Y Naik, 2000, On taking the alternative route, *The Journal of Alternative Investments* 2, 6–23.

——— , 2004, Risks and portfolio decisions involving hedge funds, *Review of Financial Studies* 17, 63–98.

Agarwal, Vikas, Vikram Nanda, and Sugata Ray, 2013, Institutional investment and intermediation in the hedge fund industry, Working paper.

Aggarwal, Rajesh K, and Philippe Jorion, 2010, The performance of emerging hedge funds and managers, *Journal of Financial Economics* 96, 238–256.

Aragon, George O, and Philip E Strahan, 2011, Hedge funds as liquidity providers: Evidence from the lehman bankruptcy, *Journal of Financial Economics*.

Arthur, David, and Sergei Vassilvitskii, 2007, K-means++: the advantages of careful seeding, in *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* pp. 1027–1035 Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Ball, Geoffrey H., and David J. Hall, 1965, Isodata: a novel method of data analysis and pattern classification, Discussion paper, Stanford Research Institute Menlo Park.

Berk, JonathanB., and RichardC. Green, 2004, Mutual fund flows and performance in rational markets, *Journal of Political Economy* 112, 1269–1295 doi: 10.1086/424739.

Böhm, Christian, Karin Kailing, Hans-Peter Kriegel, and Peer Kröger, 2004, DEnsity connected clustering with local subspace preferences, in *Fourth IEEE International Conference on Data Mining, 2004. ICDM '04.* pp. 27–34. IEEE.

Boyson, Nicole M., 2010, Implicit incentives and reputational herding by hedge fund managers, *Journal of Empirical Finance* 17, 283–299.

Brunnermeier, Markus K., and Stefan Nagel, 2005, Hedge funds and the technology bubble, *The Journal of Finance* 59, 2013–2040.

Christensen, Clayton M., Fernando F. Suárez, and James M. Utterback, 1998, Strategies for survival in fast-changing industries, *Management science* pp. 207–220.

Criton, Gilles, and Olivier Scaillet, 2011, Unsupervised risk factor clustering: a construction framework for funds of hedge funds, working paper.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, 1996, A density-based algorithm for discovering clusters in large spatial databases with noise, in *Proc. of 2nd International Conference on Knowledge Discovery and* pp. 226–231.

Fung, Bill, David Hsieh, Narayan Naik, and Melvyn Teo, 2013, Growing the asset management franchise: Evidence from hedge fund firms, Working paper.

Fung, William, and David A. Hsieh, 1997, Empirical characteristics of dynamic trading strategies: the case of hedge funds, *Review of Financial Studies* 10, 275–302.

——— , 2001, The risk in hedge fund strategies: theory and evidence from trend followers, *Review of Financial Studies* 14, 313–41.

——— , 2004, Hedge fund benchmarks: a risk-based approach, *Financial Analysts Journal* pp. 65–80.

Getmansky, Mila, 2004, The life cycle of hedge funds: Fund flows, size and performance, Working paper.

Granovetter, Mark S., 1973, The strength of weak ties, *The American Journal of Sociology* 78, 1360–1380.

Hedge Fund Research, 2012, Hedge fund capital inflows steady through volatile 2q12, url: http://www.hedgefundresearch.com/index.php?fuse=products-irglo, Accessed: 30/09/2012.

Herrera, Helios, and Enrique Schroth, 2011, Advantageous innovation and imitation in the underwriting market for corporate securities, *Journal of Banking and Finance* 35, 1097–1113.

Hoberg, Gerard, and Gordon Phillips, 2010, Product market synergies and competition in mergers and acquisitions: a text-based analysis, *Review of Financial Studies* 23, 3773–3811.

Jagannathan, Ravi, Alexey Malakhov, and Dmitry Novikov, 2010, Do hot hands exist among hedge fund managers? an empirical evaluation, *The Journal of Finance* 65, 217–255.

Kailing, Karin, Hans-Peter Kriegel, and Peer Kröger, 2004, Density-connected subspace clustering for high-dimensional data., in Michael W. Berry, Umeshwar Dayal, Chandrika Kamath, and David B. Skillicorn, ed.: *SDM*. SIAM.

Khandani, Amir E., and Andrew W. Lo, 2011, What happened to the quants in august 2007? evidence from factors and transactions data, *Journal of Financial Markets* 14, 1–46.

Lloyd, Stuart P., 1982, Least squares quantization in pcm, *IEEE Transactions on Information Theory* 28, 129–137.

Lopez, Luis E., and Edward B. Roberts, 2002, First-mover advantages in regimes of weak appropriability: the case of financial services innovations, *Journal of Business Research* 55, 997–1005.

Lounsbury, Michael, and Ellen T. Crumley, 2007, New practice creation: an institutional perspective on innovation, *Organization studies* 28, 993–1012.

MacQueen, James B., 1967, Some methods for classification and analysis of multivariate observations, in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability* pp. 281–297.

Makadok, Richard, 1998, Can first-mover and early-mover advantages be sustained in an industry with low barriers to entry/imitation?, *Strategic Management Journal* 19, 683–696.

Pastor, Lubos, and RobertF. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy* 111, 642–685 doi: 10.1086/374184.

Patton, Andrew, and Tarun Ramadorai, 2013, On the high-frequency dynamics of hedge fund risk exposures, *The Journal of Finance* Forthcoming.

Patton, Andrew J., Tarun Ramadorai, and Michael Streatfield, 2012, The reliability of voluntary disclosures: evidence from hedge funds, *SSRN eLibrary*.

Sadka, Ronnie, 2010, Liquidity risk and the cross-section of hedge-fund returns, *Journal of Financial Economics* 98, 54–71.

Steinhaus, Hugo, 1956, SUr la division des corps matériels en parties, *Bull. Acad. Polon. Sci. Cl. III. 4* pp. 801–804.

Sun, Zheng, Ashley Wang, and Lu Zheng, 2012, The road less traveled: Strategy distinctiveness and hedge fund performance, *Review of Financial Studies* 25, 96–143.

Tufano, Peter, 1989, Financial innovation and first-mover advantages, *Journal of Financial Economics* 25, 213–240.

Utterback, James M., 1971, The process of technological innovation within the firm, *Academy of management Journal* pp. 75–88.

——— , and Fernando F. Suárez, 1993, Innovation, competition, and industry structure, *Research policy* 22, 1–21.

Watts, Duncan J., and Steven H. Strogatz, 1998, Collective dynamics of 'small-world' networks, *Nature* 393, 440–442.

Appendix A. Hedge fund properties used for clustering

The binary properties in the TASS database that are used for clustering.

Assets Instruments
AE_Cash
AE_Convertibles
AE_Equities
AE_ExchangeTraded
AE_IndexFutures
AE_Options
AE_OTC
AE_PrimaryFocus
AE_Warrants
AF_Cash
AF_Convertibles
AF_ExchangeTraded
AF_FixedIncome
AF_Forward
AF_Futures
AF_Options
AF_OTC
AF_PrimaryFocus
AF_Swaps
AF_Warrants
AC_Agriculturals
AC_BaseMetals
AC_Commodity
AC_Energy
AC_ExchangeTraded
AC_Forwards
AC_Futures
AC_Indices
AC_Metals
AC_Options
AC_OTC
AC_Physical
AC_PreciousMetals
AC_PrimaryFocus
AC_Softs
ACUR_Currency
ACUR_ExchangeTraded
ACUR_Forwards
ACUR_Futures
ACUR_HedgingOnly
ACUR_Options
ACUR_OTC
ACUR_PrimaryFocus
ACUR_Spot
ACUR_Swaps
AP_OtherAssets
AP_Property
AP_PrimaryFocus

Investment Focus
SF_BioTechnology
SF_CloseEndedFunds
SF_CorporateBonds
SF_Diversified
SF_EmergingMarketBonds
SF_EmergingMarketEquities
SF_Financial
SF_Gold
SF_GovernmentBonds
SF_GrowthStocks
SF_HealthCare
SF_LargeCap
SF_MediaCommunications
SF_MediumCap
SF_MicroCap
SF_MoneyMarkets
SF_NaturalResources
SF_NewIssues
SF_OilEnergy
SF_Other
SF_PrivateEquity
SF_PureCurrency
SF_PureEmergingMarket
SF_PureManagedFutures
SF_RealEstateProperty
SF_Shipping
SF_SmallCap
SF_SovereignDebt
SF_Technology
SF_TurnaroundsSpinOffs
SF_Utilities
SF_ValueStocks
IA_Arbitrage
IA_BottomUp
IA_Contrarian
IA_Directional
IA_Discretionary
IA_Diversified
IA_Fundamental
IA_LongBias
IA_MarketNeutral
IA_NonDirectional
IA_Opportunistic
IA_Other
IA_RelativeValue
IA_ShortBias
IA_SystematicQuant
IA_Technical

Investment Focus (cont'd)
IA_TrendFollower
GF_Africa
GF_AsiaPacific
GF_AsiaPacificExcludingJapan
GF_EasternEurope
GF_Global
GF_India
GF_Japan
GF_LatinAmerica
GF_NorthAmerica
GF_NorthAmericaExcludingUSA
GF_Other
GF_Russia
GF_UK
GF_USA
GF_WesternEurope
GF_WesternEuropeExcludingUK
IF_Bankruptcy
IF_CapitalStructureArbitrage
IF_DistressedBonds
IF_DistressedMarkets
IF_EquityDerivativeArbitrage
IF_HighYieldBonds
IF_MergerArbitrageRiskArbitrage
IF_MortgageBackedSecurities
IF_MultiStrategy
IF_PairsTrading
IF_RegD
IF_ShareholderActivist
IF_SociallyResponsible
IF_SpecialSituations
IF_StatisticalArbitrage

Fund details
AcceptsManagedAccounts
CurrencyExposure
Derivatives
FXCredit
Futures
Guaranteed
HighWaterMark
InvestsInManagedAccounts
InvestsInOtherFunds
Leveraged
Margin
OpenEnded
OpenToPublic
PersonalCapital
RegisteredInvestmentAdvisor

Appendix B. The Fast Binary Clustering algorithm

The dataset has 16051 hedge funds (created after January 1994) with 144 binary variables that describe properties. The challenge for any clustering algorithm is to identify clusters based on (i) binary variables and (ii) do so in an acceptable amount of time. Existing algorithms like the k-means algorithm (Lloyd, 1982; Steinhaus, 1956; Ball and Hall, 1965; MacQueen, 1967) with smart seeding (Arthur and Vassilvitskii, 2007) and DBSCAN (Ester, Kriegel, Sander, and Xu, 1996) do not produce satisfactory results or do it in a very restrictive setting. Our Fast Binary Clustering (FBC) algorithm is a combination of the two, as each one separately is not suitable for the task, as shown in Table A1.

Insert **Table A1** here

Table A1 shows the outcomes of the two existing clustering algorithms, k-mean and DBSCAN, for a simulated clustered dataset of binary data. With the simulated data, we know the clusters beforehand so we can check the efficiency of each algorithm, in terms of the number of clusters it identifies, and whether the cluster composition is correct. From the table, we see that DBSCAN identifies at most 27% of the clusters, and the clusters it does find are of bad quality (homogeneity is low). The k-means algorithm does better, by finding close to 100% of clusters (or sometimes more, an indication of over-clustering). However, the k-means algorithm only works from the starting point of knowing in advance the number of clusters, which is not the case in the hedge fund data.

The third set of outcomes shows the performance of the FBC algorithm, that we explain in some detail below. It is a combination of approaches used in the DBSCAN and k-means algorithms and performs well: it identifies all clusters correctly in minimal time.

Fast binary clustering (FBC) is an agglomerative hybrid clustering algorithm combining hierarchical, centroid, and density-connected algorithms. It requires two parameters, the maximum distance to be considered, ε, and the amount by which distance should be incremented after each step, Δ_ε. Given the set of initial parameters

and data, FBC produces a deterministic set of clusters. Pseudo code for the FBC is presented below.

The algorithm operates as a set of two nested loops. At the outset all observations with identical characteristics are grouped into θ-clusters (temporary, θ). A θ-cluster can also be composed of only one observation.

The outer loop controls the hierarchical step by incrementing distance, ε, from 0 up. ε is used in the inner loop. The outer loop runs until any of the following is satisfied: ε is equal to its maximum allowed value, the total evaluations of the inner loop function reached its maximum allowed value, or only one cluster remains.

```
##
##FBC pseudo-code
##
#parameters
maxDistance #varepsilon
distanceIncrement #Delta_varepsilon
distanceMeasure
#algortihm
curDistance=0
clusters={}
while curDistance<=maxDistance:
    innerLoop=0
    while innerLoop==0
          or
          clusters[curDistance][innerLoop]==clusters[curDistance][innerLoop-1]:
                #Centroid step
                for cluster in clusters[curDistance[innerLoop]]:
                    clusters[curDistance][innerLoop][cluster]['archetypeGenome']=
                            average(clusters[curDistance][innerLoop][cluster]['funds'])

                #Density step
                proximityMatrix=getDistances(
                    [archetypeGenome for archetypeGenome in clusters[curDistance][
                        innerLoop][cluster].keys()],
                    distanceMeasure
                                        )
                clusterIds=0
                for cluster in cluster[curDistance][innerLoop]:
                    if clusteer['clusterId']==None:
                        clusteer['clusterId']=clusterIds
                        clusterIds+=1
                    else:
                        clusterId=clusteer['clusterId']
```

```
                    for otherFunds in neighborhood(cluster, proximityMatrix):
                        otherFunds['clusterId']=clusterId
  #hierarchical step
  curDistance+=distanceIncrement

clusters=discardSmallClusters(clusters[curDistance])

}
```

The inner loop iterates between a centroid step and a density step. In the centroid step, an archetype is assigned to each θ-cluster as a an average, possibly truncated (rounded). In the density step, all density-connected θ-clusters are merged into new θ-clusters. Two θ-clusters are said to be density-connected either if given the cosine-distance between their archetypes they are in the same ε-neighbourhood; or if there is a third θ-cluster to which they are both density-connected. The density step performs the clustering.

The inner loop is repeated until the resulting number of θ-clusters remains unchanged, i.e. a distance-equilibrium (ε-equilibrium) is attained. Once the algorithm is stopped, θ-clusters larger than a predetermined minimum value (5 in this paper) are retained as final clusters. The remaining observations are considered noise.

Each time they are applied, the centroid and density steps decreases the number of observations (and thus comparisons to be made) which increasingly speeds up the algorithm. Combined with the hierarchical nature of the algorithm and the fact it produces a deterministic partition of data, higher values of maximum distance parameters can be easily evaluated at an increasingly lower cost if results are retained after each outer loop completes.

Based on simulations, FBC performs at least on par with other clustering algorithms given the binary nature of the data. It performs well in moderate and high dimensions. As we could observe from Table A1, based on various measures of cluster quality it is evident that FBC is not affected by the curse of dimensionality. It also outperforms classical algorithms when executed with parameters which reflect reality (e.g. number of clusters in the case of k-means). FBC is also computationally more efficient than traditional algorithms. Their inferior performance is mainly

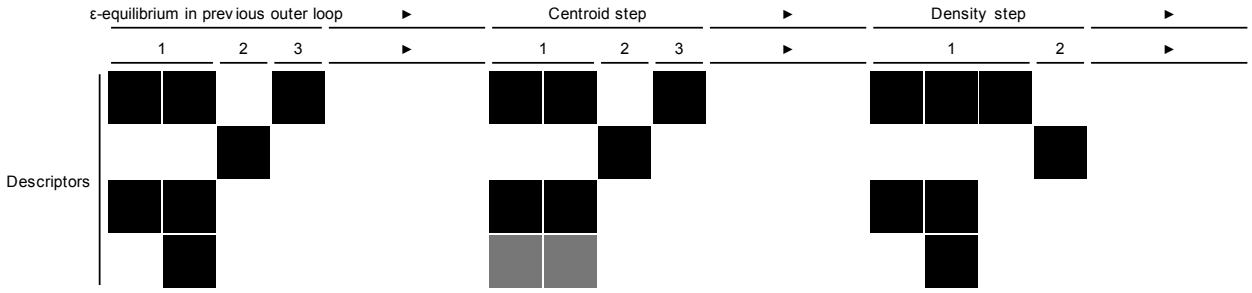
driven by the fact that they usually call for their sub-space versions, which adds a considerable computational burden as key sub-spaces need to be identified on a per observation basis, e.g. in PreDeCon (Böhm, Kailing, Kriegel, and Kröger, 2004) or SUBCLU (Kailing, Kriegel, and Kröger, 2004).

Figure 1

Illustration of Fast Binary Clustering

We show two examples of our FBC methodology for clustering hedge funds on their binary descriptors. The exact description is in Appendix B. Panel A has a four-dimensional binary example that starts from four funds in three clusters. In the first step, the centroid step, the centre of a cluster is found by eliminating some noisy descriptors (indicated by lighter shade). In the density step, the cosine distance between the central fund, unclustered candidate funds, and other clusters is computed. Funds or clusters with the lowest distance to another fund or cluster are merged into a new cluster, as long as the distance is not exceeding a threshold distance. The workings of the algorithm can also be illustrated with a two-dimensional continuous example, as in Panel B.

Panel A: Binary example



Panel B: Continuous example

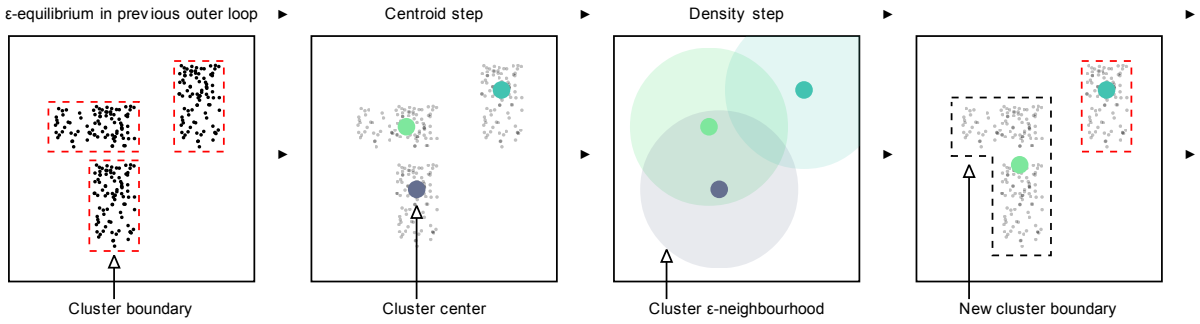


Table 1

Summary Statistics for Hedge Funds in the TASS Database

Summary statistics for hedge funds in the TASS database, per style, for the period January 1994 December 2010. The statistics are all presented as median statistics unless otherwise stated. Assets under management (AUM) is in millions of dollars, where Mean and Max are taken over the lifetime of the fund. AUM in Non-USD currencies are converted using month-end exchange rates provided by Datastream. 'Alive' is the percentage of funds that are still reporting to TASS in March 2012. The return statistics are reported for the whole sample as well as for the equally-weighted portfolios of funds per style. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

		Mean AUM [M, USD]				Fees [%]		Returns (individual)					Returns (portfolio)				
	N	Initial	Mean	Max	Alive [%]	Inc.	Manag.	Median	Mean	Std. dev.	Skewness	Kurtosis	Median	Mean	Std. dev.	Skewness	Kurtosis
All	15961	5	13	23	39	20.00	1.50	0.60	0.46	2.47	-0.44	1.57 *	0.88	0.85 ***	1.64	-0.40	2.45 ***
Fund of Funds	5791	7	14	22	41	10.00	1.50	0.51	0.27	1.98	-0.89	1.98 **	0.69	0.57 ***	1.58	-0.57	2.96 ***
Long/Short Eq. Hedge	3452	3	13	22	32	20.00	1.50	0.68	0.66 **	3.52	-0.13	1.20 **	1.19	1.18 ***	2.56	0.08	2.00 ***
Multi-Strategy	1700	5	10	16	58	15.00	1.50	0.77	0.69 *	1.99	-0.34	1.70 **	0.98	0.90 ***	1.40	-0.74	3.36 ***
Em. Markets	910	5	15	26	47	20.00	1.80	1.00	0.88 ***	4.93	-0.25	1.75 **	1.79	1.21 ***	4.24	-0.89	3.70 ***
Managed Futures	830	2	8	14	42	20.00	2.00	0.55	0.62	4.08	0.12	0.71	0.78	0.91 ***	2.51	0.24	-0.14
Global Macro	760	3	6	11	35	20.00	1.50	0.54	0.58	2.63	0.03	0.71	0.85	0.73 ***	1.65	-1.02	6.41 ***
Event Driven	717	6	32	57	26	20.00	1.50	0.79	0.70 ***	2.41	-0.38	2.29 **	1.21	0.94 ***	1.68	-1.62	6.26 ***
Eq. Market Neutral	637	4	15	26	24	20.00	1.50	0.48	0.36	2.23	-0.24	1.25 **	0.77	0.83 ***	0.98	-0.94	4.64 ***
Other	425	5	22	38	50	20.00	1.50	0.74	0.68 ***	2.27	-0.15	2.34 ***	1.09	1.04 ***	1.64	0.22	7.48 ***
Fixed Income Arb.	413	7	28	46	32	20.00	1.50	0.65	0.50 ***	1.73	-0.47	2.44 **	0.87	0.76 ***	1.15	-3.39	21.10 ***
Convertible Arb.	241	6	37	68	22	20.00	1.50	0.70	0.53	1.75	-0.54	2.12 ***	1.03	0.78 ***	2.16	-3.66	29.57 ***
Ded. Short Bias	46	4	19	37	35	20.00	1.30	0.07	0.29	5.60	0.13	0.86	-0.06	0.35	4.83	0.42	2.01 ***
Options Strategy	37	3	5	12	38	20.00	1.50	0.46	0.59	3.22	-0.09	6.98 ***	0.61	0.65 ***	0.98	0.44	1.10 **

Table 2

Summary Statistics for New Clusters Per Year

Summary statistics for newly created clusters, per year of first entry. The cluster time span denotes the number of months between the first and the last entry of a hedge fund to the cluster. ANOVA and Kruskal-Wallis are tests for the quality of clustering, under the null samples originate from the same distribution. *, ** and *** denote significance at the 90%, 95% and 99%-level, respectively.

	Number of			Cluster size		Cluster time span		ANOVA	Kruskal
	New Clusters	New Funds	New No Cluster	Avg.	Median	Avg.	Median	F-test	Wallis
1994	9	483	194	53.67	41.00	137.22	146.00	3.51 ***	49.00 ***
1995	8	734	201	91.75	12.50	124.25	143.00	4.28 ***	38.00 ***
1996	6	248	264	41.33	5.00	102.50	103.00	2.78 **	37.00 ***
1997	2	13	303	6.50	6.50	53.50	53.50	1.66	4.00 *
1998	7	88	284	12.57	8.00	115.57	132.00	5.94 ***	19.00 ***
1999	8	51	369	6.38	6.00	75.00	72.00	1.46	24.00 ***
2000	14	122	403	8.71	7.00	79.93	84.00	3.77 ***	54.00 ***
2001	11	94	469	8.55	6.00	67.18	63.00	0.96	18.00 *
2002	13	125	525	9.62	6.00	67.38	67.00	0.42	17.00
2003	12	94	621	7.83	8.00	32.00	27.00	1.06	24.00 **
2004	17	142	734	8.35	7.00	33.94	23.00	1.82 **	49.00 ***
2005	16	174	749	10.88	7.50	33.13	26.50	1.59 *	32.00 ***
2006	9	69	713	7.67	6.00	16.22	9.00	2.80 ***	17.00 **
2007	20	151	574	7.55	6.00	21.00	23.00	1.21	30.00 **
2008	8	63	484	7.88	5.50	14.75	11.50	2.81 ***	16.00 **
2009	9	55	405	6.11	6.00	9.44	5.00	2.73 ***	14.00 *
2010	3	18	268	6.00	5.00	5.00	4.00	1.89	2.00
Avg 1994-2010	10	160	444	15.84	7.00	54.47	38.00	2.61 ***	564.00 ***
Avg 2003-2010	11.75	95.75	568.50	8.15	6.00	24.20	19.00	2.21 ***	221.00 ***

Table 3

Quintiles of Hedge Funds for Different Anchor Points

This table has the summary statistics of quintiles of hedge funds from 2003–2010, based on their entry time in the cluster. Panel A describes hedge fund quintiles where quintiles are formed based on the entry time of a hedge fund relative to the inception of the cluster (the first entry). Likewise, Panel B has the quintiles formed based on the absolute time between the inception time of a clustered hedge fund and the month in which the maximum growth of the cluster is achieved. Panel C has the quintiles formed based on the absolute time between inception of a fund and the first month of negative growth. A separate sample consists of funds that are not clustered. For comparison, statistics of this sample are repeated in the last row of each panel. 'Alpha' is the average per-fund alpha from the Fung and Hsieh (2004) 7-factor model, with Newey-West corrected t-statistics in parentheses. 'Duration' is the average reporting period in months. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

Panel A: First Entry as the anchor point

	Mean	Median	Std. dev	Skewness	Kurtosis	N	Alpha	Adj. R2	Duration
Q1	0.63 ***	0.85	1.30	-9.66 ***	128.36 ***	262	0.64 ***	0.18	38.55
Q2	0.53 ***	0.59	0.72	-0.38 ***	6.26 ***	394	0.40 ***	0.21	43.17
Q3	0.45 ***	0.58	1.16	-3.01 ***	36.96 ***	556	0.35 ***	0.22	33.91
Q4	0.37 ***	0.46	1.24	-1.23 ***	10.86 ***	510	0.33 ***	0.23	28.21
Q5	0.35 ***	0.43	1.47	-5.90 ***	77.08 ***	429	0.32 ***	0.24	21.92
No Cluster	0.47 ***	0.44	1.23	1.10 ***	38.31 ***	4548	0.40 ***	0.24	36.86
Q1-Q5	0.28 ***						0.32 ***	-0.05 ***	16.63 ***
Q1-NC	0.16 *						0.24 ***	-0.06 ***	1.69
NC-Q5	0.12 *						0.09	0.00	14.94 ***

Panel B: Max Growth as the anchor point

	Mean	Median	Std. dev	Skewness	Kurtosis	N	Alpha	Adj. R2	Duration
Q1	0.43 ***	0.61	1.38	-3.02 ***	30.99 ***	360	0.43 ***	0.21	25.50
Q2	0.42 ***	0.48	1.41	-6.45 ***	85.02 ***	474	0.27 ***	0.21	36.78
Q3	0.45 ***	0.53	0.99	-0.96 ***	9.06 ***	509	0.39 ***	0.24	32.87
Q4	0.49 ***	0.61	0.95	-0.03	10.80 ***	430	0.39 ***	0.21	33.72
Q5	0.44 ***	0.69	1.27	-7.36 ***	95.16 ***	378	0.50 ***	0.19	34.81
No Cluster	0.47 ***	0.44	1.23	1.10 ***	38.31 ***	4548	0.40 ***	0.24	36.86
Q1-Q5	-0.01						-0.07	0.02	-9.31 ***
Q1-NC	-0.04						0.02	-0.03 *	-11.35 ***
NC-Q5	0.02						-0.09 *	0.05 ***	2.05

Panel C: Negative Growth as the anchor point

	Mean	Median	Std. dev	Skewness	Kurtosis	N	Alpha	Adj. R2	Duration
Q1	0.48 ***	0.78	1.66	-7.02 ***	75.28 ***	450	0.52 ***	0.22	30.88
Q2	0.43 ***	0.50	1.05	-1.19 ***	11.14 ***	572	0.34 ***	0.22	26.49
Q3	0.42 ***	0.46	1.03	-0.06	8.90 ***	558	0.31 ***	0.20	35.41
Q4	0.48 ***	0.59	1.13	-3.81 ***	48.91 ***	459	0.40 ***	0.21	40.76
Q5	0.45 ***	0.62	0.91	-1.77 ***	3.97 ***	112	0.54 ***	0.30	10.33
No Cluster	0.47 ***	0.44	1.23	1.10 ***	38.31 ***	4548	0.40 ***	0.24	36.86
Q1-Q5	0.03						-0.03	-0.08 ***	20.55 ***
Q1-NC	0.01						0.11 ***	-0.02	-5.97 ***
NC-Q5	0.02						-0.14	-0.06 **	26.52 ***

Table 4

Portfolio Results

This table has the summary statistics of portfolios of hedge funds from 2003–2010, formed by sorting hedge funds into quintile-portfolios based on their entry time in the cluster. So, Q1 represents the portfolio of hedge funds that belong to the first 20% of funds to arrive in a cluster, Q2 the following 20%, etc. A separate portfolio consists of funds that are not clustered, labelled 'No Cluster' (NC). Portfolios are equally-weighted, using the first 24 months of each fund to compute returns. The row label 'Q1-Q5' corresponds to a portfolio with a long position in Q1 and a short position in Q5. Portfolios for 'Q1-NC' and 'NC-Q5' are formed likewise. 'Alpha' and 'R2' are the portfolio alpha and adjusted R-squared from the Fung and Hsieh (2004) 7-factor model, with Newey-West corrected t-statistics in parentheses. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

	N	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis	Autocorr.	Alpha	R2
Q1	96.00	0.94 ***	1.25	-3.26	0.24	1.01	1.71	5.38	0.01	1.89 ***	0.29 ***	0.65 *** (4.56)	0.12
Q2	96.00	0.93 ***	1.28	-2.57	0.25	1.11	1.69	5.56	-0.10	1.42 **	0.23 **	0.66 *** (4.89)	0.33
Q3	96.00	0.70 ***	1.30	-3.15	0.08	0.78	1.52	3.46	-0.66 ***	0.80	0.24 **	0.43 *** (3.88)	0.29
Q4	96.00	0.47 ***	1.57	-6.38	-0.20	0.69	1.44	3.25	-1.54 ***	4.76 ***	0.35 ***	0.16 (1.31)	0.50
Q5	96.00	0.39 *	2.27	-7.02	-0.33	0.72	1.47	6.40	-0.83 ***	2.84 ***	0.18 *	0.03 (0.15)	0.25
No Cluster	96.00	0.90 ***	1.53	-4.59	0.09	1.15	1.90	4.32	-1.00 ***	1.73 **	0.29 ***	0.60 *** (7.54)	0.65
Q1-Q5	96.00	0.55 ***	2.05	-5.16	-0.41	0.28	1.47	7.94	0.41 *	2.23 ***	0.13	0.46 ** (2.31)	0.08
Q1-NC	96.00	0.04	1.28	-4.45	-0.50	0.10	0.56	3.67	-0.23	1.99 ***	0.45 ***	-0.11 (-0.87)	0.36
NC-Q5	96.00	0.52 ***	1.62	-4.85	-0.21	0.32	1.27	8.09	0.75 ***	6.74 ***	-0.05	0.41 *** (2.78)	-0.01

Table 5

Average Characteristics Per Innovation Quintile

This table shows the average of fund characteristics per quintile. Leveraged is an indicator on whether the fund uses leverage. Max. leverage is the maximum leverage used. Avg. leverage is the stated average leverage. AUM is assets under management, in millions of dollars. Redemption is the redemption frequency. Lock-up period and redemption are in months. Personal capital is a 0-1 indicator on whether principals have money invested. Hedge funds which are not assigned to any cluster are labelled 'NC'. The column labelled 'Q1-Q5' has the difference in the average statistic between funds in Q1 and in Q5; 'Q1-NC' for the difference between Q1 and the unclustered funds; 'NC-Q5' for the difference between the unclustered funds and Q5. *, ** and *** denote significance at the 90%, 95% and 99%-level, respectively.

		Q1	Q2	Q3	Q4	Q5	NC	Q1-Q5	Q1-NC	NC-Q5
	Incentive	15.25	13.89	15.24	13.74	12.84	17.72	2.41 ***	-2.48 ***	4.88 ***
Fees	Management	1.37	1.51	1.44	1.55	1.61	1.58	-0.24 ***	-0.21 ***	-0.03
	Leveraged	0.66	0.55	0.53	0.54	0.54	0.58	0.12 ***	0.08 **	0.04
	Max.	96.98	90.35	156.04	83.81	170.06	119.15	-73.08	-22.17	-50.91
Leverage	Avg.	1.67	33.56	20.19	19.52	16.91	44.30	-15.24 **	-42.63 ***	27.39 ***
	Avg. AUM	14.63	79.62	51.29	245.35	19.88	25.14	-5.25	-10.51 **	5.26
Initial AUM [M]	Median	5.90	14.95	5.75	4.84	5.77	6.96			
	Lock-up Period	0.78	0.78	1.42	1.43	0.94	3.40	-0.17	-2.62 ***	2.45 ***
	Redemption	1.57	1.35	1.69	1.68	1.59	1.76	-0.03	-0.19	0.17
	Personal Capital	0.00	0.00	0.05	0.05	0.03	0.21	-0.02 ***	-0.21 ***	0.18 ***
	Managed Accounts	0.03	0.02	0.04	0.06	0.03	0.23	-0.00	-0.19 ***	0.19 ***
	Minimum Investment	0.13	0.92	0.63	0.79	0.77	1.30	-0.64 ***	-1.17 ***	0.53
	High Water Mark	0.28	0.31	0.35	0.36	0.38	0.79	-0.10 ***	-0.51 ***	0.41 ***

Table 6

Early and Late Entry as Factors for Quintile Portfolios

This table has the regression results for the Fung and Hsieh (2004) 7-factor model, augmented with the portfolio returns from the Q1 and Q5-portfolios as risk factors. Q1 is the portfolio with early entrants in a cluster and Q5 is the portfolio with late-arriving hedge funds, as in Table 4. We report the alpha and the R2 of the regressions and loadings on the Q1 and Q5 factors. Newey-West corrected t-statistics in parentheses. *, ** and *** denote significance at the 90%, 95% and 99%-level, respectively.

Panel A: FH2004 with Q1 portfolio

	Q2	Q3	Q4	Q5	No Cluster	NC-Q5
Alpha	0.26 *	0.14	-0.14	-0.38	0.26 **	0.50 **
	(1.65)	(0.80)	(-0.76)	(-1.46)	(2.20)	(2.47)
F_Q1	0.49 ***	0.35 ***	0.37 ***	0.50 ***	0.41 ***	-0.11
	(6.16)	(2.51)	(3.03)	(2.77)	(4.41)	(-0.86)
R2	0.53	0.38	0.57	0.31	0.74	-0.02

Panel B: FH2004 with Q5 portfolio

	Q1	Q2	Q3	Q4	No Cluster	Q1-NC
Alpha	0.62 ***	0.61 ***	0.37 ***	0.10	0.55 ***	-0.09
	(4.44)	(4.96)	(3.97)	(1.06)	(8.84)	(-0.72)
F_Q5	0.18 ***	0.28 ***	0.33 ***	0.32 ***	0.26 ***	-0.09 *
	(2.77)	(4.95)	(3.40)	(3.56)	(3.14)	(-1.88)
R2	0.19	0.51	0.53	0.66	0.76	0.37

Table 7

Style Regressions With Innovation Risk Factors

Style regressions for the TASS style index returns regressed on the Fung and Hsieh (2004) 7-factor model, with the returns from the Q1 and Q5-portfolios as additional risk factors. Q1 is the portfolio with early arriving hedge funds in the cluster and Q5 is the portfolio with late-arriving hedge funds, as in Table 4. We report the alphas, R-squares of the regressions, and the exposures to the innovation factors. Estimated with Newey-West corrected t-statistics in parentheses. *, ** and *** denote significance at the 90%, 95% and 99%-level, respectively.

	Fung and Hsieh (2004)		with F_Q1			with F_Q5		
	Alpha	R2	Alpha	F_Q1	R2	Alpha	F_Q5	R2
All	0.57 *** (6.94)	0.58	0.25 ** (2.01)	0.40 *** (3.85)	0.69	0.52 *** (8.11)	0.28 *** (3.31)	0.74
Long/Short Eq. Hedge	0.52 *** (4.78)	0.64	0.15 (0.97)	0.46 *** (3.97)	0.73	0.47 *** (4.49)	0.29 *** (2.98)	0.74
Fund of Funds	0.03 (0.23)	0.50	-0.39 *** (-3.46)	0.51 *** (4.86)	0.66	-0.03 (-0.34)	0.32 *** (3.13)	0.68
Multi-Strategy	0.59 *** (8.02)	0.48	0.33 *** (2.88)	0.32 *** (3.04)	0.60	0.54 *** (9.14)	0.24 *** (3.22)	0.68
Emerging Markets	0.93 *** (4.07)	0.58	0.27 (1.20)	0.81 *** (3.78)	0.69	0.84 *** (4.20)	0.43 *** (3.12)	0.66
Managed Futures	0.54 *** (2.83)	0.20	0.12 (0.47)	0.52 *** (3.39)	0.28	0.43 *** (2.71)	0.62 *** (6.35)	0.52
Global Macro	0.60 *** (5.91)	0.21	0.31 ** (2.60)	0.36 *** (3.88)	0.35	0.56 *** (5.27)	0.22 *** (3.10)	0.37
Event Driven	0.68 *** (5.45)	0.65	0.33 ** (2.45)	0.43 *** (6.40)	0.73	0.65 *** (5.12)	0.18 ** (2.21)	0.68
Eq. Market Neutral	0.27 *** (5.06)	0.49	0.17 ** (2.33)	0.13 ** (2.53)	0.51	0.25 *** (4.85)	0.12 ** (2.31)	0.55
Other	0.70 *** (6.42)	0.46	0.63 *** (5.57)	0.09 (1.64)	0.47	0.69 *** (6.45)	0.09 ** (2.02)	0.49
Fixed Income Arb.	0.47 *** (6.58)	0.33	0.48 *** (5.03)	-0.02 (-0.39)	0.32	0.46 *** (6.54)	0.04 (1.42)	0.33
Convertible Arb.	0.29 (1.51)	0.67	-0.02 (-0.08)	0.38 *** (2.96)	0.69	0.23 (1.33)	0.28 ** (2.16)	0.71
Ded. Short Bias	0.40 ** (2.26)	0.78	0.17 (0.77)	0.28 ** (2.00)	0.78	0.36 * (1.97)	0.17 ** (2.03)	0.78
Options Strategy	0.76 *** (4.36)	0.13	1.03 *** (3.50)	-0.32 (-1.54)	0.15	0.80 *** (4.33)	-0.17 (-1.11)	0.14

Table 8

Panel and Cross-Sectional Regressions

This table has panel regressions of hedge fund abnormal returns on fund characteristics. The columns labelled 'Panel' have results from random effects GLS regressions with per fund clustered standard errors. The 'Fama-MacBeth' columns use the Fama and MacBeth (1973) methodology of estimating risk premiums in the cross-sections. Alphas are computed with a 24-month rolling window and adjusted for the Fung and Hsieh (2004) 7 factors. We control for backfill-bias. We focus on the 2003-2010 window as in Table 4. The time-varying variables: size (Log AUM), flows (Flow) and net-of-fees returns (Return) are appropriately averaged over 12-month windows and lagged by one month, denoted as (t-1). *, ** and *** denote significance at the 90%, 95% and 99%-level, respectively.

	(1)		(2)		(3)	
	Panel	Fama-MacBeth	Panel	Fama-MacBeth	Panel	Fama-MacBeth
Q1	-0.11	-0.08 ***	0.48 *	0.56 ***	0.35	0.39 ***
	(-1.01)	(-3.67)	(1.72)	(4.26)	(1.21)	(2.71)
Q1 × Age			-0.14 **	-0.14 ***	-0.10	-0.11 **
			(-2.23)	(-3.70)	(-1.57)	(-2.57)
Q1 × Flow (t-1)					0.02	0.04 ***
					(1.41)	(9.55)
Joint test for Q1 vars			*		**	
			(5.50)		(10.02)	
intercept	0.48	-1.06 ***	0.24	-1.19 ***	0.26	-1.19 ***
	(0.73)	(-4.33)	(0.35)	(-4.70)	(0.39)	(-4.68)
Age	0.07 *	0.05 ***	0.10 **	0.07 ***	0.10 **	0.06 ***
	(1.94)	(2.89)	(2.47)	(3.41)	(2.42)	(3.19)
Incentive Fee	0.02 ***	0.00	0.02 ***	0.00	0.02 ***	0.00
	(3.61)	(1.08)	(3.37)	(1.23)	(3.42)	(1.04)
Management Fee	0.01	0.01	0.01	0.01	0.01	0.02
	(0.14)	(0.25)	(0.20)	(0.44)	(0.21)	(0.80)
Log(1+Minimum Investment)	-0.03	0.04 **	-0.03	0.04 ***	-0.03	0.04 ***
	(-1.04)	(2.60)	(-0.99)	(2.72)	(-1.02)	(2.70)
Lock-up Period	0.01	0.00 *	0.01	0.00 **	0.01	0.00 *
	(0.89)	(1.87)	(0.97)	(2.23)	(0.89)	(1.81)
Redemption Frequency	0.01	-0.04 ***	0.01	-0.04 ***	0.01	-0.05 ***
	(0.90)	(-5.08)	(0.85)	(-4.75)	(0.89)	(-5.38)
Leveraged	0.06	-0.07 ***	0.04	-0.08 ***	0.04	-0.07 ***
	(0.69)	(-2.78)	(0.41)	(-3.35)	(0.46)	(-2.98)
Personal Capital	-0.07	-0.05	-0.04	-0.03	-0.05	-0.04
	(-0.41)	(-1.23)	(-0.27)	(-0.83)	(-0.29)	(-1.08)
High Water Mark	-0.21 **	-0.14 ***	-0.21 **	-0.13 ***	-0.20 **	-0.12 ***
	(-2.14)	(-4.40)	(-2.12)	(-4.35)	(-2.07)	(-4.41)
Avg 12m Flow (t-1)	0.01 **	0.01 ***	0.01 **	0.01 ***	0.01 *	0.01 ***
	(2.10)	(5.09)	(2.12)	(5.19)	(1.78)	(3.15)
Avg 12m Return (t-1)	0.14 ***	0.51 ***	0.13 ***	0.50 ***	0.13 ***	0.49 ***
	(6.06)	(16.40)	(6.13)	(16.20)	(6.05)	(16.05)
Avg 12m Log(AUM) (t-1)	-0.02	0.04 ***	-0.01	0.04 ***	-0.01	0.04 ***
	(-0.54)	(7.33)	(-0.34)	(7.20)	(-0.36)	(7.42)

Table 9

Correcting for Backfill-Bias

This table has the summary statistics of portfolios of hedge funds, with returns corrected for backfill-bias. We eliminate the first 12 months and keep the following 24 months for the portfolios. As in Table 4, portfolios are formed by sorting hedge funds into quintile-portfolios based on their entry time in the cluster. So, Q1 represents the portfolio of hedge funds that belong to the first 20% of funds to arrive in a cluster, Q2 the following 20%, etc. A separate portfolio consists of funds that are not clustered, labelled 'No Cluster' (NC). The portfolio return is equally-weighted, using months 13 to 36 of each fund to compute returns. The row label 'Q1-Q5' corresponds to a portfolio with a long position in Q1 and a short position in Q5. Portfolios for 'Q1-NC' and 'NC-Q5' are formed likewise. 'Alpha' and 'R2' are the portfolio alpha and adjusted R-squared from the Fung and Hsieh (2004) 7-factor model, with Newey-West corrected t-statistics in parentheses. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

	N	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis	Autocorr.	Alpha	R2
Q1	96.00	0.89 ***	1.01	-1.58	0.22	0.90	1.48	4.66	0.28	1.15 *	0.18 *	0.70 *** (5.90)	0.12
Q2	96.00	0.67 ***	1.29	-2.46	-0.05	0.84	1.56	2.95	-0.59 **	-0.19	0.18 *	0.45 *** (3.37)	0.29
Q3	96.00	0.51 ***	1.56	-6.60	-0.02	0.79	1.38	3.32	-1.82 ***	5.84 ***	0.40 ***	0.26 ** (2.20)	0.51
Q4	96.00	0.49 **	1.91	-9.19	-0.20	0.66	1.75	3.70	-1.95 ***	7.16 ***	0.42 ***	0.16 (1.09)	0.57
Q5	96.00	0.45 *	2.45	-7.96	-0.45	0.81	1.63	10.10	-0.48 *	3.89 ***	0.23 **	0.07 (0.36)	0.38
No Cluster	96.00	0.73 ***	1.77	-6.81	-0.03	0.99	1.94	5.11	-1.38 ***	4.06 ***	0.37 ***	0.41 *** (4.77)	0.71
Q1-Q5	96.00	0.44 **	2.15	-7.59	-0.54	0.27	1.17	7.00	0.43 *	3.18 ***	0.28 ***	0.46 ** (2.40)	0.30
Q1-NC	96.00	0.16	1.40	-3.78	-0.45	0.02	0.48	6.24	1.18 ***	5.11 ***	0.50 ***	0.12 (1.03)	0.64
NC-Q5	96.00	0.28 *	1.59	-7.44	-0.46	0.27	1.14	7.00	-0.27	7.94 ***	0.09	0.18 (1.09)	-0.05

Table 10

Portfolio Results With Additional Risk Factors

This table has portfolio results for entry-time sorted portfolios, with additional risk factors. As in Table 4, the portfolio return is equally-weighted, using the first 24 months of each fund to compute returns. Alpha' and 'R2' are the portfolio alpha and adjusted R-squared from a Fung and Hsieh (2004) 7-factor model with additional factors. The first additional factor is the Pastor and Stambaugh (2003) measure of liquidity, PS Inn. The second additional factor is the Sadka (2006) Permanent-Variable (Sadka PV) liquidity factor. The third additional factor is the return on the MSCI Emerging Market index. Newey-West corrected t-statistics in parentheses. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

Panel A: FH2004 with PS Inn

	Q1	Q2	Q3	Q4	Q5	No Cluster	Q1-Q5	Q1-NC	NC-Q5
Alpha	0.66 ***	0.66 ***	0.43 ***	0.17	0.03	0.60 ***	0.46 **	-0.11	0.41 ***
	(4.62)	(4.86)	(3.86)	(1.46)	(0.16)	(7.65)	(2.31)	(-0.88)	(2.78)
PS Inn	0.02	0.02	0.02	0.05 ***	0.02	0.03 **	0.01	-0.01	0.01
	(1.55)	(1.08)	(0.84)	(3.27)	(0.57)	(2.18)	(0.25)	(-0.50)	(0.61)
R2	0.13	0.33	0.29	0.54	0.24	0.66	0.07	0.35	-0.02

Panel B: FH2004 with Sadka PV

	Q1	Q2	Q3	Q4	Q5	No Cluster	Q1-Q5	Q1-NC	NC-Q5
Alpha	0.66 ***	0.65 ***	0.42 ***	0.13	-0.02	0.57 ***	0.52 ***	-0.07	0.43 ***
	(4.50)	(4.70)	(3.31)	(1.03)	(-0.09)	(6.31)	(2.62)	(-0.57)	(2.97)
Sadka PV	-0.07	0.03	0.07	0.22	0.31	0.17	-0.38	-0.24	-0.14
	(-0.41)	(0.22)	(0.33)	(1.08)	(0.99)	(0.84)	(-1.50)	(-1.51)	(-0.81)
R2	0.11	0.32	0.28	0.50	0.25	0.65	0.09	0.37	-0.02

Panel C: FH2004 with Em Mkt

	Q1	Q2	Q3	Q4	Q5	No Cluster	Q1-Q5	Q1-NC	NC-Q5
Alpha	0.58 ***	0.58 ***	0.32 ***	0.04	-0.13	0.47 ***	0.55 ***	-0.04	0.44 ***
	(3.83)	(3.89)	(2.89)	(0.35)	(-0.70)	(6.21)	(2.58)	(-0.34)	(2.89)
Em Mkt	0.08 **	0.09 ***	0.13 ***	0.14 ***	0.17 ***	0.14 ***	-0.10 *	-0.07 **	-0.04
	(2.32)	(3.02)	(4.05)	(5.63)	(3.27)	(5.63)	(-1.83)	(-2.41)	(-0.92)
R2	0.17	0.39	0.41	0.60	0.32	0.76	0.10	0.39	-0.02

Table 11

Zero-Distance Clustering

This table has portfolio results for entry-time sorted portfolios, with a clustering set-up that only clusters identical hedge funds, i.e., a threshold distance for clustering based on the 144 binary descriptors (see Appendix A) of 0 is used. As in Table 4, the portfolio return is equally-weighted, using the first 24 months of each fund to compute returns. Alpha' and 'R2' are the portfolio alpha and adjusted R-squared from a Fung and Hsieh (2004) 7-factor model. Newey-West corrected t-statistics in parentheses. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

	N	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis	Autocorr.	Alpha	R2
Q1	96.00	1.06 ***	1.33	-4.16	0.53	1.16	1.70	5.38	-0.34	3.00 ***	0.22 **	0.80 *** (5.48)	0.23
Q2	96.00	0.96 ***	1.41	-3.08	0.27	1.11	1.66	8.67	1.21 ***	8.95 ***	0.17 *	0.67 *** (4.51)	0.34
Q3	96.00	0.59 ***	1.23	-3.79	-0.01	0.75	1.37	3.32	-0.70 ***	1.29 **	0.23 **	0.34 *** (2.66)	0.23
Q4	96.00	0.55 ***	1.48	-6.39	-0.00	0.78	1.27	4.13	-1.52 ***	5.87 ***	0.40 ***	0.29 ** (2.24)	0.44
Q5	96.00	0.41 ***	1.50	-5.80	-0.25	0.61	1.31	3.23	-1.41 ***	4.24 ***	0.38 ***	0.14 (1.12)	0.40
No Cluster	96.00	0.89 ***	1.52	-4.44	0.03	1.11	1.92	4.14	-0.97 ***	1.51 **	0.29 ***	0.58 *** (6.89)	0.63
Q1-Q5	96.00	0.65 ***	1.49	-3.11	-0.18	0.54	1.36	5.45	0.75 ***	1.55 **	0.24 **	0.50 ** (2.28)	0.00
Q1-NC	96.00	0.17	1.19	-2.32	-0.48	0.08	0.75	4.39	0.75 ***	1.89 ***	0.32 ***	0.06 (0.49)	0.25
NC-Q5	96.00	0.48 ***	1.00	-1.98	-0.12	0.46	1.08	3.66	0.21	0.70	-0.04	0.27 ** (2.10)	0.20

Table 12

Limited Set of Clustering Variables

This table has portfolio results for entry-time sorted portfolios, with a clustering set-up that uses a subset of 129 variables out of possible 144 binary descriptors. We remove Derivatives, InvestsInManagedAccounts, OpenEnded, HighWaterMark, RegisteredInvestmentAdviser, FXCredit, Leveraged, Futures, Guaranteed, InvestsInOtherFunds, OpenToPublic, AcceptsManagedAccounts, Margin, CurrencyExposure, PersonalCapital characteristics. As in Table 4, the portfolio return is equally-weighted, using the first 24 months of each fund to compute returns. 'Alpha' and 'R2' are the portfolio alpha and adjusted R-squared from a Fung and Hsieh (2004) 7-factor model. Newey-West corrected t-statistics in parentheses. *, ** and *** denote significant differences from zero (or normality) at the 90%, 95% and 99%-level, respectively.

	N	Mean	Std	Min	25%	Median	75%	Max	Skewness	Kurtosis	Autocorr.	Alpha	R2
Q1	96.00	0.93 ***	1.25	-3.26	0.18	1.03	1.65	5.38	0.03	1.86 **	0.29 ***	0.65 *** (4.58)	0.12
Q2	96.00	0.90 ***	1.38	-3.69	-0.00	1.12	1.84	3.99	-0.55 **	0.58	0.11	0.67 *** (4.70)	0.19
Q3	96.00	0.71 ***	1.31	-3.45	0.02	0.85	1.58	3.56	-0.77 ***	1.05 *	0.23 **	0.43 *** (3.89)	0.31
Q4	96.00	0.45 ***	1.59	-6.54	-0.36	0.73	1.39	3.09	-1.67 ***	5.17 ***	0.36 ***	0.15 (1.26)	0.52
Q5	96.00	0.42 **	1.95	-6.21	-0.29	0.42	1.49	5.74	-0.76 ***	2.62 ***	0.22 **	0.08 (0.51)	0.33
No Cluster	96.00	0.92 ***	1.54	-4.55	0.10	1.20	1.89	4.42	-0.99 ***	1.75 **	0.30 ***	0.61 *** (7.73)	0.65
Q1-Q5	96.00	0.50 ***	1.73	-4.20	-0.35	0.31	1.36	5.20	0.10	1.24 **	0.19 *	0.40 ** (2.33)	0.12
Q1-NC	96.00	0.01	1.30	-4.56	-0.50	0.10	0.55	3.70	-0.20	2.06 ***	0.45 ***	-0.12 (-1.04)	0.36
NC-Q5	96.00	0.49 ***	1.31	-4.24	-0.09	0.39	1.18	5.47	-0.02	4.43 ***	-0.12	0.37 *** (3.14)	0.00

Table 13

Fund Families

This table compares the overlap between clusters as an output of the Fast Binary Clustering with maximum distance of 0.12 and classification based on fund family membership. The sample includes funds from all FBC clusters that are alive in the 2003–2010 period, as in Table 3. The column lablled 'FBC True, Family Candidate' assumes that FBC clusters are the true division of funds and is based on 2579 funds. The columns labelled 'Family True, FBC Candidate' assumes that the family membership is the true division and is based on 6141 funds in total. We report number of distinct true and candidate clusters to which all funds belong. Three entropy based measures of cluster quality are considered: homogeneity, completeness, and V-measure. Homogeneity is highest when observations from different true clusters are not grouped together by an algorithm. Completeness is highest when for each true cluster, all observations are grouped into a single cluster by an algorithm. The V-measure is a harmonic mean of the two other measures. Two normalized measures are reported: Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI). ARI measures similarity between the true and the generated partitions. AMI measures the agreement between the two labellings.

		FBC True, Family Candidate	Family True, FBC Candidate
	N True Clusters	157	905
	N Candidate Clusters	905	3719
Entropy based	Homogeneity	0.84	0.80
	Completeness	0.55	0.65
	V.	0.66	0.72
Normalized	ARI	0.05	0.03
	AMI	0.25	0.14

Table A1

Fast Binary Clustering Compared to Other Algorithms

Comparison between three clustering algorithms: DBSCAN (density based), k-Means (centroid, with k-Means++ seeding), and FBC (an agglomerative hybrid clustering algorithm combining hierarchical, centroid, and density-connected algorithms). Panels A and B differ in number of characteristics (NH) and number of clusters considered (NC). In both cases clusters comprise of the same number of observations. Each sample is randomly created by drawing a random matrix with NH rows and NC columns. In each column of this matrix a region is marked mutation-prone, on average 15% of NH characteristics are allowed to differ between each observation in the cluster and the randomness is limited to the mutation-prone region. The archetype columns are cloned 50 times each to create the cluster and for each of the 50NC columns mutation-prone regions are randomized. All observations belong to a cluster and there is no observation-noise. Where applicable different distance functions are considered. Input parameters for the clustering algorithms were chosen to reflect true, or close to true, levels. All values reported are normalized such that 1 represents the optimal level, and values farther away from unity show weakness of algorithms considered. We report the number of clusters created excluding observations considered noise as multiple of the true number of clusters. Time reported is a multiple of the minimum average time across the algorithms. Three entropy based measures of cluster quality are considered: homogeneity (homog.), completeness (compl.), and V-measure. These measures are not normalized with respect to random labelling which means they tend to be inflated for higher number of clusters. Homogeneity is highest when observations from different true clusters are not grouped together by an algorithm. Completeness is highest when for each true cluster, all observations are grouped into a single cluster by an algorithm. The V-measure is a harmonic mean of the two other measures. Two normalized measures are reported: Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI). ARI measures similarity between the true and the generated partitions. AMI measures the agreement between the two labellings. As these measures are adjusted to accommodate agreement due to chance, they are more reliable for higher number of clusters.

Panel A: 50 clusters, 2500 observations, 50 characteristics, 16348 replications

		DBSCAN						k-Means			FBC			
		city block		cosine		euclidean					city block		cosine	
		0.15	0.1	0.15	0.1	0.15	0.1	48	50	52	0.15	0.1	0.15	0.1
	N clusters	0.54 0.20	0.54 0.20	0.54 0.20	0.54 0.20	0.57 0.23	1.75 0.29	0.96 0.00	1.00 0.00	1.04 0.00	1.00 0.00	1.01 0.02	1.00 0.00	1.00 0.01
General	Time	1.13 0.64	1.13 0.64	1.21 0.61	1.18 0.64	1.15 0.65	1.14 0.64	1.11 0.64	1.11 0.61	1.17 0.58	1.04 0.60	1.00 0.67	1.11 0.61	1.05 0.60
	Homog.	0.08 0.04	0.08 0.04	0.08 0.04	0.08 0.04	0.09 0.04	0.36 0.10	0.99 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
	Compl.	0.65 0.04	0.65 0.04	0.65 0.04	0.65 0.04	0.64 0.04	0.69 0.05	1.00 0.00	1.00 0.00	0.99 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
Entropy based	V-measure	0.15 0.06	0.15 0.06	0.15 0.06	0.15 0.06	0.15 0.06	0.47 0.10	0.99 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
	ARI	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.02 0.02	0.96 0.00	0.99 0.01	0.99 0.00	1.00 0.00	0.99 0.01	1.00 0.00	1.00 0.00
Normalized	AMI	0.05 0.02	0.05 0.02	0.05 0.02	0.05 0.02	0.05 0.03	0.27 0.10	0.98 0.00	1.00 0.00	0.99 0.00	1.00 0.00	0.99 0.00	1.00 0.00	1.00 0.00

Panel B: 100 clusters, 5000 observations, 100 characteristics, 7459 replications

		DBSCAN						k-Means			FBC			
		city block		cosine		euclidean					city block		cosine	
		0.15	0.1	0.15	0.1	0.15	0.1	98	100	102	0.15	0.1	0.15	0.1
	N clusters	0.00 0.01	0.00 0.01	0.01 0.01	0.02 0.02	0.10 0.05	0.27 0.08	0.98 0.00	1.00 0.00	1.02 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
General	Time	1.02 0.58	1.03 0.58	1.02 0.58	1.02 0.59	1.01 0.58	1.02 0.58	1.00 0.58	1.00 0.59	1.02 0.58	1.02 0.59	1.01 0.59	1.02 0.59	1.01 0.58
	Homog.	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.02 0.01	0.07 0.02	0.99 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
	Compl.	0.95 0.13	0.95 0.13	0.92 0.16	0.82 0.18	0.73 0.09	0.77 0.05	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
Entropy based	V-measure	0.00 0.00	0.00 0.00	0.00 0.00	0.01 0.01	0.04 0.02	0.13 0.04	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
	ARI	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.98 0.01	0.99 0.01	0.99 0.01	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00
Normalized	AMI	0.00 0.00	0.00 0.00	0.00 0.00	0.00 0.00	0.01 0.01	0.04 0.02	0.99 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00	1.00 0.00