Hedge Fund Innovation*

Arjen Siegmann[1], Denitsa Stefanova[2], and Marcin Zamojski[3]

[1] *Department of Finance, Vrije Universiteit Amsterdam and Tinbergen Institute*

[2] *Luxembourg School of Finance, University of Luxembourg*

[3] *Centre for Finance, University of Gothenburg*

February 14, 2017

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.

*Corresponding author: Marcin Zamojski (marcin.zamojski@gu.se); Centre for Finance; University of Gothenburg; Box 640; 405 30 Gothenburg; Sweden. Marcin Zamojski thanks the Dutch National Science Foundation (NWO) for financial support. For useful comments and suggestions, we would like to thank Itzhak Ben-David, Mike Burkart, Susan Christoffersen, Joost Driessen, Petri Jylha, André Lucas, Bas Peeters, Marco Rossi, Clemens Sialm, Patrick Verwijmeren, two hedge fund insiders, and seminar participants at the Tinbergen Institute, Mathematical Finance Days 2013 at HEC Montreal, European Finance Association 2013 meeting, Luxembourg Asset Management Summit 2013, American Finance Association 2014 conference, and 2014 Yearly Cambridge, DSF-TI, Wharton Seminar.

1. Introduction

The specific fee structures and light regulation of the industry give hedge funds the opportunity to follow innovative investment strategies and to change them according to where profit opportunities arise. The excess returns and changing risk exposures as documented in the literature are a witness of their exceptional institutional structure. They can be seen as an innovative force in seeking out new profit opportunities and achieving diversification for investors. In this paper, we test whether the initially declared risk profile of a hedge fund—in terms of its investment focus and related characteristics—is a sign of innovation and an actionable piece of information for investors.

We identify hedge funds whose innovation leads to the creation of new product categories and compare them to other funds that attempt to imitate their approach. We are interested in identifying factors that motivate fund managers to innovate. We also look at whether some of the benefits of innovation are shared with investors. The benefits of early entry are a common object of study in the literature on industrial organisation. Therein, such benefits would be considered to reflect barriers to entry, see Tirole (1988).

In the case of hedge funds, no formal barriers to entry exist. However, the sophisticated nature and specialisation of hedge funds create specific *barriers to profit* from opportunities that have a tendency to diminish with the attention that they receive. We hypothesise that the type of market and investment segment cannot easily be changed by a running hedge fund and that they are in practice fixed at the inception of the fund. For instance, a combination of trading technology, local knowledge, and specific investment skills of staff are specific to certain investment styles. These create hurdles for existing hedge funds to copy from innovators. Furthermore, hedge funds operate in an environment where a high degree of connectedness and specialised knowledge of clients' needs is key. In this respect, hedge funds are similar to investment banks for which there is evidence of early-mover advantages (Tufano, 1989).

To form peer groups of hedge funds we use characteristics that are supplied by the funds when registering in the Lipper TASS database. They include information on the type of assets funds invest in (stocks, bonds, futures, etc.), their sector and investment focus (emerging markets, US equities, etc.), and other fund details (use of managed accounts,

use of leverage, etc.). Collectively, we refer to these characteristics as 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 before other funds that share a similar institutional design and risk profile. 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.

It follows that accurate identification of clusters is key. We develop an algorithm specifically for the purpose of the paper. The custom-made algorithm, which we call Fast Binary Clustering, is necessary because of high dimensionality of the dataset (144 binary variables on which to cluster). Existing algorithms are either not suitable to binary data or exhibit problems with the high dimensionality.

We look at first-mover advantages by comparing early entrants and their followers on a number of dimensions, including performance and pricing. 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. 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. We do not find evidence for benefits to late or delayed entry.

Second, we find evidence for differences in pricing between the innovators and the imitators. The earliest quintile of funds in a cluster charges significantly higher incentive fees. The effect on management fees is the opposite, innovators charge lower management fees than the followers.

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 captured by the standard risk factors. In contrast, if we extend the model by including the fifth quintile portfolio (Q5), the 'laggards', we are

not able to better explain out-performance of other hedge funds and the alphas do not decrease significantly. 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 the financial industry, such as investment banking, mutual funds and pension funds, 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 market share, but do not necessarily obtain higher margins or higher fees. Our evidence for higher incentive fees that are charged by early-movers, which is not found in the other industries, might reflect an effect of an optimal size of hedge funds, see Getmansky (2012). First-movers are aware of this effect and set their initial fees accordingly.

Another contribution of this 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.

The paper is also 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 (2015). 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 to invest

at 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 a lot of hedge funds try to imitate successful strategies and early movers are easily identifiable the systemic risk might increase. As the innovators and the imitators would hold similar positions and unwind them at the same time these risk are likely to materialise, see for example Khandani and Lo (2011), Aragon and Strahan (2012).

The remainder of the paper is structured as follows. Section 2 briefly describes the data we use. We discuss related literature and develop our hypotheses in Section 3. In Section 4, we discuss consistency of our dataset and whether it may affect our analysis. Section 5 introduces our clustering algorithm and explains entry-time variables which we use in our analysis. Section 6 presents the results and Section 7 tests for the robustness of the results. Section 8 concludes.

2. Data

We use static and monthly data from the Lipper TASS database. The database contains information for both defunct and currently operating funds. The sample period is from January 1994 to January 2012. The TASS is a commercial database to which reporting is voluntary and it is commonly used in the hedge fund literature. The sample consists of 16,051 hedge funds.

Table 1 shows summary statistics for hedge funds in the TASS database grouped by style. The largest group of hedge funds in the database, at 36%, are fund of funds. The second most popular category of funds engages in long/short equity hedging strategies (21%). Each of the remaining styles represents less than 11% of the sample. Note, that we do not include fund of funds in our analysis.

In addition to time-series of returns and values of assets under management, the TASS database provides three more tables that are relevant for our analysis. One table holds the product details of each fund, such as the name, primary category, currency, inception date, fee structure, and others. This information is used by most researchers when making selections of hedge funds or determining patterns in the cross-section of hedge funds. We use this data to compare funds that we have labelled as innovators with other funds that

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. We report median values in the cross-section. '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.

		AUM				Fees		Indvidual returns					Portfolio returns				
	N	Initial	Mean	Max	Alive	Inc.	Man.	Median	Mean	Std. dev.	Skew.	Kurt.	Median	Mean	Std. dev.	Skew.	Kurt.
All	15961	5	13	23	39	20	1.5	0.6	0.46	2.47	−0.44	1.57 *	0.88	0.85 ***	1.64	−0.4	2.45 ***
Fund of Funds	5791	7	14	22	41	10	1.5	0.51	0.27	1.98	−0.89	1.98 **	0.69	0.57 ***	1.58	−0.57	2.96 ***
L/S Eq. Hedge	3452	3	13	22	32	20	1.5	0.68	0.66 **	3.52	−0.13	1.2 **	1.19	1.18 ***	2.56	0.08	2.0 ***
Multi-Strategy	1700	5	10	16	58	15	1.5	0.77	0.69 *	1.99	−0.34	1.7 **	0.98	0.9 ***	1.4	−0.74	3.36 ***
Em. Mkts	910	5	15	26	47	20	1.8	1.0	0.88 ***	4.93	−0.25	1.75 **	1.79	1.21 ***	4.24	−0.89	3.7 ***
Managed Futures	830	2	8	14	42	20	2.0	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	1.5	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	1.5	0.79	0.7 ***	2.41	−0.38	2.29 **	1.21	0.94 ***	1.68	−1.62	6.26 ***
Eq. Mkt Neutral	637	4	15	26	24	20	1.5	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	1.5	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	1.5	0.65	0.5 ***	1.73	−0.47	2.44 **	0.87	0.76 ***	1.15	−3.39	21.1 ***
Convertible Arb.	241	6	37	68	22	20	1.5	0.7	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	1.3	0.07	0.29	5.6	0.13	0.86	−0.06	0.35	4.83	0.42	2.01 ***
Options Strategy	37	3	5	12	38	20	1.5	0.46	0.59	3.22	−0.09	6.98 ***	0.61	0.65 ***	0.98	0.44	1.1 **

we believe to be copy-cats.

The other two tables contain detailed information on a hedge fund's focus and investment approach. This data is a result of a questionnaire that hedge funds fill in when they are registering in the database. In total, there are 144 yes/no questions which hedge funds are asked to answer to help investors in narrowing down their risk profile. The questions can be grouped into thre groups: (i) details on instruments used by the funds, (ii) additional details about the investment approach, and (iii) other details. The first group is further divided based on asset type (equities, fixed income, commodities, currencies, property, etc.) and instrument type (cash, convertible, exchange traded, etc.). Within investment details, hedge funds are asked to specify their sectoral focus (biotechnology, natural resources, closed-end funds, corporate bonds, etc.), investment approach (arbi-

trage, bottum-up, contrarian, etc.), geographic focus (Africa, Asia-Pacific, India, etc.), and investment focus (bankruptcy, distressed bonds, pairs trading, etc.). The complete list of variables is in Appendix A. To our knowledge, this part of the TASS database has not been used previously in the literature. We use this data to infer both the institutional design and the initial risk profile of hedge funds. The 144 variables are then the inputs in our clustering algorithm that we use to divide hedge funds into innovators and imitators.

3. Related literature and hypothesis development

There is agreement in the literature on the fact that some funds show persistent out-performance (Jagannathan, Malakhov, and Novikov, 2010). 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). Other factors which help explain cross-sectional differences in hedge fund performance include manager characteristics (Boyson, 2010), strategy distinctiveness (Sun, Wang, and Zheng, 2012), incentives (Agarwal, Daniel, and Naik, 2009), and even the geographical location of the fund (Teo, 2009). Even though exposures of hedge funds to standard risk factors are typically time-varying, there is also evidence that the managers are consistent in their chosen investment approach and style. For fixed-income hedge funds, Fung and Hsieh (2002) find a static exposure to fixed-income related spreads. For market timing hedge funds, Chen and Liang (2007) find persistence in timing ability. Fung and Hsieh (2001) find that lookback straddle benchmarks explain trend following funds' returns. Moreover, non-linear exposures and liquidity timing effects have predictive power for individual hedge fund returns, see Agarwal and Naik (2004). This is highly suggestive of hedge funds being consistent in their investment approaches and, by extension, risk profiles.

Furthermore, from a legal perspective, hedge funds are bound to stick to an investment approach that they have described in the private placement memorandum (PPM) that is provided to potential investors. 'A PPM is a widely utilised form of disclosure which contains the type of information that would be provided by a registration statement publicly filed under section 5 of the Securities Act, along with the unique facts and circumstances surrounding the fund', see Shadab (2009) and SEC (2003).[1] The informa-

[1] Furthermore, according to SEC (2003), PPMs generally discuss broadly the fund's investment strategies and practices. It may also include a disclosure that some investments may be done outside the stated

tion used for registering with a hedge fund database, such as the TASS, is likely to be similar. Moreover, 'as a matter of law and practice, the funds typically make disclosures sufficient for investors to make informed investment decisions'. Alternatively, hedge fund managers may remain vague about their strategy and risk attracting fewer risk-averse investors. A change in an investment approach would typically be an event for which additional disclosures are warranted and require agreement from all investors. In fact, it may be easier for a hedge fund manager to open a new investment pool and possibly liquidate the old fund than to implement considerable changes in the older fund's risk profile/institutional design.

That being said, we see the changing risk profile of a representative, *newly created* fund as the first piece of evidence of the ongoing innovation and the flexible nature of the whole industry. Consider a simple breakdown of new hedge funds by their style over time in Panel A of Figure 1. We clearly see that popularity of styles varies in time. New funds with distinctive risk profiles are created in reaction to (or in anticipation of) changing market conditions. As an example, Equity Market Neutral was a popular strategy between 1998 and 2004. The Managed Futures category was popular in the 1990s and is only recently regaining the spotlight. Interestingly, Long/Short Equity Hedge, which is both the most populous and the most studied style category, has been decreasing in popularity since the turn of the century—a signal that generating high absolute returns in this crowded segment is becoming increasingly difficult. Instead, we observe a quick rise in popularity of the Multi-Strategy category which became the most commonly declared style in the late 2000s.

Style categories convey only a limited amount of information about a hedge fund's risk profile. For one, hedge funds are assigned to a single style category even if their actual approach is consistent with more than one style. Secondly, even within a single style category hedge funds were likely to pursue different strategies early on in the sample period than later on. We can get a better understanding of a hedge fund's risk profile by looking at the set of 144 other descriptors which we discussed in Section 2. In contrast to style categories, a hedge fund manger is free to select as many (or as few) descriptors as they want. In this way, hedge funds can convey a very detailed description of their risk profile or purposefully remain vague to investors. Similarly to styles, we plot popularity

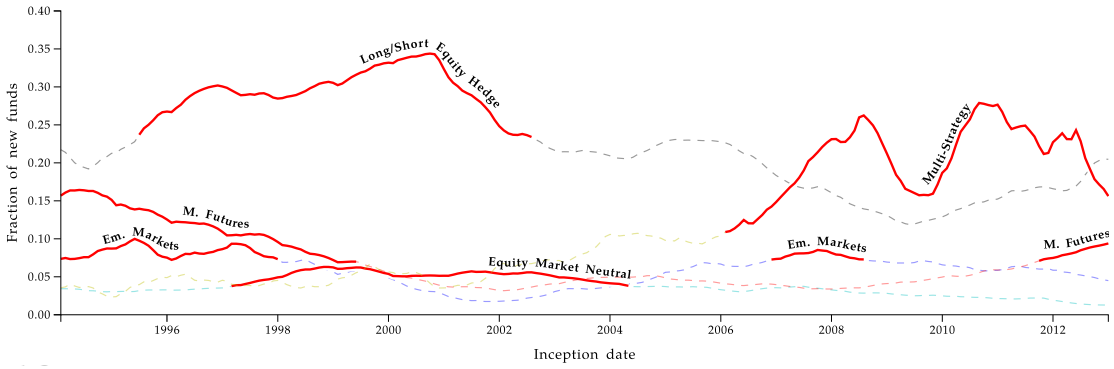
strategy at the discretion of the fund's manager.

Figure 1

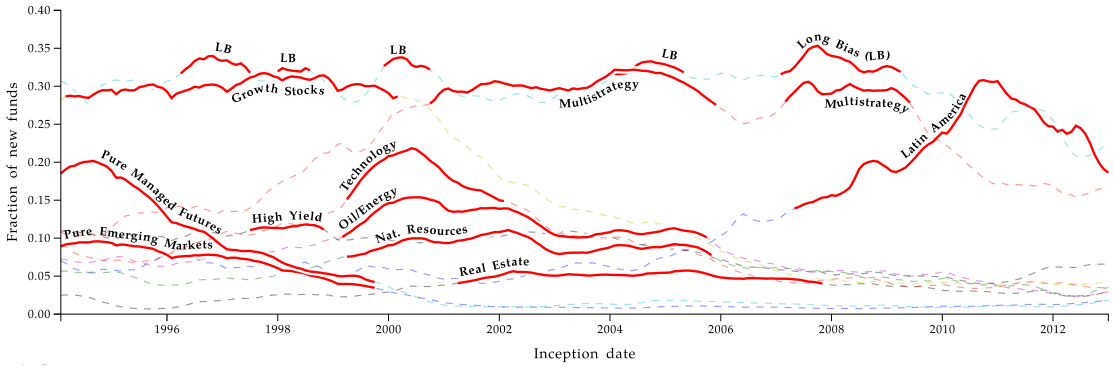
Popularity of Styles and Descriptors

In this figure we show changes in time of risk profiles of new hedge funds. In Panel A, we look at selected styles. In Panel B, we look at a selection of other descriptors (investment focus, approach, geographical focus, etc.). In our dataset, there are in total 144 such descriptors. We use them later on in our analysis to construct clusters of similar hedge funds. In each case, we plot the fraction of new funds which match a style or descriptor (dashed lines). For clarity, we highlight in red, solid lines periods in which styles or descriptors were the most popular.

Panel A: Selected styles



Panel B: Selected descriptors



of a selection of such descriptors in Panel B of Figure 1. We again find considerable fluctuations in popularity of individual descriptors.

Our main hypothesis of the benefits to innovation follows naturally. The cycles in popularity of styles and the descriptors correspond to underlying strategies gaining and losing popularity over time. Innovating and maintaining the potential for innovation usually comes at a huge cost—both material and intellectual (Cohen and Levinthal, 1990). Given that we observe signs of innovation in the industry, managers need to benefit greatly

from it. With the hedge fund industry being largely return-driven, Figure 1 is suggestive of innovators earning high returns, which brings in similar funds that are attracted by the good performance and the associated high income from incentive fees. When the performance starts decreasing, other particular strategies or setups gain the attention of investors and managers. The short lifespan and high turnover of hedge funds is one indication of such a premium to innovation. The median lifespan of a hedge fund in our dataset is approximately 5 years.[2] In the period 1994–2012, the rates of entry and exit were on average around 30% and 13% per year respectively.[3]

In this paper, we divide hedge funds into innovators and imitators based on the detailed risk profile inferred from the 144 binary descriptors. To reiterate, these include information on the funds' investment focus, geographic focus, instruments used, and investment approach. The risk profile is set at the early stage of a funds life and is reported to the data provider with the initial application. Funds only rarely make any adjustments to this part of their profile in the TASS database, see Section 4 for detailed discussion of data consistency. Before we are able to label funds as innovators or otherwise, we cluster funds into peer groups. Clustering allows us to essentially look at intersections of many characteristics rather than analyse them one-by-one.

Several other papers provide motivation for our hypothesis that profitable hedge fund innovation is the driving force of entry. Sun, Wang, and Zheng (2012) finds that funds that are more distinct in terms of returns out-perform in the subsequent year by as much as 3.5 p.p., relative to the least distinct hedge funds. Titman and Tiu (2011) find that hedge funds with a low R-squared in a factor model have a better performance. Thus, an ability to differentiate their investment approach from comparable hedge funds gives a hedge fund's manager a competitive edge. The differentiation can occur already at the moment of inception and we posit that the initial risk profile of a hedge fund is enough to classify the fund as an innovator or imitator.

Naik, Ramadorai, and Stromqvist (2007) find that hedge funds' alpha returns are

[2] Median lifespan for funds created in the 1990s, 2000–2008, and after 2008 are 9, 6, and 3 years respectively.

[3] Prior to the 2008 financial crisis these rates of entry and exit were around 30% and 6%. After the crisis, we see a considerable decrease in the inflow of new funds and considerable increase in the funds dropping out—the entry rate is at 10% while 17% of funds exit every year. Exits could be overstated by funds that stop reporting for reasons other than liquidation. The short lifespan and high incentive fees can be seen as characteristic for a highly competitive industry in which innovation and being 'first' is key.

sensitive to capital inflows in the previous month. When inflows are higher, the potential for hedge funds to create an excess performance diminishes. This provides an argument that competition within segments of hedge funds exists and that it leads to lower returns over time. In the same vein, Getmansky (2012) finds that hedge funds have a larger probability of being liquidated when the particular segment in which they are active grows. Competition drives down returns due to decreasing returns to scale, crowded trades, and the fact that inefficiencies targeted by hedge funds are only transient and disappear when more arbitrage capital is allocated to them. Innovative hedge funds, being first in a particular segment, could reap the benefits from the absence of competition. This leads to our first hypothesis:

H1: Hedge funds that come early in a cluster of similar hedge funds perform better.

We also expect the returns of hedge funds that come later not to offer the same level of performance despite being correlated with those of the innovators. The alternative hypothesis is that hedge funds are given enough flexibility and discretion to quickly change their investment strategies so that the initial risk profiles are not informative about their performance. If this is the case, we expect to find no evidence of any benefits from early entry for both the managers and the investors. Benefits to innovation may also be delayed. For example, Christensen, Suárez, and Utterback (1998) consider disadvantages of early-entry, when there is a learning window before a final specification of a product (dominant design) is determined. We test for this as well.

Furthermore, we expect that after an initial period of superior performance, the returns decline on average. This is due to the hyper-competitive nature of the industry, well-documented problems with strategy scalability (Naik, Ramadorai, and Stromqvist, 2007; Getmansky, 2012), and temporary inefficiencies that are underlying profitable hedge fund strategies. For example, a profitable opportunity in an emerging market would attract investors who trade similarly, raising prices and depressing returns. The innovative managers are unlikely to be unaffected by an increase in competition in their segment which leads to our second hypothesis:

H2: The better performance of innovators declines over time.

Innovative hedge fund managers are also likely to be affected by low familiarity of potential investors with their new strategies. As in the case of innovative investment banks, they may need to under-price new products to attract customers (Tufano, 1989). In particular, we expect the managers to lower barriers to entry and exit from the funds. This would be reflected in shorter lock-up and redemption periods. We also expect innovative managers to put more of their skin in the game and signal conviction about the profitability of their new strategy. This would correspond to lower management fees, higher incentive fees, higher personal capital invested in the fund, and lower leverage. These institutional features are not used at the clustering stage. We can, thus, test whether hedge funds that enter early are different from others in these respects. Our third hypothesis is then:

H3: Innovative hedge funds have features that are related to high-risk innovative strategies, such as lower leverage and higher personal capital, distinctive fee structure, or lower lock-up periods.

4. Data consistency

Our interpretation of the clustering results and the subsequent separation of hedge funds into innovators and imitators depend on our ability to reasonably identify initial risk profiles of the funds. By extension, we require that the 144 binary descriptors from the initial questionnaire are not restated, although small changes are not likely to affect clustering results. In this section, we briefly discuss consistency of this part of the dataset between a number of vintages.

We are aware of the fact that many data points in the TASS database have been restated by hedge funds. For instance Patton, Ramadorai, and Streatfield (2015) find that 49% of funds have changed at least one return. In this paper, we have assumed that the 144 descriptors we use to cluster hedge funds are constant and represent the initial vision of a fund's strategy. To verify if this is a reasonable assumption, we compare the data in our baseline database vintage to three older vintages, the earliest of which is from 2005. First, we are able to corroborate the findings of Patton, Ramadorai, and Streatfield (2015). In our case, a similarly large fraction of hedge funds (41%) restated at least one return. Furthermore, we find that 6.4% of funds changed their incentive fees at

least once, 2.98% made adjustments to their management fees, and 2.24% modified their payout period. Surprisingly, at least 5%[4] of hedge funds restated their primary product category, i.e., the 'style'. Note, that all these variables are routinely used in the hedge fund literature.

Although we do find that the 144 descriptors we use in clustering have also been restated, this has been done to a significantly lesser extent. For instance, the most commonly changed descriptor indicating whether a fund uses leverage was edited by only 2.1% of hedge funds. The 10th most changed descriptor (fundamental investment approach) was modified by only 1.2% of funds. Finally, any issue in consistency of the data can be attributed to a tiny fraction of hedge funds, as approximately 2% were responsible for over 60% of all changes to the 144 descriptors.

Another potential problem stems from the fact the TASS could have changed the questionnaire between 1994 and 2012. In particular, some of the 144 descriptors might not have been tracked in the earlier vintages. Fortunately, it appears that even very early in our sample period all 144 descriptors were present in the database. To verify that, we look at liquidation dates of hedge funds. For each of the descriptors, we find all defunct funds for which the particular descriptor was set to 1. We then assume that a descriptor was present in the database at the moment of the earliest liquidation. We reason that it is highly unlikely for a manager of a (long) defunct fund to adjust such information in the TASS database. Based on this analysis, we find that at least 75% of descriptors were present in the database at the beginning of 1997. By 2002, we can verify presence of 94% of the descriptors. Surprisingly, even though the currently very popular 'Multi-strategy' style was still not present in the 2005 vintage of the database, the hedge funds could indicate that they had a multi-strategy investment focus at least as far back as 1999.

We conclude that the 144 descriptors we consider can be used to infer funds' initial risk profiles with enough accuracy to use this data in our analysis.

[4] The TASS database introduced the 'Multi-strategy' style category sometime between 2005 and 2009. The fraction of hedge funds which were operating before 2009 and which changed their product category is 6.2%. Note, however, that 5.2% of hedge funds that were set up after this category was introduced have also changed their style. The descriptor 'multi-strategy investment focus' predates the style category.

5. Clustering and entry-time variables

5.1. 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. These variables are provided by the TASS database and describe the strategy and design of the hedge fund.

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 hierarchical, 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 2.

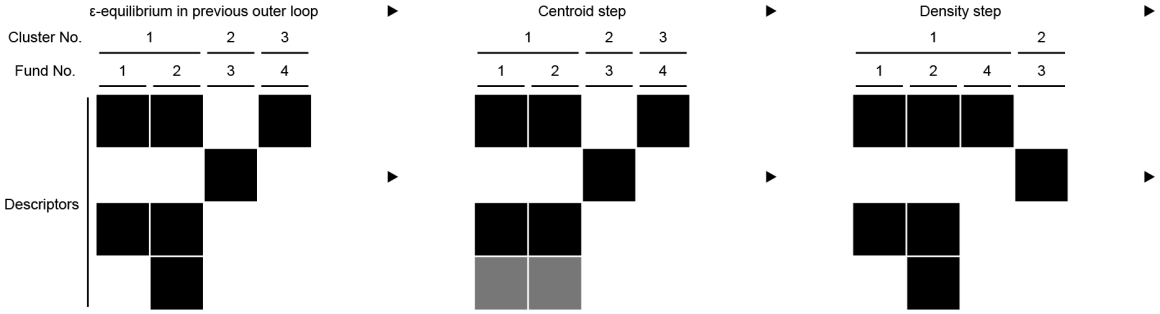
For the type of data we use, the clustering results are most affected by one input variable—the maximum distance between two clusters and/or funds to be clustered together. In the following, we work with clusters based on a maximum distance of 0.12,

Figure 2

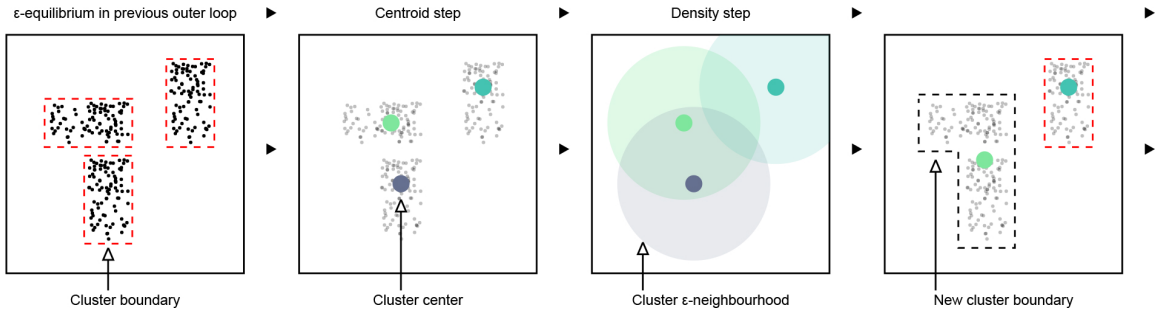
Illustration of Fast Binary Clustering

We show two examples of our FBC method 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



which leads to clusters with good properties from a clustering perspective[5]. We assess the sensitivity of our results to the distance parameter in Section 7.3.

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[6] 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.

[5]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.

[6]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.

To see more clearly how we operationalise this analysis, consider Figure 3. The clustering takes place in Step 1. We take all hedge funds available in the database and we group them into clusters based solely on the descriptors. Importantly, the clustering algorithm does not take hedge funds' entry times as an input. Consequently, the fact that we are able to obtain clusters composed of funds that entered the market at similar times is the first sign our approach is able to identify meaningful and narrow product categories. Note that it is typical in the literature on first-mover advantages to only consider markets or product categories (in our case clusters) that have already formed and where there were enough players for the analysis to be meaningful. See, for example, Tufano (1989, 2003); Utterback and Suárez (1993). In addition, we consider the unclustered funds as a special category. This allows us to take a position on whether what we observe is a story of: i) skilled hedge fund managers creating new markets (innovating) in which case the unclustered funds and first-movers should be able to generate a higher alpha. Or/and, ii) destructive impact of followers who, when they are able to replicate a strategy, through competition remove alpha from the market for all players.

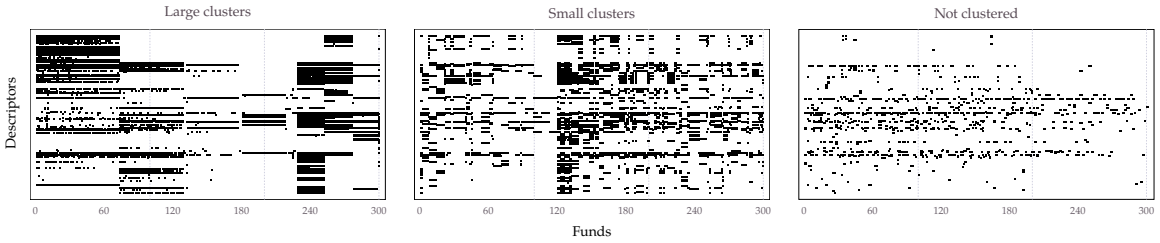
In step 2 of the analysis, we reintroduce the time dimension (see Figure 3) given the obtained clusters. This provides us with the first test of whether our approach is valid. Intuitively, if a strategy *sensu largo* is recoverable from the binary descriptors we expect the clusters to be: i) relatively short in time in the sense that entry is concentrated around a specific point in time (e.g., in 2006) and not scattered across the time line (e.g., with some funds entering in 1994 and others in 2007); and ii) clusters define meaningful markets. Table 2 has the summary statistics of the resulting clusters of hedge funds, reported at clusters' inception years. 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 which we find tight enough. Furthermore, looking at specific clusters we see clear and familiar patterns, e.g.: popularity of bio-technology and IT firms in the late 1990s, rise of the real estate markets in mid 2000s, and increasing popularity of Latin American hedge funds through the last decade.

Figure 3

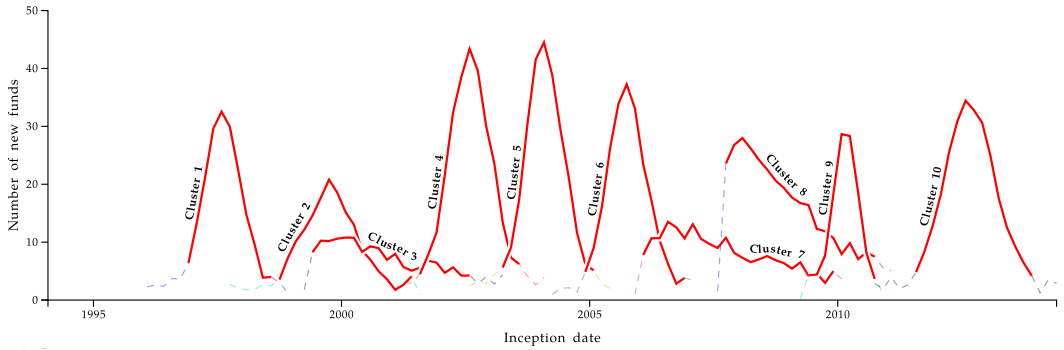
Clustering and Quintile Definition

In this figure we graphically present the steps we take to divide hedge funds into innovators and imitators. In Step 1, we group hedge funds in clusters based on the binary descriptors of their strategies. At this stage, we do not consider when a particular hedge fund was created. We present three slices of 300 funds from the TASS universe. The funds and the 144 descriptors are plotted on the horizontal and vertical axes, respectively. A black dot means that for a fund X, the descriptor Y is active/ticked off. The left-most figure contains 5 large clusters. We follow with smaller clusters, and funds which were not clustered in the right-most figure. Given the clusters, we look at the inception dates in Step 2. In Step 3, we divide hedge funds within a cluster into early- and late-movers. Panel A contains our actual clustering results. For clarity of exposition, we use simulated data in Panels B and C.

Panel A: Step 1—clusters



Panel B: Step 2—we reintroduce the time dimension



Panel C: Step 3—we divide each cluster into quintiles (here, for four clusters only)

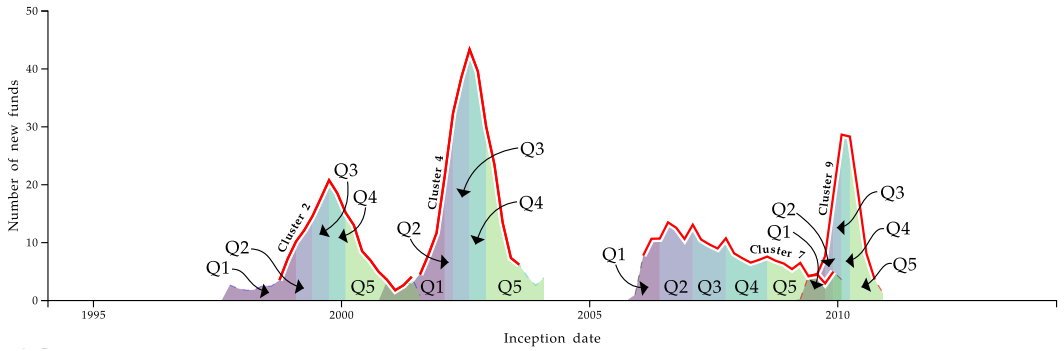


Table 2

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 new			Cluster size		Cl. time span			
	Clusters	Funds	'No Cluster'	Mean	Median	Mean	Median	ANOVA F-test	Kruskal-Wallis
1994	9	483	194	53.67	41.0	137.22	146.0	3.51 ***	49 ***
1995	8	734	201	91.75	12.5	124.25	143.0	4.28 ***	38 ***
1996	6	248	264	41.33	5.0	102.5	103.0	2.78 **	37 ***
1997	2	13	303	6.5	6.5	53.5	53.5	1.66	4 *
1998	7	88	284	12.57	8.0	115.57	132.0	5.94 ***	19 ***
1999	8	51	369	6.38	6.0	75.0	72.0	1.46	24 ***
2000	14	122	403	8.71	7.0	79.93	84.0	3.77 ***	54 ***
2001	11	94	469	8.55	6.0	67.18	63.0	0.96	18 *
2002	13	125	525	9.62	6.0	67.38	67.0	0.42	17
2003	12	94	621	7.83	8.0	32.0	27.0	1.06	24 **
2004	17	142	734	8.35	7.0	33.94	23.0	1.82 **	49 ***
2005	16	174	749	10.88	7.5	33.13	26.5	1.59 *	32 ***
2006	9	69	713	7.67	6.0	16.22	9.0	2.8 ***	17 **
2007	20	151	574	7.55	6.0	21.0	23.0	1.21	30 **
2008	8	63	484	7.88	5.5	14.75	11.5	2.81 ***	16 **
2009	9	55	405	6.11	6.0	9.44	5.0	2.73 ***	14 *
2010	3	18	268	6.0	5.0	5.0	4.0	1.89	2
All	172	2724	7560	15.84	7.0	54.47	38.0	2.61 ***	564 ***
All 2003+	94	766	4548	8.15	6.0	24.2	19.0	2.21 ***	221 ***

Finally, in step 3 of the analysis we divide funds into quintiles based on the chosen anchor point. For instance, in Panel C of Figure 3 we look at early-movers and the anchor point is chosen to be the date of first entry.

A second 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 2. For most years, the F-statistics are high and significant, which gives some indication that our clustering method 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.

6. 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.

6.1. 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 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

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. R^2	Survival
Q1	0.63 ***	0.85	1.3	−9.66 ***	128.36 ***	262	0.64 ***	0.18	38.55
Q2	0.53 ***	0.59	0.72	−0.38 ***	6.26 ***	394	0.4 ***	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	−5.9 ***	77.08 ***	429	0.32 ***	0.24	21.92
No Cluster	0.47 ***	0.44	1.23	1.1 ***	38.31 ***	4548	0.4 ***	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	0.0	14.94 ***

significance levels for differences in means from the first quintile of average returns, alpha and duration. The results are shown in Table 3.

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 (2015); Boyson

Table 3

Quintiles of Hedge Funds for Different Anchor Points (continued)

Panel B: Max Growth as the anchor point

	Mean	Median	Std. dev.	Skewness	Kurtosis	N	Alpha	Adj. R^2	Survival
Q1	0.43 ***	0.61	1.38	−3.02 ***	30.99 ***	360	0.43 ***	0.21	25.5
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	10.8 ***	430	0.39 ***	0.21	33.72
Q5	0.44 ***	0.69	1.27	−7.36 ***	95.16 ***	378	0.5 ***	0.19	34.81
No Cluster	0.47 ***	0.44	1.23	1.1 ***	38.31 ***	4548	0.4 ***	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. R^2	Survival
Q1	0.48 ***	0.78	1.66	−7.02 ***	75.28 ***	450	0.52 ***	0.22	30.88
Q2	0.43 ***	0.5	1.05	−1.19 ***	11.14 ***	572	0.34 ***	0.22	26.49
Q3	0.42 ***	0.46	1.03	−0.06	8.9 ***	558	0.31 ***	0.2	35.41
Q4	0.48 ***	0.59	1.13	−3.81 ***	48.91 ***	459	0.4 ***	0.21	40.76
Q5	0.45 ***	0.62	0.91	−1.77 ***	3.97 ***	112	0.54 ***	0.3	10.33
No Cluster	0.47 ***	0.44	1.23	1.1 ***	38.31 ***	4548	0.4 ***	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 ***

(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.

6.2. 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[7]. 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.

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 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.

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

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. dev.	Min.	25%	Median	75%	Max.	Skew.	Kurt.	Autocorr.	Alpha	Adj. R^2
Q1	96	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	0.93 ***	1.28	−2.57	0.25	1.11	1.69	5.56	−0.1	1.42 **	0.23 **	0.66 (4.89) ***	0.33
Q3	96	0.7 ***	1.3	−3.15	0.08	0.78	1.52	3.46	−0.66 ***	0.8	0.24 **	0.43 (3.88) ***	0.29
Q4	96	0.47 ***	1.57	−6.38	−0.2	0.69	1.44	3.25	−1.54 ***	4.76 ***	0.35 ***	0.16 (1.31)	0.5
Q5	96	0.39 *	2.27	−7.02	−0.33	0.72	1.47	6.4	−0.83 ***	2.84 ***	0.18 *	0.03 (0.15)	0.25
No Cluster	96	0.9 ***	1.53	−4.59	0.09	1.15	1.9	4.32	−1.0 ***	1.73 **	0.29 ***	0.6 (7.54) ***	0.65
Q1-Q5	96	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	0.04	1.28	−4.45	−0.5	0.1	0.56	3.67	−0.23	1.99 ***	0.45 ***	−0.11 (−0.87)	0.36
NC-Q5	96	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

6.3. Characteristics of Innovation Quintiles

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

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

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 fee	15.25	13.89	15.24	13.74	12.84	17.72	2.41 ***	−2.48 ***	4.88 ***
Management fee	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. leverage	96.98	90.35	156.04	83.81	170.06	119.15	−73.08	−22.17	−50.91
Avg. leverage	1.67	33.56	20.19	19.52	16.91	44.3	−15.24 **	−42.63 ***	27.39 ***
Initial AUM (mean)	14.63	79.62	51.29	245.35	19.88	25.14	−5.25	−10.51 **	5.26
Initial AUM (median)	5.9	14.95	5.75	4.84	5.77	6.96			
Lock-up period	0.78	0.78	1.42	1.43	0.94	3.4	−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	0.0	0.0	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	0.0	−0.19 ***	0.19 ***
Minimum investment	0.13	0.92	0.63	0.79	0.77	1.3	−0.64 ***	−1.17 ***	0.53
High water mark	0.28	0.31	0.35	0.36	0.38	0.79	−0.1 ***	−0.51 ***	0.41 ***

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

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: Fung and Hsieh (2004) with Q1 portfolio

	Q2	Q3	Q4	Q5	No Cluster	NC-Q5
Alpha	0.26	0.14	−0.14	−0.38	0.26	0.5
	(1.65)	(0.80)	(−0.76)	(−1.46)	(2.20)	(2.47)
	*				**	**
F_{Q1}	0.49	0.35	0.37	0.5	0.41	−0.11
	(6.16)	(2.51)	(3.03)	(2.77)	(4.41)	(−0.86)
	***	***	***	***	***	
Adj. R^2	0.53	0.38	0.57	0.31	0.74	−0.02

Panel B: Fung and Hsieh (2004) with Q5 portfolio

	Q1	Q2	Q3	Q4	No Cluster	Q1-NC
Alpha	0.62	0.61	0.37	0.1	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)
	***	***	***	***	***	*
Adj. R^2	0.19	0.51	0.53	0.66	0.76	0.37

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.

6.4. 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.

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 analysising 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.

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.

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	Adj. R^2	Alpha	F_{Q1}	Adj. R^2	Alpha	F_{Q5}	Adj. R^2
All	0.57	0.58	0.25	0.4	0.69	0.52	0.28	0.74
	(6.94)		(2.01)	(3.85)		(8.11)	(3.31)	
	***		**	***		***	***	
Long/Short Eq. Hedge	0.52	0.64	0.15	0.46	0.73	0.47	0.29	0.74
	(4.78)		(0.97)	(3.97)		(4.49)	(2.98)	
	***			***		***	***	
Fund of Funds	0.03	0.5	−0.39	0.51	0.66	−0.03	0.32	0.68
	(0.23)		(−3.46)	(4.86)		(−0.34)	(3.13)	
			***	***			***	
Multi-Strategy	0.59	0.48	0.33	0.32	0.6	0.54	0.24	0.68
	(8.02)		(2.88)	(3.04)		(9.14)	(3.22)	
	***		***	***		***	***	
Emerging Markets	0.93	0.58	0.27	0.81	0.69	0.84	0.43	0.66
	(4.07)		(1.20)	(3.78)		(4.20)	(3.12)	
	***			***		***	***	
Managed Futures	0.54	0.2	0.12	0.52	0.28	0.43	0.62	0.52
	(2.83)		(0.47)	(3.39)		(2.71)	(6.35)	
	***			***		***	***	
Global Macro	0.6	0.21	0.31	0.36	0.35	0.56	0.22	0.37
	(5.91)		(2.60)	(3.88)		(5.27)	(3.10)	
	***		**	***		***	***	
Event Driven	0.68	0.65	0.33	0.43	0.73	0.65	0.18	0.68
	(5.45)		(2.45)	(6.40)		(5.12)	(2.21)	
	***		**	***		***	**	
Eq. Market Neutral	0.27	0.49	0.17	0.13	0.51	0.25	0.12	0.55
	(5.06)		(2.33)	(2.53)		(4.85)	(2.31)	
	***		**	**		***	**	
Other	0.7	0.46	0.63	0.09	0.47	0.69	0.09	0.49
	(6.42)		(5.57)	(1.64)		(6.45)	(2.02)	
	***		***			***	**	
Fixed Income Arb.	0.47	0.33	0.48	−0.02	0.32	0.46	0.04	0.33
	(6.58)		(5.03)	(−0.39)		(6.54)	(1.42)	
	***		***			***		
Convertible Arb.	0.29	0.67	−0.02	0.38	0.69	0.23	0.28	0.71
	(1.51)		(−0.08)	(2.96)		(1.33)	(2.16)	
				***			**	
Ded. Short Bias	0.4	0.78	0.17	0.28	0.78	0.36	0.17	0.78
	(2.26)		(0.77)	(2.00)		(1.97)	(2.03)	
	**			**		*	**	
Options Strategy	0.76	0.13	1.03	−0.32	0.15	0.8	−0.17	0.14
	(4.36)		(3.50)	(−1.54)		(4.33)	(−1.11)	
	***		***			***		

6.5. 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 theorised 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 (2015) 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.

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

Table 8

Panel and Cross-Sectional Regressions

We regress hedge fund abnormal returns on fund characteristics. The columns labelled 'Panel' denote random effects GLS regressions with per fund clustered standard errors. We also run Fama and MacBeth (1973) regressions to estimate 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.

	(1)		(2)		(2)	
	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	−0.1	−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	0.1	0.07	0.1	0.06
	(1.94)	(2.89)	(2.47)	(3.41)	(2.42)	(3.19)
	*	***	**	***	**	***
Incentive Fee	0.02	0.0	0.02	0.0	0.02	0.0
	(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	0.0	0.01	0.0	0.01	0.0
	(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	−0.2	−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	0.5	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)
		***		***		***

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.

7. Robustness

7.1. 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 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[8]. Table 9 has the results for excess returns and loadings on the early entrants.

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.

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

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. dev.	Min.	25%	Median	75%	Max.	Skew.	Kurt.	Autocorr.	Alpha	Adj. R^2
Q1	96	0.89 ***	1.01	−1.58	0.22	0.9	1.48	4.66	0.28	1.15 *	0.18 *	0.7 (5.90) ***	0.12
Q2	96	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	0.51 ***	1.56	−6.6	−0.02	0.79	1.38	3.32	−1.82 ***	5.84 ***	0.4 ***	0.26 (2.20) **	0.51
Q4	96	0.49 **	1.91	−9.19	−0.2	0.66	1.75	3.7	−1.95 ***	7.16 ***	0.42 ***	0.16 (1.09)	0.57
Q5	96	0.45 *	2.45	−7.96	−0.45	0.81	1.63	10.1	−0.48 *	3.89 ***	0.23 **	0.07 (0.36)	0.38
No Cluster	96	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	0.44 **	2.15	−7.59	−0.54	0.27	1.17	7.0	0.43 *	3.18 ***	0.28 ***	0.46 (2.40) **	0.3
Q1-NC	96	0.16	1.4	−3.78	−0.45	0.02	0.48	6.24	1.18 ***	5.11 ***	0.5 ***	0.12 (1.03)	0.64
NC-Q5	96	0.28 *	1.59	−7.44	−0.46	0.27	1.14	7.0	−0.27	7.94 ***	0.09	0.18 (1.09)	−0.05

7.2. 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).

An alternative explanation of our results is that hedge fund innovators are the first

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: Fung and Hsieh (2004) 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	0.6	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: Fung and Hsieh (2004) 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	0.5	0.25	0.65	0.09	0.37	−0.02

Panel C: Fung and Hsieh (2004) 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. Mkts.	0.08	0.09	0.13	0.14	0.17	0.14	−0.1	−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	0.6	0.32	0.76	0.1	0.39	−0.02

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 timing, we include the Pastor and Stambaugh (2003) liquidity factor as well as

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. dev.	Min.	25%	Median	75%	Max.	Skew.	Kurt.	Autocorr.		Alpha	Adj. R^2
Q1	96	1.06	1.33 ***	−4.16	0.53	1.16	1.7	5.38	−0.34	3.0	0.22 ***	0.8 **	0.23 (5.48) ***	
Q2	96	0.96	1.41 ***	−3.08	0.27	1.11	1.66	8.67	1.21	8.95 ***	0.17 ***	0.67 *	0.34 (4.51) ***	
Q3	96	0.59	1.23 ***	−3.79	−0.01	0.75	1.37	3.32	−0.7	1.29 ***	0.23 **	0.34 **	0.23 (2.66) ***	
Q4	96	0.55	1.48 ***	−6.39	0.0	0.78	1.27	4.13	−1.52	5.87 ***	0.4 ***	0.29 ***	0.44 (2.24) **	
Q5	96	0.41	1.5 ***	−5.8	−0.25	0.61	1.31	3.23	−1.41	4.24 ***	0.38 ***	0.14 ***	0.4 (1.12)	
No Cluster	96	0.89	1.52 ***	−4.44	0.03	1.11	1.92	4.14	−0.97	1.51 ***	0.29 **	0.58 ***	0.63 (6.89) ***	
Q1-Q5	96	0.65	1.49 ***	−3.11	−0.18	0.54	1.36	5.45	0.75	1.55 ***	0.24 **	0.5 **	0.0 (2.28) **	
Q1-NC	96	0.17	1.19	−2.32	−0.48	0.08	0.75	4.39	0.75	1.89 ***	0.32 ***	0.06 ***	0.25 (0.49)	
NC-Q5	96	0.48	1.0 ***	−1.98	−0.12	0.46	1.08	3.66	0.21	0.7	−0.04	0.27	0.2 (2.10) **	

the permanent-variable liquidity factor of Sadka (2010) (see Panel A and B in Table 10). Portfolio results remain unchanged.

7.3. 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. Based on the whole time sample used in clustering (1994–2012) we are able, in this case, to assign 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.

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.

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

7.4. 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 (2015). 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

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, FX-Credit, 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. dev.	Min.	25%	Median	75%	Max.	Skew.	Kurt.	Autocorr.		Alpha	Adj. R^2
Q1	96	0.93	1.25 ***	−3.26	0.18	1.03	1.65	5.38	0.03	1.86	0.29 **	0.65 ***	0.12 (4.58) ***	
Q2	96	0.9	1.38 ***	−3.69	0.0	1.12	1.84	3.99	−0.55	0.58 **	0.11	0.67	0.19 (4.70) ***	
Q3	96	0.71	1.31 ***	−3.45	0.02	0.85	1.58	3.56	−0.77	1.05 ***	0.23 *	0.43 **	0.31 (3.89) ***	
Q4	96	0.45	1.59 ***	−6.54	−0.36	0.73	1.39	3.09	−1.67	5.17 ***	0.36 ***	0.15 ***	0.52 (1.26)	
Q5	96	0.42	1.95 **	−6.21	−0.29	0.42	1.49	5.74	−0.76	2.62 ***	0.22 ***	0.08 **	0.33 (0.51)	
No Cluster	96	0.92	1.54 ***	−4.55	0.1	1.2	1.89	4.42	−0.99	1.75 ***	0.3 **	0.61 ***	0.65 (7.73) ***	
Q1-Q5	96	0.5	1.73 ***	−4.2	−0.35	0.31	1.36	5.2	0.1	1.24	0.19 **	0.4 *	0.12 (2.33) **	
Q1-NC	96	0.01	1.3	−4.56	−0.5	0.1	0.55	3.7	−0.2	2.06	0.45 ***	-0.12 ***	0.36 (−1.04)	
NC-Q5	96	0.49	1.31 ***	−4.24	−0.09	0.39	1.18	5.47	−0.02	4.43	−0.12 ***	0.37	0.0 (3.14) ***	

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.

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

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 normalised 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.

		Number of clusters		Entropy based			Normalized	
True	Candidate	True	Candidate	Homogeneity	Completeness	V.	ARI	AMI
FBC	Fund family	157	905	0.84	0.55	0.66	0.05	0.25
Fund family	FBC	905	3719	0.8	0.65	0.72	0.03	0.14

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 normalised 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.

8. 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 individual clustered funds 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

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

Agarwal, Vikas, Vikram K Nanda, and Sugata Ray, 2013, Institutional investment and intermediation in the hedge fund industry, *Available at SSRN 2288102*.

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, 2012, Hedge funds as liquidity providers: Evidence from the lehman bankruptcy, *Journal of Financial Economics* 103, 570–587.

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.

Chen, Yong, and Bing Liang, 2007, Do market timing hedge funds time the market?, *Journal of Financial and Quantitative Analysis* 42, 827–856.

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

Cohen, Wesley M., and Daniel A. Levinthal, 1990, Absorptive capacity: a new perspective on learning and innovation, *Administrative science quarterly* pp. 128–152.

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.

Fama, Eugene F, and James D MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *The journal of political economy* pp. 607–636.

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.

Fung, William, and David A Hsieh, 2002, Risk in fixed-income hedge fund styles, *The Journal of Fixed Income* 12, 6–27.

Fung, William, and David A. Hsieh, 2004, Hedge fund benchmarks: a risk-based approach, *Financial Analysts Journal* pp. 65–80.

Fung, William, David A Hsieh, Narayan Y Naik, and Melvyn Teo, 2015, Growing the asset management franchise: evidence from hedge fund firms, *Available at SSRN 2542476*.

Getmansky, Mila, 2012, The life cycle of hedge funds: Fund flows, size, competition, and performance, *The Quarterly Journal of Finance* 2.

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

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.

Naik, Narayan Y, Tarun Ramadorai, and Maria Stromqvist, 2007, Capacity constraints and hedge fund strategy returns, *European Financial Management* 13, 239–256.

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 J, and Tarun Ramadorai, 2013, On the high-frequency dynamics of hedge fund risk exposures, *The Journal of Finance* 68, 597–635.

——— , and Michael Streatfield, 2015, Change you can believe in? hedge fund data revisions, *The Journal of Finance* 70, 963–999.

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

SEC, 2003, Implications of the growth of hedge funds, Discussion paper, .

Shadab, Houman B, 2009, The law and economics of hedge funds: Financial innovation and investor protection, *Berkeley Business Law Journal* 6, 240–297.

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.

Teo, Melvyn, 2009, The geography of hedge funds, *Review of Financial Studies* p. hhp007.

Tirole, Jean, 1988, *The theory of industrial organization* (MIT press).

Titman, Sheridan, and Cristian Tiu, 2011, Do the best hedge funds hedge?, *Review of Financial Studies* 24, 123–168.

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

——— , 2003, Financial innovation, in G. M. Constantinides, M. Harris, and R. M. Stulz, ed.: *Handbook of the Economics of Finance*vol. 1 of *Handbook of the Economics of Finance* . chap. 6, pp. 307–335 (Elsevier).

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.

A: Hedge fund properties 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

B: Fast Binary Clustering

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 B.1.

Table B.1 shows the outcomes of the two existing clustering algorithms, k-means 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. 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 randomised. 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 normalised 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 normalised with respect to random labelling which means they tend to be inflated for higher number of clusters. Homogene-

Table B.1

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. 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 normalised such that 1 represents the optimal level, and values farther away from unity show weakness of algorithms considered. Three entropy based measures of cluster quality are considered: homogeneity (homog.), completeness (compl.), and V-measure. Two normalised measures are reported: Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI).

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

	DBSCAN						k-Means			FBC			
	cityblock		cosine		euclidean					cityblock		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	0.54	0.54	0.54	0.57	1.75	0.96	1.0	1.04	1.0	1.01	1.0	1.0
	(0.20)	(0.20)	(0.20)	(0.20)	(0.23)	(0.29)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	0.01
Time	1.13	1.13	1.21	1.18	1.15	1.14	1.11	1.11	1.17	1.04	1.0	1.11	1.05
	(0.64)	(0.64)	(0.61)	(0.64)	(0.65)	(0.64)	(0.64)	(0.61)	(0.58)	(0.60)	(0.67)	(0.61)	0.60
Homogeneity	0.08	0.08	0.08	0.08	0.09	0.36	0.99	1.0	1.0	1.0	1.0	1.0	1.0
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.10)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	0.0
Completeness	0.65	0.65	0.65	0.65	0.64	0.69	1.0	1.0	0.99	1.0	1.0	1.0	1.0
	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	0.0
V-measure	0.15	0.15	0.15	0.15	0.15	0.47	0.99	1.0	1.0	1.0	1.0	1.0	1.0
	(0.06)	(0.06)	(0.06)	(0.06)	(0.06)	(0.10)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	0.0
ARI	0.0	0.0	0.0	0.0	0.0	0.02	0.96	0.99	0.99	1.0	0.99	1.0	1.0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	0.0
AMI	0.05	0.05	0.05	0.05	0.05	0.27	0.98	1.0	0.99	1.0	0.99	1.0	1.0
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.10)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	0.0

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

	DBSCAN						k-Means			FBC			
	cityblock		cosine		euclidean					cityblock		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	0.0	0.0	0.01	0.02	0.1	0.27	0.98	1.0	1.02	1.0	1.0	1.0	1.0
	(0.01)	(0.01)	(0.01)	(0.02)	(0.05)	(0.08)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Time	1.02	1.03	1.02	1.02	1.01	1.02	1.0	1.0	1.02	1.02	1.01	1.02	1.01
	(0.58)	(0.58)	(0.58)	(0.59)	(0.58)	(0.58)	(0.58)	(0.59)	(0.58)	(0.59)	(0.59)	(0.59)	(0.58)
Homogeneity	0.0	0.0	0.0	0.0	0.02	0.07	0.99	1.0	1.0	1.0	1.0	1.0	1.0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Completeness	0.95	0.95	0.92	0.82	0.73	0.77	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	(0.13)	(0.13)	(0.16)	(0.18)	(0.09)	(0.05)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
V-measure	0.0	0.0	0.0	0.01	0.04	0.13	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
ARI	0.0	0.0	0.0	0.0	0.0	0.0	0.98	0.99	0.99	1.0	1.0	1.0	1.0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)
AMI	0.0	0.0	0.0	0.0	0.01	0.04	0.99	1.0	1.0	1.0	1.0	1.0	1.0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

ity 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 normalised 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.

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.

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 B.1, 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).

```
##
##FBC pseudo-code
##
#parameters
maximum_distance #varepsilon
distance_increment #Delta_varepsilon
distance_measure
#algortihm
current_distance=0
```

```
clusters={}
while current_distance<=maximum_distance:
    inner_loop_counter=0
    while inner_loop_counter==0
          or
          clusters[current_distance][inner_loop_counter]==clusters[current_distance][
              inner_loop_counter-1]:
                #Centroid step
                for cluster in clusters[current_distance[inner_loop_counter]]:
                    clusters[current_distance][inner_loop_counter][cluster]['archetype_genome'
                        ]=
                            average(clusters[current_distance][inner_loop_counter][cluster]['
                                funds'])

                #Density step
                proximity_matrix=getDistances(
                    [archetype_genome for archetype_genome in clusters[current_distance][
                        inner_loop_counter][cluster].keys()],
                    distance_measure
                                                )
                cluster_ids_counter=0
                for cluster in cluster[current_distance][inner_loop_counter]:
                    if cluster['cluster_id']==None:
                        cluster['cluster_id']=cluster_ids_counter
                        cluster_ids_counter+=1
                    else:
                        cluster_id=cluster['cluster_id']
                    for other_funds in neighbourhood(cluster, proximity_matrix):
                        other_funds['cluster_id']=cluster_id
  #hierarchical step
   current_distance+=distance_increment

clusters=discardSmallClusters(clusters[current_distance])

}
```