

2 Theoretical Background

2.1 The Origins of Agglomeration Theory

The theoretical background of this study is based on several different strands of research. Therefore, this chapter will provide a comprehensive overview of the related theories. Researchers from various disciplines, such as economics, management, strategy, and economic geography have been studying one particular interorganizational process, namely the geographical agglomeration of companies in clusters. One could suppose that the research strand devoted to international trade theory (i.e., Heckscher-Ohlin, Ricardo, and their extended approaches) gives appropriate explanations. However, neoclassical trade theory treats countries as dimensionless (Fujita and Thisse, 2002: 5-6).

The relevant theories and concepts of agglomeration are strongly influenced by ideas of von Thünen (1875) and Christaller (1933). Von Thünen (1875) developed a location theory for the agricultural sector which describes the optimal use of land for an isolated city and which was formalized by the studies of Launhardt (1885) and Lösch (1940). Weber (1909) exploited a neoclassical location approach for industrial companies to find the optimal production location based mainly on regional endowments and transportation costs. Christaller's (1933) central place theory describes how the establishment of a central order system is based on the supply functions of the places involved. Lösch (1940) further develops Christaller's approach in his work on the spatial order of economies, in which he analyzes the geographical distribution of industrial locations with regional market networks. From this analysis, he derives a horizontal hierarchical spatial system. Alonso's (1964) monocentric city model is also largely based on the work of von Thünen, with commuters taking the place of farmers, and central business districts replacing the isolated city. Alonso reveals that the usage of land in central business districts is arranged in the form of concentric rings.

Since the 1950s, the new "regional science" approach has been developed based on Isard (1956) in an attempt to combine economics with geography.¹ Isard (1956) expands Weber's model by the principle of substitution of production inputs. All of these approaches assume that the regional industrial structure is predetermined by regional endowments, transportation conditions, etc. Isard mainly criticizes neoclassical theory for considering the world a "wonderland of

¹ See Roos (2002) for an overview of the different approaches in the regional science literature.

no spatial dimensions” with all inputs, outputs, producers, and consumers concentrated at a single point. He recognizes that, under the assumption of perfect competition, economic activity would be geographically evenly distributed. According to Fujita (1999: 374-375), Isard therefore sees the adoption of the monopolistic competition model as a prerequisite for explaining the spatial differences in development. Thus, his aim was to reformulate the neoclassical general equilibrium theory such that all demand, supply, and price variables could be expressed as explicit functions of the location (Scott, 2000). The general equilibrium theory of (spaceless) economics would then be a special case in which transport costs are zero and therefore disregarded and all inputs and outputs are perfectly mobile. At the time, there were no formal models of imperfect competition with increasing returns to scale. This only changed with the work of Dixit and Stiglitz in 1977 (Roos, 2002).

The earliest precise discussion of agglomeration in clusters stems from the localization analyses on industrial districts conducted by Marshall (1890 and 1921). He emphasizes that “great are the advantages which people following the same skilled trade get from near neighborhood to one another. The mysteries of the trade become no mystery: but are as it were, in the air...” (Marshall, 1890: 352). He finds positive externalities of specialized industrial locations from urban specialization.² Economies of scale can be achieved by supplier concentration and market size effects, labor market pooling and knowledge spillovers, as identified by Marshall. In contrast, Jacobs (1969) emphasizes the importance of urban diversity, which fosters the cross-fertilization of ideas. This has led to a discussion on localization versus urbanization.

Overall, past studies on agglomeration particularly emphasize cost minimization in clusters due to their proximity to inputs or to markets. Their descriptions, however, have been undercut by recent changes in globalization, technology, and mobility, which have caused a decrease in transportation and communication costs. Today, the approach of agglomeration economies has shifted from urban areas to clusters.

2.2 Insights from the Forming of Agglomeration

Duranton and Puga (2004) summarize the theoretical literature explaining agglomeration economies, i.e., the existence of urban agglomeration economies, on the basis of three general benefits: Firstly, agglomeration enables increased efficiency in the sharing of local infrastructures, more variety in intermediate

² Marshall (1921) suggests four externalities relevant for the formation of a cluster: (1) mass production (i.e., economies of scale), (2) availability of specialized input services, (3) close proximity of the labor pooling to enable face-to-face communication, and (4) the availability of modern infrastructure (see Fujita and Thisse, 2002: 8).

ancillary industries, and a larger pool of workers with similar skills. Secondly, it allows superior matching among the market participants (i.e., relationships between employers and employees as well as between buyers and suppliers), and thirdly, it facilitates knowledge spillovers.

Economists such as Kaldor (1972), Piore and Sabel (1984), as well as Krugman (1991a) identified cluster theory in the early years. At the same time, economic geographers in particular examined agglomeration driven by active collective efficiencies, such as improved access to knowledge and other intangible resources (Scott, 1988; Ratti, 1992; Morgan, 1997). Porter (1990) has built on Marshall's early insights by popularizing the cluster concept. He has also continued to develop it further in his subsequent works. His studies may be viewed as a synthesis of ideas derived from a range of social scientists in economics and economic geography. The "new economic geography" is based on Krugman (1991a and 1991b) and Venables (1996), among others. This model mainly explains clusters by using agglomeration effects and increasing economies of scale and includes potentially inefficient path dependences³ of location choices. The studies of the new economic geography contain different models of general equilibria that explain unequal spatial distribution of economic activity under the assumptions of monopolistic competition, increasing returns to scale, and the existence of transportation costs at different geographical levels: international specialization, national distribution, regional level, and city level (see Krugman, 1991a and 1991b; Krugman and Venables, 1995; Venables, 1996; Fujita et al., 1999). The models of the new economic geography are based on heterogeneous centripetal and centrifugal forces which will be explained later.

To put it simply, a cluster is a non-random geographical agglomeration of companies with similar or closely complementary capabilities (Ellison and Glaeser, 1997). Put even more simply, a cluster is a system of interconnected companies and institutions whose whole is more than the sum of its individual parts. According to the detailed definition of Porter (1998: 197), a cluster is a geographic concentration of competing and cooperating companies, related suppliers, service providers, and institutions with highly specialized skills and knowledge. Therefore, clusters encompass an array of linked industries with suppliers of specialized inputs factors. Thus, clusters include public (e.g., universities, think tanks) and private specialized service providers that provide target-oriented education, research, and technical support. Collective bodies, i.e., trade associations, are also an indication of a cluster. Foreign companies are therefore part of a cluster if they make permanent investments.

It is broadly recognized that the observed spatial configuration of economic activities is generally the outcome of a process involving two opposing types of forces. These centripetal (agglomeration) and centrifugal (dispersion) forces lead to a balance of forces that push and pull consumers and companies (Fujita and Thisse, 2002: 5). Porter (1998, 2000) shows that among individuals, geographical

³ The effects of path dependence will be specifically addressed in the following chapter 2.3.

and cultural proximity generate advantages in productivity growth and entrepreneurial activity due to, for example, special relationships with better incentives and information which is difficult to tap from a distance. The network of companies and public institutions generates many cluster advantages which have positive externalities or spillover effects across companies and industries. Moreover, the geographic proximity between two different industries leads to co-agglomeration and the growth of both industries (Ellison and Glaeser, 1997). Schmitz and Nadvi (1999: 1504) describe the process of the collaboration between cluster members as “the conscious pursuit of joint action.”

Florida and Gates (2001) examine the effects of soft location factors such as cultural diversity. They find that clusters with a culture of openness have a higher tendency towards innovation than less creative cities. Porter (1990, 2000) confirms these results, providing evidence that local agglomeration increases competition and thus encourages innovation by forcing firms to either innovate or fail. In the same vein, Glaeser et al. (1992) show that increasing competition in a cluster is positively correlated with economic growth.

The results of Maskell (2001) further verify this. He shows that, at the horizontal level, companies which are located close to their direct competitors and sell similar products enjoy several advantages. These advantages result from a superior exchange of information and from the fact that competitors’ products and strategies can be observed more closely. Proximity provides incentives to continually improve one’s products and to adapt to the ever-changing competitive conditions within the cluster. The vertical dimension shows the relationship between the individual levels of the value chain. The more diverse the levels, the greater the need for a division of labor, which in turn provides an incentive for specialized suppliers to settle in the cluster in order to enjoy specialization effects. Agglomeration and specialization processes also lead to the formation of a specific institutional environment. Market participants within a cluster share common norms and rules and establish mutual confidence and trust through intensive contact. Malmberg and Maskell (2002) draw a distinction between knowledge-driven cluster development in early and later stages. In the early stage, there is a more horizontal cluster dimension, with similar competencies, cognitive closeness, and learning mechanisms in terms of variation, observation, comparison, and rivalry. In the later stage, the cluster dimension is more vertical, i.e., the competencies are complementary, trust and social capital are the institutional basis, and the learning mechanisms are specialization, interaction, substitution, coordination, and cooperation.

Since knowledge becomes more specialized over time, a cluster-specific division of labor and institutional organization enables the emergence of distinctive approaches to learning and knowledge creation (see Bell et al., 2009: 624-625; Bathelt and Taylor, 2002: 7). Audretsch (1998) argues that due to globalization and advanced telecommunication technologies, the value of knowledge-based economic activity has encouraged the emergence of a new comparative advantage in geographical locations – an innovative, knowledge-creating culture. Since knowledge spillovers are most facilitated in spatial

proximity, knowledge-intensive industries are likely to locate themselves in a cluster.

Apart from this, dense social networks provide strong reasons for agglomeration, for example clustering in Silicon Valley. Saxenian (1994) pointed out that the region was essentially identical to Boston (Route 128) in the 1970s. However, the two locations did not have identical characteristics. Offering an entrepreneurial culture of rapid changes and quick decisions, Silicon Valley subsequently transformed into a relatively more productive environment. The author emphasizes that the success stems in particular from dense social networks and a high level of social capital over a small area (i.e., *“you can change jobs without changing the parking lots”*).

In a similar vein, Sorenson and Audia (2000) show that entrepreneurial ventures are more likely to agglomerate in environments of existing social networks despite intense competition within the cluster. This line of thought corresponds with research in organizational behavior. It views the economy as a set of interactions within interorganizational networks while conceiving all systems of interactions as networks. Therefore, relations rather than market participants constitute the focus of analysis (Baker, 1990; Gulati, 1998). A social network is defined as “a specific set of linkages among a defined set of persons, with the additional property that the characteristics of these linkages as a whole may be used to interpret the social behavior of persons involved” (Mitchell, 1969: 2). Hippel (1994) shows that high-context knowledge is best transferred through frequent face-to-face contacts which are naturally more easily achieved in clusters (i.e., “gluey knowledge”). Spatial proximity allows face-to-face contacts and facilitates a large amount of knowledge exchange at lower cost.

The analysis of Hong et al. (2005) of the impact of spatial proximity between mutual fund managers demonstrates that it influences the similarity in trades and holdings between them. They find that fund managers from the same area have greater opportunities to interact and thereby spread rumors about particular investment opportunities. In the same context, Christoffersen and Sarkissian (2009) indicate that mutual funds in financial centers perform better due to such specific information flows. Corresponding findings show that the strength – and not just the existence – of relationships between market participants in a cluster is crucial for enabling the exchange of private information and the privileged interpretation of market information to result in knowledge spillovers. Formal business collaborations enhance socialization, fostering an informal relationship between market participants (Gulati and Puranam, 2009). This can enhance the transmission of private information and interpretations in the cluster even if the formal connection is broken (e.g., change of job). However, ending the formal relationship will reduce day-to-day interaction, thereby causing a gradual decay of the relationship. At the same time, a longer prior collaboration between market participants means a stronger informal relationship and less potential for the relationship to decay over time (Burt, 2000). It is also generally expected that companies with many ties at one point of time are more likely to receive new ties in the future than those with fewer past ties (Glückler, 2007; Barabasi and Albert, 1999).

According to corroborating results presented by Gulati and Gargiulo (1999), the attractiveness of new relationships is further enhanced by existing close ties and the parties' partners. This leads to an environment of social embedding due to processes of indirect referrals and trust formation. The authors suggest that the chance of forming a tie depends on individual characteristics and the network position of a company. In the course of collaboration within the cluster, the interaction between market participants in the form of jointly attended (in-)formal events and meetings further shape their mental maps and thus their subsequent behavior (i.e., Weick et al., 2005). The authors did not explicitly consider the determinants for the willingness to modify existing mental maps based on new interactions, but there are strong reasons to suspect that senior managers would be more reluctant to change their beliefs, since they become more confident in their own beliefs as their experience increases. Conversely, this means that people earlier in their careers have less established beliefs, making them more likely to be affected by interaction within clusters (e.g., Niessen et al., 2010).

Dense clique-like local networks are strengthened over time by shared beliefs and perceptions (mental models), for example of how markets work, and enable market participants to interpret the behavior of others (Baum et al., 2003: 702). Zaheer and Bell (2005) analyze syndicate networks of Canadian mutual fund companies and find that cognitive embeddedness and the formation of mental models within clique-like, interconnected markets lead to persistent network structures. Granovetter (2005) gives further evidence that social networks affect economic outcome in general because they (1) improve the flow and quality of information, (2) facilitate reward and punishment mechanisms, and (3) foster trust among market participants.

However, there are also indications that path dependence could be involved in the process of network growth (e.g., Walker et al., 1997). This concept will be discussed in more detail in the following chapter.

2.3 Bygones Are Not Always Bygones

Several lines of thought highlight the importance of initial conditions and events for organizational development. One strand of empirical research has considered the role of "history and natural advantages" in cluster formation, explaining why certain rules of behavior come to prevail over others (e.g., Kim, 1995 and 1999; Ellison and Glaeser, 1999; Rosenthal and Strange, 2004). The findings show that agglomeration is sometimes highly dependent on natural advantages. For instance, the North American steel industry was initially concentrated in the Great Lakes region, mainly because of iron ore and coal reserves. Similarly, California's growth can be attributed to its moderate climate, which allowed employers to pay lower wages (Rosenthal and Strange, 2004).

With the "concept of institutionalization," the neoinstitutional theory stresses the relevance of symbolic-normative environments for organizations (e.g.,

Hargadon and Douglas, 2001) and examines how they formally and informally influence the structuring nexus of organizations over time (e.g., Tolbert and Zucker, 1996; Scott, 2001). A further approach within neoinstitutional theory is the “concept of imprinting.” This postulates that either initial cognitive schemes (e.g., competences of a team), or specific contextual circumstances (e.g., postwar depression, dot-com boom, financial crisis, or structure of the institution) at the time of founding leave an imprint on organizational processes at later stages (Beckman and Burton, 2008).

The two latter concepts do not address the rationale behind the escalating reinforcement of an action pattern or a course of action, i.e., a path.

However, several studies indicate that the clustering of individual institutions may also be the result of “path dependence,” which has cumulative consequences in the long run. There are various definitions of path dependence which exhibit similar characteristics. Vergne and Durand (2011) stress that, in many cases, explanations based on path dependence can be found in the organizational literature.⁴ In general, path dependence asserts that the order or sequences of events prior to the observation of the outcome have explanatory power (“history matters”) and that the underlying trend is often irreversible, so that “bygones are rarely bygones” (Teece et al. 1997: 522). Therefore, it is assumed that the ahistorical and unbounded view of rational choice theory is limited. Path dependence is an essential feature of evolutionary concepts (e.g., David, 1985 and 2001; Arthur, 1994; Bathelt and Glückler, 2003; Martin and Sunley, 2006). The usage of the term is more metaphorical than theoretical in nature, and the literature does not give a coherent definition. Hence there are no precise indicators for examining whether or not an observable process is path-dependent. In accurate terms, path dependence asserts that the state of a system is always determined by the initial point of development, i.e., that the past always exerts a certain influence on the present and is affected by any disruptions taking place over the past course of events. The order of events may also influence the state of a system. Therefore, a path-dependent process must contain at least two possible equilibria selected contingently along the path (David, 2001).

One may assume that all human activity and all institutional processes are conditioned by their history to a certain extent. However, deriving the conclusion that all institutional decisions are path-dependent would be incorrect. Path dependence means more than the existence of routines, cognitive rigidities, or structural inertia. It relates to more specific conditions that are not characteristics of decision making, such as a lock-in effect, an outcome of path dependence associated with the irreversible (and sometimes suboptimal) persistence of a particular state of affairs (see Sydow et al., 2009).

David (1985) illustrates a prominent example for the development of a path leading to a lock-in: In the US, the configuration of the letters on a keyboard

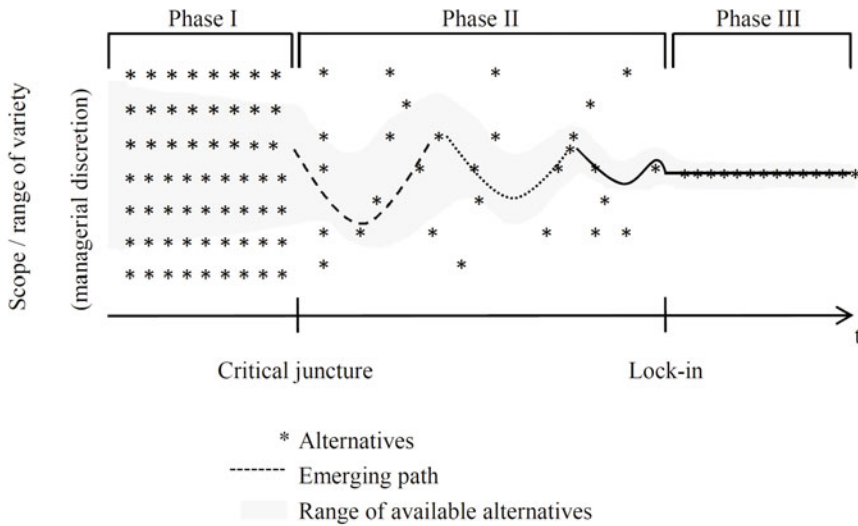
⁴ For a more detailed overview of the different related lines of thought discussed here, see Sydow et al. (2009), who cover the literature in detail and whose work is the fundamental basis for this section.

begins with QWERTY. The original reason for this configuration lies in the construction of old typewriters, yet it has remained in place until today. However, it has been shown that a change in the alignment of letters would lead to a higher typing speed. This is a technology-based example; Vergne and Durand (2011) cite studies with focal points on resource accumulation (Karim and Mitchell, 2000), innovativeness (Danneels, 2002), dynamic capabilities (Zott, 2003), cognitive processes (Lamberg and Tikkanen, 2006), institutional trajectories (Djelic and Quack, 2007), and strategic paths (Koch, 2008).

Arthur (1994: 14-15) describes the process of becoming path-dependent using four properties: (1) non-predictability (i.e., multiple equilibria, no guarantee that the decision made is the superior one in the long run); (2) non-ergodicity (i.e., historical events affect the course of the path), (3) inflexibility (i.e., a shift to another alternatives is impossible), and (4) inefficiency (i.e., inferior results due to a lock-in). Sydow et al. (2009) further establish a theory of organizational path dependence by drawing predominantly on different findings from organization studies, especially in institutional economics (i.e., North, 1990) and political science. The authors suggest a three-stage framework to distinguish the development phases of path dependence. It starts as (1) a process triggered by an event resulting in a critical juncture which (2) may transform it under certain conditions into self-reinforcing dynamics, allowing it to gain more and more predominance over alternative choices. The last potential phase (3) is an organizational lock-in, a corridor of limited scope of action which is strategically inefficient. Figure 1 illustrates this process. The implementation of the inferior result is caused by self-reinforcing events which bring about other inappropriate benefits in each step of the adopted direction. As such, path-dependent processes are characterized by the potential inefficiencies of their process results, with a general openness for future development predominating in the initial phase.

They describe the *first stage (preformation phase)* as a phase that is characterized as an open situation with a broad scope of different possible decisions. The initial situation may also be embedded in and connected with other (past) developments so that the imprinting concept described above can explain existing restrictions. The initial choice is the impetus stimulating further actions.⁵ The setting of the course of events at the beginning is established by an insignificant decision or a critical event that could occur somewhat randomly (Arthur 1989: 116; David 1985: 332). It is possible that the decision taken at the beginning is in fact inferior to other realizable courses of events.

⁵ Sydow et al. (2009: 693) refer to the prominent instance of a butterfly randomly flapping its wings, which leads to a change in the atmosphere, which in turn sets a chain of events in motion eventually causing a large-scale change in weather (e.g., a tornado).

Fig. 1. Funnel-shaped Structure of Path Dependence

Source: author's illustration, based on Sydow et al. (2009: 692).

The moment of entering into the dynamics of self-reinforcing processes marks the decisive transition to *stage two (formation phase)*. When the path is taken, it becomes further reinforced as alternative paths become relatively less attractive. The path taken progressively gains dominance to the extent that it triggers a regime of positive, self-reinforcing feedback. Self-reinforcement can be understood as a set of positive (e.g., increasing returns to scale) and negative mechanisms (e.g., negative externalities) that increases the attractiveness of a path relative to other alternatives (Vergne and Durand, 2011: 371). According to the findings of Pages' (2006: 110-112) theoretical approach, at least one negative mechanism is mandatory for the lock-in effect. For instance, in the case of "keyboard layout QWERTY," negative mechanisms occurred on an intrapersonal level in terms of lower costs and the allure of not requiring to learn to type on another keyboard, as well as on an interpersonal level, since the more users adopt the same keyboard layout, the less attractive it is for a prospective user to learn to type on a different keyboard. The positive mechanism can be found in increasing returns (e.g., economies of scale), since more typists use the same keyboard layout (Vergne and Durand, 2011: 378).

The transition to *stage three (lock-in phase)* is characterized by a further restriction of the scope of choices along the path and inherent inefficiency. A behavior of persistence could be induced either by structural inertia or the circumstance that a shift to an inherently more efficient solution has at this stage become generally more expensive, e.g., since the existing infrastructure would have to be converted at great cost. In this stage, the organization may also be affected by escalating commitments (i.e., Ross and Staw, 1993; Guler, 2007). These restrictions prevent market participants from changing their course of action with a negative feedback on the outcome, so that they replicate an inefficient solution. Sydow et al. (2009: 696) describe such events as pathological decision-making behavior based on the dynamics of self-justification and concerns of losing face. Moreover, institutions are embedded in more or less complex relationship networks. Such collaborations are also likely to become path-dependent with lock-in effects (Gulati et al., 2000). Sydow et al. (2009: 698-701) further synthesize four mechanisms which contribute to the development of self-reinforcing mechanisms in organizational path dependence:

(1) *Coordination effects* are an object of analysis in the field of institutional economics. The latter has provided evidence that interaction between market participants becomes more efficient as the number of market participants who adopt and apply a specific institutional rule increases, as their behavior can be anticipated. As a result, coordination costs decrease (i.e., North, 1990). This coincides with the economies of scale effect. Therefore, it is worthwhile to adopt these rules as long as many others follow them. Illustrative examples include deciding between left-hand and right-hand traffic and working time regimes, which enable cooperation and reduce uncertainties in interaction.

(2) *Complementary settings* allow for synergy effects from the interaction of two or more formerly separate but interrelated resources or rules, i.e., economies of scope (e.g., Stieglitz and Heine, 2007). The thought pattern of Sydow et al. (2009) confirms the results on cluster theory highlighted in the previous chapter. Combining interrelated activities produces a surplus which exceeds their mere sum ($X_{(1+2)} > X_{(1)} + X_{(2)}$).

(3) *Learning effects* potentially increase efficiency (i.e., a faster, more reliable and smooth workflow), which also causes a decrease in average costs per unit. Thus, the more attractive the chosen decision becomes due to accumulated skills and decreasing costs, the less attractive it becomes to switch to a new decision, such as a geographical relocation, where market participants have to start from scratch with a bundle of uncertainties. This behavior will typically lead to path dependence, in which market participants in the organization are more motivated to improve everyday practice (to gather legitimacy and reward in a prevailing corporate culture) than to look for alternatives and question well-established organizational structures.

(4) *Adaptive expectation effects* are derived from the assumption that the individual preferences of market participants are not fixed (as opposed to the neoclassical model) and vary in response to the expectations of other market

participants. According to Leibenstein (1950), the more market participants are expected to (informally) prefer a particular practice or service, the more attractive it becomes. To prevent uncertainty about the correct decision, market participants feel rewarded when others are likely to prefer the same. Adopting the mainstream mindset is related to seeking legitimacy and signaling. Those who defy the mainstream and follow an unsuccessful alternative become stigmatized as an “outsider” (Kulik et al., 2008).

Overall, the effects of path dependence have far-reaching consequences not just for the institution itself, but also for the social environment shared with other institutions (in the cluster) and for the (fiscal) government. The dissolution of a lock-in typically occurs through unforeseen exogenous forces, such as shocks or crises (Arthur, 1994: 118) and changes in the organizational structure introduced, for instance, when new market participants do not adopt the same rules (see Sydow et al., 2009: 701). Nevertheless, opening the window for alternatives is necessary, if insufficient. New alternatives must be superior because implementing an equal (or inferior) alternative would not be attractive in comparison with a practice that is both familiar and functional.

Macro Attractiveness and Micro Decisions in the Mutual
Fund Industry

An Empirical Analysis

Lang, G.

2014, XII, 178 p. 22 illus., Hardcover

ISBN: 978-3-642-39723-3