A MARSHALLIAN MODEL OF EFFICIENCY PROFIT DYNAMICS

Michael G. Jacobides  
Assistant Professor of Strategic and International Management  
London Business School  
Sumantra Ghoshal Fellow, Advanced Institute for Management Research  
Visiting Scholar, Harvard Business School  
Sussex Place, Regent’s Park, London NW1 4SA, United Kingdom  
mjacobides@london.edu; Tel (+ 44 20) 7706 6725; fax (+ 44 20) 7724 7875

Sidney G. Winter  
Deloitte and Touche Professor of Management  
The Wharton School of the University of Pennsylvania  
Suite 2000, SH-DH, 3620 Locust Walk, Philadelphia PA 19104-6370, USA  
winter@wharton.upenn.edu; Tel (215) 898 4140; fax (215) 898 0401

Stefan M. Kassberger  
Lecturer in Financial Mathematics  
University of Ulm, Germany  
and London School of Economics  
Helmholtzstrasse 18, 89069 Ulm, Germany  
stefan.kassberger@uni-ulm.de, Tel (+49) 731 50 23518

Working Paper  
London Business School and Advanced Institute for Management Research  
August 7, 2006 (v 1.6)
A MARSHALLIAN MODEL OF EFFICIENCY PROFIT DYNAMICS

Abstract

In the literature of strategic management, most of the research on profitability focuses on conditions under which a firm can enjoy extraordinary profits in the long run, despite the efforts of its competitors. This paper extends this line of thinking by considering not only the role played by idiosyncratic resources in the final equilibrium, but also the dynamic process by which the sector approaches that equilibrium. Firm financial performance is assessed not only in terms of long run equilibrium levels of flow profitability, but also in terms of the present value the firm realizes over the trajectory as a whole. Firms are assumed to be price takers, so market power and strategic output restraint play no role in profitability – this is the sense in which the analysis relates to “efficiency profits.” The model analyzed is an elaboration of a simple Marshallian representation of competition in an industry. Each period is a Marshallian “short run,” over time the industry approaches some long run equilibrium. The main themes are familiar ones of evolutionary economics; profits drive growth and more efficient firms drive out less efficient ones; imitation may permit the inefficient to become efficient. A key feature of the present model is that firms compete not only in the market for final goods, but also for resources; so both input and output prices are endogenously determined. A key assumption – departing from textbook economics but converging with the “resource based view” (RBV) in strategic management – is that firm heterogeneity centrally involves idiosyncratic firm resources that are not simply the inputs that are available in the market. Resources are purchased in the market, but are specialized to the firm (and become “attached” to it) by virtue of sunk investments in some costly process of customization. Profits emerge (as a function of capability heterogeneity), and erode over time (as a function of the speed of capacity expansion and imitation, or of rules of endogenous capability development). The model shows how this happens, and also illustrates how profitability co-evolves with input prices, output prices and quantities. By computing model trajectories under a wide range of conditions, we explore how the evolutionary path and asymptotic level of profitability depends on a wide range of parameters. We show that the relative importance of one-off investments in resources directly affects the ability to sustain profits, as well as to the extent to which firms of differential efficiency can co-exist in equilibrium. We also show that sustainable profits are relatively limited parts of total industry profitability and that adjustment paths of industry evolution affect both profit NPV and the equilibrium profits in a sector.
The question of how profitability emerges and evolves has been a central concern in both strategic management and Industrial Organization (IO) research. Recognizing that asking “how do profits emerge, and when are they sustained?” may be too grand and too ambitious a question, researchers have tackled more focused questions – in the hope that, taken together, the answers to these will allow us to construct the bigger picture. In the process of focusing attention and dividing labor to yield useful insights, existing research has followed a number of trajectories. Within each of these (e.g., analysis of the oligopolistic conditions leading to profitability; consideration of the nature of competitive games and how it shapes the profits of a small number of firms; and, more recently, the analysis of resource or capability heterogeneity, and the discussion of how this affects profits), much progress has been made. Yet, as any strategic management textbook discussion will show (see, e.g., Besanko et al. 2004; Grant 2005; Saloner et al. 2001), our view of the big picture remains as fragmented as the diverse trajectories that have been followed. More importantly, this fragmentation may have left some important ground not covered. In this paper we go “back to basics” and explore the profitability issues in the framework of an analysis of a competitive industry constructed in the spirit, and with many of the tools, of Alfred Marshall.

By virtue of its explicit concern with dynamics, our analysis is also much in the spirit of Joseph Schumpeter’s work (1911/1945; 1934). Schumpeter stressed the role of profit as a motivator for economic activity and entrepreneurial action, whether by aspiring entrants (1911) or established firms (1945). More importantly, though, he stressed that profits are often temporary; they erode over time, and they should be expected to do so, since imitation, excessive capacity growth and competition, or substitution, do not allow success to be particularly long-lived. Thus, even apparently stable firm-level (or industry-level) profits, Schumpeter argued, should be seen as subject to the continuing hazard of “waves of creative destruction”. While a large stream of research has built on Schumpeter, most of it focuses on understanding innovation and its determinants, or examines entry and exit over time (e.g. Jovanovic 1982; Winter 1984; Klepper 1997; Pakes and Ericson 1998). It has attempted to explain why firms differ, and why such differences can persist over time (Nelson 1991; Nelson and Winter 1982). But the question of how profit evolves over time in industries of dissimilar firms, which compete for markets and resources alike, has received limited attention (Lipmann and Rumelt, 1982; 2003a; Winter 1995).

The importance of understanding how profit emerges in the presence of heterogeneous firms has only been heightened by recent emphasis on firm heterogeneity. In the field of strategic management, the widely accepted “Resource-Based View” (RBV) starts from the hypothesis that firms are different in ways that derive in large part from the fact that they draw on different resources, which makes them differentially effective (Barney 1986; Wernerfelt 1984). This long-standing emphasis on heterogeneity has become ever more credible with the continuing accumulation of evidence that firms in the same line
of business do, indeed, typically display wide and persistent differences in their practices, productivity and efficiency (Lieberman and Dhawan 2001). Therefore, the need to incorporate these inter-firm differences in analyses of firm profitability becomes paramount. Yet, to date, most of the theoretical research on how firm-level differences translate into profitability differences has focused on how, in equilibrium, after the hand of selection has eroded temporary profits, some firms may still be more profitable than others. This research has concluded that a fairly stringent set of criteria needs to be fulfilled for firms to be able to be profitable in equilibrium (see Peteraf and Barney 2003). It also sheds some important light on some of the dynamic processes through which firms, in equilibrium, might be able to sustain profitability, e.g., by investing in assets that have “time compression dis-economies” and thus yield superior returns when accumulated slowly over time (Dierickx and Cool 1989; Pacheco de Almeida and Zemsky 2006). Nevertheless, this research is predicated on the goal of understanding the conditions under which profits will persist in equilibrium, and identifying strategies that could be associated with this result.

This still leaves a substantial gap, which this paper aspires to fill. While we offer a significant elaboration of the equilibrium analysis, we also address the process of adjustment from an initial profitable position towards equilibrium. By explicitly assuming firm heterogeneity; introducing resource pricing into our analysis, and considering how these factors affect competitive dynamics, we bridge some of the gap between the “positioning school” (influenced by IO economics) and the “resource school”, while focusing on the analytics of the latter. More importantly, we shift attention away from the conditions for persistence of flow profitability (the focus of both the established camps) and toward the dynamic mechanisms of wealth creation.

To achieve this, the paper offers a systemic model of industry evolution. It allows for a market for resources, in which firms compete for resources that are not in fully elastic supply to the industry. Firms also compete in the market for final product, facing a price-sensitive demand. They have varying degrees of capability; and make idiosyncratic investments to implement and extend their distinctive approaches to the competitive struggle. Capacity levels are fixed in the short term (a single period), but firms respond to profit signals by adjusting capacity over time. The extension of capacity by more capable firms is the central mechanism driving (Darwinian) selection. Imitation is also considered as an extension.

Our setup, in special cases, allows us to compute long-run equilibria and the relationships between our key variables. Wherever possible, we do so, highlighting, e.g., the factors that determine when and why firms of differential efficiency can co-exist in equilibrium. Yet the main contribution of this paper is to provide a setup that allows us to examine the complicated, path-dependent dynamics of industry and profitability evolution with firms of differential efficiency. Our set-up provides a theory-informed
experimental platform, from which we can explore computationally a wide variety of cases defined by alternative values of the parameters of the system. In this paper, we vary in particular the parameters affecting the degree of heterogeneity, the cost of the “customizing” investments that give resources their firm-specific character, the speed of imitation, and the interest rate.\footnote{These explorations do not by any means exhaust the interesting questions the model can help us with. We leave the others, such as the role of learning in the evolutionary process, on the agenda for future work. Some preliminary results on the role of demand and supply structure, as it interacts with capability heterogeneity to affect profits, resource prices and quasi-rents, are reported in the concluding section.} We examine how these parameter changes affect firm profits and their distribution, the prices and quantities of both resources and final goods, and the present values of profits at both the firm and industry level.

The model also enables us to sort out the conundrum of the relationship of firm profits to resource rents, which is central to the logic of the resource-based view. If it is resource rents that account for sustainable profits, as is often maintained, why is it that the firm can collect these returns rather than the resource owner? Or does the firm benefit only when it is the resource owner – and perhaps only to the extent of the normal returns on such ownership? What are the key factors that determine when firms capture the rents, and how does this change as the industry evolves? Our contribution focuses on the analysis of the “resource customization cost”, which is shown to lead to sustainable profits.

We also show that sustainability per se has limited impact, when viewed in the context of “total wealth creation”, and the associated profit NPV calculations. More important, we show that the dynamic adjustment paths affect both long-run equilibrium values, and profit NPV; and show that factors such as imitation, when seen in an appropriate, dynamic context, produce different effects than those provided by the familiar comparative static equilibrium analysis. This dynamic analysis highlights the competitive logic of wealth creation in the sector as it evolves.

The following section addresses the theoretical background of the investigation, with particular reference to the central concept of “profit.” We then describe in turn the motivation for the model, its detailed workings, and the partial analytical solutions that can be derived. Next we turn to the description of the range of situations considered in the illustrative calculations. The final three sections present the results, their interpretations, and the broader implications.

**Theoretical Background: Profits, Rents and Resources**

For all of its centrality, the conceptualization and explanation of profitability remains a complex and challenging piece of intellectual territory (Lippman & Rumelt, 2003a). The answer to the question “what is profit, and how does it emerge?” yields multiple answers even within the discipline of economics, and a
A cacophony of opinion emerges when one looks at the uses of the concept in accounting, finance, and strategic management. This persistent diversity is perhaps attributable to the fact that clarity on what “profit” means is typically needed only “locally,” i.e., for the purposes of a particular investigation, and a locally satisfactory solution can often be framed without addressing questions that would arise in a broader context. Unfortunately, for many key analytical tasks in the field of strategic management, such “local” resolutions of the conceptual issues tend to be inadequate. We argue, more pointedly, that a wealth-creation orientation provides the best intuitive guidance for strategic analysis. The right top-level question is, “what course of action will enhance the wealth of the firm’s principals?” To address this question satisfactorily, a broad view of the relevant mechanisms is needed – broader, in particular, than the textbook-level economic analysis of “profit” typically provides.

To address this question, it is useful to sort out the key issues and mechanisms while taking note of some relevant antecedents in the literature. To do so, we consider different “types” of profit, as they emerge: First, we can distinguish between equilibrium and disequilibrium conditions, and the different sorts of returns that accrue under each. Second, we can distinguish between profit that emerges when firms use “informational” goods, which are “non-rivalrous” (such as a concept, an idea or a method) and those where firms can only draw on resources which are rivalrous in use, and scarce (location, capital, labour). And third, we can distinguish between competitive and non-competitive situations (the latter focusing on how individual firms can influence prices and as such obtain profits).

**Profitability and non-competitive behaviour**

Of these three distinctions, the last deserves particular mention, since the analysis of non-competitive situations has played a big, even dominant, role in discussions of business strategy. From the time of Adam Smith and even before, economists have been fascinated by the efforts of firms to profit through price manipulation, collusion, entry restriction and the like, as well as by the efforts of governments to develop policies to check (or sometimes facilitate) such abuses. Michael Porter (1980) famously adapted

---

2 For instance, the standard textbook analysis of the firm declares that some input levels and some costs are “sunk” or fixed in the “short run”; the fact that they are fixed makes them irrelevant to the short-run decision calculus, if not in the definition of profit (though students often resist that point). This familiar set-up is highly tolerant of ambiguity concerning which costs are properly reflected in the “true bottom line,” i.e., what the conceptually correct value is for the constant that (at least in the specific short-run context) doesn’t matter anyway.

3 Institutional devices such as patents create property rights in information and thus a form of scarcity that is artificial (in the sense that it is dependent on the institutional context, which is in substantial part the product of intentional human design efforts.) We consider the implications of this later in this section.

4 This distinction has several important aspects, but at the conceptual level the fundamental issue is one of measurement. Whereas the amount of an ordinary good is presumed to be ratio-scale measurable, independent of other considerations, information is measured only by its economic consequence and its value is intrinsically dependent on the context.
for strategy analysis the structure-conduct-performance paradigm of IO economics, turning profitability from a symptom of threats to the public interest to an indicator of business success. In this perspective, the strategy problem largely reduces to (a) how the members of the oligopolistic club can maintain a mutually beneficial order, keeping price competition in check, (b) how they can prevent others from joining the club, (c) how they can defend against diverse hazards from other sources. Subsequent analyses, often with game-theoretic tools, have explored many details of this basic framework, sometimes with great subtlety. While the central and classic problem of price determination under oligopoly still lacks a general and satisfying solution, a great deal of insight has been obtained into situations involving important asymmetries among the players, such as the entry deterrence problem. Following leads in pioneering work of Schelling, themes of *credibility* and of *mixed motives* have inspired substantial literatures. The latter theme has been particularly prominent recently as the methods of cooperative game theory have been applied to understand how the returns from an economic opportunity are divided among self-interested players who must cooperate to create value (Dixit and Nalebuff 1991; Lippman and Rumelt 2003).

It is well beyond the scope of this paper to present a thorough review of the understandings achieved in the non-competitive branch of profitability analysis. Substantial as those achievements are, our current concerns lie elsewhere. The present paper follows the lead of the many authors who have seen business competition more in terms of a struggle to achieve a competitive advantages that derive from superior efficiency (in some sense). In Williamson’s (1991) terms, our analysis relates to “economizing” as opposed to “strategizing.” (See also Teece, Pisano and Shuen, 1997; Peteraf and Barney, 2003). The division of opinion on this point in the strategic management literature has important antecedents in the IO literature. While prices remain close to the center of our story, as they obviously must in any analysis of profitability, we adopt here the presumption that is standard in competitive economic analysis: the *individual* firm does not consider itself to have influence over the prices at which it transacts. Firms collectively do of course have such influence, as our analysis describes.

Having clarified that in this paper we will not concern ourselves with the profits resulting from collusion and oligopoly, we can focus on how profit emerges in the context of a competitive setting. First, we

---

5 Demsetz (1973), for instance, suggested that the “non-competitive” approach did not often test its own premises, and highlighted the risk of *assuming* that profit emerges as the result of setting prices, without investigating the possibility that profit may, e.g., be due to differentially capable firms. He famously noted that the observed correlation between profitability and share of the market may be due to the fact that more efficient firms grow larger *and* are more profitable. This hypothesis, known as the *Differential Efficiency Hypothesis*, led to some interesting empirical research, largely in the 1980s; oddly enough, little work in the field of strategy picked up on this line of research. See Schmalensee (1987).
consider the equilibrium profits under competition (and consider how different types of resources or production factors relate to them); and then move to dis-equilibrium profits, broadly speaking, focusing in particular on the profits that emerge as an industry gradually converges to its equilibrium.

Equilibrium profits under a competitive regime

In the competitive branch of the analysis, the case of competitive equilibrium returns to ordinary resources is by far the most studied and best understood. The familiar tendency is for the profits of the firm itself, i.e., apart from any ownership stake it may have in inputs, to be zero. Theoretically, a sufficient condition for this result is that production occurs under constant returns to scale and productive knowledge is a non-excludable public good, i.e., fully imitable. The assumption of constant returns to scale can itself be derived from an axiomatic analysis of production that includes the axiom of additivity (Koopmans 1957; Mas-Colell, et. al, 1995).

Of course, the conclusion of a zero return evaporates if the firm is considered inseparable from some input; for example, if the farmer and the farmer’s land together constitute an enterprise, “the farm.” In this case, the equilibrium land rent on the (scarce) farm land obviously accrues to the farm-firm (along perhaps with farmer wages), but there is an equally obvious conceptual question as to whether such a return is properly called a “profit.” Simple as this point is, the issue continues to create obfuscation to this day in the context of the RBV, where firm “profits” are unapologetically identified with “rents” attributed to ownership of ordinary resources (see Lippman and Rumelt (2003) for a discussion and apologia concerning this issue).

A related relevant case arises if returns to scale are strictly diminishing rather than constant. In this case, as we know from standard competitive analysis, positive profits persist. The intuition for this is simple; in the presence of decreasing returns to scale, each additional unit of output becomes more expensive. So firms produce until their last unit of output is equal to their marginal revenue. But if there are decreasing returns (at the level of the sector, and in the long run), then the infra-marginal units will cost less, yielding a profit. (Though, strictly speaking, such a profit could also be considered to be a particular type of rent).6

6 Indeed, it is not entirely clear whether this profit should unambiguously be attributed and paid to the firm, or if it should accrue to what gives rise to the diminishing returns in the first place. Suppose, for instance, that there is a limited number of inimitable production functions, each controlled by a single entrepreneur. Such a return is often called an “entrepreneurial rent,” but it may also be called profit. Here, both the terminology and the reality are indeterminate on the question as to whether this return is analogous to the land rent of the farm – reflecting the input of the time of the entrepreneur as an ordinary resource – or whether the return is primarily attributable to the entrepreneur’s private information, a fully replicable, non-rivalrous “recipe” that self-interest leads her to keep secret. If the recipe is not institutionally separable from the entrepreneur’s labor time, neither the appropriate terminological choice nor the structure of the reality can be resolved.
Finally, another possibility is for profits to emerge concerns the dynamics of time compression diseconomies (e.g., Dierickx & Cool, 1989; Pacheco de Almeida & Zemsky, 2006). In this case, firms, through their constant investment to a resource or capability that can only happen over time, create a source of profit. However, it is not entirely clear whether this should be attributed as a profit or as a rent to a non-tradeable resource, since the benefit that the firm enjoys is based on its ownership of a subsidized resource (subsidized by virtue of consistent, path-dependent actions that made this either more valuable and scarce, or cheaper for the firm to use). In our model, we will consider similar dynamics that enable firms to enjoy sustainable profitability in equilibrium, without requiring inter-temporal consistency.

*From Long-Run equilibrium profits to the Short-Run profits and adjustment paths*

In addition to considering what can lead to profits “in the long run”, however, we should consider the profits that emerge over time, as an industry evolves. The case where production methods are replicable (constant returns), but not imitable, is of particular interest as a reference point and is often considered in evolutionary economics. That setting can help us bridge the long-run equilibrium analysis with the short-run dynamics, as well as to the *paths of selection and the profits that emerge along these paths*. In the simplest evolutionary story, individual firms create production methods by processes of path-dependent learning. Although an underlying fund of public knowledge may provide a common starting point for this process across all firms, the contingencies of path-dependent learning in complex situations make the operational methods significantly different. Once having achieved their idiosyncratic solutions to the production problem, firms settle on their methods permanently and seek to apply them on a larger scale when this is profitable.\(^7\) An “industry” of firms facing the same input and output markets then becomes the arena for the process of economic natural selection. If several firms have overall efficiency levels that are effectively identical and superior to those of other firms, then these efficient firms survive in the long run and the ultimate equilibrium is the familiar competitive one with zero profits.

However, while in equilibrium, we could have (in this simple scenario) a set of identically efficient firms with zero-profits, there are some important differences from the textbook economic picture of the process in which each and every firm reaches into the public grab-bag of methods and cleverly pulls out the best technique, so that the most efficient survive. The most important implication of us looking at the *path* that leads to even a fully competitive equilibrium is that we have to consider the speed and extent of the workings of the selection process. And, the selection process is a time-consuming one, and along the way

\(^7\) The reason for inimitability, i.e., whether natural or artificial, is not consequential for this analysis as long as there is nothing that can be done to change the situation. Firms also can protect their secrets, if that is the basis of the inimitability (if not, then they just add to the “pool of the common knowledge” of the industry, advancing welfare but undermining their profitability.) Firms may protect their advantage by denying “access to the template,” i.e., by not letting others closely inspect what they are doing. (Nelson and Winter 1982; Rivkin, 2003).
the efficient firms are typically making profits. Therefore, *in present value terms*, though *not* in terms of flow profitability in the long run, *differential efficiency pays (and creates wealth)*. Note another important difference to the analysis of the preceding section: The returns (as the industry converges from dis-equilibrium to equilibrium) are not rents to ordinary resources; they accrue to the superior informational resources, i.e., to the superior achievements in the initial learning process. (It is optional to attribute such superiority to the efforts of a particularly competent entrepreneur, engineer or inventor; profits can be simply attributed to the gradual workings of the selection process.)

Since our interest now shifts to the profits that emerge as the industry moves from dis-equilibrium (e.g., the early stages of a new sector, where demand, as it often does, vastly outstrips supply), we need to consider the length of the path. So some explanation should be offered for why the selection process takes time, and such an explanation can have the incidental benefit of suggesting how much time it might take. There are several possible answers to this. First, consistent with evolutionary theory, replication may not be costless (Winter & Szulanski, 2002); the services of the resources of the existing enterprise at a point of time are needed as inputs to the replication process (Rubin, 1973), and this creates adjustment costs and time compression diseconomies that check the rate of growth (Dierickx & Cool, 1989). Second, financing growth may be an issue, and capital markets may not be so perfect as to make retained earnings irrelevant to the investment rate, especially considering that the nature of the superiority enjoyed by a particular firm cannot be readily ascertained. A third potential driver for the slow operation of the hand of selection, which has not been explored in depth but we feel is quite important, is that, especially in early stages of an industry, many firms of widely varying efficiencies enter (cf. Jovanovic, 1982; Lippmann & Rumelt, 1988). And it may well be the case (as we assume in our model) that several firms that exist at the onset of a sector may not be viable in the long run. To the extent that the relevant investments are sunk or only slowly depreciating, the capacity created by such inferior firms hangs over the market and checks the profitability and growth of the superior firms. Indeed, it is entirely possible for non-viable firms not only to survive for a long time, but to exit in the end with positive present values, having made a positive contribution to the process of wealth creation.

---

8 This tale is perhaps the simplest theoretical representation for the appealing idea that it is the competitive (selection) process that ultimately gets the answers right, and what it needs to work with is “variation” and “retention.” It does not require optimization at the firm level, but only the basic motivation of profit seeking and the idea of “sticking with a winner” or “leveraging success.” It does reward competence if indeed it is differential competence that lies behind “variation”

9 A modest extension of the classic “uncertain imitability” model of Lippman and Rumelt (1982) could generate this unconventional result in an elegant way. The crucial element of sunk costs of technique acquisition is there. But in the mathematics of that analysis, the new industry jumps immediately to long run equilibrium; as a result, the empirically plausible patterns of early profitability, subsequent shakeout and slow convergence to equilibrium are not captured.
From conceptual discussion to our analytical objectives: The dynamics of efficiency profits

To summarize, research so far has focused primarily on non-competitive conditions that allow firms to profit, by controlling prices (or quantities). In addition to the non-competitive branch, and the related analysis of strategic interaction, research in the “competitive branch” has focused on the conditions that allow firms to make profit (or appropriate rents) in equilibrium. This provides us with an opportunity to consider the non-equilibrium (or, strictly speaking, the short-run-equilibrium) profits in an industry, as they evolve.

Our analysis looks both at the long-run equilibrium competitive profits (and the resulting quasi-rents), and, more important, the (potentially transient) profits in a sector as it evolves. Yet we do not consider non-competitive behaviour – although we do acknowledge that if one or a few firms are particularly efficient, their increasingly dominant role should enable them to affect prices. Thus, the hand of selection, by increasing concentration, may bring oligopoly or monopoly. In that case the assumption of price-taking becomes increasingly dubious as time passes, and the mode of analysis should jump to the non-competitive branch. For simplicity’s sake, we have not integrated this likely pattern in our study, focusing on competitive (efficiency) profits instead.

The Model: Motivation and Conceptual Issues

Model Structure

In many respects, the industry model developed here follows the modeling tradition of evolutionary economics (Winter 1964, 1971; Nelson and Winter, 1982; Dosi, Malerba, et. al., 1999). It is a “temporary equilibrium” model in which an individual period represents a Marshallian “short run.” In each such period, the interaction of the firms determines key features of their shared environment at that time, particularly prices. The environment then shapes the change of the industry state to the next period, by determining the values of firm state variables for the following period, one firm at a time. In particular, prices determine profits, and profits in turn are a major influence on firm investment behavior. The equations determining investment are understood as behavioral rules that are responsive to the firm’s prevailing state and realized values of prices; we do not consider firms to be endowed with the extensive structural knowledge that would permit them to anticipate correctly (even in probabilistic terms) the long run evolution of the industry.10 The iteration of the single period equilibrium and change processes generates the dynamic path of the industry.

10 This is the key point that distinguishes this type of model from full rationality (rational expectations) models of industry evolution such as those put forward by Jovanovic, Erickson and Pakes, Hopenhayn, McDonald and others.
The industry’s development is an evolutionary process in which selection plays a key role. Firms are heterogeneous (except in one set of scenarios included for reference purposes). Firms operate under long-run constant returns to scale, and they are capable of replicating their distinctive practices as they grow. In the long run, the more efficient firms expand and put competitive pressure on the less efficient ones, causing them to shrink. We assume that firms behave as price takers both in the short run and in their investment decisions, so the expansion of the more successful firms is not checked by voluntary restraint derived from perception of the market power associated with a large market share. Given the constant returns assumption, a legitimate interpretation of the model is that each recognized “firm” actually corresponds to a “firm type” involving many atomistic firms with identical characteristics and behavior. This optional interpretation may serve to alleviate any discomfort with the idea that a firm with a substantial market share does not recognize its market power.

In any model with constant returns to scale, price-taking behavior and otherwise standard assumptions, economic profits are zero in long run equilibrium, and this conclusion holds independent of the number of firms. When our specific assumptions are indeed “standard,” the analytical solution for this zero-profit equilibrium is straightforward. Given this familiar foregone conclusion, the long run equilibrium analysis holds little interest per se if the inquiry is motivated by questions about wealth creation and/or the nature of competitive advantage. The illumination that our analysis casts on the latter issues, which are indeed our concern, derives in part from its explicit dynamics, in part from a key departure from standard assumptions, and in part from our attention to present value outcomes as opposed to single-period profit measures. Numerical computation of the dynamic paths of the model, with comparisons across time and across parameter settings, is our primary tool in this respect. The analytical computation of the long run equilibrium does, however, play a valuable role: it provides a reference baseline for the other runs, and it makes it easy to re-normalize the model for some parameter changes, in such a way as to hold the long-run equilibrium picture constant as we examine the impact of the parameter changes on the dynamics.

Modelling (idiosyncratic) Resources

A distinctive feature of the model is its treatment of idiosyncratic resources. To motivate this treatment, we take note of some intrinsic and formidable difficulties that lie in the way of any effort to elucidate the insights of resource-based theory in a formal economic model. First of all there is the theory’s emphasis on resource heterogeneity, or in other formulations, the idiosyncratic or firm-specific character of resources. This immediately warns of complications affecting the market–level relationships of resources of similar type. Consider, for example, the category of retail locations in a major city – plainly

For a rare attempt to anchor a full rationality analysis in the details of a real situation, see Porter and Spence, and the comment by Winter.
heterogeneous, clearly in the “same market” in some sense, obviously carrying different prices. How does one characterize the price relationships in a simple way appropriate to a formal model, capturing the implications of diverse site attributes? Second, there is the question of the mechanism(s) by which a firm’s resources are “semi-permanently attached” to it (Wernerfelt 1984). Ownership by the firm is one important form of attachment, and it is one that is relatively easy to understand.\textsuperscript{11} It is clear, however, that RBT would be quite impoverished if that form of attachment were the only one recognized, and nobody seems to think of the theory in that way. It is likewise clear that the actual mechanisms of attachment are quite diverse. If one thinks, for example, of the mechanisms that keep an R&D scientist attached to her job in a particular firm, it is quite clear that these include some quite subtle considerations affecting job satisfaction, personal or family mobility, and the character of the potential alternative offers.

Third, a typical case involves numerous resources, and the market processes affecting their valuations outside the firm are complex, interwoven and shaped by idiosyncratic history. Consider, for example, a product development team with an impressive track record, and suppose a rival firm undertakes an effort to hire away the whole team. Suppose this effort falls short of its original objective because of family considerations affecting the location choices of a single team member. Considering that the effectiveness of a team derives partly from shared experience in highly complementary roles, what does the unavailability of a single member imply for the value of the team? Fourth, partly because of some of the foregoing considerations, it is quite common that actual market quotations on particular resources are rare events – a phenomenon that is familiar enough in the context of the personnel issues of academia. In the absence of competitive price quotations that pin down the opportunity cost, how do we get any clue as to whether the resource owner or the employer is capturing the rents?

Given these challenges and the hazardous analytical terrain they present, our modeling choices are guided by the objective of making the key economic issues stand out as clearly as possible in a scheme that is as simple as possible. Thus we consider a single focal resource with a price that is endogenously determined in the market every period (not a rare event). There is a second input, which we conceive as being merely “supportive” to the focal one. A possible image is that the focal resource is talented professionals of some sort, and the other input is simply the office space – of a generic type, not owned by the firm, but available every period at a given market price.\textsuperscript{12} Third, the nature of “attachment” and its consequences is modeled after the notion of firm-specific training familiar in labor economics. A firm makes a costly investment to customize the resource, i.e., to convert the generic resource to its own idiosyncratic type,

\textsuperscript{11} It must be said, however, that the task of understanding its relevance to strategy has been seriously neglected. (Jacobides and Winter 2006).

\textsuperscript{12} Because the supporting input lacks any attachment to the individual firm beyond a single period, it would not be constituent with RBV thinking to call it a “resource.”
and the benefits of this investment are realizable only within that firm. Thus the conversion does not confer bargaining power on the resource that could be used to extract a significant fraction of the benefit the investment produces. On the other hand, by conceding a trivial fraction of the rent (modeled as zero), the firm can make its offer superior to the market alternative indefinitely, thus assuring the attachment and avoiding the need to repeat the conversion investment with a resource newly acquired from the market place. Fourth, we abstract from the problems of finite lives and turnover, and make the attachment potentially permanent. We simply assume that the customization investment is a one-time cost, which remains effective so long as the resource is required by the firm. A key implication is that the costs of customization investments are borne only when the firm is growing.\textsuperscript{13}

Considered at the level of an individual employee or piece of equipment, the assumption that customizing investments are infinitely durable is extreme. The assumption serves, however, to lay bare a logic that still operates as long as these investments, made by a firm as it grows, do not have to be fully repeated every time a piece of equipment wears out or an individual leaves the firm. There are good reasons, and understandable mechanisms, supporting the claim that this is the case. The idea that firms “remember by doing” (Nelson and Winter 1982) is consistent with the emphasis in cognitive psychology on “situated” and “distributed” cognition, which underscores the fact that human cognition is fundamentally rooted in a context of artefacts and relationships. Isolated, sequential replacements of machines and personnel can be made with the support of a continuing supportive context that is largely unchanged, which greatly reduces the cost of continuing to operate an organization effectively at a given scale – compared to the cost of initially achieving that effectiveness. This is the realistic and important point whose consequences are highlighted by our extreme assumption.

\textit{Firm-level attributes: Capabilities, firm choices, and competitive interaction}

Heterogeneity in production methods is formally represented in a simple, familiar way, by a firm-specific multiplicative constant affecting the firm’s production function. Thus, firms are (allowed to be) differentially capable; capability being defined as the efficiency with which a firm turns a set of (customized) resources into outputs. This allows us to distinguish analytically between resources (traded in the market), idiosyncratic resources (special to the firm) and capabilities. It also suggests capabilities’ link to resources: To replicate its distinctive method at a larger scale, a firm needs to make the customization investment for additional units of the focal resource. It also needs an amount of the supporting input that is appropriate to the chosen scale. Although the supporting input is hired for a

\textsuperscript{13} There are potential complications (with real world significance) if a firm is fluctuating in its scale of operations and thus has changing needs for its idiosyncratic resources. Can a firm, for example, rehire its former workers after a business cycle slump or a seasonal lull? Our modeling efforts do not actually involve such contingencies, so we can sidestep this issue with a minor technical assumption.
single period, the commitment to its level must be made before prices and actual production levels are known. Hence, the supporting input acts as an input that is fixed in the short run of the individual period, implying diminishing returns to the focal resources, and forms a part of the background determining other firm choices for that period. We assume that, in setting the amount of the supporting input, the firm has in view an anticipated level of the focal resource and chooses the support quantity that would be optimal for that. In terms of the example given previously, the amount of office space hired for the period is the amount that would be long-run optimal for the anticipated level of activity in the periods, although that anticipation can only be based on previous prices rather than the actual prices of the period. This produces the familiar, analytically desirable and empirically plausible condition of short-run dis-economies of scale, with long-run constant returns to scale.

Firms also decide their level of output, and react to prices as they form in both the input and the output market. A point of contrast with many evolutionary models is that we here represent the firm as responding optimally to the price situation that prevails within the individual time period. Collectively, firm choices interact in the market with resource supply and output demand conditions to determine, jointly, market prices and the specifics of all individual firm decisions. While this formulation is in some tension with the reliance on a behavioural approach that is generally advocated in evolutionary economics, the tension is less severe than it might appear. First, from the evolutionary viewpoint, a (myopically) “optimal” rule is just another rule or routine, and deliberate optimization is just another candidate process, occasionally a realistic one, that can generate such rules (Nelson and Winter 1982). Second, prevailing market prices provide a type of guidance for behaviour that is realistically available to firms (in sharp contrast to knowledge of the future, which is not realistically available). Third, the model’s acknowledgment of heterogeneity in production methods permits the interpretation that the “optimization” is entirely subjective in any case. There is no competitive process in the model that provides an independent check on the accuracy of the optimization as distinguished from the efficiency of the learned production methods; shortfalls in the one domain are entangled with shortfalls in the other.14

14 To put it another way, suppose purely subjective constraints limit a firm to choosing within a proper subset of its objectively available possibilities; e.g., its decision makers are simply prejudiced against certain ranges of techniques. An attempted optimization that is flawed by virtue of such constraints is indistinguishable from the case of a perfectly optimizing firm that is objectively constrained to the smaller set – unless inter-firm comparisons are accepted as informative regarding the objective possibilities, or direct confirmation of flawed decision processes is available and accepted as relevant.
Model: Analytical Structure

Firm decisions and market equilibrium in the short run.

We begin with the development of the more conventional portion of the model framework. Final output, \( QF \), is produced according to a production function that is a constrained version of the Cobb-Douglas form,

\[
QF = a \cdot R^b \cdot K^{1-b}
\]  

(1)

Here, \( R \) is the amount of the focal resource and \( K \) is the quantity of the supporting input. Firms have different values of the efficiency parameter, \( a \), but the same value of \( b \). For the time being, we analyze the “representative firm” and suppress the firm indexes.

We define \( R_{int} \) (\( R \) intended) as the amount of the resource \( R \) the firm expects to be using in a period, which we will subsequently assume to be determined by the firm’s situation and results of the previous period. Denote by \( K \) the quantity of the input that optimally supports \( R_{int} \). Denoting by \( PR_e \), the anticipated price of one unit of \( R \) and by \( v \) the price of one unit of \( K \), the amount of the supporting input that is optimal given the intended activity level \( R_{int} \) satisfies

\[
PR_e \cdot R_{int} \cdot \frac{K}{v} = \frac{b}{1-b}
\]

(2)

This is the familiar condition for cost minimization when the production function is Cobb-Douglas. We assume that the price \( v \) is given exogenously, and without loss of generality that its value is one – the unit of \( K \) is “a dollar’s worth”. This yields

\[
K = PR_e \cdot R_{int} \cdot \frac{1-b}{b}.
\]  

(3)

If a firm expects to use \( R_{int} \) amount of the resource \( R \) in the period, and expects \( R \) to trade at a price of \( PR_e \) per unit, it will purchase/rent \( K \) units of the input. We modify the production function by not allowing \( R \) to surpass the intended level \( R_{int} \) by more than a certain fraction, \( Resrv \), reflecting the fact that investment is required to customize the resource, and that such investment should respond to long run considerations. \( R_{max} \) is thus defined as \( R_{int} \) multiplied by \( 1+Resrv \) (The decision rule governing the investments is specified below.)

With \( K \) and \( R_{max} \) determined, the firm faces in period \( t \) the following the optimization problem:

\[
\text{Maximize } \Pi_{op}(t) \quad \text{subject to } 0 \leq R(t) \leq (1+Resrv) \cdot R_{int}(t) = R_{max}(t)
\]

(4)

where

\[
\Pi_{op}(t) = PF(t) \cdot a \cdot R(t)^b \cdot K(t)^{1-b} - PR(t) \cdot R(t) - l \cdot K(t)
\]

(5)
The final constant does not, of course, affect the maximizing choice, but it is by definition part of the maximand. The optimization yields the short run input demand curve, relating the firm’s resource use to prevailing prices and the extent of input usage within that period,

$$ Ropt = \min \left\{ \left( a \cdot b \cdot \frac{PF}{PR} \right)^{\frac{1}{1-b}} \cdot K, R_{max} \right\} $$ \hspace{1cm} (6)

Collectively, the firms of the industry face a supply curve relating the price $PR$ and the quantity $QR$ of the resource:

$$ QR = S \cdot (PR - PR0)^{\varepsilon} $$ \hspace{1cm} (7)

Parameter $PR0$ is introduced to allow for a positive intercept of the supply curve at the price axis, which is realistic, and reflects the resource’s reservation price (or price in an alternate sector).

On the output side, the firms face the market demand curve

$$ QF = D \cdot PF^{-\varepsilon_0} $$ \hspace{1cm} (8)

Market clearing in the short run involves equating supply and demand in the two markets, one for the resource and one for final product. At this point, clarity is served by making firm-specificity explicit:

$$ S \cdot (PR - PR0)^{\varepsilon} = \sum_{j} Ropt_j(t) $$ \hspace{1cm} (9a)

$$ \sum_{j} a_j \cdot Ropt_j(t)^{b} \cdot K_j(t)^{1-b} = D \cdot PF^{-\varepsilon_0} $$ \hspace{1cm} (9b)

Here, as shown in (6) above,

$$ Ropt_j(t) = Ropt_j(t, PF, PR) = \min \left\{ \left( a_j \cdot b \cdot \frac{PF}{PR} \right)^{\frac{1}{1-b}} \cdot K_j(t), R_{max}(t) \right\} $$ \hspace{1cm} (6')

After substituting in from (6’), equations (9a-b) become two equations in the two prices $PF$ and $PR$; the solutions are $PF(t)$ and $PR(t)$.

To recapitulate the story to this point, the logic of each period starts from the firms’ intentions regarding the level of resource $R$ to be used in that period, intentions which are based on expectations derived from the experience of the past and constraints that limit possible changes. On the basis of these intentions and the expected price $PR_e$, firms commit to levels of supporting input $K$. The two equations above then determine the input and output prices for the period, and also individual firm resource levels, hence outputs and profits.
Model Dynamics.

The linkage between periods is provided by assumptions regarding expectations and decision rules affecting the choice of $R_{int}(t+1)$. We assume first that the market price of the generic version of the resource is expected to remain constant:

$$PR_{e}(t+1) = PR(t)$$ (10)

For every additional unit of the resource that the firm wishes to have available in $t+1$, beyond the amount available in $t$, the firm incurs a “capability cost” in the amount $C$. This cost is considered to be incurred at the start of a period, contemporaneously with the commitment to $K$ for that period. Thus the net cash flow of period $t$ is

$$\Pi_n(t) = \Pi_{op}(t) - C \cdot \text{Max}(0,R_{max}(t)-R_{max}(t-1))$$ (11)

The determination of $R_{max}(t+1)$ is based on $\Pi_n(t)$, representing both financing and growth incentive considerations, and an investment profitability test – but we also impose bounds on the rate of percentage change in both upward and downward directions. Specifically, the investment profitability test based on the data of time $t$ is based on the sign of the expression

$$T_{j}(t) = PF(t) \cdot a_j \cdot \left(PR(t) \cdot \left(1 - \frac{b}{b}\right) - \frac{PR(t)}{b} - i \cdot C\right)$$ (12)

That is, $T_{j}(t)$ is the excess of the quasi-rent per unit earned by the resource, in firm $j$, over the interest charge for the customizing investment. A positive value of $T(t)$ indicates an attractive investment in present value terms, assuming the current level of quasi-rents is projected into the future. So as to smooth inter-temporal investment behaviour, it is assumed that firms take their investment decision in the previous period into account when deciding on the investment for the following period. The actual increase in capacity is assumed to be a weighted average of the desired change in capacity calculated on the basis of the investment profitability test (12) (subject to upper and lower bounds), and the capacity change in the preceding period. More precisely,

$$R_{max}(t+1)-R_{max}(t) = \begin{cases} 
\rho \cdot \text{Max}\left(0, \text{Min}\left(\lambda \cdot \frac{\Pi_n(t)}{C}, \mu \cdot R_{max}(t)\right)\right) + (1-\rho) \cdot (R_{max}(t)-R_{max}(t-1)) & \text{if } T(t)>0 \\
\rho \cdot \text{Min}\left(0, \text{Max}\left(\lambda \cdot \frac{\Pi_n(t)}{C}, -\mu \cdot R_{max}(t)\right)\right) + (1-\rho) \cdot (R_{max}(t)-R_{max}(t-1)) & \text{if } T(t) \leq 0 
\end{cases}$$ (13)
where $0 \leq \rho \leq 1$. In the limiting case $\rho = 1$, investment decisions are made solely on the basis of the profitability test (12). Since $R_{\text{max}}(0)$ is undefined, the case $t=1$ is not covered by (13) and needs to be handled separately:

$$R_{\text{max}}(2)-R_{\text{max}}(1) = \begin{cases} 
\max \left( 0, \min \left( \lambda \cdot \frac{\Pi(1)}{C}, \mu \cdot R_{\text{max}}(1) \right) \right) & \text{if } T(1) > 0 \\
\min \left( 0, \max \left( \lambda \cdot \frac{\Pi(1)}{C}, -\mu \cdot R_{\text{max}}(1) \right) \right) & \text{if } T(1) \leq 0
\end{cases}$$

(13')

**Analytic Results: The Quasi-Rent Schedule, and Co-existence of Differentially Capable Firms**

Before turning to computational experiments with the model, we examine special cases in which the character of the long run equilibrium can be explored analytically. Generally speaking, we have exact analytical results only for the cases in which all firms are identical. However, when capability cost $C$ is zero, we know that the only firms that survive in the long run are ones that all have the highest efficiency level; when $C=0$ regardless of the initial heterogeneity, for selection always entails homogeneity in the long run. By contrast, when $C$ is positive, firms with differing efficiencies can coexist in the long run (the reason will become apparent later in this section). The mix of firms that co-exist in the long run is a path-dependent, historical phenomenon. Our analysis can illuminate the extent of the potential indeterminacy in the mix of survivors, but no general conclusions beyond that are possible.

*The “demand rent quasi-schedule”*

To illustrate the logic of our model, it may be instructive to start with a convenient analytical exercise that illustrates how the sector converges to equilibrium. We will consider the case of homogeneous firms, and ask some questions at the level of the industry. In particular, we want to ask “what is the quasi-rent that the firms in the sector receive as a function of increasing levels of resource use, given all other parameters?” As it may be intuitively obvious, the smaller the current level of production, the lower the resource usage (and as such the lower the price of the resource); and the higher the final good price, and as such the higher the return from using that resource. So in low output levels, there will be a positive quasi-rent – if capacity in the sector is limited (for reasons we will consider in detail later), then firms will earn a quasi-rent on the resource. We can thus create a “schedule”, which explains what the quasi-rent value is, for any given use of resource $R$, and given all other parameters.

The demand quasi-rent schedule construction is straightforward. Given a value of $R$, we consult the supply curve of the resource to identify the corresponding $PR$. That value, in conjunction with the
exogenously given price of the supporting input, determines optimal input proportions between \( R \) and \( K \). With those proportions known, the production function tells us the output quantity \( Q \) that corresponds to the initial \( R \). Consulting the demand function, or rather its inverse, we find the output price \( PF \). The value of output and its cost are now determined, the difference divided by \( R \) is the quasi-rent value for that \( R \). So, first, we consider the supply-price curve \( s(R) \)

\[
PR = s(R)
\]  

(14)

Invoking equations (3) and (1) above, we find that long-run quantity is related to \( R \) by

\[
Q = A(R) \cdot R
\]  

(15a)

where

\[
A(R) = a \cdot \left( s(R) \cdot \frac{1-b}{b} \right)^{1-b}
\]  

(15b)

Note that a relationship of the form of (15a) would prevail for any production function displaying constant returns, but the particular Cobb-Douglas specification is obviously involved in (15b).

Let the demand-price function be \( h(Q) \):

\[
PF = h(Q)
\]  

(16)

Then the desired result, the demand quasi-rent per unit (\( PRV \)), is obtained as

\[
PRV = \left( Q \cdot h(Q) - PR \cdot R - 1 \cdot PR \cdot R \cdot \frac{1-b}{b} \right) / R = \frac{Q \cdot h(Q)}{R} - PR - PR \cdot \frac{1-b}{b}
\]  

(17a)

Here, equation (3) has been invoked again. And this gives us the quasi-rent value. Now if there would be nothing to halt firms from expanding, this expression would be zero; at which case we can substitute zero at the left hand-side, and define the relationship between \( Q \) and \( R \), for given levels of \( b \).

However, we know that there is a cost to expanding; and this cost is the function of the expansion cost \( C \) and of the interest rate \( i \). So, barring any other forms of output growth (such as those that would emerge through collusion or other non-competitive behaviour in the long run), we can match the “demand rent quasi-schedule” (the benefit from the use of resource) to the “cost” of customizing this and putting it to use within the company. We can do so analytically (provided all firms are identical). So, invoking (15a), \( R \) conveniently divides out, and (17a) can be re-written as

\[
PRV = A(R) \cdot h(A(R) \cdot R) - s(R) / b
\]  

(17b)
For given values of interest rate $i$ and capability cost $C$, the equation

$$A(R) \cdot h(A(R) \cdot R) - s(R) / b = i \cdot C$$

(18)

thus determines the equilibrium value of $R$ (and hence all other variables) in the sense described above. Note the particularly simple form that appears if $s(R)$ is a constant, with the result that $A(R)$ is as well. This formula can be a convenient benchmark for some comparative statics (under certain dynamic conditions)\(^{15}\); it also explains the dynamic rationale with which firms converge to equilibrium.

**Co-existence of differentially capable firms in equilibrium**

While the demand quasi-rent schedule provides an analytical solution for identical firms, it cannot tell us where the system will equilibrate when firms do not have homogeneous capabilities. However, it can provide a different, and quite valuable insight, by helping us understand when differentially efficient firms might co-exist in equilibrium. To do so, we have to consider a situation with any two firms with unequal capability (where $a > a'$) and see the conditions under which firm $a$ would not want to expand, allowing for firm $a'$ to co-exist, if it is in the sector already. (For instance, if $a'$ had already expended its “capacity expansion cost” in the past, then it would be pushed out of the market if firm $a$ would find it profitable to expand its capacity). So the analysis described above will still hold, and will focus on the most efficient firm, with capability $a$. What we know for sure is that the most efficient firm cannot be in equilibrium if its prevailing quasi-rent rate exceeds the value that would give it an incentive to expand its capability. We also know that none of the less efficient firms will have an incentive to invest when the most efficient firm is getting close to zero incentive. In effect, then, there is a locus in the $(PR, PF)$ space where equilibrium might occur, if it is a case where the asymptotic $R$ value of the most efficient firm is approached (gradually) from below. This locus can be obtained by modifying (18), replacing $h(\cdot)$ by $PF$ and $s(\cdot)$ by $PR$, and taking the $a$ value corresponding to the most efficient firm:

$$a \cdot \left( PR \cdot \frac{1-b}{b} \right)^{1-b} \cdot PF - \frac{PR}{b} = i \cdot C$$

(19)

A less efficient firm that has already sunk an investment in capability will not have the incentive to contract as long as its $a$ parameter, $a'$, satisfies

\(^{15}\) This is a handy form for comparative statics analysis of the long run system. For example, the response of $R$ to a change in $C$ can be determined in the usual fashion by treating (18) as the definition of a function relating $R$ to $C$, and differentiating the identity with respect to $C$. However, since we have a path-dependent dynamical system in the background and not just an equilibrium condition, the result of such “comparative statics” analysis need not describe the actual result of a perturbation of the dynamical system at a point near its equilibrium! In fact, it does not describe it if the perturbation is an increase in $C$. In that case, any remaining expansion incentives are eliminated, and equilibrium is established at a higher level of activity than would have occurred if the higher $C$ value had been prevailing all along, or the increase had been foreseen.
Thus the “coexistence condition” for $a'$ is

$$a' > \frac{PR}{PR + i \cdot b \cdot c}$$

(21)

We can further simplify this expression; given that we have expressed $C$ as an absolute cost value, it is easy to scale it as a fraction of the prevailing price for the resource $PR$, and set $c$ as the “cost to invest in a unit of resource as a fraction of the cost of the resource itself” (which is a more intuitive measure). Since $C = PR \cdot c$, $PR$ then conveniently simplifies from this fraction, yielding

$$a' > \frac{1}{1 + i \cdot b \cdot c}$$

(22)

This equation presents a fairly important result. It shows that given the efficiency level of the most efficient firm, survival prospects for less efficient firms tend to be an increasing function of the capability cost, the interest rate, and the resource elasticity of output. In other words, it suggests that firms of heterogeneous skills may well persist, even in long-run equilibrium (as long as they happened to “be there” early on), thus providing an evolutionary explanation for heterogeneity in capabilities. It also shows that the extent of dispersion that can be sustained in equilibrium is a function of the extent to which it is difficult to convert “generic” into “idiosyncratic” resources, as well as of the interest rate, and of the relative role of these “idiosyncratic” resources in total output.

Yet, while this provides an interesting link between RBV theorizing and equilibrium analysis, it does not offer a full analysis of how exactly $C$, or capability dispersion affects not only the co-existence of firms, but also profits, resource prices, and quantities. More important, this calculus focuses on what happens in

---

16 This result is somewhat over-simplified, as $PR$ is endogenous to our model, whereas $C$ has been set as a fixed cost, not dependent on price; so $c$ cannot be thought of as a true constant, since it depends on the value of $PR$. So in that sense, (21) allows us to capture some further intricacies of our systemic setup that are not visible in (22). An instructive thought experiment is to consider what happens when the resource supply is infinitely elastic, so $PR$ is a given constant. Assume that there is an equilibrium of the kind characterized above, with one or more less efficient firms along with the most efficient one, and the less efficient ones not at the edge of being selected out. Consider a small shift of some share of resource use from the most efficient firm to a less efficient one, such as would appear if the initial size of the two firms had differed a bit in the indicated direction. The effect of this is to reduce output and raise the output price, since production is now less efficient. The most efficient firm is therefore out of equilibrium and will expand, driving price back to its initial level. The less efficient firm will not contract in this process; it remains with augmented resource use. Therefore total resource use increases, though to something less than the initial value in the most efficient firm, while output and price return to the previous values. The output consequence of the reduction in productivity is just compensated by the expansion of the most efficient firm. The result is a quasi-rent increase for the less efficient firm, per unit as well as in total, and a larger decrease for the most efficient firm.
equilibrium, leaving the important issues of paths of adjustment, of the net profits given the adjustment paths and of the industry dynamics we discussed earlier, entirely in the dark. For us to consider them, we need to shift from closed-form solution, to numerical investigation, treating our model as an “experimental design”, and then considering the “results” of the model and their theoretical implications.

Experimental Design

To better understand the workings of the model and the determinants of profitability, we now turn to the calculation of illustrative adjustment paths. We vary the experimental conditions in three different ways. There are three values for capability cost $C$, of which the lowest is effectively zero and the largest produces an equilibrium quasi-rent of approximately the same magnitude as the price of the generic resource. We examine both the model as described above and a modified version that includes the following imitation mechanism: each period, firms improve in efficiency by closing a fraction $\beta$ of the gap between their current efficiency and that of the most efficient firm. Our “base case” is characterized by the intermediate $C$ value, and no imitation. Within each of the six parameter settings thus defined, there are eleven heterogeneity “scenarios”. These range from no heterogeneity to a case where the efficiency ratio between the best and worst firms is three to one. The specific formulation is that there are eleven firms (or types) spread uniformly across an interval centered on $a = 1$, and ranging from $1 - x/20$ to $1 + x/20$, for $x = 0, 1, 2, ..., 10$.

In all cases, we posit an initial state representing an early stage in the process of industry evolution. Firms have already developed their production methods, but industry resource use is a modest fraction of the equilibrium values. The initial resource levels (determined by $R_{int(1)}$) are constant across settings and scenarios. Initial resource use is the same across cases, and in the base case is 32% of its equilibrium value; increasing $C$ makes equilibrium output smaller and in this sense implies that the initial condition is closer to equilibrium. We compute the model outcomes for 50 years, so that even the slowest-moving equilibration processes have time to produce something close to their asymptotic results. (For example, in the scenarios with low but positive heterogeneity, it takes a long time for the modest efficiency differences to express themselves in firm-size differences.) The profitability conditions of the final period are then extrapolated to infinity for the purposes of present value calculations, assuring that there are no distortions arising from arbitrary truncation of the cash flow series. At reasonable discount rates,

---

17 Mathematically, this problem consists of a set of $n$ individual optimizations, linked through two global conditions (the markets for final and intermediate resource, that need to equilibrate). This yields a Mixed Complementarity Problem (MCP) structure, which can be solved using Dynamic Non-Linear Programming. (This model draws on Jacobides, 2006, where methods are discussed in greater detail). We used GAMS, and employed the PATH algorithm (see Ferris & Munson, 2000).
however, this adjustment is of negligible consequence. As we discuss below, this observation reflects a fundamental qualification to the strategic significance of “sustained competitive advantage.” For our discount rate we use 7.5%, which reasonably approximates the long run nominal rate of return on U.S. equities. At that rate, the time zero present value of 100 dollars in year 51 is $2.50.

For this first set of experiments with the model, we chose some key parameters with an eye to the interpretability of results, and also to assure computational stability. The elasticity of demand is set to one, which means that revenue is a constant and thus not an active factor in the determination of the industry profitability picture. The supply curve of the generic version of the resource is chosen to be highly elastic, though not perfectly so. The consequence is that the effects of changes in the market price of the resource are visible in the results, but are not powerful determinants. We constrain firm growth rates (of $R_{max}$) to a maximum of 20% on the upside and a minimum of -20% on the downside.

The Cobb-Douglas exponent on the resource is 0.8, reflecting our intention to give the resource the dominant role in the picture, relative to the generic “supporting” input. We set the upper limit on resource use in the short run to be 10% above the level corresponding to minimum of short run average cost, given the chosen level of the supporting input (and assuming the prevailing input price of the previous period). (i.e., in equation (4) above, $Resrv = 0.1$) Thus the short run behaviour at high output prices is similar to what would happen if the production function displayed fixed coefficients; the output supply curve turns vertical. At low output prices, however, the Cobb-Douglas form implies that very high resource productivity is achieved at very low utilization rates of the supporting input. The result is that a firm with positive capacity – and all firms are always in that condition given the bounded shrinkage rate – always chooses a positive output level. Thus inefficient firms “fade away” from the market place; they do not disappear abruptly. This formulation is not recommended on the grounds of realism, but it is convenient in an exploratory exercise that does not allow for continuing entry and thus has no mechanism to balance the effects of permanent exit on the industry dynamics.

Also highly relevant to the model results is our specific assumption concerning investment finance. When capability increase passes the profitability test (eqn. (12), it is constrained to something less than the allowed maximum only by the level of net profit, i.e., the current operating profit less the previous amount of capability expenses (eqns, 11, 13a-b). Firms might obviously spend more than that, by borrowing. On the other hand, given a restriction to internal financing, they would spend less if capability increase involved costs other than those of customizing the resource – e.g., if the supporting input had to be owned rather than rented – or if there were adjustment costs of some sort, or if there were a dividend payout. We consider that interpretation of the time scale of events in the model is facilitated by our assumption, which is on an interpretable dividing line between considerations pointing to higher
investment levels and ones pointing in the opposite direction. We leave to the future the exploration of alternative assumptions either way. We note, however, that the effect of this assumption is particularly powerful when capability cost is negligible, for in that condition the maximum growth limits are the effective control on investment rates in almost all circumstances. Since it is fundamental to our approach that capability cost is not zero in interesting and relevant cases, the distorting effect when C=0 does not represent a significant flaw in our results.

Our setup also allows us to track imitation – the ability of firms to learn from and emulate superior ways of organizing production. Specifically, we model imitation through the following equations:

$$a_{i}(t+1) = (1 - \beta) \cdot a_{i}(t) + \beta \cdot \max_j a_{j}(t)$$  \hspace{1cm} (23)

with $0 \leq \beta \leq 1$. The parameter $\beta$ describes the speed of at which the less capable firms catch up with the most advanced firm. However, our baseline results do not consider imitation; we look at imitation as a treatment, in our result section. Tables 1 and 2 summarize the parameter settings of our baseline case, and the experimental design.

<table>
<thead>
<tr>
<th>S</th>
<th>D</th>
<th>E_S</th>
<th>E_D</th>
<th>MU_G</th>
<th>MU_D</th>
<th>Resv</th>
<th>V</th>
<th>PR0</th>
<th>L_+</th>
<th>L_-</th>
<th>Rint(l)</th>
<th>K(l)</th>
<th>B</th>
<th>C</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1050</td>
<td>4</td>
<td>1</td>
<td>0.15</td>
<td>0.15</td>
<td>0.1</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5.25</td>
<td>0.8</td>
<td>10</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 1: Parameter settings of the baseline case

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Imitation settings</th>
<th>Capability Costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 &quot;scenarios&quot; of firm heterogeneity per run, with a ranging from 0.5 to 1.5 (a variance range: 0 to 0.5)</td>
<td>“Off” and “On”, with $\beta = 0.05$</td>
<td>CC= 1, 10 or 100</td>
</tr>
</tbody>
</table>

Table 2: Experimental Design: Within and across Scenarios

Note that for each “scenario run”, we ran 11 optimization problems (11 scenarios) for 50 periods. So for the results presented in this paper, we ran 550 x 3 (values of cc) x 2 (settings of imitation), i.e. solved 3,300 linked optimization problems, which were compiled from GAMS to Excel and reported graphically.
Results

We rely on graphical methods to present our results, since these convey a strong sense of the qualitative features that are the main interest. Since our model is deterministic, there is no need for multiple runs to provide a clear picture of the model’s logic. We rely particularly on a 3-D display with the value for a particular variable shown over a plane; the two dimensions of the plane are the time period and the heterogeneity scenario (1 no heterogeneity, 11 maximum). The results thus displayed include some key industry aggregates and descriptive statistics as well as variables that actually enter the model logic.

Industry evolution: Profits, Prices, Quantities, Dispersion

The familiar logics of industry evolution are visible in common features of the computed industry trajectories; these form a qualitative picture that is the common background for the experimental effects that are of central interest. Figures 1a – 1e show the trajectories, in the baseline case, for output price, total operating profit, the average (share-weighted) capability in use, the variance of capability, and the resource price. The average capability in use is the weighted average across firms, with output share weights. The variance referred to is the corresponding Gini concentration coefficient of productivity.

Figure 1a: Output price for C=10

It should be noted that while the model is deterministic, it is a dynamic model with a lot of non-linearity in it. It can therefore generate visible “pseudo-random” effects in the results, but these do not appear to be substantial enough to be worthy of attention. The possibility of a substantively significant “bifurcation” (or “butterfly effect”) cannot be ruled out, but we have seen no evidence of such and have no reason to expect it.
Figure 1b: Total operating profit for C=10

Figure 1c: Average capability in use for C=10
In the first two panels, the most obvious feature is the decline in price and profitability associated with the approach to equilibrium from an initial condition at far below equilibrium capacity. When firms are identical, there is an abrupt transition over three periods from a maximum growth condition to virtual equilibrium. This transition is rounded off to an increasing extent with increasing heterogeneity, because
firms are differentially slowed by the costs of capability investments. Note that the operating profit total is clearly approaching an asymptotic value that is above zero; this is the quasi-rent consequence of the capability cost and receives more extensive comment below. In panel c, the effects of heterogeneity and selection are portrayed. The average (share-weighted) resource productivity increases with the heterogeneity, and to an increasing extent over time as selection has time to enhance the market shares of the most efficient firms. The final panel shows the decline of the variance of productivity as share becomes increasingly concentrated in the most productive firms. Note however that the variance does not appear to be approaching zero. This is a consequence of the long-term coexistence of firms of different efficiency, a possibility discussed above in abstract terms (see eqn. 21 and related discussion).

These basic features are accompanied by others, in a reasonably straightforward way. Total resource use rises steeply in the initial phase, but then stabilizes as total output growth slows. While output grows and price declines continue as a result of relatively slow selection effects, resource use is effectively constant. This is because, with unitary demand elasticity, the decline of output price exactly offsets the effect or rising productivity on input demand. Figure1e shows the amount of total resource used, that confirms this. (We leave the analysis of how higher or lower elasticity of demand affects industry profits (and resource prices) for future research; suffice it to say, it does matter, and interacts with firm heterogeneity in interesting, and under-appreciated ways).

Sustaining Profitability: Causes and Consequences

As implied by our theoretical discussion above, the profitability test for capability investment ultimately checks the growth of capacity and output, and produces an industry equilibrium situation in which price exceeds the continuing production costs for the most efficient firms. The effects of increasing capability cost are seem most clearly in the contrast between our $C=10$ baseline case and the $C=100$ case. The final value of output price is increased by over 75% in all heterogeneity conditions.
Equilibrium output is correspondingly reduced, and the time required for equilibrium to be reached is reduced by about 25 years, or 50%. Equilibrium operating profit accounts for about 26% of revenue, as against 1% in the baseline case (both because it is much higher, and because it “starts” earlier on). Also, in contrast to the baseline case, the levels the generic resource price is substantially lower, as can be seen by comparing Figures 3a and 3b. The reason for this price discrepancy is that high $C$ means that firms do not wish to expand, which in turn means they will use less of the resource. So while firms are expending
more on “training” and turning their resources to idiosyncratic uses, the resource owners are likely to suffer – they do create value they do not capture, and this unfortunate predicament is caused by the fact that they become “sticky”, attached to a particular firm. This is precisely the rationale we identified in our discussion of the demand quasi-rent in the previous section. Quasi-rents in 3b are particularly high, but firms compare this quasi-rent with the value of $C^*_t$ as we saw earlier on in our analytical exposition. This suggests that in settings where profits are more sustainable, resources may lead to quasi-rents because of the dynamic considerations of training and adapting them to the firm.

Also, it is interesting to note that in 3b generic resource price are visibly (though not substantially) increasing across heterogeneity scenarios – which is to say that the higher $C$ value depresses these by greater amounts at lower heterogeneity. This interaction effect emerges because of the path-dependent competitive dynamics. In high-$C$ sectors, firms have limited reasons to expand; they are all “average”, and face high costs of expanding. However, if there are heterogeneous firms, then the best of them will still expand, even though it costs quite a bit. (We discuss this effect in the following section as well, suggesting that whereas in low-$C$ sectors capability heterogeneity helps profits, in high-$C$ sectors it eventually hurts it- it creates more competition and leads to higher capacity, thus pushing resource prices and output up and demand quasi-rents and profits down.

Figure 3a: Resource price for $C=10$
Also as analyzed above, positive capability costs promote the long-term coexistence of firms of different efficiencies. When such costs are negligible (\(C=1\)), the familiar selection logic is fully in effect. The output share of the most efficient firm in the highest heterogeneity scenario has reached 97.6% by period 50 and is headed for 100%. In the corresponding scenario for baseline case, the top 2 firms are headed for long-term survival, with output shares of 70% and 26%. And in the high capability cost case, it is 7 firms with shares from 32% down to 5%.

Figures 4 shows measures of the cross-sectional dispersion (Gini coefficients) of operating profit in period 50. Panel 4a shows the result across firms for the three capability levels in terms of dollar amount; 4b shows the corresponding picture for operating profit as a fraction of revenue. As we would expect, in all values of capacity cost, the higher the heterogeneity, the higher the difference between firms – even in the future. What is interesting is that the greater the cost for expanding, the higher the dispersion in terms of total profits (figure 4a). This is because in higher-capacity expansion cost scenarios, the most profitable firms have managed to be quite profitable, and maintain the level of profitability- capacity expansion “protects them” and their profits. However, if we look at profitability (figure 4b), the picture is different. True, as we would expect, heterogeneity in firms leads to greater dispersion in terms of profitability; yet when it comes to rates of profitability, as opposed to total profits, higher CC costs tend to reduce the dispersion between firms. This is because in the high capability dispersion scenarios, even less efficient firms have managed to maintain a respectable level of profitability, as they cannot be “pushed out” of the market; as such profitability rates are more similar than they would have been with stronger selection.

Figure 3b: Resource price for C=100

Co-existence of firms in equilibrium, and dynamic paths
The important role of capability costs in affecting long-term industry dynamics can also be confirmed by looking at the concentration ratios in the two polar cases - CC=1 and CC=100. In the latter, the most effective firm expands, taking the entire market (but driving profits down, not allowing for any sustainable profits); and in the former, concentration remains low, as even the most efficient firms do not find it profitable to expand; this yields an interesting case of a highly fragmented market, with a number of quite (occasionally, very) profitable firms. This could be likened to some of the law or consulting firms, which do have very high resource customization costs and which, even without any collusion in prices, manage to keep high profits. One could even extend this logic to business schools and their executive education divisions.
Another interesting thing to note is that, from the wealth creation viewpoint, sustainable profitability is not as interesting as net present value. For instance, consider the proportion of the equilibrium profits (here, considered in all scenarios to be period 50, which is a somewhat arbitrary, even if not unreasonable cut-off point) of sustainable profits, when divided by all the profits (seen in terms of Net Present Value). Figure 6a provides this percentage for CC=1, 10 and 100. As we can see, the percentage is pretty negligible; even for CC=100, the proportion of equilibrium profits, which we have extended into infinity (a fairly generous assumption) is below 3.5%.
Yet someone might argue that this is simply the result of discounting returns far off in the future. To such an objection, we would retort that this is precisely the point – it seems that strategists, excited as they are for some ultimately sustainable returns “sometime in the future” miss the point that it is the near-term and the adjustment path that matters. Yet we can complement this analysis further. So Figure 6b provides a much more stringent test for the temporary profits. As such, in Figure 6b, we provide an “artificial” picture, which considers what would have happened if all industry evolution happened overnight, and the equilibrium profits started flowing from day 1; and compares this to the actual profits in the sector. As we can see, even with this criterion, only in CC=100 do sustainable profits seem to be important. (To give a sense of proportion, note that the price for the resource, in our simulations, is around 10 to 14; so that a cost of 100, even if amortized, is a quite substantial fraction).

![Figure 6a](image-url): Sustainable profits as a percentage of total profits, in NPV terms

![Figure 6b](image-url): Hypothetical Sustainable profits from period 1, as a percent of total real profits, in NPV terms
However, our analysis offers us more insights than just demonstrating that the infatuation with sustainable profits may be a bit misplaced. In particular, it provides an appreciation of competitive dynamics over time. Indeed, it is particularly instructive to consider industry and firm-level profit NPV. In this paper, we only consider two cases – the case with low CC, in Figure 7a, and the high CC case, in Figure 7b. As we can see, for low CC, the more heterogeneous the industry, the greater the NPV; the reason is that high capability differentials allow the most efficient firms to make good profits for a while (as we saw in Figure 1b) and as such this supports profitability.

Interestingly, though, in higher levels of CC, this relationship is inversed, for lower values of capability heterogeneity, so that two forces are at play: one that accounts for the benefits of heterogeneity in generating profits; and the other eroding profits as heterogeneity increases, creating this curvilinear pattern (see Figure 7b). The logic of this “opposing force” has to do with the dynamics at the firm level (even without any strategic interaction). We should note that higher capability dispersion is beneficial to the most efficient firms even with high CC; but it is proportionately less beneficial. In addition, if there is a high CC, then more equal firms simply don’t have much incentive to grow; and there is no efficient firm pushing them out. So the industry stabilizes in a comfortable equilibrium. On the other hand, if there are particularly efficient firms, they have to expend the costs for expanding, as they will; and there will be more (and more efficient) capacity in the industry, driving prices down (as compared to the baseline of all firms being identically efficient). So, perhaps counter-intuitively, in an industry with its own, “built-in” sustainability protection, firm heterogeneity may end up depressing profits at the level of the sector. This is an insight which we can observe only by focusing on NPV – an approach we hope will start being considered in academe, as, we believe, it is used in practice.

![Figure 7a: Net Present Values for Profits, per capability dispersion scenario, CC=1](image-url)
Imitation: from “erosion of advantage” to “impact on NPV”- and sustainability

As we noted in the model setup, we consider firms with potentially heterogeneous capabilities in our sector, so as to track the role of firm heterogeneity in competitive dynamics and profitability. Our setup also allows us to track imitation, for which we report the results.

As it may be evident, imitation serves to endogenously limit the extent of dispersion between the most and least efficient firms – and as such, in all scenarios, speeds up the improvement of the average capability-in-use. Essentially, imitation means that in addition to the Darwinian process of capability improvement through capacity expansion of the most efficient firms, there is a Lamarckian process whereby capabilities of all firms are improved.

As imitation reduces the differences between the most and the least profitable firms, it always reduces the profits of the most efficient firms, both in equilibrium and as the industry progresses, since these firms “give up” part of their advantage, through imitation. Yet while the better firms are worst off, the worst performers are better off, as they benefit from learning from their more effective competitors; if we look at the firm-level profit NPV, we see that the worse performers gain from imitation as they are given the opportunity to join in the feast, especially early on in an industry’s life-cycle. True, this hurts the sector later on; but benefits (profits) have been created at the short term.

Figures 8a and 8b show these results. Figure 8a shows profits under imitation in our base-case scenario (cc=10) – this is the equivalent of Figure 1b. What is noteworthy is that it looks pretty similar to Figure 1b, and this suggests that imitation does not instantaneously wipe out profits – nor does it reduce the firms’ profit NPV. Figure 8b provides a clearer image of this, by showing the “treatment effect” of
imitation on profits: It shows how the “base-case” results (i.e. industry evolution without imitation) differ from the “treatment effects” (i.e. industry evolution with imitation). Thus Figure 8b shows how profit changes over time as a result of imitation (treatment inverted, for visual reasons, so that this shows how much greater profits would be without imitation). Two things stand out. First, imitation “hurts” after capacity has grown enough to cover demand; before that, everyone can turn a profit, so imitation is not so detrimental. In other words, when it comes to profits, Darwinian and Lamarckian selection mechanisms seem to be substitutes, rather than complements. Second, we see that under some conditions, profits may even increase as a result of imitation, at the level of the sector (see the brown values). Figure 8c and d provides another treatment effect – it shows how the NPV of the total and equilibrium profits changes as a result of imitation, over the 11 scenarios of capability dispersion, for C values of 1 and 100.

Figure 8a: Operating Profits, CC=10, allowing for imitation

19 Note that the use of such “treatment effects” opens the way to a series of exciting questions, some of which we consider in our companion paper. For instance, this analysis of “treatment effects” shows that imitation’s impact on profits crucially depends on the elasticity of demand, which here has been conveniently set to 1. We find that when demand is elastic, imitation generally leads to greater profits. The strategic rationale is compelling: imitation helps reduce the total cost in the industry; and with elastic demand, cost reductions lead to substantial additional demand, thus leading to greater profitability through the dynamics which we term “market deepening”. This finding alone qualifies the general fear of imitation as a factor that is sure to erode industry profits.
An interesting mediating factor of the impact of imitation on profits (and their distribution) is the capability cost, $C$. The higher the level of $C$, the greater the “beneficial” aspect of imitation. This fairly counter-intuitive result is due to our model dynamics and parallels our discussion of the way $C$ interacts with capability heterogeneity to drive profits. In the presence of high-$C$, imitation allows firms to “catch up”, and as such retain their profitability, which may well become sustainable; more efficient firms do not need to expend costly training as they push their rivals away, since their rivals are becoming better. And
the loss of NPV of profits from the better firms is more than compensated by the gained NPV from the firms that imitate, so that *imitation leads to greater aggregate sector-level profits*.

Figure 8c is particularly interesting, as it shows that the equilibrium values (the NPV of sustainable profits) are affected by imitation – and that imitation can also *increase the NPV of the sustainable profits*, the increase being fairly uniform in high-$C$ settings. This result is, *prima facie*, surprising. Why would imitation affect equilibrium values of NPV – much more so, why would it increase them? Since imitation acts so as to reduce the capability variation (but cannot affect the upper bound, i.e. the best capability in our sector), why wouldn’t we get the same sustainable profits in equilibrium, whether the convergence happens through Darwinian or Lamarckian mechanisms, i.e. from selection or imitation? The answer is that equilibrium profits depend on the adjustment paths – they depend on the decision calculus of firms that consider when the quasi-rent is high enough to motivate training more resources, and converting them into firm-specific idiosyncratic inputs. In high $C$ settings, imitation acts as a “check” on the very aggressive growth plan of the effective firms; as the industry gradually becomes homogeneous, effective firms have less reason to expand (the other competitors are not quite as inefficient), and as such the industry equilibrates at a smaller capacity level, which also means higher prices and lower resource costs. So a different path can allow the industry to equilibrate to different levels of sustainable profits – and imitation may well end up helping increase sustainable profits in NPV terms.

This last point has some important conceptual ramifications. It shows that not only the NPV of a sector, but even the long-run equilibrium values depend on the adjustment path. And it suggests that familiar factors such as imitation, when seen in the appropriate, dynamic context, act in ways which are subtler than those implied by simple equilibrium reasoning. It thus underscores the need to shift both to an explicit dynamic analysis, and to the exploration of a sector’s NPV of profits, rather than flow profitability at any one time.

**Discussion**

Our model formulation centered on capability cost seems, on the one hand, to be highly consistent with the theory of sustainable advantage presented in the RBV literature, and on the other, to resolve some outstanding puzzles about the relationship of that theory to standard microeconomics. We show how profitability that is *sustainable in long run equilibrium even under competitive conditions* may derive from the possession of idiosyncratic rent-earning resources. In our formulation, “idiosyncrasy” matters not just because it might have something to do with the feasibility of imitation by rivals, but because it has everything to do with the input market “imperfections” that allow a firm to trap the rent stream from a resource that it does not necessarily own – a resource that, in its generic form, is priced in the market every period. In the interest of framing this conceptual issue in the sharpest possible terms, we have
posited that the specialized version of the resource that is effective in an individual firm is literally unique in the sense that the customizing investments that create it produce no value in any other firm. This extreme formulation is only a guidepost, but a valuable one, for understanding the complex reality in which the degree of idiosyncrasy varies with the routines of individual firms. A related feature of that reality is that the precision of market pricing of resources varies markedly with the level of aggregation—being generally higher for substantial packages of complementary resources (whole firms or business units) than it is for the constituent elements of those packages (individual workers with firm-specific skills, customized machines and structures, etc.). Our formulation is consistent with this point as well, in the sense that our image of how firms make money from the idiosyncratic resources that serve their routines does not in any way imply that investors can capture superior returns by simply buying firms that possess such resources (Barney 1986).

While the heritage of Ricardian rent theory remains highly visible, our picture of the origins of rents departs from that tradition in ways that are plainly congenial to RBV thinking. Far from being “the original and indestructible powers of the soil” (with access thereto being acquired simply by a transaction) the ultimate source of the value that resources contribute is something that the firm creates through a time-consuming process—and creates to an extent that is guided and constrained by market competition. Here again we have suppressed important aspects in the interest of simplicity. The only resource “leveraging” that is reflected in the model is the simple kind featured in basic evolutionary models—the profit-motivated effort to do more of the same, which is accomplished by replicating routines. We have consigned to the pre-history of our process the creation of the idiosyncratic routines themselves, which are the typically the fruit a complex interweaving of creativity, learning and investment over extended periods (Dierickx and Cool 1989, Montogomery, 1995, Winter 1995).

As our introductory discussion emphasizes, the interpretation of any “rent-as-profit” story must address the key question, “is this return legitimately called profit?” While there can be no definitive resolution of this question—because the legitimacy criterion is ultimately a matter of terminological taste—we argue that the persistent returns in our model are reasonably called profits. As a first observation, of considerable practical significance, we note that such returns reflect continuing benefits to which no continuing outlay corresponds, and it is unquestionably the case that standard accounting will show such a return as a part of net income. Similarly, there is little question but what the types of costs involved in customizing a generic input are typically expensed rather than capitalized; the books of real firms show no asset corresponding to most capability investments. Thus, to the extent that propositions about sustained profitability have a connection to empirical evidence that is based on accounting measures, our

---

20 There is the possible exception of goodwill, when control of the asset has been achieved via purchase.
candidate for a profit concept serves the purpose. Second, we note that such profits are indeed the sort of net return that firms pursue, i.e., strive to increase through effort, investment and the exercise of strategic judgment (though much of that process is suppressed in the model, as we have conceded). Third, we can respond to the possible objection that our profits “merely” reflect the normal return on investment. Actually, “normal return” plays a narrow but critical role in our model; it controls the margin of the process of capability creation for the most capable firm. On infra-marginal units in the resource creation process, the most capable firm typically makes above-normal returns, and other firms may benefit from their capability investments as well, even when though they fail to yield quasi-rents corresponding to normal returns in the long run. Here and generally, the industry evolution framework serves to underscore the point that economic competition is typically analogous to athletic competition in a track event, not a field event. In the broad jump or the pole vault, your medal chances may be quite insensitive to precisely when you start your run at the pit or the bar – but that is not true at the start of the 100 meters. There, timing is of the essence.

In the existing literature, the “uncertain imitability” model of Lippman and Rumelt (1982) offers an analysis of sustained profitability with key similarities to the present treatment. Like our own, that model examines competition among heterogeneous firms. The ex ante perfect symmetry among potential firms is broken by a chance mechanism that assigns different, and permanent, efficiency levels to different firms; in our case this mechanism is implicit in the learning contingencies that give rise to the differing routines and efficiency levels at the start of our dynamic process. There are costs that must be incurred to establish a potential firm as an actual firm; in their case this cost is the one-time purchase of the information underlying efficiency, information that can be assessed by the purchasing entrepreneur only in probabilistic terms. Since these are not continuing production costs, yet are voluntarily incurred only in anticipation of returns, the equilibrium picture is one of firms earning differential rents. Further entry is blocked by the combination of the investment requirement and the uncertain outcome of an effort to imitate existing success.

Our analysis also differs from Lippman and Rumelt in a number of respects, which mostly have to do with its grounding in evolutionary economics. First, like most evolutionary models, ours respects the additivity axiom of production theory. By contrast, the production functions assumed by Lippman and Rumelt are not merely inimitable, they are also non-replicable by the firm itself. Returns to scale are diminishing, and this is the basic explanation for the fact that firms of different efficiencies can coexist in the equilibrium they describe. The question of why the scale of application of information should be

---

21 The critic might concede that capability creation is a relatively unfamiliar category of investment.
limited *even for its initial possessor* is left dangling.\(^{22}\) Second, our model shares the common evolutionary premise that mistaken decisions not only happen but have lasting consequences in the evolution of the system – in fact, these negative consequences can be viewed as the obverse side of the benefits that behavioral variety brings to the system. Lippman and Rumelt, by contrast, adhere closely to the full rationality paradigm of mainstream economics, and manage with great analytical skill to describe an equilibrating process in which all of the worst mistakes (i.e., the investments with no returns) perish promptly and without consequence for other actors. Third, and closely related to the foregoing, we view the competitive process as occurring in real time and compute the implications of behavioral rules embedded in time. Behavior is represented as driven by information that actors might plausibly know.

As previously suggested, we do not find it plausible to assume that the individual actor is richly endowed with information that would support accurate inference about the situations of *other* actors. (Of course, we make no claim that our assumed rules are unique or even prominent in the class of rules with the requisite plausibility.)

Our perspective also has some commonalities with Sutton (1991), who shows that “endogenous sunk costs” (i.e. the need to irreversibly commit funds into features such as advertising, or R&D that can attract consumers) can account for patterns of industry structure, and for sustained profitability. We too consider the strategic ramifications of costs the firm extends, though our interests lie with the supply side and with a behaviourally plausible conceptualization of production, rather than issues of demand. Also, unlike Sutton, we do not assume that firms can have access to any appropriate technologies, neither do we focus on equilibrium conditions. Instead, we focus on the “dis-equilibrium-to-equilibrium” adjustment, since we believe that the majority of profits in the economy reflect these transient, temporary profit streams, the results of successive waves of industry creation and re-generation.

More generally, our paper shows that we need to complement the analysis of flow profitability at one point in time (and the analysis of the long-run equilibrium) with the analysis of the Net Present Value of profits, since what we should be concerned with is the total wealth that is created in a sector. We also show how shifting the focus to NPV and to the adjustment paths helps us revisit received wisdom – such as the perception that imitation will always be harmful for industry profits.

\(^{22}\) We do not mean to imply that this question has no plausible answer, only that Lippman and Rumelt do not offer one. Indeed, our own formulation of capability cost illustrates one path to an answer: it is possible to treat information as non-rivalrous without necessarily assuming that replication is costless. See the discussion of “non-standard examples of information economics” in Winter and Szulanski (2002).
Limitations

Our exploration of the dynamic consequences of efficiency differences has not, of course, produced a “rule for riches” – or even a promising set of hints. We have not accounted for the mechanisms that separated the best firms from the worse in our arrays. The ultimate sources of superior efficiency – whether in creativity, persistent effort, or luck – remain beyond the scope of our analysis. The very diversity of these potential sources suggests that the prospects for future insights through the application of the sorts of techniques used in this paper should be assessed with restrained optimism. Other approaches may hold greater promise.

The simplifications we have employed to clarify the economic logic tend to obscure other highly relevant aspects – the time-consuming nature of capability building, the consequences of secondary markets for partially customized assets (which may exist, even though thin). Most importantly, our effort to clarify the subtle issues involving pricing of resources that are “semi-permanently attached” to the firm led us to exclude from the analysis the more substantial and straightforward “attachment” of asset ownership. In so doing, we also excluded from the present value analysis the wealth changes attributable to changes in the market prices of owned assets – which in many cases form an important part of the picture. Whereas the positioning school has tended to discuss timely purchase of key resources in terms of the creation of entry barriers, we would argue that this view has, at a minimum, to be complemented by acknowledgment of the straightforward consequences of changing market valuations – changes that the combination of input scarcity and industry growth will very commonly produce. (Winter 1995, Jacobides and Winter, 2006).

Conclusions and Future Agenda

This paper has provided a bridge between some of the tenets of the RBV and micro-economic theory, as well as evolutionary theorizing. Doing so, it considered the factors that explain when differentially capable firms can co-exist in equilibrium; focused on the role and nature of capability development as it affects short-run and long-term profits. It also considered the nature of profits, as distinct from returns to resources, and looked at how the perspective changes by shifting the focus from flow profitability to the NPV of profits. By proposing a systemic model with endogenous markets for resource and for the final good, and heterogeneous goods, we advocate a dynamic mode of analysis, which can be extended and built upon.

Our approach provides a template, a theoretically motivated experimental design that allows us to look into a number of strategic issues that cannot be traced through a conventional model.
Specifically, we can explore the role of endogenous capability development (driven by learning curves or profit reinvestment), as it combines with capability distribution to affect profits, their NPV, and their distribution. Another important feature of this modeling platform is that it considers the *entire industrial system as it evolves*. This allows us to ask some important questions such as, how does increased elasticity of demand or a demand shock (akin to increased power of buyers) affect profits and the resource payments? When do pressures hit the bottom-line of firms in a sector, and when do they hit the upstream resource-holders, e.g., when should a reduction in the demand for oil hit the profitability of owning oil reserves, and when will it also dent the profits of refiners? Inversely, how do changes in the condition of the market for resources (e.g. a shortage in oil supplies) affect profits, and prices and quantities downstream? How do capability differences affect these results, in the short or the long run? In other words, how do the forces in an industry intertwine? And how do strategic dynamics along a value chain relate to firm differences and resources? Our preliminary findings suggest that both capability heterogeneity in a sector, and the extent of resource customization cost play an important role. We thus look forward to extending these results and hope that, building on this platform, we will be able to advance our understanding of profit evolution and strategic dynamics.
References


Argote, Linda, Sara L. Beckman, and Dennis Epple (1990), "The persistence and transfer of learning in industrial settings," Management Science, 36 (2), 140-54.


