Chapter 35

MARKET STRUCTURE: THEORY AND EVIDENCE

JOHN SUTTON
London School of Economics

Contents

Handbook of Industrial Organization, Volume 3
Edited by M. Armstrong and R. Porter

DOI: 10.1016/S1573-448X(06)03035-4

Abstract

This chapter reviews the literature which has developed around the 'bounds approach' to market structure over the past fifteen years. The focus of this literature lies in explaining cross-industry differences in concentration, and in the size distribution of firms. One of the main ideas involved is that a study of these cross-industry differences offers a powerful way of uncovering the operation of some key competitive mechanisms.

Keywords

Oligopoly, Industrial structure, Manufacturing, Advertising, R&D, Size distribution, Endogenous sunk costs, Bounds approach

JEL classification: L13, L16, L60, M37

1. Introduction

Why are some industries dominated worldwide by a handful of firms? Why is the size distribution of firms within most industries highly skewed? Questions of this kind have attracted continued interest among economists for over half a century. One reason for this continuing interest in 'market structure' is that this is one of the few areas in economics where we encounter strong and sharp empirical regularities arising over a wide cross-section of industries. That such regularities appear in spite of the fact that every industry has many idiosyncratic features suggests that they are molded by some highly robust competitive mechanisms – and if this is so, then these would seem to be mechanisms that merit careful study. If ideas from the IO field are to have relevance in other areas of economics, such as International Trade or Growth Theory, that relevance is likely to derive from mechanisms of this robust kind. Once we ask, "what effect will this or that policy have on the economy as a whole?", the only kind of mechanisms that are of interest are those that operate with some regularity across the general run of markets.

The recent literature identifies two mechanisms of this 'robust' kind. The first of these links the nature of price competition in an industry to the level of market concentration. It tells us, for example, how a change in the rules of competition policy will affect concentration: if we make anti-cartel rules tougher, for example, concentration will tend to be higher. (A rather paradoxical result from a traditional perspective, but one that is quite central to the class of 'free entry' models that form the basis of the modern literature.)

The second mechanism relates most obviously to those industries in which R&D or Advertising play a significant role, though its range of application extends to any industry in which it is possible for a firm, by incurring additional fixed and sunk costs (as opposed to variable costs), to raise consumers' willingness-to-pay for its products, or to cut its unit variable cost of production. This mechanism places a limit on the degree to which a fragmented (i.e. low concentration) structure can be maintained in the industry: if all firms are small, relative to the size of the market, then it will be profitable for one (or more) firm(s) to deviate by raising their fixed (and sunk) outlays, thus breaking the original 'fragmented' configuration.

In what sense can these mechanisms be said to be 'robust'? Why should we give them pride of place over the many mechanisms that have been explored in this area? These questions bring us to a central controversy.

1.1. The bounds approach

The first volumes of the *Handbook of Industrial Organization*, which appeared in 1989, summed up the research of the preceding decade in game-theoretic IO. In doing so, they provided the raw materials for a fundamental and far-reaching critique of this research program. In his review of those volumes in the *Journal of Political Economy*, Sam Peltzman pointed to what had already been noted as the fundamental weakness of the

project [Shaked and Sutton (1987), Fisher (1989), Pelzman (1991)]: the large majority of the results reported in the game-theoretic literature were highly sensitive to certain more or less arbitrary features of the models chosen by researchers.

Some researchers have chosen to interpret this problem as a shortcoming of game-theoretic methods per se, but this is to miss the point. What has been exposed here is a deeper difficulty: many outcomes that we see in economic data are driven by a number of factors, some of which are inherently difficult to measure, proxy or control for in empirical work. This is the real problem, and it arises whether we choose to model the markets in question using game-theoretic models or otherwise [Sutton (2001c)]. Some economic models hide this problem by ignoring the troublesome 'unobservables'; it is a feature of the current generation of game-theoretic models that they highlight rather than obscure this difficulty. They do this simply because they offer researchers an unusually rich menu of alternative model specifications within a simple common framework. If, for example, we aim to model entry processes, we are free to adopt a 'simultaneous entry' or 'sequential entry' representation; if we want to examine post-entry competition, we can represent it using a Bertrand (Nash equilibrium in prices) model, or a Cournot (Nash equilibrium in quantities) model, and so on. But when carrying out empirical work, and particularly when using data drawn from a cross-section of different industries, we have no way of measuring, proxying, or controlling for distinctions of this kind. When we push matters a little further, the difficulties multiply: were we to try to defend any particular specification in modeling the entry process, we would, in writing down the corresponding game-theoretic model, be forced to take a view (explicitly or implicitly) as to the way in which each firm's decision was or was not conditioned on the decisions of each rival firm. While we might occasionally have enough information about some particular industry to allow us to develop a convincing case for some model specification, it would be a hopeless task to try to carry this through for a dataset which encompassed a broad run of industries. What, then, can we hope to achieve in terms of finding theories that have empirical content? Is it the case that this class of models is empirically empty, in the sense that any pattern that we see in the data can be rationalized by appealing to some particular 'model specification'?

Two responses to this issue have emerged during the past decade. The first, which began to attract attention with the publication of the *Journal of Industrial Economics* Symposium of 1987, was initially labeled 'single industry studies', though the alternative term 'structural estimation' is currently more popular. Here, the idea is to focus on the modeling of a single market, about which a high degree of information is available, and to 'customize' the form of the model in order to get it to represent as closely as possible the market under investigation. A second line of attack, *which is complementary to* (rather than an alternative to) the 'single industry approach',[2] is offered by the bounds approach developed in Sutton (1991, 1998), following an idea introduced in Shaked and Sutton (1987). Here, the aim is to build the theory in such a way as to focus attention

[2] On the complementarity between these two approaches, see Sutton (1997a).

on those predictions which are robust across a range of model specifications which are deemed 'reasonable', in the sense that we cannot discriminate a priori in favor of one rather than another on empirical grounds (Sutton, 2000).

A radical feature of this approach is that it involves a departure from the standard notion of a 'fully specified model', which pins down a (unique) equilibrium outcome. Different members of the set of admissible models will generate different equilibrium outcomes, and the aim in this approach is to specify *bounds* on the set of observable outcomes: in the space of outcomes, the theory specifies a *region*, rather than a point. The question of interest here, is whether the specification of such bounds will suffice to generate informative and substantial restrictions that can be tested empirically; in what follows, it is shown that these results: (i) replicate certain empirically known relations that were familiar to authors in the pre-game theory literature; (ii) sharpen and re-specify such relations; and (iii) lead to new, more detailed empirical predictions on relationships that were not anticipated in the earlier literature.

1.2. Scope and content

The literature on market structure is extensive, and the present chapter does not offer a comprehensive overview. Rather, it focuses heavily on two leading strands in the literature, in which it has proved possible to bring together a robust theoretical analysis with sharp empirical tests. The first of these relates to the cross-industry studies pioneered by Bain (1956) which lie at the heart of the structure–conduct–performance tradition (Section 2). The second relates to the size distribution of firms, first studied by Gibrat (1931) (Section 3). In Section 4, we look at the area of market dynamics, where it has proved much more difficult to arrive at theoretical predictions of a robust kind, but where a substantial number of interesting empirical regularities pose a continuing challenge for researchers.

Two notable literatures that lie beyond the scope of this review are the Schumpeterian literature, and the organizational ecology literature. On the (close) relations between the bounds approach and the Schumpeterian literature, see Sutton (1998, pp. 29–31) and Marsili (2001). A good overview of current work in the organizational ecology literature will be found in Carroll and Hannan (2000). Critiques of the bounds approach will be found in Bresnahan (1992), Schmalensee (1992) and Scherer (2000).

2. The cross-industry literature

2.1. Background: the Bain tradition

The structure–conduct–performance paradigm, which began with Bain (1956), rested on two ideas. The first idea involved a one-way chain of causation that ran from structure (concentration) to conduct (the pricing behavior of firms) to performance (profitability).

High concentration, it was argued, facilitated collusion and led to high profits. To explain why these high profits were not eroded by entry, the second idea came into play: it was argued that high levels of concentration could be traced to the presence of certain 'barriers to entry'.

In Bain's 1956 book, these barriers were associated with the presence of scale economies in production, a factor that can be taken as an exogenous property of the available technology. Attempts to account for observed levels of concentration by reference to this factor alone, however, were clearly inadequate: many industries, such as the soft drinks industry, have low levels of scale economies in production, but have high levels of concentration. This prompted a widening of the list of candidate 'barriers' to include inter alia levels of advertising and R&D spending. The problem that arises here, is that these levels of spending are not exogenous to the firms, but are the outcomes of the firms' choices. It is appropriate, therefore, to model these levels as being determined jointly with the level of concentration as part of an equilibrium outcome; this is a central feature of the modern game-theoretic literature. To appeal to observed levels of advertising or R&D as an 'explanation' for high concentration levels is a mistake.

The central thrust of the structure–conduct–performance literature lay in relating the level of concentration to the level of profitability (profits/fixed assets, say) across different industries. Here, it is necessary to distinguish two claims.

The first relates to the way in which a fall in concentration, due for example to the entry of additional firms to the market, affects the level of prices and so of price–cost margins. Here, matters are uncontroversial; that a fall in concentration will lead to a fall in prices and price–cost margins is well supported both theoretically and empirically. [While theoretical counter-examples can be constructed, they are of a rather contrived kind; see Rosenthal (1980).] To test this idea it is appropriate to look at a number of markets for the same product, which differ in size (the number of consumers), so that larger markets support more sellers. It can then be checked whether prices and so price–cost margins are lower in those larger markets which support more sellers. The key body of evidence is that presented in the collection of papers edited by Weiss (1989). For a comprehensive list of relevant studies, see Schmalensee (1989, p. 987).

A second, quite different (and highly controversial) claim relates to the net profit of firms (gross profit minus the investment costs incurred in earlier stages), or their rates of return on fixed assets. In the 'free entry' models used in modern game-theoretic literature, entry will occur up to the point where the gross profits of the marginal entrant are just exhausted by its investment outlay. In the special setting where all firms are identical in their cost structure and in their product specifications, the net profit of each firm will be (approximately)[3] zero, whatever the level of concentration. This symmetric setup provides a useful point of reference, while suggesting a number of channels through which some relationship might appear between concentration and profitability. [For a discussion on this issue see Sutton (2001c); on the current dubious status of this

[3] I.e. up to an integer effect, which may in practice be substantial [Edwards and Starr (1987)].

concentration/profitability relationship, see Schmalensee's contribution to volume II of this Handbook Schmalensee (1989).]

A separate strand of this early literature focused in explaining concentration by reference to the 'barriers' just mentioned. Notwithstanding the objections noted above, it is of interest that regressions of this kind generated one rather robust statistical regularity, and one long-standing puzzle.

The statistical regularity is one that appears in cross-industry regressions between concentration, and a series of explanatory variables that include: (i) some measure of scale economies relative to market size (usually the cost of a single plant of 'minimum efficient scale' – the 'set-up cost' – divided by total industry sales revenue), and (ii) a measure of advertising intensity (usually the ratio of total industry advertising to industry sales revenue). Regressions of this kind suggest a clear positive relation in respect of setup costs/market size, and a rather weak positive relation in respect of the advertising–sales ratio [Sutton (1991, p. 124)]. One of the implications of the models considered below is that these relations should indeed be observed in such a (misspecified) regression [Sutton (1991, pp. 127–128)]. The puzzle that appears in regard to such regressions relates to the results obtained when the industry-wide R&D/sales ratio is included as an additional explanatory variable; typically, it turns out to be uncorrelated with concentration. A large literature developed during the 1970s and '80s in which concentration was regressed on the R&D/sales ratio (without including such additional factors as the ratio of setup costs to market size, or the advertising/sales ratio). This literature generated no generally agreed conclusions; in volume II of this Handbook, Cohen and Levin (1989, p. 1075) note that most papers on the question report a positive relation, though some find a negative relation, and others argue for a non-monotonic relation.[4] Results change substantially when industry-specific effects are controlled for, but there is no general agreement on what kind of control variables are appropriate though many authors favor including some index of "technological opportunity". Most tellingly, once such control variables are included, the partial correlation between R&D intensity and concentration is extremely weak. The authors cite the example of Scott's (1984) study, which found that "line of business concentration and its square explained only 1.5 percent of the variance in R&D intensity across 3388 business units, whereas two-digit industry effects explained 32 percent of this variance". This suggests that the cloud of observations on which such regressions are being run is so diffuse as to cast doubt on the usefulness of the exercise.

One of the central themes in what follows relates to a simple resolution of this issue, following Sutton (1998). It is argued that there are two problems with the idea of examining simple correlations between R&D and concentration: (a) it is vital to control for the fact that some markets, as conventionally defined in this literature, are single well-defined markets in the economic sense, while others are more complex, incorporating

[4] This claim should be distinguished from the claim for a U-shaped relation (across firms rather than industries) between R&D intensity and an index of the intensity of competition based on price–cost margins, posited by Aghion et al. (2005), as shown in their Figure 1.

various clusters of substitute goods ['competing groups' in Caves's (1986) terminology], and this must be controlled for, and (b) there are many further factors that impinge on the relationship between R&D and concentration, some of which are difficult to control for, so that the appropriate specification is a 'bounds' relationship rather than a conventional regression relationship. Once these two issues are addressed, a clear and straightforward picture emerges (Section 2.6).[5]

2.2. *Some preliminary examples*

The analysis developed below is based on multi-stage game models of a standard kind; before turning to formalities, we begin with a few elementary examples. The simplest setup involves a two-stage game of the following form. There are N_0 ($\geqslant 2$) firms. At stage 1, each firm chooses an action 'enter' or 'don't enter'. A firm choosing not to enter receives a payoff (profit) of zero. At stage 2, all those firms who have entered compete for consumers.[6] The firms offer a homogeneous product, which is produced by all firms at the same constant level of marginal cost $c \geqslant 0$. The payoff of a firm is given by the profit earned in stage 2, minus a sunk cost $\varepsilon > 0$ associated with the firm's entry at stage 1.

To complete the specification of the model, we model demand for the product by assuming that all consumers have a utility function of the (Cobb–Douglas) form,

$$U = x^{\delta} z^{1-\delta}$$

defined over two goods, the good x that is the focus of our analysis, and some outside good z. It follows from the form of the utility function that consumers spend a constant fraction δ of their incomes on good x, independently of the price of x. To avoid technical problems in the case where only one firm enters, assume that some 'imported' good is available at some (high) price p_0 that is a perfect substitute for x, so that consumers make no purchases of x if $p > p_0$. The price p_0 now serves as a monopoly price in the model.

Now denote total consumer expenditure on x as S, where S serves as a measure of the size of the market. We can now write the market demand schedule as

$$X = S/p,$$

where p denotes market price and $X \equiv \sum x_j$ is the total quantity sold by all firms.

We characterize equilibrium as a perfect Nash equilibrium of the two-stage game. Taking as given the number N of firms who have entered at stage 1, we solve for a (symmetric) Nash equilibrium in quantities at stage 2 (Cournot equilibrium). A routine

[5] For an alternative view that proposes the degree of appropriability of R&D returns as the relevant 'missing variable'; see Lee (2005).

[6] Thus a strategy takes two forms: either 'don't enter', or 'enter; and choose an action in the second stage as a function of the decisions taken by firms at the first stage (in effect, as a function of the number of entrants)'.

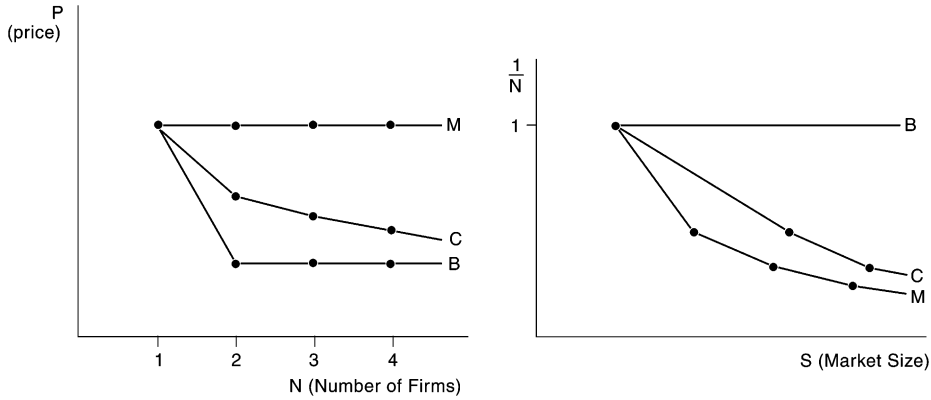

Figure 35.1. Equilibrium price as a function of the number of entrants, N, and equilibrium concentration $(1/N)$ as a function of market size, for three simple examples (B = Bertrand, C = Cournot, M = joint profit maximization).

calculation leads to the familiar result that, at equilibrium, price falls to marginal cost as N rises. (Figure 35.1, left panel, schedule C; the calculation is set out in Appendix A.) The equilibrium profit of firm i in the second stage subgame is given by S/N^2. Equating this to the entry fee $\varepsilon > 0$ incurred at stage 1, we obtain the equilibrium number of entrants as the largest integer satisfying the condition $S/N^2 \geqslant \varepsilon$. As S increases, the equilibrium number of firms rises, while the 1-firm concentration ratio $C_1 = 1/N$ falls monotonically to zero (Figure 35.1). It is also worth noting that output per firm rises with S, and that the number of firms rises less than proportionally with S; these observations are central to the literature on 'market size and firm numbers' (see below, footnote 17).

Now consider an alternative version of this model, in which we replace the 'Cournot' game by a Bertrand (Nash equilibrium in prices) model at stage 2. Here, once two or more firms are present, equilibrium involves at least two firms setting $p = c$, and all firms earn zero profit at equilibrium. Here, for any size of market S that suffices to support at least one entrant, the only (pure strategy) equilibrium for the game as a whole involves exactly one firm entering, and setting the monopoly price (Figure 35.1, schedule B).[7]

Finally consider a third variant of the model, in which we replace our representation of the second stage subgame by one in which all firms set the monopoly price.[8] Now, for any number N of firms entering at stage 1, we have that price equals p_0, and each firm receives a fraction $1/N$ of monopoly profit; the number of entrants N is the number which equates this to ε, as before (Figure 35.1, schedule M).

[7] Mixed strategy equilibria are discussed in Section 3.

[8] This can be formalized by replacing stage 2 by an infinite horizon stage game, and invoking the basic 'Folk theorem' result [see, for example, Tirole (1990)].

Figure 35.2. Alternative equilibria in a Hotelling-type model of horizontal production differentiation.

The results illustrated in Figure 35.1 serve to introduce an important result. We can interpret a move from the monopoly model to the Cournot model, and then to the Bertrand model, as an increase in the 'toughness of price competition', where this phrase refers to the functional relationship between market structure, here represented by the 1-firm concentration ratio $C_1 = 1/N$, and equilibrium price.[9] An increase in the toughness of price competition (represented by a downward shift in the function $p(N)$ in the first panel of Figure 35.2), implies that for any given level of market size, the equilibrium level of concentration $C_1 = 1/N$ is now higher (Figure 35.2, second panel). This result turns out to be quite robust, and it will emerge as one of the empirically testable predictions of the theory in what follows (Section 2.4).

All the cases considered so far have involved firms that produce a homogeneous product. We may extend the analysis by considering products that are ('horizontally') differentiated, either by geographic location, or by way of product characteristics that cause some consumers to prefer one variety, while others prefer a different variety, their prices being equal. When we do this, a new feature appears, since models of this kind tend to exhibit multiple equilibria. For any given market size, we will in general have a set of equilibrium outcomes; the case in which N firms each offer a single variety arises as one possible outcome, but there will usually be additional equilibria in which a smaller number of firms each offers some set of products.

The simplest way to see this point is by thinking in terms of the classic Hotelling model, in which products are differentiated by their locations along a line [Figure 35.2; see Sutton (1991, pp. 38–39)]. Imagine a 'single product firm' equilibrium in which firms occupy a set of discrete locations A, B, C, D, E, etc. We can construct, for example, a new equilibrium in which every second location is occupied by a single ('multiproduct') firm. There will now be an equilibrium in which prices are the same as before. This firm's profit function is additively separable, into functions which represent the separate contributions from each of its products, and the first-order condition for profit maximization coincides with the set of first-order conditions for the firms owning products A, C, E, etc. in the original setup. If the original 'single-product firm' configuration constituted an equilibrium of the two-stage game, so too will this 'high concentration' configuration in which our multi-product firm owns every second product.

[9] This phrase refers, therefore, to the 'form of price competition' which is taken as an exogenously given characteristic of the market. It will depend on such background features of the market as the cost of transport of goods, and on such institutional features as the presence or absence of anti-trust laws. [On the determinants of the 'toughness of price competition' see Sutton (1991, ch. 6).] In particular, it does not refer to the equilibrium level of prices, or margins, which, within the two-stage game, is an endogenous outcome. For a novel application of this concept, see Raith (2003).

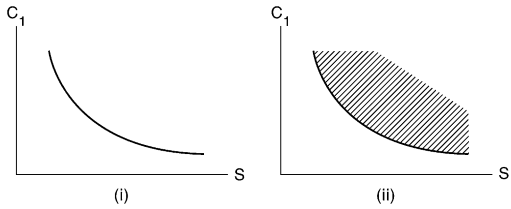

Figure 35.3. The relation between market size S and concentration C_1, in (i) a homogeneous goods model, and (ii) a model of horizontal product differentiation. In the latter case, the only restriction that we can place on the concentration–size relationship is a 'bounds' relationship.

Now in this setting, the functional relationship between concentration and market size, illustrated in Figure 35.1, must be replaced by a lower bound relationship. The lower bound is traced out by a sequence of 'single product firm' equilibria; but there are additional equilibria lying above the bound (Figure 35.3).[10]

This Hotelling example illustrates a general problem in relation to modeling the final-stage subgame. Here we face two questions: (a) what is the best way to represent price competition (à la Cournot, à la Bertrand, or otherwise)? (b) if products are differentiated, are they best modeled by means of a Hotelling model, or a non-locational model that treats all varieties in a symmetric fashion,[11] or otherwise? This is the first problem that motivates the introduction of a Bounds approach.

A second problem that strengthens the case for such an approach relates to the form of the entry process. In the above examples, we have confined attention to the case of 'simultaneous entry'. If we replace this by a 'sequential entry' model, the (set of) equilibrium outcomes will in general be different. For example, in the setting of (horizontal) product differentiation, there may (or may not) be a tendency in favor of 'more concentrated' equilibria, in which the first mover 'pre-empts' rivals by introducing several varieties [Schmalensee (1978); for an extended example, see Sutton (1998, ch. 2 and Appendix 2.1.2)].

The main burden of the analysis developed in the next section is concerned with framing a theory that gets around these two sources of difficulty. Before proceeding to a general treatment, however, there is one final example that is worth introducing.

[10] One comment sometimes made about this setup is that we might hope, by introducing a 'richer' model specification, or by appealing to characteristics of individual firms, to arrive at a model that led to some particular (i.e. unique) equilibrium. However, to justify any specific model of this kind – since many such models might be devised – we would need to appeal to information about the market and the firms that we would be unlikely to have access to, at least in a cross-industry study. Thus we are led back once again to the problem of 'unobservables'.

[11] Examples of such 'symmetric' models include the models of Dixit and Stiglitz (1977), and the 'linear demand model' discussed in Shubik and Levitan (1980), Deneckere and Davidson (1985), Shaked and Sutton (1987) and, in a Cournot version, in Sutton (1998).

2.2.1. A quality choice model

A key feature that has not arisen in the examples considered so far relates to the possibility that firms might choose to incur (additional) fixed and sunk costs at stage 1 with a view to improving their competitive positions in the final-stage subgame. This kind of expenditure would include, for example, outlays on R&D designed to enhance the quality, or technical characteristics, of firms' products; it would include advertising outlays that improve the 'brand-image' of the product; and it would include cost-reducing 'product innovation' in which R&D efforts are directed towards the discovery of improved production methods.

We can illustrate this kind of situation by extending the simple Cournot model introduced above, as follows. Suppose all consumers have the same utility function of the form

$$U = (ux)^{\delta} z^{1-\delta}$$

defined over two goods, where u represents an index of perceived quality for good x. Increases in u enhance the marginal utility derived from this good. We will refer to this first good x as the "quality" good, in order to distinguish it from the "outside" good, z.

Rival firms are assumed to offer single goods of various qualities. Let u_i and p_i denote the quality and price respectively of firm i's offering. Then, the consumer's decision problem can be represented as follows: given the set of qualities and prices offered, it is easy to show that the consumer chooses a product that maximizes the quality–price ratio u_i/p_i; and the consumer spends fraction δ of his or her income on this chosen quality good, and fraction $(1-\delta)$ on the outside good. Total expenditure on the quality goods is therefore independent of their levels of prices and perceived qualities and equals a fraction δ of total consumer income. Denote this level of total expenditure on quality goods by S.

The first step in the analysis involves looking at the final stage of the game. Here, the qualities are taken as given (having been chosen by firms at the preceding stage). Equilibrium in the final stage of the game is characterized as a Nash equilibrium in quantities (Cournot equilibrium). A feature of this equilibrium is that, since each consumer chooses the good that maximizes u_i/p_i, the prices of all those products enjoying positive sales at equilibrium must be proportional to their perceived qualities, that is, $u_i/p_i = u_j/p_j$ for all i, j.

The calculations are set out in Appendix B; here, we summarize the relevant properties of the solution. In the final stage subgame, some number of products survive with positive sales revenue; it may be the case that products with qualities below a certain level have an output level of zero, and so profits of zero, at equilibrium. Denoting by N the number of firms that enjoy positive sales ('survive') at equilibrium, the final stage profit of firm i is given by

$$\left\{1 - \frac{N-1}{u_i}\frac{1}{\sum(1/u_j)}\right\}^2 \cdot S. \tag{2.1}$$

Associated with this equilibrium is a threshold level of quality $\underline{u}$; all 'surviving' products have $u_i > \underline{u}$ and all products with $u_j \leqslant \underline{u}$ have an output of zero at equilibrium. The sum in the above expression is taken over all 'surviving' products, and N represents the number of such products. The threshold $\underline{u}$ is defined by adding a hypothetical $(N+1)$th product to the N surviving products, and equating the profit of good $N+1$ to zero, viz. $\underline{u} = u_{N+1}$ is implicitly defined by[12]

$$\frac{1}{\underline{u}} = \frac{1}{u_{N+1}} = \frac{1}{N}\sum_{1}^{N+1}\frac{1}{u_j} \quad \text{or equivalently} \quad \frac{1}{\underline{u}} = \frac{1}{N-1}\sum_{1}^{N}\frac{1}{u_j}.$$

Now consider a 3-stage game in which each of the N_0 potential entrants decides, at stage 1, to enter or not enter, at cost $F_0 > 0$. At stage 2, the N firms that have entered choose a quality level, and in so doing incur additional fixed and sunk costs. Denote by $F(u)$ the total fixed and sunk cost incurred by an entrant that offers quality u, where u lies in the range $[1, \infty)$ and

$$F(u) = F_0 u^{\beta}, \quad u \geqslant 1.$$

Thus the minimum outlay incurred by an entrant equals F_0 (>0).

Given the qualities chosen at stage 2, all firms now compete à la Cournot in stage 3, their gross profit being defined as above. A firm's payoff equals its net profit (gross profit minus the fixed and sunk outlays incurred).

A full analysis of this model will be found in Sutton (1991, ch. 3). Here, we remark on the key feature of the relationship between market size and concentration. At equilibrium, N firms enter and produce a common quality level u. For small S, the level chosen is the minimum level $u = 1$, and the size–structure relation mimics that of the basic Cournot model. But once a certain critical value of S is reached, the returns to incurring fixed outlays on quality improvement rise, and the level of u rises thereafter with S. The number of firms N, on the other hand, remains constant: the 'convergence' effect, whereby the (lower bound to the level of) concentration falls to zero as $S \to \infty$, breaks down. Increases in market size are no longer associated with a rise in the number of firms; rather, the expenditures incurred by each firm rise, while the number of firms remains unchanged (Figure 35.4).[13] It is this breakdown of the convergence property that will form the central theme of Section 3.

[12] The number of products that survive can be computed recursively by labeling the products in descending order of quality, so that $u_1 \geqslant u_2 \geqslant u_3 \geqslant \cdots$ and considering successive candidate sets of surviving products of the form $(u_1, u_2, \ldots, u_k)$. For each such set there is a corresponding value of $\underline{u}$; the set of surviving products is the smallest set such that the first excluded product has a quality level $u_{k+1} < \underline{u}$.

[13] Chapter 3 of Sutton (1991) analyses a wider range of cost functions of the form $a + bu^{\beta}$, which illustrate a number of different forms that the concentration–size relationship can take. Here, I have confined attention to the simplest case, in order to provide a preliminary illustration of the 'non-convergence' result. The only new feature arising when we move to this wider set of cost functions is that the right-hand segment of the lower bound need not be flat; it may rise or fall towards its asymptotic level (which depends on β alone and not on F_0).

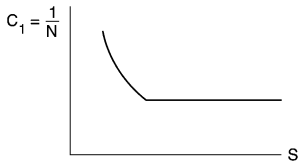

Figure 35.4. Market size and the (one-firm) concentration ratio in the 'quality competition' example.

2.2.2. A limiting case

An interesting 'limiting case' of this example is obtained by letting $\beta \to \infty$. Here, the effectiveness of fixed outlays in raising product quality is arbitrarily low. In the limit, such spending has no effect, and the optimal choice for all firms is to set $u = 1$, its threshold value. Here, the equilibrium collapses to that of the simple Cournot example considered earlier, in which firms paid an exogenously given entry fee, here equal to F_0, to enter. It is natural from the point of view of the theory to interpret the 'exogenous sunk cost' model considered above as a limiting case arising within the general ('endogenous sunk cost') model.

2.2.3. Extensions

The idea embodied in the 'quality competition' example is more general than might appear to be the case at first sight. The key idea lies in the notion that a firm can increase its gross (i.e. final stage) profit by incurring additional fixed and sunk costs at stage 1. This idea carries over immediately to a setting in which firms offer a homogeneous product, and where each firm can reduce its unit cost of production at stage 2 by incurring additional fixed outlays at stage 1 ['process innovation'; see Dasgupta and Stiglitz (1980), and Sutton (1998, ch. 14)].

More generally, we may combine both these channels in a single model, by characterizing each firm by a pair of numbers (c_i, u_i) denoting its unit cost of production, and its 'perceived quality', which we can label as the firm's 'capability' in some particular market. We can then model firms as 'competing in capabilities' [Sutton (2001b)].

Finally, these ideas can be extended in a straightforward way to the analysis of 'learning by doing' and 'network externalities', as discussed in Section 4 below.

2.3. A theoretical framework

In this section, we move to a general treatment. We specify a suitable class of multi-stage games, and consider a setting in which the fixed and sunk investments that firms make are associated with their entering of products into some abstract 'space of products'. This setup is general enough to encompass many models used in the literature. For example, in a Hotelling model of product differentiation, the (set of) action(s) taken by a firm would be described by a set of points in the interval [0, 1], describing the

location of its products in (geographic) space. In the 'quality choice' model considered above, the action of firm i would be to choose a quality level $u_i \geqslant 1$. In the model of 'competing in capabilities', the firm's action would be to choose a pair of numbers (u_i, c_i) and so on (Sutton, 2001a).

2.3.1. A class of stage-games

We are concerned with a class of games that share the following structure: There are N_0 players (firms). Firms take actions at certain specified stages. An action involves occupying some subset, possibly empty, of "locations" in some abstract "space of locations", which we label A. At the end of the game, each firm will occupy some set of locations.

The notation is as follows: a location is an element of the set of locations A. The set of locations occupied by firm i at the end of the game is denoted $\mathbf{a}_i$, where $\mathbf{a}_i$ is a subset of A viz. $\mathbf{a}_i \subset A$. If firm i has not entered at any location then $\mathbf{a}_i$ is the empty set, i.e. $\mathbf{a}_i = \emptyset$. Associated with any set of locations is a fixed and sunk cost incurred in entering at these locations. This cost is strictly positive and bounded away from zero, namely, for any $\mathbf{a}_i \neq \emptyset$, $F(\mathbf{a}_i) \geqslant F_0 > 0$. The outcome of the game is described by an N_0-tuple of all locations occupied by all firms at the end of the game. Some of the entries in this N_0-tuple may be null, corresponding to firms who have not entered the market. In what follows, we are concerned with those outcomes in which at least one firm has entered the market and are interested in looking at the locations occupied by the firms who have entered (the "active" firms). With that in mind, we label the number of active firms as N ($\geqslant 1$), and we construct an N-tuple by deleting all the null entries, and re-labeling the remaining firms from 1 to N. The N-tuple constructed in this way is written as $(\mathbf{a}_i) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ and is referred to as a "configuration".

The payoff (profit) of firm i, if it occupies locations $\mathbf{a}_i$, is written

$$\Pi\big(\mathbf{a}_i \mid (\mathbf{a}_{-i})\big) - \mathbf{F}(\mathbf{a}_i),$$

where $(\mathbf{a}_{-i})$ denotes $(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_N)$. The function, $\Pi(\mathbf{a}_i \mid (\mathbf{a}_{-i}))$ which is non-negative everywhere, is obtained by calculating firms' equilibrium profits in some final stage subgame (which we refer to as the "price competition subgame"), in which the $\mathbf{a}_i$ enter as parameters in the firms' payoff functions. It is defined over the set of all configurations. The second argument of the profit function is an $(N-1)$-tuple $(\mathbf{a}_{-i})$, which specifies the sets of locations occupied by each of the firm's rivals. Thus, for example, if we want to specify the profit that would be earned by a new entrant occupying a set of locations $\mathbf{a}_{N+1}$, given some configuration $\mathbf{a}$ that describes the locations of firms already active in the market, we will write this as $\Pi(\mathbf{a}_{N+1} \mid \mathbf{a})$. If, for example, only one firm has entered, then $(\mathbf{a}_{-i})$ is empty, and the profit of the sole entrant is written as $\Pi(\mathbf{a}_1 \mid \emptyset)$, where $\mathbf{a}_1$ is the set of locations that it occupies. A firm taking no action at any stage incurs zero cost and receives payoff zero. In writing the profit function $\Pi(\cdot)$ and the fixed cost function $F(\cdot)$ without subscripts, we have assumed that all firms face the same profit and cost conditions, i.e. a firm's payoff depends only on its actions, and those of its rivals; there are no 'firm-specific' effects. (Since our focus is on looking at

a lower bound to concentration, it is natural to treat firms as symmetric; asymmetries between firms (i.e. firm-specific effects) will tend to lead to levels of concentration that are above the bound we specify here.)

We are interested in examining the set of configurations that satisfy certain conditions. These conditions will be defined in a way that does not depend upon the order of the entries in **a**. Two configurations that differ only in the order of their elements are equivalent, in the sense that each one satisfies the conditions if and only if the other does.

2.3.2. Assumptions

We introduce two assumptions on the games to be considered. The first assumption relates to the payoff function Π of the final-stage subgame, on which it imposes two restrictions. Restriction (i) excludes 'non-viable' markets in which no product can cover its entry cost. Restriction (ii) ensures that the number of potential entrants N_0 is large. (The role of this assumption is to ensure that, at equilibrium, we will have at least one inactive player, so that $N < N_0$.) Denote the configuration in which no firm enters as $\emptyset$.

ASSUMPTION 1. (i) There is some set of locations $\mathbf{a}_0$ such that

$$\Pi(\mathbf{a}_0 \mid \emptyset) > F(\mathbf{a}_0).$$

(ii) The sum of final stage payoffs received by all agents is bounded above by $(N_0 - 1)F_0$, where N_0 denotes the number of players and F_0 is the minimum setup cost (entry fee).

The second assumption relates to the rules specifying the stages at which firms may enter and/or make investments:

ASSUMPTION 2 (extensive form). We associate with each firm i an integer t_i (its date of 'arrival'). Firm i is free to enter any subset of the set of products A at any stage t, such that $t_i \leqslant t \leqslant T$.

This assumption excludes a rather paradoxical feature that may arise in some basic 'sequential entry' models, where a firm would prefer, if allowed, to switch its place in the entry sequence for a later position [Eaton and Ware (1987)]. In practice, firms are always free to delay their entry; this assumption avoids this anomalous case by requiring that a firm arriving in the market at stage t is free to make investments at

stage t and/or at any subsequent stage, up to some final stage T (we exclude infinite horizon games).[14,15]

2.3.3. *Equilibrium configurations*

The aim of the present exercise is to generate results that do not depend on (a) the details of the way we design the final-stage subgame, or (b) the form of the entry process. To handle (a), we work directly in terms of the 'solved-out profit function' of the final-stage subgame, introduced as our profit function $\Pi(\cdot)$ above. To deal with (b), the entry process, we introduce an equilibrium concept that is defined, not in the space of strategies (which can only be specified in the context of some particular entry process), but in the space of outcomes, or – more precisely – configurations. The key idea here is this: the set of 'equilibrium configurations' defined below includes all outcomes that can be supported as a (pure strategy, perfect) Nash equilibrium in any game of the class defined by Assumptions 1 and 2 above. In what follows, we develop results which show that certain ('fragmented') market structures cannot be supported as equilibrium configurations – and so they cannot be supported as (pure strategy, perfect) Nash equilibria, irrespective of the details of the entry process.

Now there are two obvious properties that must be satisfied by any pure strategy, perfect Nash equilibrium within this class of models. In what follows, we define the set of all outcomes satisfying these two properties, as follows:

DEFINITION. The N-tuple $\mathbf{a}$ is an *equilibrium configuration* if:

(i) Viability[16]: For all firms i,

$$\Pi\left(\mathbf{a}_i \mid (\mathbf{a}_{-i})\right) - F(\mathbf{a}_i) \geqslant 0;$$

(ii) Stability: There is no set of actions $\mathbf{a}_{N+1}$ such that entry is profitable, viz. for all sets of actions $\mathbf{a}_{N+1}$,

$$\Pi(\mathbf{a}_{\mathrm{N}+1} \mid \mathbf{a}) - F(\mathbf{a}_{N+1}) \leqslant 0.$$

[14] The intuition underlying this assumption is worth noting: it says that *any* firm can enter any set of products at stage T of the game, taking as given all products entered by itself and its rivals at earlier stages. It is important to note that the actions involved here are ones that involve firms' incurring sunk costs (irreversible investments). This should be distinguished from the notion used in the 'contestability' literature, where an entrant is allowed to take incumbents' current *prices* as given [Baumol, Panzar and Willig (1982)]. This is not assumed here; price competition is modeled by reference to some weak restrictions on the solved-out profit function for the final stage subgame.

[15] The question of whether the results obtained in this multistage game setting will hold good in a dynamic game setting, in which there is no 'final' stage after which no further investments occur, is considered is Section 6.

[16] It is worth noting that the viability condition has been stated in a form appropriate to the 'complete information' context in which we are working here, where exit is not considered. In models where exit is an available strategy, condition (i) must be re-stated as a requirement that profit net of the *avoidable* cost which can be saved by exiting should be non-negative.

PROPOSITION 1 (inclusion). *Any outcome that can be supported as a (perfect) Nash equilibrium in pure strategies is an equilibrium configuration.*

To see why Proposition 1 holds, notice that 'viability' is ensured since all firms have available the action 'do not enter'. Assumption 1(ii) ensures that there is at least one firm that chooses this action at equilibrium; while if the stability condition does not hold, then a profitable deviation is available to that firm: given the actions prescribed for its rivals by their equilibrium strategies, it can profitably deviate by taking action $\mathbf{a}_{N+1}$ at stage T.

2.4. The price competition mechanism

We can now formalize the discussion of the 'price competition' mechanism introduced in the examples of Section 2.2 above, following Selten (1983) and Sutton (1991, ch. 2).[17] We confine attention, for ease of exposition, to the class of 'symmetric' produce differentiation models.[18] In these models, each firm chooses some number n of distinct product varieties to offer, and incurs a setup cost $\varepsilon > 0$ for each one. The profit of firm i in an equilibrium of the final stage subgame can be written as

$$S\pi\big(n_i \mid (n_{-i})\big).$$

Now consider a family of such models, across which the form of price competition in the final stage subgame differs. We consider a one-parameter family of models that can be ranked in the following sense: we define a family of profit functions parameterized by θ, denoted by

$$\pi\big(n_i \mid (n_{-i}); \theta\big).$$

An increase in θ shifts the profit function downwards, in the sense that, for any given configuration we have that if $\theta_1 > \theta_2$, then

$$\pi\big(n_i \mid (n_{-i}); \theta_1\big) < \pi\big(n_i \mid (n_{-i}); \theta_2\big).$$

[17] These ideas are also developed in the literature on 'market size and firm numbers' pioneered by Bresnahan and Reiss (1990a, 1990b) and Berry (1992), which is very closely related to the bounds approach; for a discussion see Sutton (1997a) and the contribution of Berry and Reiss to this volume. Specifically, this literature holds constant the nature of the price competition regime, and focuses on how an increase in market size affects outcomes. It is a generic property of single-product firm models with exogenous sunk costs, that an increase in market size leads to a fall in price–cost margins, and so, in a symmetric equilibrium, to a rise in the output of each firm, as the product of output and the price–cost margin must suffice to allow recovery of the sunk cost (as in the Cournot example of Section 2.1). A recent paper by Campbell and Hopenhayn (2005) examines this effect empirically. When we turn to a multi-product firm setting, the issues of interest relate to the range of market sizes over which alternative configurations can be supported as equilibria [for example, Shaked and Sutton (1990)]. Recent empirical studies by Mazzeo (2002) on motels and Manuszak (2002) on the evolution of the U.S. brewing industry explore these issues.

[18] Such models include for example the linear demand model [for a Bertrand version see Shubik and Levitan (1980), Shaked and Sutton (1982); for a Cournot version see Sutton (1998)], and the model of Dixit and Stiglitz (1977).

The parameter θ denotes the 'toughness of price competition' in the sense that an increase in θ reduces the level of final stage profit earned by each firm, for any given form of market structure (i.e. configuration).

We now proceed as follows: for each value of S, we define the set of configurations satisfying the viability condition, viz.

$$S\pi\left(n_i \mid (n_{-i}); \theta\right) \geqslant \varepsilon \quad \text{for all } i. \tag{2.2}$$

For each configuration, we define an index of concentration. For concreteness, we choose the 1-firm sales concentration ratio C_1, defined as the share of industry sales revenue accounted for by the industry's largest firm. We now select, from the set of configurations satisfying (2.2), the configuration with the lowest (or equal lowest) value of C_1, and we define this level of concentration as $\underline{C}_1(S;\theta)$. This construction defines the schedule $\underline{C}_1(S;\theta)$, which forms a lower bound to concentration as a function of market size. Assuming that π is increasing in C_1, then it follows immediately from Equation (2.2) that an increase in θ shifts this schedule upwards.

Say we begin, then, with an equilibrium configuration in some market. Holding the size of the market constant, we introduce a change in the external circumstances of the market which implies a rise in θ; for example, this might be a change in the rules of competition policy (a law banning cartels, say), or it might be an improvement in the transport system that causes firms in hitherto separated local markets to come into direct competition with each other (as with the building of national railway systems in the nineteenth century, for example).

If the associated shift in θ is large enough, then the current configuration will no longer be an equilibrium, and some shift in structure must occur in the long run.

At this point, a caveat is in order: the theory is static, and we cannot specify the dynamic adjustment path that will be followed once equilibrium is disturbed.

All that can be said is that restoration of the stability and viability conditions requires a rise in concentration.[19] We may distinguish two candidate mechanisms that may bring this about: the exit of some firm(s), and/or the consolidation of others via mergers and acquisitions. This argument relies upon the link between concentration and price (and so gross profit per firm), whose theoretical and empirical status was noted in Section 2.1 above.

2.4.1. Empirical evidence

The most systematic test of this prediction is that of Symeonidis (2000, 2001), who takes advantage of an unusual 'natural experiment' involving a change in competition

[19] The speed of adjustment by firms will be affected inter alia by the extent to which the setup cost ε is sunk, as opposed to fixed. If ε is a sunk cost, then a violation of the viability constraint will not require any adjustment in the short run; it is only in the long run, as the capital equipment needs to be replaced, that (some) firms will, in the absence of any other structural changes, no longer find it profitable to maintain their position, and will exit. If ε is partly fixed, rather than sunk, then exit is likely to occur sooner.

law in the UK in the 1960s. As laws against the operation of cartels were strengthened, a rise in concentration occurred across the general run of manufacturing industries. Symeonidis traces the operation of these changes in detail, and finds a process at work that is consistent with the operation of the response mechanisms postulated above.

Sutton (1991) reports some 'natural experiments' affecting particular industries in the wake of the spread of the railways in the late nineteenth century. The salt industry, both in the U.S. and Europe, went through a process of consolidation in the wake of these changes. First prices fell, rendering many concerns unviable. Attempts to restore profitability via price coordination failed, due to 'free riding' by some firms. Finally, a process of exit, accompanied by mergers and acquisitions, led to the emergence of a concentrated industry [Sutton (1991, ch. 6)].

The history of the sugar industry offers some interesting illustrations of the way in which differences in the competition policy regime affected outcomes. In the U.S., it follows a similar pattern to that of the salt industry over the same period. In Continental European countries, on the other hand, a permissive competition policy regime allowed firms to coordinate their prices, thus permitting the continuance of a relatively fragmented industry into the twentieth century. The Japanese market provides an unusually informative natural experiment, in that it went through three successive regimes in respect of competition policy. A tight cartel operated in the period prior to the First World War, and concentration was low. In the inter-war years, the cartel broke down and concentration rose. In the years following the Second World War, however, the authorities permitted the industry to operate under a permissive 'quota' regime; and this relaxation in the toughness of price competition encouraged new entry, and a decline in concentration [Sutton (1991, ch. 6)].

2.5. The escalation mechanism

We now turn to a general statement of the 'non-convergence' result introduced in the 'quality choice' example of Section 2.1. The analysis is developed in two steps. In this section, we consider a 'classical' setting in which each firm offers a (single) product within the same market, and all these products are substitutes. In Section 2.6 we will turn to a more complex setting in which the market comprises several distinct product groups, or 'submarkets'. Here, then, each firm's action takes one of two forms, 'do not enter' or 'enter with quality u_i', where u_i is chosen from the interval $[1, \infty)$.

The outcome of firms' actions is described by a configuration

$$\mathbf{u} = (u_1, \ldots, u_i, \ldots, u_N).$$

We associate with every configuration $\mathbf{u}$ a number representing the highest level of quality attained by any firm, viz.

$$\hat{u}(\mathbf{u}) = \max_i u_i.$$

We summarize the properties of the final-stage subgame in a pair of functions that describe the profit of each firm and the sales revenue of the industry as a whole. Firm i's

final stage profit is written as

$$\Pi\big(u_i \mid (\mathbf{u}_{-i})\big) \equiv S\pi\big(u_i \mid (\mathbf{u}_{-i})\big) \geqslant 0,$$

where $\mathbf{u}_{-i}$ denotes the $N-1$ tuple of rivals' qualities, and S denotes the number of consumers in the market.[20] Total industry sales revenue is denoted by

$$Y(\mathbf{u}) \equiv Sy(\mathbf{u}).$$

It is assumed that any firm entering the market incurs a minimum setup cost of F_0 and that increases in the quality index above unity involve additional spending on fixed outlays such as R&D and advertising. We choose to label this index so that the fixed outlay of firm i is related to the quality level u_i according to

$$F(u_i) = F_0 u_i^{\beta}, \quad \text{on } u_i \in [1, \infty), \text{ for some } \beta \geqslant 1.$$

We identify the level of spending on R&D and advertising as

$$R(u_i) = F(u_i) - F_0.$$

The economics of the model depends only on the composite mapping from firms' fixed outlays to firms' profits, rather than on the separate mappings of fixed outlays to qualities and from qualities to profits. At this point, the labeling of u is arbitrary up to an increasing transformation. There is no loss of generality, therefore, in choosing this functional form for $R(u_i)$.[21] (The form used here has been chosen for ease of interpretation, in that we can think of β as the elasticity of quality with respect to fixed outlays.[22])

To avoid trivial cases, we assume throughout that the market is always large enough to ensure that the level of sales exceeds some minimal level for any configuration, and that the market can support at least one entrant. With this in mind, we restrict S to the domain $[1, \infty)$, and we assume, following Assumption 1 above.

ASSUMPTION 3. The level of industry sales associated with any non-empty configuration is bounded away from zero; that is, there is some $\eta > 0$ such that for every configuration $\mathbf{u} \neq \emptyset$, we have $y(\mathbf{u}) \geqslant \eta > 0$ for all $\mathbf{u} \neq \emptyset$.

[20] The motivation for writing the profit function (and the industry sales revenue function) in this form (i.e. multiplicative in S), derives from an idea which is standard throughout the market structure literature: firms have flat marginal cost schedules, and increases in the size of the market involve an increase in the population of consumers, the distribution of consumer tastes being unaltered. Under these assumptions, a rise in the size of the population of consumers S shifts the demand schedule outwards multiplicatively and equilibrium prices are independent of S.

[21] There is, however, a (mild) restriction in writing $F(u_i)$ as $F_0 u_i^{\beta}$ rather than $F_0 + b u_i^{\beta}$, as noted in footnote 14 above. See Sutton (1991, ch. 3) for details.

[22] Rather than represent F as a single function, it is convenient to use a family of functions parameterized by β, since we can then hold the profit function fixed while varying β to capture changes in the effectiveness of R&D and advertising in raising final stage profits.

This assumption, together with Assumption 1(i), implies that the level of industry sales revenue $Sy(\mathbf{u}) \geqslant S\eta$ in any equilibrium configuration increases to infinity as $S \to \infty$.

2.5.1. A non-convergence theorem

In what follows, we are concerned with examining whether some kinds of configuration $\mathbf{u}$ are unstable against entry by a 'high-spending' entrant. With this in mind, we investigate the profit of a new firm that enters with a quality level k times greater than the maximum value $\hat{u}$ offered by any existing firm. More specifically, we ask: What is the minimum ratio of this high-spending entrant's profit to current industry sales that will be attained *independently* of the current configuration $\mathbf{u}$ and the size of the market?

For each k, we define an associated number $a(k)$ as follows.

DEFINITION. $a(k) = \inf_{\mathbf{u}} (\pi(k\hat{u} \mid \mathbf{u}))/(y(\mathbf{u}))$.

It follows from this definition that, given any configuration $\mathbf{u}$ with maximal quality $\hat{u}$, the final-stage profit of an entrant with capability $k\hat{u}$, denoted $S\pi(k\hat{u} \mid \mathbf{u})$, is at least equal to $a(k)Sy(\mathbf{u}) = a(k)Y(\mathbf{u})$, where $a(k)$ is independent of $\mathbf{u}$ and S.[23]

The intuition is as follows: k measures the size of the quality jump introduced by the new 'high spending' entrant. We aim to examine whether such an entrant will earn enough profit to cover its fixed outlays, and so we want to know what price it will set, and what market share it will earn. This information is summarized by the number $a(k)$, which relates the gross (final-stage) profit of the entrant, $S\pi(k\hat{u} \mid \mathbf{u})$, to pre-entry industry sales revenue, $Sy(\mathbf{u}) = Y(\mathbf{u})$. Since we wish to develop results that are independent of the existing configuration, we define $a(k)$ as an infimum over $\mathbf{u}$.

We are now in a position to state:

THEOREM 1 (non-convergence). *Given any pair $(k, a(k))$, a necessary condition for any configuration to be an equilibrium configuration is that a firm offering the highest level of quality has a share of industry sales revenue exceeding $a(k)/k^{\beta}$.*

PROOF. Consider any equilibrium configuration $\mathbf{u}$ in which the highest quality offered is $\hat{u}$. Choose any firm offering quality $\hat{u}$ and denote the sales revenue earned by that firm by $S\hat{y}$, whence its share of industry sales revenue is $S\hat{y}/SY(\mathbf{u}) = \hat{y}/Y(\mathbf{u})$.

Consider the net profit of a new entrant who offers quality $k\hat{u}$. The definition of $a(k)$ implies that the entrant's net profit is at least

$$aSy(\mathbf{u}) - F(k\hat{u}) = aSy(\mathbf{u}) - k^{\beta}F(\hat{u}),$$

where we have written $a(k)$ as a, in order to ease notation.

[23] It is worth noting that the above definition implies that $a(1) = 0$. To see this, notice that we can choose a configuration in which all n firms, and our new entrant, offer quality 1, so that all have the same profit. The ratio between the profit of any firm, and total industry sales revenue, can now be made arbitrarily small by letting $n \to \infty$. Since $a(k) \geqslant 0$ is defined as the infimum over all configurations u, it follows that $a(1) = 0$.

The stability condition implies that this entrants' net profit is non-positive, whence

$$F(\hat{u}) \geqslant \frac{a}{k^{\beta}} Sy(\mathbf{u}).$$

But the viability condition requires that each firm's final-stage profit must cover its fixed outlays. Hence the sales revenue of the firm that offers quality $\hat{u}$ in the proposed equilibrium configuration cannot be less than its fixed outlays:

$$S\hat{y} \geqslant F(\hat{u}) \geqslant \frac{a}{k^{\beta}} Sy(\mathbf{u})$$

whence its market share

$$\frac{S\hat{y}}{Sy(\mathbf{u})} \geqslant \frac{a}{k^{\beta}}.$$

This completes the proof. □

The intuition underlying this result is as follows: if the industry consists of a large number of small firms, then the viability condition implies that each firm's spending on R&D is small, relative to the industry's sales revenue. In this setting, the returns to a high-spending entrant may be large, so that the stability condition is violated. Hence a configuration in which concentration is "too low" cannot be an equilibrium configuration. This result motivates the introduction of a parameter, which we call alpha, as the highest value of the ratio a/k^{β} that can be attained by choosing any value $k \geqslant 1$, as follows.[24,25]

DEFINITION. $\alpha = \sup_k (a(k))/k^{\beta}$.

We can now reformulate the preceding theorem as follows: since the one-firm sales concentration ratio C_1 is not less than the share of industry sales revenue enjoyed by the firm offering quality $\hat{u}$, it follows from the preceding theorem that, in any equilibrium configuration, C_1 is bounded below by α, independently of the size of the market, viz.

$$C_1 \geqslant \alpha. \tag{2.3}$$

[24] The reason for introducing the supremum over k is as follows: some 'sizes of jump', measured by k, may be profitable for the deviant, while other are not. In seeking to characterize a lower bound to concentration, we seek to eliminate configurations that can be broken by a high-spending entrant using *any* value of k.

[25] In defining alpha, we have for convenience taken the limits in a particular order: we first seek a pair k, $a(k)$ which hold for *all* configurations; alpha is then defined by taking the supremum over k. This begs the question: what if, as the quality level(s) of firms rise(s), we could always find a suitable pair k, $a(k)$, but only by choosing a different (larger) value of k, as quality levels increase? It is possible to construct an example of the kind, in which there is a lower bound to concentration which is strictly positive – even though there is no single pair k, a (k) with $a(k) > 0$ as defined above. This indicates that there is a (slight) restriction introduced in defining alpha in the present manner. (In other words, the restriction stated in (2.3) below always holds, but in some (rather special) circumstances a tighter version of the restriction is valid.)

INTERPRETING ALPHA

In an industry where alpha is strictly positive, a high-spending entrant can achieve a profit exceeding some fixed proportion of current industry sales *independently of the number of low-spending rivals*. If the industry consists of a large number of firms, all with a small market share, then this arrangement can be disrupted by the arrival of a single 'high spender'; the profits of such a high spender cannot be eroded to zero by the presence of low spenders, however many are present. Even if the prices of the low quality products fall to the unit (variable) cost of production, at least some fraction of consumers will be willing to pay a price premium for the high-quality product.[26]

The interpretation of alpha hinges on the question: can the profit of a high-spending firm be diluted indefinitely by the presence of a sufficiently large number of low-spending rivals?

A loose but useful analogy is provided by thinking of a lottery in which N players buy tickets costing \$1, and one winner draws a prize of predetermined value Y. The expected payoff to a high-spending individual who buys k tickets while the remaining $(N-1)$ players buy one ticket each is equal to $kY/[k+(N-1)]$. For any k, this can be made arbitrarily close to zero by choosing N sufficiently high. This captures the nature of an industry where alpha equals zero: the returns to a high-spending firm can be diluted indefinitely by the presence of many

Equation (2.3) constitutes a restatement of the basic non-convergence result developed in the preceding theorem. In the light of this result, we see that alpha serves as a measure of the extent to which a fragmented industry can be destabilized by the actions of a firm who outspends its many small rivals in R&D or advertising. The value of alpha depends directly on the profit function of the final stage subgame, and on the elasticity of the fixed cost schedule. Hence it reflects both the pattern of technology and tastes and the nature of price competition in the market. We noted earlier that the results do not depend on the way we label u, but only on the composite mapping from F to π. To underline this point, we can re-express the present result as follows: increasing quality by a factor k requires that fixed outlays rise by a factor k^β. For any given value of β, write k^β as K. We can now write any pair $(k, a(k))$ as an equivalent $(K, a(K))$ pair. Alpha can then be described as the highest ratio $a(K)/K$ that can be attained by any choice of $K \geqslant 1$.

[26] It is worth emphasizing that our assumption on the cost structure states that a higher value of u involves an increase in fixed (and sunk) outlays; it does not involve a rise in the unit variable cost of production. It is natural in the present context to ask: what if a rise in quality involves both a rise in fixed outlays, and a rise in unit variable cost. The answer is: if the latter effect is small, there will still be an $a(k)$, k pair as defined above, and the 'non-convergence' result still holds. But if the rate at which unit variable cost rises with u is sufficiently steep, then $a(k)$ will fall to zero for all k. This idea lies at the heart of the literature on vertical product differentiation [Shaked and Sutton (1982)]. [For an overview of the main ideas, see Sutton (1991, pp. 70–71) and the references cited therein.]

low-spending rivals. It is possible, to in this setting, to have an equilibrium configuration in which a large number of firms each purchase a single ticket – so that if we measure concentration in terms of the number of tickets purchased, we have a fragmented industry.

In contrast to this, consider the Cournot model with quality described in Section 2.2 above. Here, if we begin from a configuration in which N firms have qualities not exceeding $\hat{u}$, and introduce an $(N+1)$th firm with quality $k\hat{u}$, then the profit of this entrant can be computed from Equation (2.1) of the text as

$$\pi(k\hat{u} \mid u) = \left\{1 - \frac{N}{k\hat{u}} \frac{1}{\sum_1^N \frac{1}{u_j} + \frac{1}{k\hat{u}}}\right\}^2 \cdot S.$$

Now for any N and any set of qualities $u_1, \ldots, u_N$ none of which exceed $\hat{u}$, the expression on the r.h.s. cannot be less than

$$\left\{1 - \frac{1}{k}\right\}^2 \cdot S$$

whence, noting that industry sales revenue equals S, we have

$$a(k) = (1 - 1/k)^2 > 0 \quad \text{for } k > 1.$$

2.5.2. *An ancillary theorem*

Within our present context of a classical market in which all goods are substitutes, the interpretation of alpha is straightforward. The parameter $a(k)$ measures the degree to which an increase in the (perceived) quality of one product allows it to capture sales from rivals. Thus the statement, within this context, that there exists some pair of numbers k and $a(k) > 0$ satisfying the above conditions requires only very weak restrictions on consumer preferences [for a detailed justification of this remark, by reference to a specific representation of consumer preferences, see Sutton (1991, pp. 74–76)].[27] The question of interest is: how costly is it, in terms of fixed outlays, to achieve this k-fold increase in u? This is measured by the parameter β. With this in mind, we proceed to define a family of models, parameterized by β, as follows: we take the form of the profit function, and so the function $a(k)$, as fixed, while allowing the parameter β to vary. We assume, moreover, that for some value of k, $a(k) > 0$. The value of $\alpha = \sup_k a(k)/k^\beta$

[27] What is required is that at least some fraction of consumers will be willing to pay some price in excess of unit variable cost c for the good of quality $k\hat{u} > \hat{u}$, whatever the prices ($\geqslant c$) of all rival goods. It is intuitively clear that this will be the case once some fraction of consumers are willing to switch from rival substitute goods, in response to a quality increase. In a simple 'vertical product differentiation' model such as the Cournot model with quality, this result is immediate (see Box). A formal statement of the (weak) restrictions on consumer preferences required to ensure this result in the more complex setting that combines vertical and horizontal product attributes is set out in Sutton (1991, p. 75).

varies with β. (The case $\alpha = 0$ can be treated as a limiting case as $a(k) \to 0$ or $\beta \to \infty$.)

We are now in a position to develop an ancillary theorem whose role is to allow us to use the observed value of the R&D and/or advertising to sales ratio to proxy for the value of β. The intuition behind the ancillary theorem is this: if the value of β is high, this implies that the responsiveness of profit to the fixed outlays of the deviant firm is low, and under these circumstances we might expect that the level of fixed outlays undertaken by all firms at equilibrium would be small; this is what the theorem asserts.

We establish this by showing that certain configurations must be unstable, in that they will be vulnerable to entry by a low-spending entrant. The idea is that, if spending on R&D and advertising is ineffective, then a low spending entrant may incur much lower fixed outlays than (at least some) incumbent firm(s), while offering a product that is only slightly inferior to that of the incumbent(s). The ancillary theorem allows us to fix some threshold level for the ratio of R&D plus advertising to sales, and consider the set of industries for which the ratio exceeds this threshold level: we may characterize this group as being 'low β' and so 'high alpha' industries, as against a control group of industries in which R&D and advertising levels are (very close to) zero. It is this result which leads to the empirical test of the non-convergence theorem.

Before stating the theorem however, some preliminary development is necessary, since the proof of the ancillary theorem rests on an appeal to the entry of a low-spending firm. This raises a technical issue: suppose, for the sake of illustration, that the underlying model of the final stage subgame, whose properties are summarized in the profit function $\pi(\cdot)$, takes the form of the elementary 'Bertrand model'. In this setting, if all firms offer the same quality level, once one firm is present in the market, no further entry can occur; for any entry leads to an immediate collapse of prices to marginal cost, and so the entrant can never earn positive margins, and so cover the sunk cost incurred in entering. In what follows, we will exclude this limiting case. (To exclude it is harmless, relative to the theory, since the theory aims to place a lower bound on the 1-firm concentration ratio; and if we are working in this 'Bertrand limit', then the 1-firm concentration ratio is necessarily unity, as we saw in Section 2.2.)

To define and exclude this limiting case, we need to specify the relationship between the profit earned by an entrant, and the pre-entry profit of some active firm (the 'reference firm').

Consider an equilibrium configuration in which the industry-wide R&D (or advertising) to sales ratio is x (>0). Within this industry, we select some reference firm whose R&D and advertising outlays constitute a fraction x (or greater) of its sales revenue. There must be at least one such firm in the industry, and since this firm must satisfy the viability condition, it must earn a gross profit of at least fraction x of its sales revenues in order to sustain its level of R&D and advertising.

Now consider an entrant that offers the same quality level as the reference firm. Insofar as entry reduces prices, this entrant will enjoy a lower price–cost margin than that earned by the reference firm in the pre-entry situation. But, for a sufficiently high value of x, we assume that the entrant will enjoy some strictly positive price–cost mar-

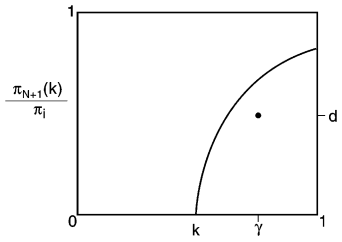

Figure 35.5. The relative profit of a low-quality entrant. The incumbent firm, labeled i, offers quality u_i and earns (pre-entry) profit π_i. The entrant, labeled firm $N+1$, offers quality ku_i and earns profit $\pi_{N+1}(k)$.

gin, so that its final stage profit is strictly positive (this is what fails in the 'Bertrand limit').

This is illustrated in Figure 35.5. The horizontal axis shows the ratio between the quality of the entrant's product, and that of the reference firm; k varies from 0 to 1, with a value of 1 corresponding to the entrant of equal quality. Our assumption states that for $k = 1$, the entrant's post-entry profit is strictly positive. On the vertical axis, we show the ratio of the entrant's profit to the pre-entry profit of the reference firm.[28] Our exclusion of the Bertrand limit states that, for $k = 1$, this ratio is strictly positive. We further assume that the entrant's profit varies monotonically with its quality, and so with k. It then follows that we can depict the entrant's profit as a function of k as a curve; the assumption states that this curve does not collapse inwards to the bottom right-hand corner of the diagram (the 'Bertrand limit'). Specifically, it says that there is some value of x, such that if the pre-entry price–cost margin exceeds x, then there is some point in the interior of the square such that the curve we have just described lies above this point.

We state this formally as follows.

ASSUMPTION 4. There is some triple (x, γ, d) with $0 < x, \gamma < 1$, and $0 < d < 1$, with the following property: suppose any firm i attains quality level u_i and earns final-stage profit π_i that exceeds a fraction x of its sales revenue. Then an entrant attaining a quality level equal to $\max(1, \gamma u_i)$ attains a final-stage profit of at least $d\pi_i$.

The ancillary theorem linking the R&D (or advertising)/sales ratio to the parameter β now follows.

THEOREM 2. *For any threshold value of the R&D (or advertising)/sales ratio exceeding* $\max(x, 1-d)$*, where x and d are defined as in Assumption* 4*, there is an associated*

[28] It is convenient to work in terms of this ratio, so as to state the assumption in a way that does not involve the size of the market S.

value of β^* such that for any $\beta > \beta^*$, no firm can have an R&D/sales ratio exceeding this threshold in any equilibrium configuration.[29]

A proof of this theorem is given in Sutton (1998, ch. 4). An implication of Theorem 2 is that an industry with a high R&D (or advertising)/sales ratio must necessarily be a high-alpha industry. With this result in place, we are now in a position to formulate an empirical test of the theory: Choose some ('sufficiently high') threshold level for the R&D (or advertising)/sales ratio (written as R/Y in what follows), and split the sample of industries by reference to this threshold. All industries in which alpha is close to zero will fall in the low R/Y group, and so for this group the lower bound to the cloud of points in (C, S) space should extend to zero as $S \to \infty$. For all industries in the group with high R/Y, on the other hand, the value of β will lie below β^*, and so the lower bound to concentration will be bounded away from zero by $C_1 \geqslant a(k)/k^{\beta}$.

In pooling data across different industries, it is necessary to 'standardize' the measure of market size by reference to some notion of the minimum level of setup cost F_0, which we write as ε. A practical procedure to represent this as the cost of a single plant of minimum efficient scale, and to write the ratio of annual industry sales revenue to minimum setup cost as S/ε. This leads to the prediction illustrated in Figure 35.6; tests of this prediction are reported in the next section.[30,31]

It is interesting to consider the relationship between this prediction and the traditional practice of regressing concentration on a measure of scale economies to market size (essentially S/ε), together with a measure of advertising intensity and R&D intensity. Such regressions indicated that concentration fell with S/ε, and rose (weakly) with the advertising–sales ratio [Sutton (1991, p. 124)]. It can be shown that, under the present theory, these results are predicted to emerge from the (misspecified) regression relationship. [For a full discussion, see Sutton (1991, Annex to ch. 5).]

[29] It may be helpful to illustrate the ideas of Assumption 4 and Theorem 2 by reference to a numerical example: suppose $x = 0.04$ and $d = 0.95$ (intuitively: entry reduces prices only slightly). Say we select all those industries with R&D sales ratios exceeding $\max(x, 1 - d) = 0.05$, whence we can find at least one incumbent firm i, that spends at least 5% of its sales revenue Sy_i on R&D or advertising. Now suppose we let $\beta \to \infty$, so that these fixed outlays become completely ineffective. Then an entrant to this industry can, by spending nothing on such fixed outlays, enjoy a positive net profit. Its final-stage profit falls short of the pre-entry final-stage profit of the incumbent by at most $0.05S\pi_i$, but its saving on fixed outlays relative to the incumbent is at least $0.05Sy_i > 0.05S\pi_i$, whence its net profit exceeds the pre-entry net profit of the incumbent, which is non-negative. It follows that, once β is 'sufficiently large', the pre-entry configuration is not an equilibrium configuration.

[30] For a discussion of alternative measures of setup cost, see Sutton (1991, pp. 93–99).

[31] A further practical issue arises in relation to the use of the (theoretically appropriate) 1-firm concentration ratio C_1. Since official statistics never report this, for reasons of confidentiality, it has long been customary in IO to use a readily available measure such as C_4. The prediction shown in Figure 35.7 still applies, here, of course.

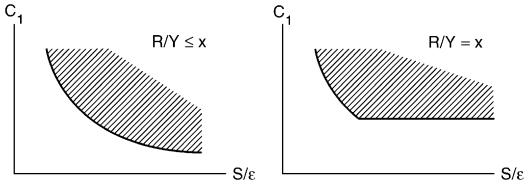

Figure 35.6. The 'bounds' prediction on the concentration–market size relationship.

2.5.3. *Empirical evidence*

We have developed this first version of the non-convergence theorem in a context in which the classical market definition applies, i.e. the market comprises a single set of substitute goods, so that an increase in fixed and sunk outlays enhances consumers' willingness-to-pay for all its products in this market. Now this condition will apply strictly only in rather special circumstances. One setting in which it applies to a good approximation is that of certain groups of advertising-intensive industries. Here, even though the market may comprise a number of distinct product categories, the firm's advertising may support a brand image that benefits all its products in the market. (A similar argument applies in R&D intensive industries when scope economies in R&D operate across different groups of products in the market; see below, Section 2.6.)

The first test of the non-convergence theorem [Sutton (1991, ch. 5)] was carried out on a dataset for 20 industries drawn from the food and drink sector, across the six largest Western economies. The industries were chosen from a single sector so as to keep constant as many extraneous factors as possible. The food and drink sector was chosen because, alone among the basic 2-digit SIC industry groups, it is the only one in which there is a nice split between industries that have little or no advertising (sugar, flour, etc.) and industries that are advertising-intensive (breakfast cereals, petfood, etc.).

Data was compiled from market research reports, combined with company interviews. The industry definitions used are those which are standard in the market research literature, and these correspond roughly to 4-digit SIC definitions. All industries for which suitable data could be assembled were included. The size of each market was defined as the number of 'minimum efficient scale' plants it would support, where the size of a m.e.s. plant was measured as the median plant size in the U.S. industry.

The sample was split on the basis of measured advertising sales ratios into a control group ($A/S < 1\%$) and an experimental group ($A/S \geqslant 1\%$); though the large majority of industries in the latter group had advertising–sales ratios that were very much higher than 1%.

The data from the study is illustrated in Figure 35.7, which shows the scatter of observations for the control group (upper panel) and the experimental group (lower

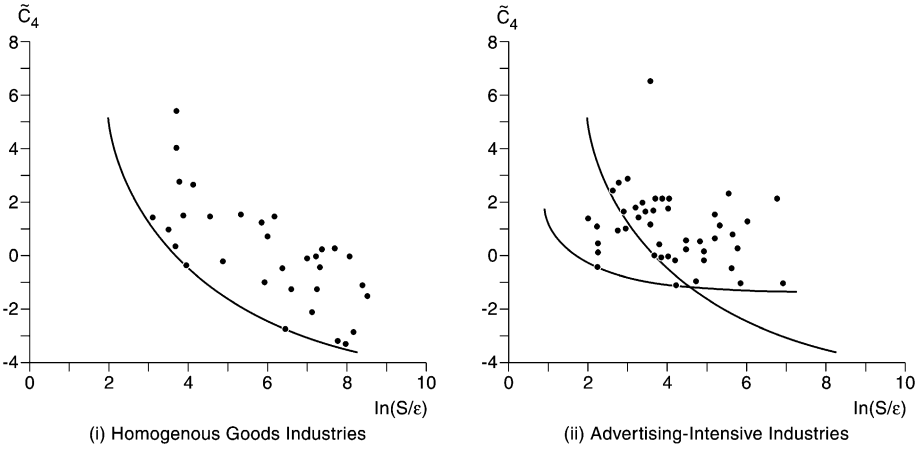

Figure 35.7. A plot of $\tilde{C}_4$ versus S/ε for advertising-intensive industries (ii) and a control group (i). The fitted bound for the control group is reproduced in panel (ii) for comparison purposes.

panel) on plots of (a logit transformation[32] of) the 4-firm concentration ratio. A fitted lower bound[33] for the control group indicates an asymptotic value for $\underline{C}_4(S)$ in the limit $S \to \infty$ which is 0.06%; the corresponding lower bound for the experimental group is 19%, which is significantly different from zero at the 5% level.[34]

The non-convergence property has also been investigated by Robinson and Chiang (1996), using the PIMS data set, a dataset gathered by the Strategic Planning Institute

[32] The logit transformed value $\tilde{C}_4 = \ln(C_4/(1 - C_4))$ is defined on $(-\infty, +\infty)$ rather than [0, 1] and this may be preferred on econometric grounds.

[33] Following Smiths' (1985, 1988, 1994) maximum likelihood method. Techniques for bounds estimation are discussed in Sutton (1991, ch. 5), and Sutton (1998, ch. 4). An alternative method which has some attractive features from a theoretical viewpoint, but which has less power than that of the maximum likelihood methods described by Smith, is that of Mann, Scheuer and Fertig (1973); see Sutton (1998) for details. Both these approaches are sensitive to the presence of outliers, and for this reason some authors, including Lyons, Matraves and Moffat (2001), favor alternative methods that have proved useful in the estimation of frontier production functions. A very simple method of attack is provided by quantile regression methods, which Giorgetti (2003) has recently applied, in combination with maximum likelihood methods, to examine the lower bound to concentration for a sample of manufacturing industries.

[34] These findings have been questioned for the case of the U.S. food and drink sector by Rogers (2001). Rogers reports a regression of concentration against market size/setup costs, for 40 4-digit SIC industries in this sector, over five census years and finds that advertising raises the level, but not the slope of this relationship. He interprets this result as reflecting claims by Rogers and Ma (1994) and Rogers and Tockle (1999) that food and drink advertising in the U.S. has been losing its effectiveness over time and/or that merger activity in non-advertising intensive food and drink industries has led to a narrowing of the difference in concentration between these two groups of industries. Rogers also notes that a possible reason for the difference in findings relative to Sutton (1991) lies in problems of market definition for some of the 4-digit SIC industries in the sector. [Such problems arise when SIC industries comprise both advertising intensive and non-advertising intensive submarkets; see, for example, Giorgetti (2003).]

representing a wide range of firms drawn mostly from the Fortune 1000 list. Firms report data for each of their constituent businesses (their operations within each industry, the industry being defined somewhat more narrowly than a 4-digit SIC industry); for a discussion of the PIMS dataset, see, for example, Scherer (1980).

The unit of observation here is the individual business, and the sample comprises 1740 observations. The sample is split into a control group (802 observations) in which both the advertising–sales ratio and the R&D–sales ratio for the business lie below 1%. The remaining ('experimental') groups comprise markets where one or both ratios exceed 1%.

Within the control group, the authors set out to test whether an increase in the 'toughness of price competition' raises the lower bound $\underline{C}_k(S)$. They do this by using three proxies for the 'toughness of price competition': price competition is tougher if (1) the product is standardized rather than customized, (2) the product is a raw or semi-finished material, or (3) buyer orders are infrequent. The findings of the study are:

(i) the 'non-convergence' property is confirmed for all 'experimental' groups,
(ii) the asymptotic lower bound for the control group converges to zero, but
(iii) when the control group is split into the 'tough' and 'non-tough' price competition sub-groups, it is found that tougher price competition shifts the bounds upwards (as predicted by the theory), but the asymptotic lower bound to concentration for the 'tough price competition' group is now strictly positive, i.e. it does not converge to zero asymptotically, contrary to the predictions of the theory. Instead, the (3-firm) concentration ratio converges to an asymptotic value of 10%, intermediate between that for the 'weak price competition' control group, and the values found for the 'experimental groups' (15.8%–19.6%). (The authors add a caveat to this conclusion, noting that this finding may reflect data limitations in their sample.)

An investigation of the non-convergence property by Lyons and Matraves (1996) and Lyons, Matraves and Moffat (2001) uses a data set covering 96 NACE 3-digit manufacturing industries for the four largest economies in the European Union, and a comparison group for the U.S. Splitting the sample by reference to observed levels of the advertising–sales ratio and R&D–sales ratio as in Robinson and Chiang, the authors estimate a lower bound to concentration for each group.

A key novelty of this study is that it attacks the question of whether it is more appropriate to model the concentration–size relationship at the E.U. level, or at the level of national economies (Germany, UK, France, Italy). The authors construct, for each industry, a measure (labeled 't') of intra-EU trade intensity. They hypothesize that, for high (respectively low) values of t, the appropriate model is one that links concentration in the industry to the size of the European (respectively national) market. They proceed to employ a maximum likelihood estimation procedure to identify a critical threshold t^* for each country, so that according as t lies above or below t^*, the concentration of an industry is linked to the size of the European market, and conversely. Within this setting, the authors proceed to re-examine the 'non-convergence' prediction. They find

that 'a very clear pattern emerges, with ... the theoretical predictions ... receiving clear support' [Lyons, Matraves and Moffat (2001)].

The key comparison is between the asymptotic lower bound to concentration for the control group versus that for the experimental groups. Over the eight cases (4 countries, E.U. versus National Markets) the point estimate of the asymptotic lower bound for the control group lies below all reported[35] estimates for the three experimental groups, except in two instances (advertising-intensive industries in Italy, advertising and R&D intensive industries in France); in both these cases the reported standard errors are very high, and the difference in the estimated asymptotic value is insignificant.

2.6. *Markets and submarkets: the R&D vs concentration relation*

The theory set out above rests on the classical definition of a market as comprising a set of goods, all of which are substitutes. We may reasonably apply this model to, for example, a narrowly defined market in which firms' advertising outlays create a 'brand image' that benefits all the firms' offerings in the market. But once we turn to the case of R&D intensive industries, the formulation of the theory developed above becomes inadequate. For in this setting, once we define the market broadly enough to incorporate all substitute goods, we may, for example, be left with various sets of products, each of which requires some distinct technical know-how. Here, each firm must choose not only its level of R&D spending, but the way in which its R&D efforts should be divided among the various product groups ('submarkets'). These different R&D programs may, or may not, contain common elements, leading to 'economies of scope' in R&D across different submarkets. On the demand side, too, there may be linkages across submarkets: it may, for example, be the case that products within each sub-group are close substitutes, but some products from different subgroups are weak substitutes. It is tempting to dismiss all such problems as a 'question of aggregation' by suggesting that we should analyze competition and market structure at the level of the submarket. However, the logic of partial equilibrium analysis lies in defining a market broadly enough to justify taking as given what is going on in other markets; this was the idea behind Joan Robinson's classic definition of a market by reference to a 'break in the chain of substitutes'. Here, however, firms' actions in one submarket will have an effect on the profits of firms in other submarkets, and so on their strategic choices.

The idea of responding to these problems by working at a lower level of aggregation becomes increasingly unattractive as we move to the context of markets where the pattern of linkages across submarkets is relatively complex; for a discussion of these difficulties, see Sutton (1998, pp. 14–16, 165). The only satisfactory way forward in this setting lies in building these features into the theory. In what follows, we extend the model of the preceding section by introducing the notion of a set of 'technological trajectories', and their associated 'submarkets' as follows.

[35] Some cases are unreported due to lack of a sufficient sample size.

The capability of firm i is now represented by a set of quality indexes, its quality index on trajectory m (equivalently, in submarket m), being denoted by $u_{i,m}$, where m runs from 1 to M. A firm's capability is represented by the vector

$$\mathbf{u}_i = (u_{i,1}, \ldots, u_{i,m}, \ldots, u_{i,M})$$

and a configuration is written as $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_N)$.

The firm's fixed cost can now be written as the sum of the costs incurred on each trajectory,[36] viz.

$$\sum_m F_0 u_{i,m}^{\beta}.$$

A full discussion of the theory, within this more complex setting lies outside the scope of this chapter; for details, the reader is referred to Sutton (1998, ch. 3). Here, attention is confined to an informal account of the key idea, which concerns the way in which we represent the idea of 'linkages' across submarkets.

To motivate ideas, we begin by raising some questions. We begin with linkages on the demand side. Suppose, firstly, that the products in the different submarkets are relatively close substitutes. We might expect intuitively, in this setting, that as firms advanced along one trajectory, they might 'steal' market share from firms operating along other trajectories. Can a process of this kind lead to the emergence of a single 'dominant trajectory', and so to high concentration at the level of the market as a whole?

At the other extreme, suppose the products in the different submarkets are poor substitutes. Here, an obvious limiting case arises, in which the market becomes separable into a number of independent submarkets – and we might expect that, even if each of these constituent submarkets is concentrated, the fact that different firms may be present in different submarkets makes possible an outcome in which the market as a whole is highly fragmented.

Similar questions arise in respect of linkages on the supply side, i.e. when there are economies of scope in R&D across the different submarkets. A simple way of introducing such scope economies into the analysis is to replace the above additive cost function by a sub-additive function. For example, we may suppose that a firm's quality index on trajectory m is a function of its spending both on trajectory m, and – to some degree – on its spending on other trajectories. As with linkages on the demand side, the presence of such scope economies can influence the degree to which a concentrated outcome emerges in the market as a whole.

It will be clear at this point that a general depiction of the nature of the linkages that may be present between submarkets would be rather complicated. It turns out, however, that for our present purposes a very simple representation proves to be adequate. This

[36] For simplicity, I will confine attention to a setting in which all the submarkets are treated symmetrically. The additive form of the cost function implies that there are no economies of scope in R&D; the introduction of such scope economies is considered below.

involves introducing a new parameter σ, defined on the interval [0, 1], which represents the strength of linkages between submarkets. Our 'class of models', which was parameterized by β above, will now be parameterized by the pair (β, σ). The focus of analysis will lie in distinguishing between two cases: where σ is 'large', and where σ becomes close to zero, the latter case being identified with the limiting case where the market consists of a set of 'independent submarkets'.

Before introducing this new parameter, we first pause to define precisely what is meant by a 'submarket' and its associated 'technological trajectory', in the context of the present theory. We do this by asserting the existence of some pair of numbers k_0 and $a(k_0)$ which play the role of the k and $a(k)$ pair in our preceding discussion – but which relate, not to the market as a whole, but to any specific submarket. In other words, we assume that a firm that raises its capability along some technical trajectory, i.e. raises the quality u of its product(s) in some associated submarket, will thereby steal sales from other products *in the same submarket*; but we leave open the question of what happens in respect of products in other submarkets. This captures the idea that products within the same submarket are close substitutes, and incorporate the same technology.

ASSUMPTION 5. There is a pair (a_0, k_0) with $a_0 > 0$, $k_0 > 1$ such that in any configuration $\mathbf{u}$ with maximum quality $\hat{u}$ attained along trajectory m, an entrant offering quality $k_0\hat{u}$ along trajectory m will achieve a final-stage profit of at least $a_0 Sy_m(\mathbf{u})$, where $Sy_m(\mathbf{u})$ denotes the (pre-entry) total sales revenue in submarket m.

We augment the set of assumptions introduced in the preceding section by two additional assumptions, whose role is to introduce the parameter σ, and to pin down the distinction between the two cases just described.

Assumption 6 introduces the substitution parameter. For each configuration $\mathbf{u}$ define the configuration $\mathbf{u}^{(m)}$, in which firms' capabilities on trajectory m are as in $\mathbf{u}$, but all other trajectories are unoccupied (so that goods in category m face no competition from goods in other categories). The following intuition motivates Assumption 6: removing substitute products does not decrease the demand for products of any category[37]; and in the special case where goods in different groups are poor substitutes, the demand for products in any particular group is unaffected by the prices and qualities of goods in other groups.

ASSUMPTION 6. (i) For any $\sigma \geqslant 0$

$$y_m\left(\mathbf{u}^{(m)}\right) \geqslant y_m(\mathbf{u}),$$

whereas for $\sigma = 0$, this relation holds as an equality.

(ii) As $\sigma \to 0$, the ratio

$$\frac{y_m(\mathbf{u})}{y_m(\mathbf{u}^{(m)})}$$

converges to 1, uniformly in $\mathbf{u}$.

[37] We ignore any demand complementaries throughout.

Part (i) of the assumption says that removing products in other submarkets does not diminish the sales of products in submarket m. Part (ii) of the assumption says that when σ is close to zero, the removal of products in other submarkets has a negligible effect on the sales of products in submarket m.

The next assumption constitutes the key step. What it does is to pin down the concept of σ as a measure of the strength of linkages between trajectories, and in particular to identify the limiting case where $\sigma \to 0$ as that of independent submarkets (or trajectories). We introduce this idea by re-examining the case of a low-quality entrant. We now ask: how low can the quality ratio fall before this entrant's final-stage profit becomes zero? Here it is appropriate to consider the cases of entry both along trajectory m, and along a different trajectory. In the case of entry along the same trajectory, we might expect that there is some quality ratio $\gamma_0 > 0$ sufficiently low that a product of quality less than $\gamma_0 \tilde{u}$ could not earn positive profit in competition with a product of quality $\tilde{u}$. In the case of entry along a different trajectory, this should still be true if the products associated with different trajectories are substitutes. However, if $\sigma = 0$, so that demand for each product category is independent of the prices and qualities of products in other categories, then this would no longer be so. This motivates:

ASSUMPTION 7. For any $\sigma > 0$, there exists a quality ratio $\gamma_0 \in (0, 1]$ such that a product of quality $\gamma \tilde{u}$, where $\gamma \leqslant \gamma_0$, cannot command positive sales revenue if a rival firm offers a product of quality $\tilde{u}$ on any trajectory.

REMARK. Assumption 7 is the only assumption that restricts the way in which the profit function $\pi(\cdot)$ varies with the parameter σ. The restriction is very weak; in particular, it imposes no monotonic relationship between σ and $\pi(\cdot)$. Rather, it simply imposes a certain restriction for any strictly positive value of σ, thereby leaving open the possibility that this restriction breaks down in the limit $\sigma \to 0$.

The intuition behind this assumption may be made clearer by noting what would follow if γ_0 were equal to zero (i.e. if Assumption 7 could not be satisfied for any strictly positive γ_0). This would imply that we could find some pair of products whose qualities were arbitrarily far apart, both of which could command positive sales at equilibrium. Assumption 7 states that this can happen only if σ is close to zero.[38]

The intuition behind the assumption is clear, in regard to the case where the linkages are on the demand side, i.e. where the products are substitutes. The formulation of the assumption is designed, however, to capture both these demand–side linkages, and supply–side linkages operating via scope economics in R&D. To see the intuition regarding these latter linkages, consider a high-spending firm operating in another submarket, whose R&D spending in that submarket enhances its product quality in submarket m. Once again, the low-quality firm in submarket m may be unable to achieve

[38] This assumption can be illustrated using Figure 35.6 above, as follows: it states that for $\sigma > 0$, the curve showing the relative profit earned by a new (low quality) entrant will meet the horizontal axis at some strictly positive value of γ. For $\sigma = 0$, it may meet at the origin.

positive profit if this rival's relative spending, and so its quality level, in its primary submarket is sufficiently high. If the strength of these (scope economy) linkages is vanishingly small however, this is no longer the case; an arbitrarily wide gap between a low-quality product in submarket m, and a high-quality product in some other submarket is consistent with the former product's viability.

Within this framework, we can now develop a version of the non-convergence theorem appropriate to the setting of markets that contain many submarkets. To do this, we need to extend the set of 'observables' R/Y and C_1 used in the preceding section, to incorporate a third parameter, labeled h, which measures the degree to which the equilibrium outcome is seen to involve a breaking up of the market into a greater or lesser number of submarkets.

We define a 'homogeneity index', labeled

$$h = \max_m \frac{y_m(\mathbf{u})}{y(\mathbf{u})}.$$

Here, h represents the share of industry sales revenue accounted for by the largest product category. If all products are associated with the same trajectory, then $h = 1$. If there are many different trajectories, each associated with a small product group, then h is close to zero. We now state the reformulated version of the non-convergence theorem.

THEOREM 3. *In any equilibrium configuration, the one-firm sales concentration ratio satisfies*

$$C_1 \geqslant \frac{a_0}{k_0^{\beta}} h.$$

The proof of this theorem mimics that of Theorem 1 above, and is omitted here.

It is worth noting that, while β and σ are exogenous parameters that describe the underlying pattern of technology and tastes in the market, h is an endogenous outcome. The intuition is as follows: if σ is very high, so that submarkets are very closely linked, the process of competition among firms on different trajectories will lead to the emergence of a single dominant trajectory, so h will be high, and C_1 will be high also. But if σ is close to zero, firms in one submarket have little or no influence on those in another. One possible form of equilibrium is that in which a different group of firms operate in each submarket, so that h is low, and C_1 is low also. This is not the only possible outcome: another equilibrium involves having the same group of firms in each submarket, so that h is low but C_1 is high.

2.6.1. Some illustrations

The new idea that arises when we move beyond the classical market to this more complex setting is that two polar patterns may emerge in high-technology industries. The first is the pattern of 'R&D escalation' along a single technical trajectory, leading to a high level of concentration – this was the pattern explored in the preceding section.

The second is a pattern of 'proliferation' of technical trajectories and their associated submarkets. The key point to note is that the structure of submarkets emerges endogenously: specific illustrations may be helpful here.

The history of the aircraft industry from the 1920s to the end of the pre-jet era in the late 1950s illustrates the first pattern. The industry of the 1920s and early 1930s featured a wide variety of plane types: monoplanes, biplanes and triplanes; wooden planes and metal planes; seaplanes and so on. Yet buyers were primarily concerned with one key attribute: the "cost per passenger per mile". So once one design emerged which offered the best prospects for minimizing this target (the DC3), the industry quickly converged on a single technical trajectory [the details of this case are set out in Sutton (1998, ch. 16)].

The other polar pattern is illustrated by the Flowmeter industry, an industry characterized by a high level of R&D intensity, which supports a large number of firms, many of whom specialize in one, or a few, of the many product types (submarkets) that co-exist in the market. Different types of flowmeter are appropriate for different applications, and the pattern of 'substitution' relationships among them is complex and subtle [see Sutton (1998, ch. 6)]. The focus of R&D spending in the industry is associated with the introduction of new basic types of flowmeter, which offer advantages to particular groups of buyers. Thus the pattern here is one of 'proliferation' rather than escalation; the underlying pattern of technology and tastes is such that the industry features a large number of submarkets.

2.6.2. *Empirical evidence II*

Theorem 3, together with Theorem 2 above, implies an empirical prediction regarding the joint distribution of concentration, R&D intensity, and market segmentation (Figure 35.8). Suppose we take a group of industries within some large economy for which the R&D/sales ratio lies above some (high, though unspecified) cutoff value. Theorem 2 implies that associated with the cutoff level of R&D intensity, there is some associated value of β^* such that for all industries in this group, $\beta \leqslant \beta^*$. Theorem 3 then implies that for all industries in this group

$$C_1 \geqslant \frac{a_0}{k_0^{\beta}} \cdot h.$$

If we define a control group of industries for which R&D intensity is low, this group should contain some industries for which the value of β is high. Here, according to the theory, the lower bound to concentration converges to zero as the size of the economy becomes large,[39] independently of the degree of market segmentation, as measured by h. Hence if we examine such a group, for a large economy, we expect to find that concentration can be close to zero independently of h (Figure 35.8).

[39] Recall that $k_0 > 1$ and $a(k_0) > 0$; the fact that $a(1) = 0$ was noted in footnote 23 above.

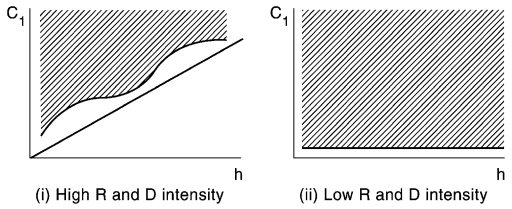

Figure 35.8. The empirical prediction.

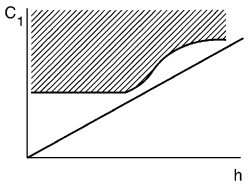

Figure 35.9. The effect of scope economies in R&D.

There is one important caveat, however. Linkages between submarkets are of two kinds: those on the demand side (substitution) and those on the supply side (scope economies in R&D). Our empirical measure of the parameter h relates only to demand side effects; the identification and measurement of scope economies in R&D across submarkets would not be feasible in practice, and so the above test has been formulated in a way which neglects these supply-side linkages. But if such linkages are present, how is the prediction illustrated in Figure 35.8 affected? It is easy to show that the presence of such linkages will lead to an upward shift in the lower bound in the region of the origin, as illustrated in Figure 35.9, so that we will no longer find points in the bottom left-hand corner of the diagram [for details, see Sutton (1998, ch. 3)].

Sutton (1998) reports empirical evidence on the C_4, h relationship for U.S. 5-digit manufacturing industries in 1977.[40] The control group consists of the 100 5-digit industries for which the combined advertising and R&D to sales ratio was least ($\ll 1\%$). The experimental group consists of all industries with an R&D/sales ratio exceeding 4% (46 industries). The value of h is computed as the ratio of the sales of the largest 7-digit product group to the sales of the industry.[41] The results are illustrated in Figure 35.10,

[40] This is the only census year that coincides with the short period for which the Federal Trade Commission's Line-of-Business program was in operation, so that figures for R&D intensity computed at the level of the business, rather than the firm, are available, albeit only at the 4-digit level. It is also, fortunately, the case that for that year, sales by product group at the 7-digit level were reported in the Census of Manufactures, thus allowing h to be computed for each 5-digit industry.

[41] The level of aggregation at which submarkets should be defined in estimating h should be low enough to ensure that the firms in that submarket offer competing (groups of) products. Working at the lowest available (7-digit) level seems appropriate, on this criterion. In defining the market, it is appropriate to work at a level of aggregation corresponding to a 'break in the chain of substitutes', and here it is probably best to use the 4-digit or 5-digit SIC level as the best available approximation in official statistics.

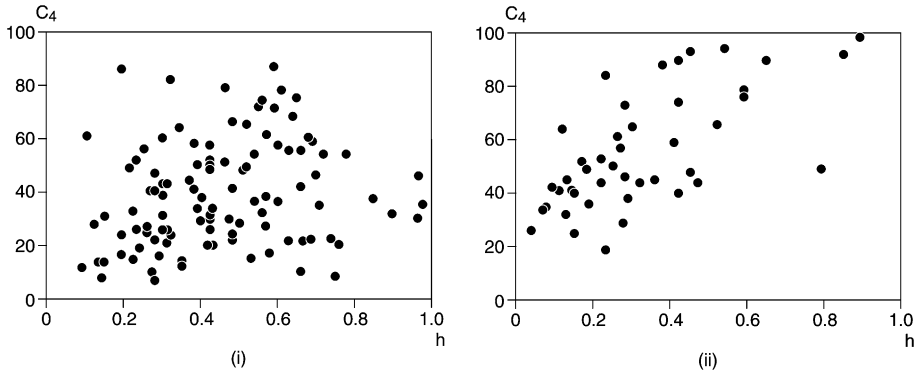

Figure 35.10. The (C_4, h) relationship for R&D-intensive industries (ii) and for a control group (i).

and they show a clear pattern of the form predicted.[42] A test of the same form as that reported above, following Smith's (1985, 1988) maximum likelihood method, indicates that the slope of the ray bounding the observations from below is significantly different from zero at the 5% level; the same results holds for the logit-transformed measure $\tilde{C}_4 = \ln(1/(1 - C_4))$.

A recent study investigating this relationship is that of Marin and Siotis (2001), who use the Chemintell data base to build up a dataset covering 102 markets in the chemicals sector in Europe (Germany, UK, Italy, France and Spain). Taking the European market as the relevant market, and splitting the sample into a 'low R&D' control group[43] of 42 industries and a 'high R&D' group of 60 industries, the authors examine the (C_1, h) relationship by carrying out a 'frontier analysis' exercise, within which they examine whether h affects the lower bound to concentration differently as between the control group and experimental group. They show that the lower bound rises with h in the experimental group, but is independent of h in the control group, consistently with the prediction illustrated in Figures 35.8 and 35.10[44].

As in the scatter diagrams from Sutton (1998) shown in Figure 35.10, the scatters shown by Marin and Siotis show an absence of points in the region of the origin for

[42] The outlier to the center right in the lower panel of Figure 35.10 is SIC 35731 (Electronic Computers). The recorded h index for this industry may be anomalous. The reported four-firm concentration ratio for SIC 35731 (Electronic Computers) fell from 75 percent in 1972 to 49 percent in 1977. The U.S. Department of Commerce's *Industrial Outlook* for 1977 noted that this industry's product lines had already fragmented into mainframes, minicomputers, etc. This was not reflected in the seven-digit product listings until the classification was revised in 1987. From that date, the measured h-index is much lower.

[43] The cutoff level chosen for the R&D/Sales ratio is 1.8%; the authors choose this level on the basis of a detailed examination of R&D spending figures, together with patent data.

[44] They also replicate the form of test used in Sutton (1998) by examining whether the ratio C_1/h is bounded away from zero for the experimental group; here, they report that the lower bound to the ratio is significantly different to zero at the 1% level.

the R&D intensive group. As noted above, this is consistent with the presence of scope economies across (at least some) submarkets in low-h industries; and in the presence of such scope economies, the asymptotic lower bound to $C_1(S)$ will lie above zero, in datasets collected at the level of the 'market'; nonetheless, as Marin and Siotis show, there is an important loss of information involved in applying the $C_1(S)$ relation at this level (i.e. without controlling for h).

2.6.3. Some natural experiments

It is of interest, as before, to look to some case histories of industries in which an exogenous shock led to a rise in the lower bound to concentration. As emphasized already, the theory does not specify any dynamic adjustment path to a new equilibrium. It does, however, specify the form of the 'profitable deviation' that becomes available to firms once the exogenous shock arises. Here, we expect to see a process of escalating R&D outlays among the firms. Checking that the exogenous shock sparks off such a process provides us with an ancillary test of the theory; if such a process is not in evidence, then the explanation offered by the theory for any subsequent rise in concentration is implausible.

The theory indicates that the (asymptotic) lower bound to concentration depends on two parameters, β, measuring the effectiveness of R&D (or advertising) in raising technical performance (or perceived quality), and σ, which measures the strength of linkages between submarkets. It is of interest, therefore, to investigate natural experiments which are driven by each of these parameters.

The photographic film industry affords a nice example of a shift in β, associated with the advent of color film in the 1960s. Up to the early 1960s, black and white film was dominant, and the technology of production was well established. There was little incentive to spend on R&D, as the quality of existing film was high, and few consumers were willing to pay a premium for higher quality film. Color film, on the other hand, was in its infancy, and quality was poor. As quality began to rise, its share of the market increased, and it became clear that it would in due course come to dominate the market. As this became apparent to firms, their R&D efforts escalated, and following a process of exit, and consolidation by merger and acquisition, the global industry came to be dominated by two firms (Kodak and Fuji).

A natural experiment involving a shift in σ is provided by the telecommunications sector. Up to the end of the 1970s, it had been standard practice for each major producer of switching systems to operate domestic procurement policies that strongly favored local firms. A shift in the U.S. policy marked by the breakup of AT&T in 1982, however, was seen – both within the U.S. and elsewhere – as signaling a likely move towards more deregulated markets, in which domestic procurement would no longer be the rule. In terms of the theory, this is equivalent to the joining-up of hitherto separated submarkets (and so to a rise in σ). The aftermath of these events saw an escalation of R&D efforts in the next generation of switching devices, and a rise in global concentration levels,

to the point where the world market came to be dominated by five firms by the early 1990s. [The details of both these cases will be found in Sutton (1998, ch. 5).]

2.6.4. Case histories

A series of recent studies have examined the evolution of structure in particular industries by reference to the 'endogenous sunk costs' framework. Bresnahan and Greenstein (1999) apply the model to the computer industry, Motta and Polo (1997) to the broadcasting (television) industry, while Matraves (1999) explores the pharmaceutical industry as an example of a 'low-alpha' industry [see also Sutton (1998, chs. 8 and 15) for the pharmaceuticals and computer industries, respectively]. Bakker (2005) explores the early history of the European film ('movie') industry, and explains the decline of the industry at the advent of the 'talkies' era by reference to an 'escalation effect' in which the U.S. film makers took the lead. The early history of the aircraft industry, explored in Sutton (1998, ch. 16), offers a nice example of 'escalation and shakeout', which was first documented systematically in the classic analysis of Almarin Phillips (1971).

3. The size distribution

A curious feature of all the models considered so far lies in the role played by outcomes in which all firms are of the same size. The bounds defined in Section 2 were generated by outcomes of this type, yet it is rare to encounter such an outcome in practice. Most industries are characterized by a fairly skew distribution, with a small number of relatively large firms. It is this observation that motivates the traditional growth-of-firms literature that began with the seminar contribution of Gibrat (1931).

In contrast to the game-theoretic models considered so far, which place strategic interactions at the heart of the analysis, the growth-of-firms tradition begins by abstracting from all such effects. Implicitly or explicitly, it works in a framework in which the market consists of a number of independent 'island' submarkets, each large enough to support a single production plant. It models the evolution of market structure by looking at a population of firms, each of which grows over time by taking up a succession of these discrete 'investment opportunities'. At the heart of the analysis lies a simple point: if firms enter a market over time, the recent arrivals will have had fewer of these opportunities, and will on average be smaller. In other words, the mere disparity in firms' ages is a source of size inequality. It will turn out in what follows that the size inequality deriving from this point alone will place sharply defined limits on the size distribution. The degree to which this mechanism induces inequality in firms' sizes turns on the question: how will the current size of an already active firm (as measured by the number of investment opportunities it has already take up on different islands) affect the likelihood that it will be the one to take up the next 'investment opportunity'?

3.1. Background: stochastic models of firm growth

The 'growth-of-firms' tradition generated a major literature during the 1950s and '60s which crystallized in the work of Simon and his several co-authors [see Ijiri and Simon (1964, 1977), Sutton (1997b)]. The Simon model provides a useful point of departure in assessing the later literature. It assumes a framework in which the market consists of a sequence of independent opportunities, each of size unity, which arise over time. As each opportunity arises, there is some probability p that it will be taken up by a new entrant. With probability $(1 - p)$ it will be taken up by one of those firms already in the market ('active firms'). The size of any (active) firm is measured by the number of opportunities it has already taken up. There are two assumptions:

(i) Gibrat's Law: the probability that the next opportunity is taken up by any particular active firm is proportional to the current size of the firm.

(ii) Entry: the probability that the next opportunity is taken up by a new entrant is constant over time.

Assumption (ii) is rather arbitrary, though it may be a reasonable empirical approximation. Simon regarded it merely as providing a useful benchmark, and presented various robustness tests showing that 'reasonable' departures from the assumed constancy of p would have only a modest effect on the predictions of the model. The predictions are driven crucially by assumption (i) (Gibrat's Law). What this leads to is a skew distribution of the Yule type, and Simon presented various empirical studies for the U.S. which suggested that it provided a good approximation to the size distribution of large manufacturing firms.

The goodness of fit of the size distribution provides only indirect evidence for Gibrat's Law. A second strand of the literature of the 1950s and '60s focused on the direct investigation of Gibrat's Law, by looking at the relation between firm size and growth over successive years in a panel of firms. While various studies of this kind cast doubt on the idea that proportional growth rates were independent of firm size, no clear alternative characterization emerged.

These questions were re-explored during the 1980s, when researchers obtained access to U.S. census data which allowed them to examine the growth of manufacturing establishments (plants) as a function of both size and age.[45] These new studies suggested a more subtle characterization, which involves two statistical regularities. The first regularity relates to survival rates: it was found that the probability of survival increases with firm (or plant) size, while the proportional rate of growth of a firm (or plant) conditional on survival is decreasing in size. The second regularity relates to age and size: for any given size of firm (or plant), the proportional rate of growth is smaller according as the firm (or plant) is older, but its probability of survival is greater.

These new findings prompted new interest in theoretical models of firm growth. An obvious candidate model was the recently published 'learning' model of Jovanovic

[45] Evans (1987a, 1987b), Hall (1987), Dunne, Roberts and Samuelson (1988).

(1982). In the Jovanovic model, a sequence of firms enters the market. Each firm has some level of 'efficiency' (its unit cost of production), but it does not know what its relative efficiency is prior to entering. Over time, the profits it achieves provide information on its relative efficiency. More efficient firms grow and survive. Less efficient firms 'learn' of their relative inefficiency, and (some) choose to exit.

This model provides a qualitative description of a process of excess entry followed by some exit, and this was the aspect of the model which made it attractive as a vehicle for discussing the new empirical results. As to the size distribution of firms, the model said little: it would depend on inter alia on unobservables, such as the initial distribution of 'efficiency levels'.

Other attempts to model from growth and the size distribution using strategic models led to a similar conclusion: results depended delicately on industry-specific features that would be difficult to control for in cross-industry studies [Hjalmarsson (1974), Selten (1983)].

In parallel with these theoretical developments, new appraisals on the empirical evidence on the size distribution led to a complementary conclusion. In the second volume of this Handbook, Schmalensee (1989) concluded that attempts to fit the data on size distributions for different industries led to the conclusion that 'no one form of distribution fits all industries well'.

It is these findings which motivate the search for a weaker, bounds type, characterization in what follows.

3.2. A bounds approach to the size distribution

The approach introduced in Sutton (1998) proceeds in two steps. The first step, set out in this section, remains within the traditional growth-of-firms framework, and explores the consequences of replacing Gibrat's Law with a weak inequality restriction on the size–growth relationship (labeled the 'provisional hypothesis'). In the second step, set out in the next section, we return to a game-theoretic setting, and show how this restriction arises naturally from a more fundamental 'symmetry principle' within the special context of a market that comprises many (approximately) independent submarkets.

The traditional literature began with the question: how does the size of a firm affect the likelihood that it will be the one to enter the next 'island' market? Rather than offer a direct answer, we take a different approach. Since we aim to find a bound in the space of outcomes, corresponding to the least unequal distribution of firm size, it is natural to ask what would happen if we treated all firms symmetrically, i.e. if we supposed, in particular, that a firm's past history on other islands had no effect on its future prospects. Now this is clearly not a good description of what will happen *on average*, since in many if not most markets the firm that is already operating on other islands will have an advantage, via economies of scale or learning effects, say, over its smaller rivals. To allow for this, however, we introduce our replacement for Gibrat's Law in the form of an inequality constraint.

Gibrat's Law states that if there are two incumbent firms, A and B, whose sizes (as measured by the number of opportunities they have taken up so far) are denoted n_A and n_B, then the probability that firm A (respectively B) takes up the opportunity is proportional to the current size of firm A (respectively B). In what follows, this assumption is replaced by the restriction:

(i)a The provisional hypothesis: the probability that the next market opportunity is filled by any currently active firm is non-decreasing in the size of that firm.

Stated in terms of growth rates, Gibrat's Law assumes that a firm's *proportional* growth rate is independent of its size; the present 'provisional hypothesis' states that a firm's *absolute* growth rate is non-decreasing in firm size.

It is shown in Sutton (1998, ch. 10) that this modified Simon model leads, in the limit where the number of opportunities becomes large, to a size distribution which features a certain minimum degree of inequality in the size distribution of firms. Specifically, it leads to the prediction that the Lorenz curve must lie farther from the diagonal than a limiting 'reference curve', which is defined by the relationship

$$C_k \geqslant \frac{k}{N}\left(1 - \ln \frac{k}{N}\right), \tag{3.1}$$

where C_k is the k-firm concentration ratio (which here represents the fraction of all opportunities taken up by the k largest firms in the industry), and N is the number of firms. (The case of equal sizes would correspond to $C_k = k/N$, and here the Lorenz curve lies on the diagonal.)

This result has two interesting features:

1. The lower bound to concentration is *independent* of Simon's entry parameter p which represents the probability that any opportunity will be taken up by a new entrant. Here, this parameter affects average firm size but not the shape of the size distribution or the associated concentration measures. This contrasts sharply with the traditional literature on the size distribution of firms, which led to a family of size distributions of varying skewness, parameterized by p. Simon's work linked this parameter to empirical estimates of the entry rate of new firms. Other early models also led to a *family* of size distributions; in Hart and Prais (1956), for example, the lognormal distribution's variance could be linked to the variance of the distribution of shocks to firm size between successive periods. The present setup contains no free parameters whose measurement might be subject to error; it leads to a quantitative prediction regarding the lower bound to concentration, conditional only on the assumed constancy of the entry rate (condition (ii)).
2. Various countries publish data on k-firm concentration ratios for several different values of k. The present result implies that the various k-firm ratios are all bounded below by a curve which approximates the above reference curve. In what follows, we take advantage of this in pooling data for various reported k-firm concentration ratios.

One final comment is in order. So far, we have confined attention to a setting in which all opportunities are identical, so that a firm's size can be measured by the number of opportunities that it captures. What if opportunities differ in size?

Suppose that the size of each opportunity is given by an independent random draw from some distribution, and consider two distributions in which the size of a firm is measured (a) by a count of the number of opportunities that it has taken up (i.e. the distribution considered in the preceding discussion), and (b) by the sum of the sizes of the opportunities it has taken up.[46]

It can be shown [Sutton (1998, Appendix 10.4)] that the Lorenz curve associated with distribution (b) lies further from the diagonal than that of distribution (a). In other words, the heterogeneity of opportunities simply adds an additional component to the inequality in the size distribution. This will cause a greater degree of inequality in firm sizes; it will not lead to a violation of the bounds specified by the above formula.

The results described in this section emerge from a re-working of the traditional growth-of-firms literature, in which Gibrat's Law is replaced by a weak inequality constraint on the size-growth relationship (condition (i)a). How does this relate to a game-theoretic analysis? This is the subject of the next section.

3.3. The size distribution: a game-theoretic approach

The bridge from the 'stochastic process' model just explored, to a game-theoretic analysis, rests on the idea of 'markets and submarkets', developed in Section 2.6 above. Here, we focus on the limiting case of a market that contains many independent submarkets, between which there are no linkages, either on the demand side or the supply side. In this setting, all strategic interactions occur *within* submarkets, rather than across submarkets.

The simplest context of this kind is the one used in the growth-of-firms approach, where each submarket consists of a single island market supporting one plant. The results for this case generalize immediately to the setting in which each submarket supports several plants. We begin, then, with the single plant case, and suppose that the island submarkets open up in succession over time. We distinguish between firms that already operate one or more plants ('active firms') and those who do not. The size of an 'active' firm is measured by the number of plants it already operates.

The key idea relates to the analysis of entry to a single island market, where our pool of potential entrants consists of all the currently active firms. Loosely, what we want to explore is the idea that each of these active firms has the same probability of occupying

[46] For the sake of concreteness, we might consider each opportunity to involve an investment of one unit, and to generate a level of sales revenue that was described by a independent random draw from some distribution. We can then interpret distribution (i) as the 'size distribution by assets', and distribution (ii) as the 'size distribution by sales'.

the next island submarket, independently of their histories in other submarkets, and so independently of their sizes.[47]

This idea mirrors the 'provisional hypothesis' of the preceding section, where the minimal degree of size inequality occurred when the probability of taking up the next opportunity was independent of firm size.

In a game-theoretic setting, we can generate the appropriate entry probabilities directly by focusing on symmetric equilibria; here the probabilities emerge naturally as part of the symmetric (mixed strategy) equilibrium. To see how this works, recall the (static) game-theoretic analysis of a single island market large enough to support exactly one firm: if we confine attention to pure strategy equilibria, there are several asymmetric equilibria, in which firm 1 (or 2, or 3) enters while all other firms do not enter. There is also one symmetric equilibrium, in which each firm uses the same mixed strategy (of the form 'enter with probability p, do not enter with probability $(1 - p)$'). Building on this idea, it is straightforward to generate mixed strategy equilibria, in a suitably constructed dynamic entry game, which has the feature that exactly one firm enters, and where the probability of being the entrant is the same for all firms [Sutton (1998, ch. 11)].

The key novelty of the game-theoretic approach is that it allows us to examine situations in which each island submarket can support several firms between which there may be various kinds of strategic interactions. For example, along the equilibrium path of the game, it may be that the first firm to enter 'pre-empts' further entry (either permanently or temporarily). It may be that a sequence of firms enter, each with a different number of products, these products being entered at a series of discrete times [see Sutton (1998, ch. 11) for examples of this kind]. Once a firm has entered its first product, then, it is no longer symmetric with firms that have not yet entered on this island, and it may (and in general will) play a pure strategy in the subgame that follows its entry. But there is one generic feature that must always hold, which relates to the set of firms that have not yet entered. It must be the case, at any decision point at which entry occurs, along the equilibrium path of the game, that each of these firms has the same probability of being selected as the entrant.

This suggests a simple nomenclature: we can think of the potential entrants as taking up 'roles' in the game, and we can name these roles as 'first entrant', 'second entrant' and so on. Along the equilibrium path of the game, the firm filling a particular role ('first entrant' say) may for example enter a higher number of plants than a firm playing a different ('second entrant') role, so that roles differ in size. The key assumption, labeled the 'symmetry' principle, relates to the allocation of roles to new entrants: all potential entrants are treated equally in role assignment.[48]

[47] Since we are looking at a lower bound to concentration we are, as before, abstracting from all other forms of asymmetry between the firms; as noted earlier, any such 'firm specific-effects' tend to lead to a higher level of concentration, and so on in characterizing a lower bound, it is natural to abstract from such effects.

[48] Formally this requirement is expressed in a manner analogous to the Selten–Harsanyi restriction for games with affine subgames: we require that the strategy used by each firm induces the same strategy for each submarket (independently of the history of actions in other submarkets) [Harsanyi and Selten (1988)].

In Sutton (1998, ch. 13), a formal model is presented in which firms enter a series of submarkets, subject to the above principle. It is shown that, irrespective of the nature of (the game played in) each submarket, the limiting form of the size distribution, where a firm's size is measured on the number of roles it occupies, converges to a geometric distribution. Where all the roles are identical, as in the 'single plant per island' setting, then the limiting Lorenz curve satisfies (3.1) above as an equality. Once roles differ in size, however, the Lorenz curve moves further from the diagonal, and (3.1) is satisfied as an inequality.

We have, therefore, two ways of arriving at this limiting Lorenz curve, which derive respectively from (i) the 'provisional hypothesis' of the preceding section, and from (ii) the 'symmetry' principle for markets with independent submarkets. The two lines of attack are complementary, and it is useful in empirical investigations to bear in mind the intuitions underlying each approach. The game-theoretic model pertains only to the context of markets with many (approximately) independent submarkets, and in this context it provides a vehicle for investigating the size distribution both *within* submarkets (where the bounding Lorenz curve does not apply) and for the market as a whole (where it does). This leads to some sharp tests of the theory, as we note below.

One advantage of combining the first ('growth of firms') treatment of this issue with the game-theoretic model, is that it focuses attention on an alternative interpretation of what drives the result on the limiting Lorenz curve. In order to violate the 'inequality' relationship that replaces Gibrat's Law, and so this limiting Lorenz curve, we need to have a setting in which large firms suffer a systematic disadvantage relative to smaller rivals, in respect of their absolute growth rates. Even in markets that do *not* contain 'many independent submarkets', we would not *normally* expect to see a violation of this kind; such violations might be expected only in rather special circumstances (see below).

One final technical remark is in order, regarding the role of 'independence effects' in the above analysis. We have worked here in terms of a setup containing many independent submarkets. Yet it is rare in economics to encounter a market in which the submarkets are strictly independent. Nonetheless, across the general run of 4- or 5-digit SIC industries in the U.S., it is in most cases easy to identify various submarkets (in product space, or in geographic space) which are approximately independent [for a definition of the concept of 'approximate independence', see Sutton (1998), Barbour, Holst and Janson (1992)].[49] It is of considerable relevance in the present context to note that the results developed above do *not* require that the submarkets be independent, but only that they be approximately independent.

[49] A simple illustration may be helpful: let $\ldots, x_{-1}, x_0, x_1, x_2, \ldots$ be independent random variables. Define the set of random variables θ_i as follows: for a fixed $\tau > 1$, let θ_i be a linear combination of $x_{i-\tau}, x_{i-\tau+1}, \ldots, x_{i+\tau}$. Now the θ_i are approximately independent. (Note that θ_1 is not independent of θ_2 as both depend on x_1; but θ_i is independent of all θ_j, where $j \leqslant i - 2\tau$ or $j \geqslant i + 2\tau$.) This example is of particular economic interest in the context of geographically separated submarkets.

3.4. The size distribution: empirical evidence

The empirical predictions may be summarized as follows:

1. A reference curve exists which bounds the Lorenz curve away from the diagonal ((3.1) above). There are two alternative sufficient conditions[50] for this curve to apply:
   (a) The *absolute* growth rate is non-decreasing in firm size;
   (b) The market comprises many (approximately) independent submarkets.
2. The Lorenz curve will lie strictly further from the diagonal than the reference curve, if either of two conditions hold:
   (a) If the *absolute* growth rate is strictly increasing in firm size. (The presence of economies of scale or scope will lead to an effect of this kind.)

   The second condition relates to the 'independent submarkets' model:
   (b) If different roles within submarkets (i.e. 'first mover', 'second mover', etc.) are associated with different sizes of businesses *within* each submarket, then again the Lorenz curve for the market as a whole will lie strictly beyond the reference curve.

In the context of markets containing many submarkets, an additional and more powerful test of the theory can be formulated. This depends on identifying conditions under which a game-theoretic model would predict that, within each individual submarket, the bound defined by the reference curve would be violated (in the sense that the Lorenz curve should lie on, or close to, the diagonal). In Sutton (1998, ch. 2), a sufficient set of conditions for this to hold is developed.[51] In this context we have the twin prediction that:

3. (a) The Lorenz curves for individual submarkets lie close to the diagonal;
   (b) The Lorenz curve for the market as a whole lies at or beyond the reference curve.

The above 'bounds' prediction has been tested using data for manufacturing industries in the U.S. and Germany [Sutton (1998, ch. 13)]. A comparison of the U.S. and German cases is of particular interest in testing a 'bounds' prediction, since it is well known that, among those countries that produce high quality census of manufactures data, the average level of industrial concentration is relatively high in the U.S., and relatively low in Germany. Hence if we pool all observations for (C_k, N), where N denotes

[50] Assuming, following Simon's 'benchmark case' assumption, that the fraction of opportunities filled by new entrants is constant over time; if this fraction is decreasing over time, the bound may be violated (Section 3.1).

[51] Specifically, the results are developed for the class of 'symmetric' product differentiation models (Dixit–Stiglitz model, Linear demand model, etc.). These models can be interpreted as pertaining to markets where the submarkets are small in the sense that all firms' market areas are overlapping. Within this class of models, it is shown that if (a) the products are close substitutes, and (b) the toughness of price competition is low, then irrespective of the form of the entry process, the only form of equilibrium is one where each entrant offers a single product.

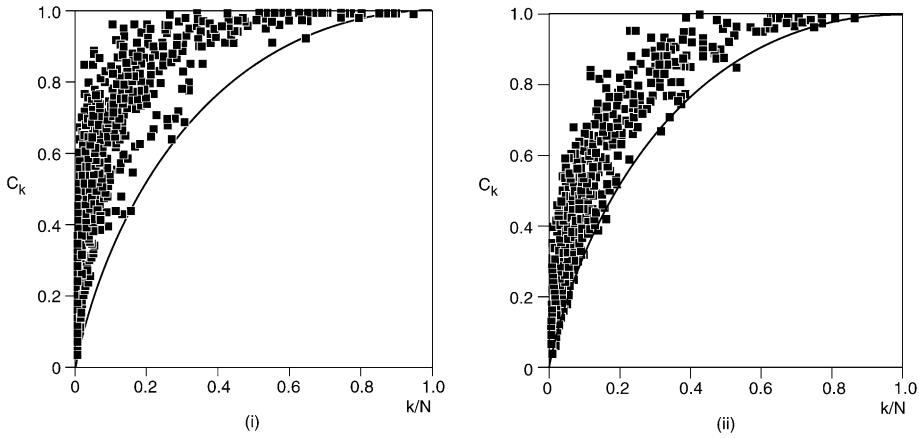

Figure 35.11. Panel (i) shows the scatter diagram of C_k against k/N for pooled data ($k = 4$, 8 and 20) for the United States 1987, at the four-digit level. The Lorenz curve shown on these figures is the reference curve (3.1). Panel (ii) shows data for Germany, 1990 ($k = 3$, 6, 10 and 25).

the number of firms, and C_k is a k-firm concentration ratio, then the resulting 'cloud' of points in (C_k, N) space for the U.S. will lie much farther above the diagonal than will the cloud for German data. Yet if a 'bounds' approach is appropriate, then we should find that the edge of the two clouds should lie on the predicted reference curve above. From Figure 35.11, which shows the data for the U.S. and Germany, it is clear that this is the case.

A second way of testing the prediction is by examining the induced relationship between C_k and C_m, where $m > k$, as described in Sutton (1998, ch. 10, Proposition 10.3). This test has the advantage of relying only on the (relatively easy to measure) concentration ratios, and not on the (much more problematic) count of firm numbers. The test procedure involves substituting m and C_m into the reference curve (3.1) to infer a corresponding value of N; then inserting this value of N, and k, in (3.1) we obtain a lower bound to C_k conditional on C_m, which we label $\underline{C}_k(C_m)$.

Results for this conditional prediction are shown in Figure 35.12. An interesting feature of these results appears when the residuals, $C_4 - \underline{C}_4(C_{50})$, are plotted as a histogram (Figure 35.13). It is clear that the histogram is strong asymmetrical, with a sharp fall at zero, where the bound is reached. Such a pattern of residuals can be seen, from a purely statistical viewpoint, as a 'fingerprint' of the bounds representation, suggesting that on statistical grounds alone, this data would be poorly represented by a conventional 'central tendency' model which predicted the 'center' of the cloud of points, rather than its lower bound.

These predictions on the 'reference curve' bound can be derived from a modified model of the traditional kind, without reference to game-theory, or to the 'independent submarkets' model as we saw above. A more searching test of the 'independent submarkets' model developed above is provided by focusing on a market that comprises

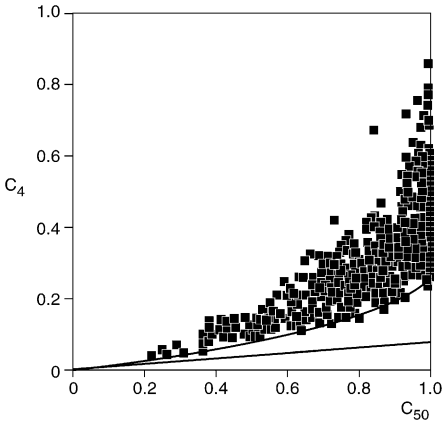

Figure 35.12. An alternative test of the bounds prediction, using data for U.S. manufacturing at the five-digit level, 1977 by reference to a scatter diagram of C_4 versus C_{50}. The solid curve shows the lower bound $C_4(C_{50})$ predicted by the theory. The ray shown below this curve corresponds to the symmetric equilibrium in which all firms are of equal size.

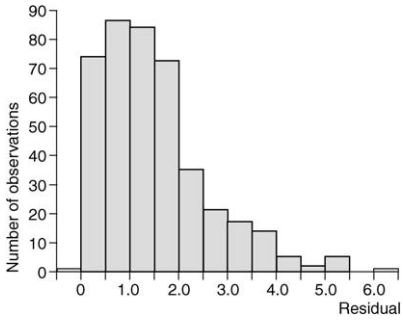

Figure 35.13. A histogram of differences between the actual concentration ratio C_k and the predicted lower bound $C_k(C_{50})$ for the data in Figure 35.12.

many submarkets, and which also satisfies a set of 'special conditions' under which a game-theoretic analysis predicts that the Lorenz curves for individual submarkets must lie close to the diagonal.[52] The U.S. cement market, which satisfies these conditions well, is examined for 1986 in Sutton (1998, ch. 13). It is found that, of 29 states[53] having more than one plant, all but one have Lorenz curves lying closer to the diagonal than the reference curve; yet at the aggregate level, the Lorenz curve for the U.S. as a

[52] See above, footnote 51.

[53] In Sutton (1998), data at state level for the U.S. was chosen as the size of the typical state is of the same order as the typical 'shipping radius' for cement plants.

whole lies almost exactly on the reference curve. These results are consistent with the predictions of the 'independent submarkets' model.

A number of recent studies have re-examined these predicted relations on the size distribution. De Juan (2002, 2003) examines the retail banking industry in Spain, at the level of local (urban area) submarkets (4977 towns), and at the regional level. As in the case of the cement market, described above, conditions in the retail banking industry appear to satisfy the special conditions under which individual submarkets will have Lorenz curves close to the diagonal. A question that arises here is, how large a town can be considered as an (independent) submarket in the sense of the theory? A large city will presumably encompass a number of local submarkets. Rather than decide a priori on an appropriate size criterion for the definition of a submarket, the author sets out to 'let the data decide' on the threshold size of a submarket. With this in mind, she carries out a regression analysis across all towns with population sizes in the range 1000–5000 inhabitants, distinguishing between two sets of explanatory variables: decomposing the number of branches per town into the product of the number of branches per submarket, and the number of submarkets per town, she postulates that the number of branches per submarket depends on population density and the level of per-capita income, while the number of submarkets per town depends (non-linearly) on the population of the town. The results of this regression analysis are used to fix a threshold that determines a sub-set of 'single submarket towns'. For this sub-set of towns, she finds that 96% are 'maximally fragmented' (i.e. the Lorenz curve lies on the diagonal).

The analysis is then extended to larger towns; using a map of branch locations, cluster analysis methods are used to identify alternative 'reasonable' partitionings of the area of the town into 'local submarkets'. Examining one typical medium-size town in this way, she finds that 71% of those local submarkets are maximally fragmented. Finally, the analysis is extended to the level of major regions, each of which comprises many towns; here, the Lorenz curves for each region are found to lie farther from the diagonal than the reference curve.

Buzzacchi and Valletti (2006) examine the motor vehicle insurance industry in Italy. Here, the institutional framework of the industry is such as to produce an administratively determined set of submarkets: each vehicle is registered within one of 103 provinces, and owners are required to buy insurance within their own province; moreover, over nine-tenths of premia are collected through local agents within the province. The authors develop a model of strategic interaction to model competition within submarkets. At the submarket level they find, consistently with their model, that Lorenz curves for the 103 provinces lie almost wholly between the reference curve and the diagonal. At the national level, however, the Lorenz curve lies farther from the diagonal than the reference curve defined by Equation (3.1), as predicted.

In an interesting extension of the analysis, the authors examine, for a set of 13 European countries, the conditional prediction for the lower bound to C_5 as a function of C_{15}; the results show that the cloud of observations lies within, but close to the predicted (conditional) lower bound (as in Figure 35.12).

While all the empirical tests considered so far deal with geographic submarkets, Walsh and Whelan (2001) deal with submarkets in product space: specifically, they look at retail market shares in carbonated soft drinks in Ireland. Within this market, they identify 20 submarkets; and they justify this level of definition of submarkets by reference to an estimation of cross–price elasticities, by reference to an estimated model of demand.[54]

In this market, the special conditions set out above do not apply; and so the expectation under the theory is that the Lorenz curves for submarkets should not stand in any special relationship to the reference curve; in fact, these submarket level Lorenz curves lie in a widely dispersed cloud that extends from the diagonal to well beyond the reference curve. Here, the authors make a important distinction, relative to the theory, in distinguishing between the Lorenz curve based on a count of roles, versus the curve based on sales data. The latter incorporates, as noted earlier, an additional variance component associated with the dispersion in role size, and is expected to lie farther from the diagonal than the reference curve. One unusual feature of this study is that the authors can follow year-to-year fluctuations in role size. It is shown that

(i) the Lorenz curve based on role size is stable from year to year and is very close to the reference curve defined by (3.1);
(ii) the Lorenz curve based on sales data lies farther from the diagonal, and is relatively volatile over time. This finding is consistent with prediction 2(b) above.

In two recent papers, Ellickson (2006, 2007) presents a wide-ranging analysis of the U.S. supermarket industry. The point of departure of the analysis lies in his finding that supermarket concentration in local markets (defined by reference to their range of distribution to stores) exhibits the 'non-convergence' property. The author considers several alternative models to explain observed structure (horizontal product differentiation, capacity competition, product proliferation models, etc.), and rejects each of these in favor on an 'endogenous sunk cost' model. He conjectures that the important element in these costs relates to the development of more efficient (chain-level distribution systems based in part of information technology).

Ellickson then goes on to distinguish two groups of firms in the industry: those (large firms) which compete in such investments at the chain-level, and the 'fringe' of (smaller) firms that do not. He argues that the latter group of firms should be well represented by the 'independent submarkets' model, but the former group should not (as their firm-level investments create economies of scope across local submarkets). He then examines the Lorenz curves for each group, while using the characteristics of the empirical Lorenz curves to determine the dividing line between the two groups. It is found that the dividing line corresponds well to the distinction between firms that operate distribution networks organized at the national chain level, and those which do not.

[54] One caveat is in order here, insofar as many of the firms involved here are foreign firms, and so it is less easy to imagine the entry process in terms of the model set out above.

4. Dynamics of market structure

The two literatures considered up to this point have been concerned with explaining cross-sectional regularities of two different kinds. Each one leads to a constraint on the pattern of outcomes seen in cross-industry data sets. These constraints can be meshed together in a straightforward way, to provide a unified analysis of cross-industry patterns [Sutton (1998, ch. 14)].

In this section, we turn to questions of dynamics. Here, the link between theory and evidence is much less tight. This reflects the fact that the problems posed by unobservables (such as the beliefs of firms and the way this impinges on entry decisions, etc.) pose much more serious problems, and it is notoriously difficult to arrive at robust theoretical results that place testable restrictions on the data.

4.1. Dynamic games

To what extent can the results of the multi-stage game models developed in Section 2 above be carried over to a 'dynamic games' framework, in which firms spend on fixed outlays in each successive period, while enjoying a flow of returns in each period that reflects the current quality and cost levels of the firm and its rivals? This is a difficult question, which has been explored from two angles. In Sutton (1998, ch. 13), a dynamic game is set out in which an exogenously fixed 'imitation lag' T is imposed, in the sense that if a firm raises its R&D spending at time t, then the resulting quality jump occurs only at time $t + T$; rivals do not observe the expenditure at time t, but do observe the quality jump at time $t + T$. The point of this model is to expose the nature of the implicit assumption that is made in choosing a multi-stage framework: the results of such a framework can be mimicked in the dynamic game set up by letting the lag T become large. In other words, the multi-stage game framework implicitly excludes certain kinds of equilibria that may arise in dynamic games. It is therefore of particular interest to ask: to what extent might ('new') forms of equilibria appear in an unrestricted dynamic game, that could undermine the non-convergence results developed above? The issue of interest turns on the appearance of 'underinvestment equilibria', as described in Sutton (1998). Here the idea is that firms 'underspend' on R&D, because rivals strategies prescribe that any rise in R&D spending by firm 1 at time t will result in a rise by its rivals at period $t + 1$. It is shown in Nocke (2006) that this kind of equilibrium can indeed appear in a suitably specified dynamic game in which firms can react arbitrarily quickly to rivals' actions. This leads to a reduction in the lower bound to concentration relative to the equivalent multi-stage game model; but the 'non-convergence theorem' developed above in the multi-stage game setting continues to hold good in the dynamic game.

A different approach to modeling industry equilibrium as a dynamic game has been introduced by Ariel Pakes and his several coauthors [see in particular Ericson and Pakes (1995)], which provides inter alia an alternative vehicle within which to explore the evolution of structure in a dynamic setting. The examples based on this approach which have appeared in the published literature to date employ profit functions that, in the

language of the present review, have a value of alpha equal to zero. In some recent work, however, Hole (1997) has introduced the simple 'Cournot model with quality' example of Section 2.2 above into the Pakes–Ericson framework. The results show, in the context of a 3-firm example, as the parameter β falls (so that alpha rises), the outcomes cluster in a region in which two firms account for an arbitrarily high fraction of sales. These results provide an analog of the 'non-convergence theorem' in a stochastic, dynamic setting. (For a discussion of the approach, see the contribution by Ariel Pakes and Uli Doraszelski to this volume.)

4.2. Learning-by-doing models and network effects

Two special mechanisms often cited in the literature as leading to a concentrated market structure are the learning-by-doing mechanism, and the network effects mechanism. [See, for example, Spence (1981), Fudenberg and Tirole (1983, 1985), Cabral and Riordan (1994), Gruber (1992, 1994).] It is shown in Sutton (1998, chs. 14 and 15) that these two mechanisms can be represented in simple 2-stage game models that are identical in structure ('isomorphic'). The idea is that the firms plays the same 'Cournot' game in each period, but each firm's second period cost function (in the learning-by-doing model) or the perceived quality of its product (in the network effects model) is affected by its level of output (sales) in the first period. This induces a linkage between the profit earned in the first period, and in the second. This effect is precisely analogous to the 'escalation mechanism' described above; we can think of the profit foregone in period 1 (as the firm raises its output beyond the level that maximizes first period profit) as an 'opportunity cost' analogous to the fixed cost $F(u)$ in the 'quality competition' model of Section 2.2 above. [This analogy is spelt out precisely in Sutton (1998, ch. 14, footnote 1, p. 351).] Beyond this simple 2-stage game characterization, however, the analysis of these games becomes more complex. The contribution of Cabral and Riordan (1994) is noteworthy, in that it gives a full characterization of the dynamics of the process in a learning-by-doing context.[55]

A central theme in the associated literature relates to the idea that small changes in initial conditions may have large effects on outcomes. ('History dependence.') This theme has been extensively explored by David (1975), David and Bunn (1990) and Arthur (1989). From an analytical viewpoint, this phenomenon can be seen as pertaining to a wide rate of games featuring an 'escalation' mechanism of the kind explored in Section 2.[56] A crisp theoretical characterization of this idea comes from the dynamics of patent–race games, of the form developed by Harris and Vickers (1985, 1987).

[55] Albeit at the cost of working in a setting in which a single indivisible unit is sold in each period, and attention is confined to symmetric equilibria.

[56] For some cases of 'history dependence' in this setting, see Sutton (1991, ch. 9).

4.3. Shakeouts

One of the most striking features of industry dynamics is the occurrence of 'shakeouts', a phenomenon documented and characterized in considerable depth by Klepper and his several co-authors [Klepper and Graddy (1990), Klepper and Simons (2005)]; as a new industry develops, the number of producers tends first to rise to a peak and later falls to some lower level. The extent and timing of this 'shakeout' varies widely across product markets. In some cases, it comes early in the life of the product, and is very sharp. In others, it is relatively muted, or does not occur at all. For example, the market for lasers, which is characterized by a large number of submarkets corresponding to lasers designed for different applications, shows no 'shakeout'; rather, the number of firms rises steadily over time. In the terminology of Section 2 above, this is a low-alpha industry. By contrast, the early history of the 'high-alpha' aircraft industry was marked by a very sharp shakeout [Sutton (1998, ch. 15)]. It seems the 'shakeout' process can plausibly be seen as part of a dynamic adjustment process associated with the evolution of concentration, and models of 'shakeout effects' can by seen as dynamic counterpoints of the static models of the 'escalation effect' of Section 2 above.

Two types of models have been postulated for shakeouts. The first is due to Jovanovic and MacDonald (1994) who begin by stating that Klepper's data on shakeout cannot be accounted for by appealing to the 'learning' model of Jovanovic (1982). Instead, the authors postulate a model in which early entrants employ a common technology which after some time is superseded by a new technology. The new technology offers low unit costs, but at a higher level of output per firm (scale economies). The transition to the new technology involves a shakeout of first generation firms, and the survival of a smaller number of firms who now employ the new large-scale technology. By calibrating the model against the data for the U.S. tire industry, the authors can simulate successfully the number of firms, and the movement of stock prices over time.[57]

The model of Klepper (1996) combines a stochastic growth process for firms, who enter by developing some new variant ('product innovation'), with the idea that each firm may spend some fixed costs to lower its unit cost of production ('process innovation'). Assuming some inertia in sales, and some imperfection in capital markets, those firms whose current sales are larger find it profitable to devote more fixed costs to process innovation (because the fixed costs incurred are spread over a larger volume of sales). As the larger firms cut their unit production costs, some smaller firms are no longer viable, and these exit, generating the 'shakeout'.

4.4. Turbulence

A striking feature of industry dynamics is that, across different industries, there is a positive correlation between gross entry rates, and gross exit rates, i.e. the 'churning' of

[57] A recent model which focuses on explaining the temporal pattern of hazard rates for exit in this industry is that of Tong (2000).

the population of firms is greater in some industries than others. However, most of this entry and exit has little effect on the largest firms in the industry.[58]

Within any one country, quite a strong correlation usually exists between entry and exit rates by industry. Geroski (1991), for example, reports a correlation coefficient of 0.796 for a sample of 95 industries in the UK in 1987. The most comprehensive data on this issue comes from a compilation of country studies edited by Geroski and Schwalbach (1991). The cross-country comparisons afforded by this study indicate that there is at least a weak correspondence between the ranking of industries by turbulence in different countries. This is important in that it suggests that there may be some systematic, industry-specific, determinants of turbulence levels.

These results have prompted interest in the determinants of turbulence (defined conventionally in this literature as the sum of gross entry and gross exit rates) across different industries. At least four types of influence are likely to be involved:

(a) Underlying fluctuations in the pattern of demand across product varieties or plant locations.
(b) The displacement of existing technologies (modes of production) by alternatives.
(c) The displacement of existing products by new and superior substitutes.
(d) Fluctuations in relative efficiency (productivity) levels across firms.

Of these, the first factor may be of primary importance, but while it is easy to model, it is very difficult to measure or control for empirically. The second and third factors pose some interesting questions in terms of modeling. Some new models have been developed recently, but these have not yet led to empirically tested claims regarding the influence of industry characteristics on the degree of turbulence. The effect of a displacement of production technologies has been modeled by Lambson (1992), who considers an industry facing exogenous shocks to relative factor prices, which occur at infrequent intervals. Firms incur sunk costs in building a plant using a given technology, and when factor prices change, an entrant – knowing that factor prices shift rarely – may find it profitable to enter the industry and displace incumbents. In this kind of model, the level of sunk costs incurred by firms will influence entry and exit rates, conditional on the volatility of industry demand.

The third factor listed above relates to the idea that (some) exit may be induced by entry, as new and superior product varieties displace existing products. This is a basic idea discussed in the vertical product differentiation literature. The key theoretical question is why the old varieties cannot continue to retain a positive market share at some price, given that their costs of product development are sunk. Such varieties would indeed continue to survive in a 'horizontal' product differentiation model, but this is not generally true in a 'vertical' product differentiation model [Gabszewicz and Thisse (1980), Shaked and Sutton (1982)].

[58] The volatility of market shares among large firms has been less widely studies. An important early study was that for Caves and Porter (1978), which used the PIMS data-set. See also the studies by Steven Davies (1991) and David and Haltiwanger (1992).

The mechanism that has been explored most fully in the literature is that involving shocks to the relative efficiency (productivity) levels across firms. This is a central feature of the "passive learning" version of the Ericson–Pakes model discussed above. It is also the mechanism underlying the model of Hopenhayn (1992); a recent extension of the Hopenhayn model has been used by Asplund and Nocke (2006) to examine, both theoretically and empirically, the way in which changes in market size affects the rate of firm turnover.

5. Caveats and controversies

5.1. Endogenous sunk costs: a caveat

One procedure that has become common in the literature, following Sutton (1991), is to treat industries as falling into two discrete groups, those in which advertising and R&D are unimportant, and those in which they play a substantial role. Schmalensee (1992), in reviewing Sutton (1991), referred to these as type I and II industries, respectively.

In tandem with this nomenclature, it has become common to identify these two groups of industries as being represented by the 'exogenous sunk cost' model and the 'endogenous sunk cost model,' respectively. This leads to some confusion, since it begs the question: are not all sunk costs endogenous? (A firm can decide, for example, on its level of plant capacity, or its number of manufacturing plants.) While it is helpful in empirical testing to split the sample into two groups, it is worth noting that the underlying theoretical model is one of 'endogenous sunk costs'; and that the 'exogenous sunk cost model' is just a simplified representation of a special limiting case of the endogenous sunk cost model, corresponding to the limit $\beta \to \infty$ as noted in the text [Shaked and Sutton (1987)]. What matters to the level of concentration is not the 'endogeneity of sunk costs', but the value of α, which may be zero *either* because β is high, *or* because σ is low.

5.2. Can 'increasing returns' explain concentration?

The appeal of 'increasing returns' as a 'general' explanation for observed levels of market concentration is highly problematic, since different authors use this term in different ways. At one extreme, the term is used in its classic sense, to refer to the idea that the average cost curve is downward sloping. This feature holds good in all the models described above, including those cases where the lower bound to concentration falls to zero in large markets. It follows that an appeal to 'increasing returns' in this sense does not provide an explanation for high levels of concentration in large markets.

Another sense in which the term is used arises in empirical work, where it is said to be important to discover whether there are increasing returns to R&D. The implication behind this concern seems to be that the presence or absence of increasing returns could

carry implications for market structure, with increasing returns being linked to high concentration.

What does "increasing returns" mean in this context? This is rarely spelled out, but what is often measured is a technical relation between R&D spending and some output measure, such as a count of patents. In terms of the present theory, diminishing returns in this sense are consistent with any value of alpha – and indeed, in the examples used above, we used a diminishing returns form for the function linking R&D spending to product quality.

Another sense in which we might interpret "increasing returns to R&D" would be to look, in the spirit of the present theory, at the relation between a firm's R&D spending and the gross profit it earns as a result of that spending. Within the theory, this relation can take either a diminishing returns form, or an S-shaped form (first increasing, then diminishing). Either shape is consistent with both low and high values of alpha. It would seem, then, that looking to increasing returns as an explanation for high concentration levels is not a helpful way forward.

5.3. *Fixed costs versus sunk costs*

It has been suggested that many of the features of the models described above should carry over to a setting in which costs are fixed but not sunk [Schmalensee (1992), Davies and Lyons (1996)]. It is not clear that any *general* claim of this kind can be supported by reference to a formal analysis, so long as we identify the '2-stage' (or multistage) game framework with the 'sunk cost' interpretation. [For a discussion of this point see Sutton (1991), and for a response, see Schmalensee (1992).] If costs are fixed but not sunk, it seems appropriate to model firms' actions by reference to a 1-shot game in which firms take simultaneous decisions on entry and prices. This captures the notion introduced in the Contestability literature by Baumol, Panzar and Willig (1982). It is crucial to results of this literature that sunk costs be *exactly* zero. The Bertrand example of Section 2 illustrates how an arbitrarily small departure from this assumption can change the qualitative features of equilibrium outcomes.[59] In practice, it seems to be extremely difficult to find any industry in which sunk costs are zero; for a recent attempt to quantify the extent to which fixed outlays are sunk, see Asplund (2000).

6. Unanswered questions and current research

In the light of the preceding discussion, there are four areas that are worth noting as being potentially fruitful areas for future research:

[59] If the sunk cost of entry is exactly zero in the Bertrand example, then any number $n \geqslant 2$ of firms will enter, and price will coincide with marginal cost.

Bounds and 'single industry studies' It was noted above that there is a deep complementarity between the bounds approach, and the single industry studies (or 'structural estimation') approach. The first aims at a low level characterization of some mechanisms that operate in a more or less uniform way across a wide range of industries. The latter approach focuses on 'model selection', its aim being to arrive at a richly specified model that captures various industry-specific factors. Building a bridge between the two levels offers some interesting challenges. This can in principle be worked upon from either end: by adding structure to a 'bounds model' or by uncovering, through the accumulation of evidence from different industries, some new candidate generalizations. One strand of current research in this area involves the study of 'limits to monopolization'. This line of inquiry is motivated by the empirical observation that we rarely see industries in which a single firm has a market share close to unity in large markets; the gradual 'fade-out' of the scatter diagrams in Figure 35.7 above as we move to the top right of the diagram illustrates this point. Research to date has been limited, but suggests that the mechanisms involved here are of a relatively delicate (industry-specific) kind [Nocke (2000), Vasconcelos (2002)].

The variance of growth rates While the growth-of-firms literature has focused considerable attention to the relation between the size of a firm and its *expected* growth rate, it was not until quite recently that attention was directed to the dispersion (variance) of firms' growth rates, and the way this varied across different size classes. A seminal paper appeared in *Nature* in 1996 which drew attention to a striking empirical regularity, in the form of a simple 'power law' relationship [Stanley et al. (1996)]. The interpretation of these results remains controversial, and deserves further scrutiny [for one candidate explanation, see Sutton (2001c), and for a dissenting view, Wyart and Bouchard (2002)]. A second empirical regularity reported in Stanley et al. (1996) relates to the shape of the distribution of proportional growth rates. The authors reported, for the Compustat data-set on U.S. corporations, a distribution of the double-exponential type. A recent contribution by Fu et al. (2005) combines the notion of 'independent submarkets' with Gibrat's Law to develop a candidate explanation for the distribution of firm growth rates.[60] It is perhaps because virtually all papers on these topics have appeared in physics, rather than economics journals, that this important strand in the recent literature has received less attention in the IO literature than it merits.[61]

[60] Firms are represented as a collection of business units operating in different sub-markets; and Gibrat's Law is applied both to the sales of each individual business, and to the firm's introduction of new businesses. The pharmaceutical industry offers a unique context for testing such a model, since (approximately) independent business units can be identified with different 'therapeutic groups' within the industry [Sutton (1998)]. The authors show that the model leads to a form of distribution of firm growth rates that is double-exponential close to the origin, but has power-law ('fat') tails; and this predicted form fits the data very closely.

[61] Another, less justifiable, reason for this lack of impact may be that the best candidate models are 'statistical', rather than ones based on (profit maximizing) firm behavior. But there is no reason why some economically interesting relationships should not derive from primitive and robust features of markets, in-

The size distribution revisited Notwithstanding the fact that the size distribution of firms varies widely across industries, there is continuing interest in characterizing the shape of the aggregate distribution of firm size for the economy as a whole. Axtell (2001) uses comprehensive data from the U.S. Census to show that the aggregate distribution conforms well over its entire range to a Pareto (i.e. power law or scaling) distribution with an exponent slightly above unity. While, as Axtell notes, there are various stochastic growth processes based on Gibrat's Law that converge to a distribution of this form, a deeper challenge lies in reconciling the apparent uniformity of the aggregate distribution with the fact that quite different patterns are observed within different industries.[62]

Market dynamics I: turbulence In this area, our knowledge remains quite limited. The key empirical finding is that the ranking of industries by the degree of (entry–exit) turbulence is broadly similar across countries. This strongly suggests that there are industry-specific factors at work in molding this pattern; the elucidation of the factors driving this pattern is one of the most intriguing challenges for future research.

Market dynamics II: market shares and market leadership A second aspect of market dynamics relates to fluctuations over time in the pattern of market shares within an industry. While a considerable literature has been devoted to developing stochastic models of market share dynamics, the main challenge lies in uncovering statistical regularities that can provide a focus for the interplay of theory and evidence in this area. One unresolved debate of long standing relates to the 'persistence of leadership' question: to what extent should we expect a market share leader to retain the leadership position over time? To what degree does leadership persist in practice? On what factors does the persistence of leadership depend? For a review of these issues in the context of an empirical investigation, see Sutton (2007).

Mergers One topic that has not been covered explicitly in this chapter is 'mergers and concentration'.[63] The reason for this is that robust results of the kind emphasized here remain elusive in regard to the motives and mechanisms underlying merger activity. This longstanding area of investigation continues to pose challenges in respect of the

dependently of whether or not firms are profit maximizers. Nonetheless, it is all the more interesting, against this background, to probe the status of Gibrat's Law, and alternative postulates of this kind, relative to models of profit maximizing firms (see footnote 64 below).

[62] Against this background, a recent contribution by Cabral and Mata (2003) is of particular interest. Using data for the Portuguese manufacturing sector, they explore the way in which the size distribution evolves over time within various industries. A further strand in the recent literature relates to the use of new data sources to examine the size distribution of firms in developing countries; see, for example, Van Biesenbroeck (2005).

[63] Mergers enter the picture developed above in two ways: (a) as one of adjustment mechanisms driving a rise in concentration when exogenous influences shift the lower bounds upwards, and (b) as a factor leading to outcomes 'inside the bound' (see the discussion relating to Figures 35.2 and 35.3 above).

characterization of mechanisms that operate in a systematic way both over time and across the general run of industries.

Capabilities A new strand in the literature seeks to relate the market structure literature to the notion of firms' 'capabilities' [Nelson and Winter (1982)]. It was noted in Section 2.2 that we can think of a firm's capability as being represented, in one sense, by its levels of productivity and product quality in each market in which it operates. More fundamentally, the term 'capability' relates to the set of 'shared know-how' embodied in a set of individuals within the firm, from which these levels of productivity and quality derive. In the language of the present chapter, this raises the challenge of opening the 'black box' represented by the fixed cost schedule $F(\cdot)$ that maps a firm's quality and productivity levels into its fixed (R&D) outlays. One payoff from moving to this deeper level of analysis is that we might arrive at a better understanding of the problems of 'markets dynamics' discussed above.[64] A firm's growth and survival, in a world in which demand and supply conditions fluctuate across the several markets in which it may operate, will depend not only on its observed levels of productivity and quality in the sub-markets or product groups in which it currently operates, but on the underlying know-how that will determine its levels of productivity and quality in other (new) product groups to which it may move. Developing a satisfactory theoretical and empirical analysis of these issues would seem a natural next step relative to the current literature.

Acknowledgements

I would like to thank Volker Nocke, Rob Porter, Michael Raith, and Tomasso Valletti for their extremely helpful comments on a preliminary draft.

Appendix A: The Cournot example

The profit of firm i in the second stage subgame is

$$(p - c)x_i = \left(S \Big/ \sum x_j - c\right)x_i. \tag{A.1}$$

Differentiating this expression w.r.t. x_i we obtain the first-order condition,

$$-\frac{S}{(\sum x_j)^2} \cdot x_i + \frac{S}{\sum x_j} - c = 0. \tag{A.2}$$

[64] An interesting implication of a 'capabilities' view is that it suggests a robust and natural naturalization of (a weak form of) Gibrat's Law as an outcome of profit maximization: for if the firm's depth and breadth of know-how is both a driver of its current range of activities, and of its comparative advantage across the range of new market opportunities that arise over time, then the expansion of activities taken in equilibrium by profit maximizing firms may show the rough proportionality to their current sizes that is observed in practice.

Summing Equation (A.2) over i, and writing $\sum x_j$ as X, we obtain

$$\sum x_j \equiv X = \frac{S}{c}\frac{N-1}{N}. \tag{A.3}$$

It follows from (A.2), (A.3) that all the x_i are equal, whence $x_i = X/N$, whence

$$x_i = \frac{S}{c}\frac{N-1}{N^2} \quad \text{and} \quad p = c\left\{1 + \frac{1}{N-1}\right\} \quad \text{for } N \geqslant 2. \tag{A.4}$$

Substituting (A.4) into (A.1) and rearranging, it follows that the profit of firm at equilibrium equals S/N^2.

Appendix B: The Cournot model with quality

The profit function may be derived as follows. The profit of firm i is

$$S\pi_i = p_i x_i - cx_i = \lambda u_i x_i - cx_i, \tag{B.1}$$

where

$$\lambda = S\Big/\left(\sum_j u_j x_j\right). \tag{B.2}$$

To ease notation it is useful to express the first-order condition in terms of λ. With this in mind, note that

$$\frac{\mathrm{d}\lambda}{\mathrm{d}x_i} = -\frac{S}{(\sum_j u_j x_j)^2}\frac{\mathrm{d}}{\mathrm{d}x_i}\left(\sum_j u_j x_j\right) = -\frac{Su_i}{(\sum_j u_j x_j)^2} = -\frac{u_i}{S}\lambda^2. \tag{B.3}$$

Now the first-order condition is obtained by differentiating (B.1), viz.

$$\frac{\mathrm{d}\pi_i}{\mathrm{d}x_i} = \lambda u_i + u_i x_i \frac{\mathrm{d}\lambda}{\mathrm{d}x_i} - c = 0.$$

On substituting for $\frac{\mathrm{d}\lambda}{\mathrm{d}x_i}$, from (B.2) and (B.3), and rearranging, this becomes

$$u_i x_i = \frac{S}{\lambda} - \frac{cS}{\lambda^2}\frac{1}{u_i}. \tag{B.4}$$

Summing over all products, we have,

$$\sum_j u_j x_j = \frac{NS}{\lambda} - \frac{cS}{\lambda^2}\sum_j (1/u_j).$$

But from (B.2) we have $\lambda = S/(\sum_j u_j x_j)$ whence $\sum_j u_j x_j = S/\lambda$ so that

$$\frac{S}{\lambda} = \frac{NS}{\lambda} - \frac{cS}{\lambda^2}\sum_j (1/u_j)$$

whence

$$\lambda = \frac{c}{N-1}\sum_j (1/u_j). \tag{B.5}$$

Substituting this expression for λ into (B.4) we have on rearranging that

$$x_i = \frac{S}{c}\cdot\frac{N-1}{u_i\sum_j(1/u_j)}\left\{1-\frac{N-1}{u_i\sum_j(1/u_j)}\right\}. \tag{B.6}$$

Setting the expression in brackets equal to zero leads to a necessary and sufficient condition for good i to have positive sales at equilibrium, as described in the text. By ranking firms in decreasing order of quality, and considering successive subsets of the top $1, 2, 3, \ldots,$ firms, we can apply this criterion to identify the set of products that command positive sales at equilibrium. Denoting this number by N henceforward, we can now solve for prices, using $p_i = \lambda u_i$, whence from (B.5) we have

$$p_i - c = \left\{\frac{u_i}{N-1}\sum_j(1/u_j) - 1\right\}c. \tag{B.7}$$

Inserting (B.6) and (B.7) into the profit function

$$\pi_i = (p_i - c)x_i$$

and simplifying, we obtain

$$\pi_i = \left\{1 - \frac{N-1}{u_i}\frac{1}{\sum_j(1/u_j)}\right\}^2 S. \tag{B.8}$$

References

Aghion, P., Bloom, N., Blundell, R., Griffith, R., Howitt, P. (2005). "Competition and innovation: An inverted-U relationship". Quarterly Journal of Economics 120, 701–728.

Arthur, W.B. (1989). "Competing technologies, increasing returns, and lock-in by historical events". Economic Journal 99, 116–131.

Asplund, M. (2000). "What fraction of a capital investment is sunk costs?". Journal of Industrial Economics 48, 287–304.

Asplund, M., Nocke, V. (2006). "Firm turnover in imperfectly competitive markets". Review of Economic Studies 73, 295–327.

Axtell, R. (2001). "Zipf distribution of U.S. firm sizes". Science 293, 1818–1820.

Bain, J.S. (1956). Barriers to New Competition. Harvard Univ. Press, Cambridge, MA.

Bakker, G. (2005). "The decline and fall of the European film industry: Sunk costs, market size and market structure, 1890–1927". Economic History Review 58, 310–351.

Barbour, A.D., Holst, L., Janson, S. (1992). Poisson Approximation. Clarendon Press, Oxford.

Baumol, W.J., Panzar, J.C., Willig, R.D. (1982). Contestable Markets and the Theory of Industry Structure. Harcourt Brace Jovanovich, San Diego.

Berry, S. (1992). "Estimation of a model of entry in the airline industry". Econometrica 60, 889–917.

Bresnahan, T.F. (1992). "Sutton's sunk costs and market structure: Price competition, advertising and the evolution of concentration". RAND Journal of Economics 23, 137–152.
Bresnahan, T.F., Greenstein, S. (1999). "Technological competition and the structure of the computer industry". Journal of Industrial Economics 47, 1–40.
Bresnahan, T.F., Reiss, P.C. (1990a). "Entry in monopoly markets". Review of Economics Studies 57, 531–553.
Bresnahan, T.F., Reiss, P.C. (1990b). "Do entry conditions vary across markets?". Brookings Paper on Economic Activity 3, 833–881.
Buzzacchi, L., Valletti, T. (2006). "Firm size distribution: Testing the 'independent submarkets model' in the Italian motor insurance industry". International Journal of Industrial Organization 24, 809–834.
Cabral, L., Riordan, M. (1994). "The learning curve, market dominance and predatory pricing". Econometrica 62, 1115–1140.
Cabral, L.M.B., Mata, J. (2003). "On the evolution of firm size distribution: Facts and theory". American Economic Review 93, 1075–1089.
Campbell, J.R., Hopenhayn, H.A. (2005). "Market size matters". Journal of Industrial Economics 53, 101–122.
Carroll, R., Hannan, M.T. (2000). The Demography of Corporations and Industries. Princeton Univ. Press, Princeton.
Caves, R.E. (1986). "Information structures of product markets". Economic Inquiry 24, 195–212.
Caves, R.E., Porter, M.E. (1978). "Market structure, oligopoly and the stability of market shares". Journal of Industrial Economics 26, 289–313.
Cohen, W.M., Levin, R.C. (1989). "Innovation and market structure". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 1059–1107.
Dasgupta, P., Stiglitz, J. (1980). "Industrial structure and the nature of innovative activity". Economic Journal 90, 266–293.
David, P. (1975). "Clio and the economics of QWERTY". American Economic Review Proceedings 75, 332–337.
David, P.A., Bunn, J.A. (1990). "Gateway technologies and the evolutionary dynamics of network industries: Lessons from electricity supply history". In: Heertje, A., Perlman, M. (Eds.), Evolutionary Technology and Market Structure: Studies in Schumpeterian Economics. University of Michigan Press, Ann Arbor, pp. 121–156.
Davies, S.W. (1991). "The dynamics of market leadership in UK manufacturing industry, 1976–86". Report Series. London Business School, Centre for Business Strategy.
Davies, S.W., Lyons, B.R. (1996). "Industrial Organization in the European Union: Structure, Strategy and the Competitive Mechanism". Oxford Univ. Press, Oxford.
Davis, S.J., Haltiwanger, J. (1992). "Gross job creation, gross job destruction and employment reallocation". Quarterly Journal of Economics 57, 819–863.
De Juan, R. (2002). "Entry in independent submarkets: An application to the Spanish retail banking market". Economic and Social Review 33, 109–118.
De Juan, R. (2003). "The independent submarkets model: An application to the Spanish retail banking market". International Journal of Industrial Organization 21, 1461–1487.
Deneckere, R., Davidson, C. (1985). "Incentives to form coalitions with Bertrand competition". RAND Journal of Economics 16, 473–486.
Dixit, A.K., Stiglitz, J.E. (1977). "Monopolistic competition and optimum product diversity". American Economic Review 67, 297–308.
Dunne, T., Roberts, M.J., Samuelson, L. (1988). "Patterns of firm entry and exit in U.S. manufacturing industries". RAND Journal of Economics 19 (4), 495–515.
Eaton, B.C., Ware, R. (1987). "A theory of market structure". RAND Journal of Economics 18, 1–16.
Edwards, B.K., Starr, R.M. (1987). "A note on indivisibilities, specialization and economies of scale". American Economic Review 93, 1425–1436.
Ellickson, P. (2006). "Quality competition in retailing: A structural analysis". International Journal of Industrial Organization 24, 521–540.

Ellickson, P. (2007). "Does Sutton apply to supermarkets?". RAND Journal of Economics. In press.
Ericson, R., Pakes, A. (1995). "Markov-perfect industry dynamics: A framework for industry dynamics". Review of Economic Studies 62, 53–82.
Evans, D.S. (1987a). "The relationship between firm growth, size and age: Estimates for 100 manufacturing industries". Journal of Industrial Economics 35 (4), 567–581.
Evans, D.S. (1987b). "Tests of alternative theories of firm growth". Journal of Political Economy 95 (4), 657–674.
Fisher, F.M. (1989). "Games economists play: A noncooperative view". RAND Journal of Economics 20, 113–124.
Fu, D., Pammolli, F., Buldyrev, S.V., Riccaboni, M., Matia, K., Yamasaki, K., Stanley, H.E. (2005). "The growth of business firms: Theoretical framework and empirical evidence". Proceedings of the National Academy of Sciences of the United States of America 102, 18801–18806.
Fudenberg, D., Tirole, J. (1983). "Learning-by-doing and market performance". Bell Journal of Economics 14, 522–530.
Fudenberg, D., Tirole, J. (1985). "Preemption and rent equalization in the adoption of new technology". Review of Economic Studies 52, 383–402.
Gabszewicz, J.J., Thisse, J.F. (1980). "Entry (and exit) in a differentiated industry". Journal of Economic Theory 22, 327–338.
Geroski, P. (1991). Market Dynamics and Entry. Basil Blackwell, Oxford.
Geroski, P., Schwalbach, J. (1991). Entry and Market Contestability. Basil Blackwell, Oxford.
Gibrat, R. (1931). "Les inégalités économiques ; applications : aux inégalités des richesses, à la concentration des enterprises, aux populations, des villes, aux statistiques des familles, etc., d' une loi nouvelle, la loi de l' effet proportionnel". Librairie du Recueil Sirey, Paris.
Giorgetti, M.L. (2003). "Quantile regression in lower bound estimation". Journal of Industrial Economics 51, 113–120.
Gruber, H. (1992). "Persistence of leadership in product innovation". Journal of Industrial Economics 40, 359–375.
Gruber, H. (1994). Learning and Strategic Product Innovation: Theory and Evidence from the Semiconductor Industry. North-Holland, London.
Hall, B. (1987). "The relationship between firm size and firm growth in the U.S. manufacturing sector". Journal of Industrial Economics 35 (4), 583–606.
Harris, C., Vickers, J. (1985). "Perfect equilibrium in a model of a race". Review of Economic Studies 52, 193–209.
Harris, C., Vickers, J. (1987). "Racing with uncertainty". Review of Economic Studies 54, 1–21.
Harsanyi, J.C., Selten, R. (1988). A General Theory of Equilibrium Selection in Games. MIT Press, Cambridge, MA.
Hart, P.E., Prais, S.J. (1956). "The analysis of business concentration: A statistical approach". Journal of the Royal Statistical Society (Series A) 119, 150–181.
Hjalmarsson, L. (1974). "The size distribution of establishments and firms derived from an optimal process of capacity expansion". European Economic Review 5 (2), 123–140.
Hole, A. (1997). "Dynamic non-price strategy and competition: Models of R&D, advertising and location". Unpublished Ph.D. Thesis. University of London.
Hopenhayn, H.A. (1992). "Entry, exit and firm dynamics in long run equilibrium". Econometrica 60, 1127–1150.
Ijiri, Y., Simon, H. (1964). "Business firm growth and size". American Economic Review 54, 77–89.
Ijiri, Y., Simon, H. (1977). Skew Distributions and the Sizes of Business Firms. North-Holland, Amsterdam.
Jovanovic, B. (1982). "Selection and the evolution of industry". Econometrica 50 (3), 649–670.
Jovanovic, B., MacDonald, G.M. (1994). "The life cycle of competitive industry". Journal of Political Economy 102 (2), 322–347.
Klepper, S. (1996). "Entry, exit, growth and innovation over the product life cycle". American Economic Review 86 (3), 562–583.

Klepper, S., Graddy, E. (1990). "The evolution of new industries and the determinants of market structure". RAND Journal of Economics 21 (1), 27–44.
Klepper, S., Simons, K. (2005). "Industry shakeouts and technological change". International Journal of Industrial Organization 23, 23–43.
Lambson, V. (1992). "Competitive profits in the long run". Review of Economics Studies 59, 125–142.
Lee, C.-Y. (2005). "A new perspective in industry R&D and market structure". Journal of Industrial Economics 53, 1–25.
Lyons, B.R., Matraves, C. (1996). "Industrial concentration". In: Davies, S.W., Lyons, B.R. (Eds.), Industrial Organization in the European Union: Structure, Strategy and the Competitive Mechanism. Oxford Univ. Press, Oxford.
Lyons, B.R., Matraves, C., Moffat, P. (2001). "Industrial concentration and market integration in the European Union". Economica 68 (269), 1–26.
Mann, N., Scheuer, E., Fertig, K. (1973). "A new goodness-of-fit test for the two-parameter Weibull or extreme-value distribution with unknown parameters". Communications in Statistics 2 (5), 383–400.
Manuszak, M.D. (2002). "Endogenous market structure and competition in the 19th century American brewing industry". International Journal of Industrial Organization 20, 673–692.
Marin, P., Siotis, G. (2001). "Innovation and market structure: An empirical evaluation of the 'Bounds approach' in the chemical industry". Working Paper. Universidad Carlos III de Madrid and CEPR.
Marsili, O. (2001). The Anatomy and Evolution of Industries. Edward Elgar, Cheltenham.
Matraves, C. (1999). "Market structure, R&D and advertising in the pharmaceutical industry". Journal of Industrial Economics 47, 169–194.
Mazzeo, M.J. (2002). "Product choice and oligopoly market structure". RAND Journal of Economics 33, 221–242.
Motta, M., Polo, M. (1997). "Concentration and public policies in the broadcasting industry: The future of television". Economic Policy 25, 295–334.
Nelson, S., Winter, D. (1982). "An Evolutionary Theory of Economic Change". Harvard Univ. Press, Cambridge, MA.
Nocke, V. (2000). "Monopolization and industry structure". Working Paper. Nuffield College, Oxford.
Nocke, V. (2006). "Collusion and dynamic (under)investment in quality". RAND Journal of Economics. In press.
Pelzman, S. (1991). "The Handbook of Industrial Organization: A review article". Journal of Political Economy 99, 201–217.
Phillips, A. (1971). Technology and Market Structure: A Study of the Aircraft Industry. D.C. Heath, Lexington, MA.
Raith, M. (2003). "Competition, risk and managerial incentives". American Economic Review 93, 1425–1436.
Robinson, W., Chiang, J. (1996). "Are Sutton's predictions robust?: Empirical insights into advertising, R&D and concentration". Journal of Industrial Economics 44 (4), 389–408.
Rogers, R. (2001). "Structural change in U.S. food manufacturing, 1958–1977". Agribusiness 17, 3–32.
Rogers, R.T., Ma, Y.R. (1994). "Concentration change in an area of lax antitrust enforcement: A comparison of two decades: 1967 to 1977 and 1977 to 1987, evidence from the food processing industries". Paper presented at the Northeast Regional Research Project NE-165 Research Conference, Montreal, Quebec.
Rogers, R.T., Tockle, R.J. (1999). "The effects of television advertising on concentration: An update". New York Economic Review 30, 25–31.
Rosenthal, R. (1980). "A model in which an increase in the number of sellers leads to a higher price". Econometrica 48, 1575–1579.
Scherer, F.M. (1980). Industrial Market Structure and Economic Performance, second ed. Rand McNally, Chicago.
Scherer, F.M. (2000). "Professor Sutton's technology and market structure". Journal of Industrial Economics 48, 215–223.
Schmalensee, R. (1978). "Entry deterrence in the ready-to-eat breakfast cereal industry". Bell Journal of Economics 9, 305–327.

Schmalensee, R. (1989). "Inter-industry differences of structure and performance". In: Schmalensee, R., Willig, R. (Eds.), Handbook of Industrial Organization, vol. 2. North-Holland, Amsterdam, pp. 951–1009.
Schmalensee, R. (1992). "Sunk costs and market structure: A review article". Journal of Industrial Economics 40, 125–133.
Scott, J.T. (1984). "Firm versus industry variability in R&D intensity". In: Zvi, G. (Ed.), R&D Patents and Productivity Change. University of Chicago Press for the National Bureau of Economic Research, Chicago.
Selten, R. (1983). "A model of oligopolistic size structure and profitability". European Economic Review 22 (1), 33–57.
Shaked, A., Sutton, J. (1982). "Natural oligopolies". Econometrica 51, 1469–1484.
Shaked, A., Sutton, J. (1987). "Product differentiation and industrial structure". Journal of Industrial Economics 36, 131–146.
Shaked, A., Sutton, J. (1990). "Multiproduct firms and market structure". RAND Journal of Economics 21, 45–62.
Shubik, M., Levitan, R. (1980). Market Structure and Behavior. Harvard University Press, Cambridge, MA.
Smith, R.L. (1994). "Nonregular regression". Biometrika 81, 173–183.
Smith, R.L. (1985). "Maximum likelihood estimation in a class of non-regular cases". Biometrika 72, 67–90.
Smith, R.L. (1988). "Extreme value theory for dependent sequences via the Stein–Chen method of Poisson approximations". Stochastic Processes and Their Applications 30, 317–327.
Spence, A.M. (1981). "The learning curve and competition". Bell Journal of Economics 12, 49–70.
Stanley, M.R., Nunes Amaral, L.A., Buldyrev, S.V., Harlin, S., Leschorn, H., Maass, P., Salinger, M.A., Stanley, H.E. (1996). "Scaling behaviour in the growth of companies". Nature 319 (29), 804–806.
Sutton, J. (1991). Sunk Costs and Market Structure. MIT Press, Cambridge, MA.
Sutton, J. (1997a). "Game theoretic models of market structure". In: Kreps, D., Wallis, K. (Eds.), Advances in Economics and Econometrics, Proceedings of the World Congress of the Econometric Society. Tokyo, 1995. Cambridge Univ. Press, Cambridge, pp. 66–86.
Sutton, J. (1997b). "Gibrat's legacy". Journal of Economics Literature 35, 40–59.
Sutton, J. (1998). Technology and Market Structure. MIT Press, Cambridge, MA.
Sutton, J. (2000). Marshall's Tendencies: What can Economists Know?. MIT Press, Cambridge, MA.
Sutton, J. (2001a). "Rich trades, scarce capabilities: Industrial development revisited" (Keynes Lecture, 2000). In: Proceedings of the British Academy, vol. III (2000 Lectures and Memoirs), pp. 245–273. Reprinted in: Economic and Social Review 33 (1) (2002) 1–22.
Sutton, J. (2001b). "The variance of firm growth rates: The scaling puzzle". Physica A 312, 577–590.
Sutton, J. (2001c). "Market Structure and Performance". In: International Encyclopaedia of the Social and Behavioral Sciences. Elsevier, Amsterdam.
Sutton, J. (2007). "Market share dynamics and the "persistence of leadership" debate". American Economic Review 97, 222–241.
Symeonidis, G. (2000). "Price competition and market structure: The impact of restrictive practices legislation on concentration in the U.K.". Journal of Industrial Economics 48, 1–26.
Symeonidis, G. (2001). The Effects of Competition: Cartel Policy and the Evolution of Strategy and Structure in British Industry. MIT Press, Cambridge, MA.
Tirole, J. (1990). Theory of Industrial Organization. MIT Press, Cambridge, MA.
Tong, J. (2000). "Submarkets, shakeouts and industry life cycle". Sticerd Discussion Paper EI/26. London School of Economics.
Van Biesenbroeck, J. (2005). "Firm size matters: Growth and productivity growth in African manufacturing". Economic Development and Cultural Change 53, 545–583.
Vasconcelos, H. (2002). "Three essays on collusion and mergers". Unpublished Ph.D. Thesis. European University Institute, Florence.
Walsh, P.P., Whelan, C. (2001). "Portfolio effects and firm size distribution: Carbonated soft drinks". Economic and Social Review 33, 43–54.
Weiss, L. (Ed.) (1989). Concentration and Price. MIT Press, Cambridge, MA.
Wyart, M., Bouchard, J.P. (2002). "Statistical models for company growth". Available at SSRN: http://ssrn.com/abstract=391860 or doi:10.2139/ssrn.391860.