The Size Distribution of Businesses*
Part I: A Benchmark Case

by

John Sutton
London School of Economics and Political Science

The Toyota Centre
Suntory and Toyota International Centres for
Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel.: (020) 7955 6674

Discussion Paper
No. EI/9
December 1995

---

* I would like to thank Alison Hole, Reinout Koopmans, Michael Raith, Mark Schankerman and Hugh Wills for t heir very helpful comments on an earlier draft and Javier Hidalgo, Martin Knott and Jan Magnus for their helpful advice. The financial support of the Economic and Social Research Council and the Leverhulme Trust are gratefully acknowledged.

Abstract

This paper examines the evolution of a skew distribution of firm sizes from the viewpoint of the 'Bounds' approach to market structure. It confines attention to the role played by non-strategic factors (statistical independence, and cost side-effects). A model is proposed, which leads to a prediction regarding the least skew size distribution which is likely to be observed. This distribution provides a benchmark relative to which the impact of strategic effects on the form of the size of distribution may be assessed.

Keywords: Skew size distribution; firms; market structure; 'Bounds' approach; strategic effects; benchmark.

Contact address: Professor John Sutton, STICERD, London School of Economics and Political Science, Houghton Street, London WC2A 2AE. email: j.sutton@lse.ac.uk

Abstract

This paper examines the evolution of a skew distribution of firm sizes from the viewpoint of the 'Bounds' approach to market structure. It confines attention to the role played by non-strategic factors (statistical independence, and cost side-effects). A model is proposed, which leads to a prediction regarding the least skew size distribution which is likely to be observed. This distribution provides a benchmark relative to which the impact of strategic effects on the form of the size of distribution may be assessed.

Keywords: Skew size distribution; firms; market structure; 'Bounds' approach; strategic effects; benchmark.

Contact address: Professor John Sutton, STICERD, London School of Economics and Political Science, Houghton Street, London WC2A 2AE. email: j.sutton@lse.ac.uk

1. INTRODUCTION

The typically skewed size distribution of firms within any industry ('businesses') has for long been seen as one of the most salient features of market structure. Attempts to explain the form of the size distribution date back to Gibrat (1931), a paper that initiated a line of research on the 'Growth of Firms' in which purely stochastic influences played a key role. A central aim of this literature was to show how the size distribution would converge over time to some specific form[2]. This literature has been revived recently with an emphasis on the role of 'economic' factors. As this literature has developed, it has become clear that the form of the size distribution depends delicately on the details of the model employed, including many details that must be treated as 'unobservables' in empirical applications[3]. Meanwhile, it has become accepted in empirical work that no single form of distribution can be regarded as 'standard' or typical[4].

This paper proposes a new approach to the analysis of the size distribution. The motivation for this approach lies in the idea that while strategic interactions (and other 'economic' factors) have an important influence on equilibrium structure, they are not the whole story. Any industry will contain clusters of products or

---

[2]Lognormal, Yule, etc. A full review of this literature will be found in Sutton (1995a).

[3]See, for example, Jovanovic (1982), Ericson and Pakes (1988) and Roberts and Samuelson (1988). This point also holds for 'non-strategic' models such as Lucas (1978), where the size distribution depends on an unobservable distribution of managerial abilities.

[4]Richard Schmalensee's (1989) survey in the Handbook of Industrial Organisation concludes that 'all families of distributions so far tried fail to describe at least some industries well'. These twin themes, of delicate theoretical results and a lack of any 'sharp' statistical regularities, are also central themes in the other major literature on market structure, the 'stage-game' literature that follows the Bain tradition of analysing how industry-specific factors influence cross-industry differences in concentration (Sutton (1991)).

plants that compete closely. But an industry, as conventionally defined in official statistics, will usually contain more than one such cluster; it will be possible to identify pairs, or sets, of products that do not compete directly. In other words, most conventionally defined industries exhibit both some strategic interdependence, and some degree of independence across submarkets[5].

The task of combining these two features requires a fairly lengthy analysis. If the theory is to lead to testable predictions as to the effect of 'strategic factors', it is necessary to set up a 'null hypothesis' based on a description of what happens when such strategic factors are absent, i.e. where the only sources of skewness lie in statistical 'independence effects' and in (absolute or size-related) cost differences between firms. The aim of this paper is to analyse this 'benchmark' case.

In order to develop the main idea in the simplest possible way, we set up the benchmark model using the simple framework employed in the Growth of Firms

---

[5]It might seem attractive to respond to this problem by insisting on an 'appropriately narrow' definition of an industry. This however, is not practicable. What is at issue is that a firm's profit function may be additively separable into contributions deriving from a number of 'remote' products. Consider, for example, the standard Hotelling model where products are placed along a line. A firm offering a set of non-neighbouring products has, at equilibrium, a profit function which is additively separable into contributions from each product. Any real market in which products are spread either over some geographic space, or some space of attributes, will tend to exhibit this feature.

It might also seem that the problem of 'independence' might be easy to deal with in the standard game-theoretic models, since all that is involved in principle is a special (additively separable) form of the profit function. This is not so, however. The standard program of 'listing all the perfect Nash equilibria' leads us, in this kind of setting, to have some equilibria in which all firms have the same size. Such equilibria play a key role whenever this body of theory is applied to discussions of market structure. It is the fact that this kind of outcome is rarely, if ever, observed in practice that motivates the 'size distribution' literature.

literature by Herbert Simon and his collaborators[6]. A companion paper (Sutton (1995b)) re-casts the present analysis in the form of a game, introduces 'strategic' factors, and examines their effects.

In defining this 'benchmark' model, we need to ask about its domain of application: to what set of industries does this model refer? To answer this, we need to ask what kinds of strategic effects are potentially important, and in what kinds of industries will the influence of these effects be small. It is well known that at least two kinds of strategic effects may be important. The first is associated with 'escalations' of Advertising or R&D spending, and the second is associated with the externalities that each plant (or product) has on its near neighbours in geographic (or product) space. It is proposed in what follows that the model be applied to markets in which Advertising and R&D are negligible, and in which the range of products and/or plants are widely dispersed in geographical or product space.

II. THE SIMON MODEL REVISITED

Imagine a process by which the activities of firms within some specific industry ("businesses")[7] grow over time as the industry expands. Imagine that a discrete sequence of 'investment opportunities' become available to firms. These opportunities may be thought of as involving the opening of a new plant, the establishment of an outlet in a new area, or the introduction of a new product

[6]Simon and Bonnini (1958), Ijiri and Simon (1977).

[7]This distinction between "business" and "firm" becomes important once we turn to empirical applications; all empirical data reported below relate to "businesses". In presenting the theory, we deal always with a single industry so that "firms" and "businesses" are synonymous.

variety. What considerations will determine how the size distribution of firms evolves over time?

However we choose to model this situation, the shape of the size distribution will turn upon the answers to two questions:

(i) Is there any systematic bias in favour of 'large' firms (those that have already entered many products), or 'small' firms? In other words, is the next product which is introduced by some currently active firm more likely to be introduced by a larger, or by a smaller, firm?

(ii) How likely is it that the next product will be introduced by a new entrant, as opposed to a firm that is already active; and how does this likelihood change over the course of time?

The traditional 'Growth of Firms' literature dealt with question (i) by postulating Gibrat's Law, i.e. a larger firm was more likely to fill the next opportunity, in proportion to its current size. This may appear reasonable, but it is certainly a rather arbitrary hypothesis to introduce[8]. Here, in eschewing all attempts to say what is 'likely', we avoid taking any position on this issue. Instead, we aim to explore the implications of the following condition:

Condition 1: The probability that the next market opportunity is filled by any currently active firm is nondecreasing in

[8]Recent empirical studies have suggested that the best simple generalization is that, on average, smaller firms that survive grow proportionately faster than large firms, but the probability of survival is lower for smaller firms (Evans (1987), Dunne Roberts and Samuelson (1988)). The real problem lies not in characterising what happens 'on average', but in the fact that a wide range of different patterns occur across different markets, so that it is difficult to make any generalisations as to what is the 'normal' size/growth relation, or the 'typical' shape of the size distribution (for a review of these issues see Sutton (1995a)).

the size of that firm.

Consider two businesses of different sizes. Condition 1 is violated if the smaller business is more likely to take up the next market opportunity than is the larger one. This might happen, for example, if the incremental profit realised from the new investment was smaller for the larger firm. This supposed disadvantage to the larger firm could derive either from the cost side or through 'strategic effects' on the demand side.

As to the cost side, a larger business may enjoy an advantage through economies of scope in offering several products, or in operating many plants. On the other hand, a traditional argument suggests that it will not suffer any cost disadvantage; for, if an integrated business of larger size had higher unit costs, then it should be possible to split the business into completely independent and separately managed units under single ownership, so that any such disadvantage is eliminated[9,10]. This is the standard 'replication' argument for non-diminishing returns, and it is a very appealing one. Can an analogous argument be offered on the demand side?

The answer is 'no'. The game-theoretic literature has afforded us a rich menu of examples in which the larger firm suffers a disadvantage in the sense that the profit per product (or plant, or unit capacity) is *decreasing* in the number of

---

[9]Such an argument supposes that 'managerial diseconomies' can be avoided over the empirically relevant range, whether by divisionalization or otherwise.

[10]The relationship between the postulate of non-decreasing returns and Condition 1 is worth noting. A wide range of assumptions regarding the link between firm size and expected growth are consistent with constant returns: Simon, for example, appealed to the presence of constant returns to motivate Gibrat's Law. Condition 1 is consistent with either constant returns or increasing returns. On the other hand, if diminishing returns are present, so that large firms are disadvantaged relative to small, then Condition 1 would no longer be tenable.

products (or plants, or units of capacity) operated by the firm. This effect has a simple intuitive interpretation: if the multi-product or multi-plant firm expands output or cuts price in order to improve the profit of one of its plants, it generates a negative externality for the other plants. In maximizing its total profit, the firm seeks to 'internalise' this externality. This leads to higher prices and lower profits on each product or plant.

Nonetheless, empirical evidence on size-profitability relationships across businesses of different sizes within an industry suggests that the rate of return (profit) is nondecreasing in the size of the business[11]. This suggests that firms may have some way of circumventing such strategic disadvantages where they arise. This will be the case whenever market opportunities are dispersed either geographically or in some space of 'product attributes'. If a firm that owned a number of closely clustered plants were to earn lower profit per plant, then that firm could simply expand by opening a sequence of plants in dispersed locations, thereby avoiding the strategic disadvantage[12]. It is this argument which motivates the claim that the present model may reasonably be applied to the general run of manufacturing industries defined at the 4- or 5-digit SIC level[13].

We now turn to the issue of entry, as posed in question (ii) above. Here, this paper follows Simon in noting that no particular hypothesis suggests itself on

---

[11]The FTC Line of Business data and the PIMS dataset are the standard sources (Scherer (1980)).

[12]This argument is made precise in the companion paper, Sutton (1995b), where it is shown that the results of the present paper hold in a setting where there are a large number of similar submarkets, whatever the nature of strategic interactions within each submarket.

[13]It is not difficult to identify certain narrowly defined markets in which Condition 1 seems likely to fail. It is argued in the companion paper that the behaviour of very narrowly defined markets provides one useful test of the validity of the present analysis.

a priori grounds. What is at issue here is the fraction of new products or plants introduced by new entrants, as opposed to incumbents. What matters, as will be shown, is not whether this fraction if high or low -the results of interest turn out to be independent of this - but whether this fraction varies over time, and in what manner. Fortunately, this is something which can be checked directly. It turns out that Simon's simple assumption that this fraction remains constant over time provides a natural benchmark case, and it will be shown that the empirical predictions of the theory are fairly robust to empirically reasonable deviations from this case[14].

Condition 2: The probability p that the next market opportunity is filled by a new entrant is constant over time.

III. THE BASIC PROCESS

The following process is identical to that used by Ijiri and Simon (1977) apart from the replacement of Gibrat's Law by Condition 1:

A sequence of discrete and independent investment opportunities arise over time. Each opportunity is of the same size, in terms of the sales revenue and profit it yields to any single firm which takes it up; and each opportunity would be

[14]It is worth noting that the rate of capture of opportunities by new entrants depends inter alia on the number of potential entrants available. This introduces an exogenous influence that cannot be removed by appealing to 'optimizing behaviour'. Any attempt to do so will merely push back the arbitrariness by introducing some new exogenous influence, such as the (probably unmeasurable) distribution of entrepreneurial talent. It seems preferable to develop predictions that are condtioned directly on the rate of capture of opportunites by entrants, since this is directly measurable, allowing the robustness of predictions to be examined.

unprofitable if more than one firm took it up. We label these opportunities by an index t = 1, 2, 3, ...,T. The size of a firm is measured by the number of opportunities it has taken up. Firms that have already taken up at least one opportunity are referred to as active.

We denote by $n_{i,t}$ the number of firms of size i at stage t, and by N_t the number of active firms at time t, whence $N_t = \sum_{i=1}^{T} n_{i,t}$. The process begins at stage t = 1, when the first opportunity is taken up by some firm, whence $N_1 = 1$. What we aim to examine is the evolution of the number of firms N_t and their size distribution, described by the vector $\{n_{i,t}\}$.

The evolution of N_t can be analysed independently of $n_{i,t}$, and is driven only by Condition 2. This implies that the total number of firms entering between stage 2 and stage t, which by definition equals N_t - 1, is described by a binomial distribution with density

$$\text{Prob}(N_t = N) = \binom{t-1}{N-1} p^{N-1}(1-p)^{t-N} \qquad (1)$$

Equation (1) defines Prob (N_t = N) for all N = 1, 2, 3, ... and t = 2, 3, ... The number of firms N_t takes values 1, 2, ..., t and has mean 1 + p (t - 1), where p denotes the probability that the new opportunity arising in any period is captured by an entrant.

The evolution of $n_{i,t}$ is more complex. Our aim is to characterize the *least skew* distribution permitted by Condition 1. It is intuitively clear that this corresponds to the sub-case of Condition 1 under which each active firm has the *same* probability of taking up the next opportunity; this is established in Appendix 2. This remark motivates:

Condition 1': The probability that the next opportunity is taken up by any active firm is independent of the size of that firm.

Conditions 1' and 2 imply that each of the N_{t-1} firms active at stage t - 1 has an equal probability $(1-p)/N_{t-1}$ of taking up the opportunity which arises at stage t.

We aim to describe the distribution of firm size $\{n_{i,t}\}$ conditional on N_t. At t = 1, we have $N_1 = 1$ and $\{n_{i,1}\} = \{1, 0, 0 \ldots\}$.

For values of t less than 4, it is easy to see that there is only one possible size distribution vector $\{n_{i,t}\}$ for each value of N_t. For $t \geq 4$, some values of N_t are supported by two or more size distribution vectors. We denote the expected value of $n_{i,t}$ conditional on N_t as $E(n_{i,t}|N_t)$.

The initial condition for t = 1 is:

$$N_1=1;\ E(n_{1,1}|1) = 1;$$
$$E(n_{2,1}|1) = E(n_{3,1}|1) = \ldots = 0 \qquad (2)$$

and for the special case $N_t = 1$ in which no entry occurs after stage 1, so that there is a single firm of size i = t,

$$E(n_{i,t}|1) = 1 \quad \text{for } i=t;$$
$$= 0 \quad \text{otherwise} \qquad (3)$$

Our aim is to analyse the behaviour of $E(n_{i,t}|N_t)$ as t increases.

The evolution of N_t is fully described by equation (1). Given N_{t+1}, the number of firms in the preceding period, N_t, takes values N_{t+1} and $N_{t+1} - 1$. Moreover, it follows from (1) that[15]

$$\text{('No entry')} \quad \text{Prob}(N_t=N \mid N_{t+1} = N) = 1 - (N_t-1)/t$$
$$\text{('Entry')} \quad \text{Prob}(N_t=N-1 \mid N_{t+1} = N) = (N_t-1)/t \qquad (4)$$

[15]see Appendix 1, Note 1.

Now consider $n_{1,t+1}$. This takes three values, according as the (t + 1)th opportunity is taken by (a) an entrant, (b) a firm of size 1, or (c) a firm of size $i \neq 1$. If entry occurs, (case (a)), n_1 rises by 1 unit at stage t. If no entry occurs, then the firm capturing the new opportunity is a firm of size 1 with probability $n_{1,t}/N_t$, and here n_1 falls by 1 unit (case (b)); otherwise n_1 remains unchanged (case (c)).

It follows that, for all $t \geq 2$ and $2 \leq N_t \leq t$,

$$E(n_{1,t+1}|N_t) = \frac{N_t - 1}{t}\left\{E(n_{1,t}|N_t - 1) + 1\right\} + \left(1 - \frac{N_t - 1}{t}\right)\left\{E(n_{1,t}|N_t) - \frac{E(n_{1,t}|N_t)}{N_t}\right\} \qquad (5)$$

To interpret (5), note that the first term on the r.h.s. corresponds to the case where the opportunity is taken by a new entrant. Conditional on entry, n_1 rises by one unit, with probability 1. The second term corresponds to the case where no entry occurs. Conditional on this, n_1 falls by one unit with probability $n_{1,t}/N_t$.

A similar argument applies when $i > 1$. If entry occurs, n_i remains unchanged. If no entry occurs, then the firm capturing the new opportunity is a firm of size i with probability $n_{i,t}/N_t$, and here n_i falls by 1 unit; while the firm capturing the new opportunity is of size (i-1) with probability $n_{i-1,t}/N_t$, and here n_i rises by one unit. It follows that:

For $i \geq 2$ and for all $t \geq 2$ and $2 \leq N_t \leq t$,

$$E(n_{i,t+1}|N_t) = \frac{N_t-1}{t}\left\{E(n_{i,t}|N_t-1)\right\}$$

$$+\left(1-\frac{N_t-1}{t}\right)\left\{E(n_{i,t}|N_t) - \frac{E(n_{i,t}|N_t)}{N_t} + \frac{E(n_{i-1,t}|N_t)}{N_t}\right\} \qquad (6)$$

Given the boundary conditions (2) and (3), equations (5) and (6) uniquely define the function $E(n_{i,t} \mid N_t)$ on the domain $t \geq 2$, $2 \leq N_t \leq t$. The solution is:

$$E(n_{i,t}|N_t) = N_t \cdot \frac{\binom{t-i-1}{N_t-2}}{\binom{t-1}{N_t-1}} \qquad (7)$$

This can be checked by direct substitution of equation (7) in equations (5) and (6)(see Appendix 1, Note 2). For i = 1 equation (7) reduces to

$$E(n_{1,t}|N_t) = N_t \cdot \frac{N_t-1}{t-1} \qquad (7a)$$

and for $i \geq 2$ it becomes

$$E(n_{i,t}|N_t) = N_t\frac{N_t-1}{t-1} \cdot \left\{\frac{t-N_t}{t-2} \cdot \frac{t-N_t-1}{t-3} \cdot \cdots \cdot \frac{t-N_t-(i-2)}{t-i}\right\}$$

$$= E(n_{1,t}|N_t)\left\{\frac{t-N_t}{t-2} \cdot \frac{t-N_t-1}{t-3} \cdot \cdots \cdot \frac{t-N_t-(i-2)}{t-i}\right\} \qquad (7b)$$

The total number of firms N_t is binomially distributed with mean $1 + p(t-1)$, and the ratio $(N_t-1)/(t-1)$ converges in probability to p as $t \to \infty$. For any fixed i, the expected number of firms of size i increases to infinity as $t \to \infty$. In what follows, we focus on the behaviour of

$$\frac{1}{1 + p(t-1)} E(n_{i,t}|N_t)$$

as a function of i.

From (7a) we have

$$\frac{1}{1+p(t-1)} E(n_{1,t}|N_t) = \frac{N_t}{1+p(t-1)} \cdot \frac{N_t-1}{t-1}$$

$$= \frac{t-1}{1+p(t-1)}\left(\frac{N_t-1}{t-1}\right)^2 + \frac{1}{1+p(t-1)}\left(\frac{N_t-1}{t-1}\right)$$

It is convenient to define the random variable $(N_t-1)/(t-1) = \theta_t$, whence

$$E\left(\frac{n_{1,t}}{1+p(t-1)} | \theta_t \right)$$

$$= \frac{1}{1+p(t-1)} E(n_{1,t}|\theta_t) = \frac{t-1}{1+p(t-1)}\theta_t^2 + \frac{1}{1+p(t-1)}\theta_t$$

The unconditional expectation may be obtained by taking the expectation of this expression over θ_t. Since $\theta_t \xrightarrow{p} p$, however, the limit of the unconditional expectation may be obtained by substituting p for θ_t on the right hand side (by the Helly-Bray theorem (Rao (1973)). We have,

$$\lim_{t\to\infty} E\left(\frac{n_{1,t}}{1+p(t-1)} \right) = p$$

It follows in the same way from (7b), on writing each term in $\{\cdot\}$ in the form

$$\frac{t-N-(i-2)}{t-i} = 1 - \frac{t-1}{t-i}\theta_t - \frac{1}{t-i}$$

that

$$\lim_{t\to\infty} E\left(\frac{n_{i,t}}{1+p(t-1)} \right) = p(1-p)^{i-1} \qquad (8)$$

Hence the size distribution tends to a geometric distribution[16] with parameter p.

---

[16]This characterisation of the size distribution for large T is a weak one. It would be nice to establish a stronger property, and simulations of the process suggest that a stronger characterisation may be available. Simon's method (Ijiri

It will be convenient in empirical applications to treat the size of the firm as a continuous variable x, replacing the geometric distribution (8) by the corresponding exponential and expressing the size distribution by the density

$$f(x) = p \exp(-px) \tag{9}$$

Two features of this result are worth noting:

(i) The mean size of firm converges to a constant, 1/p. (Increases in T, and so in N_T, raise the size of the largest firms, but only in the sense that the size of the largest draw among a total of n draws from a given distribution rises with the number of draws, n.)

(ii) The exponential size distribution (9) may be thought of as the envelope of a set of size distributions of different age cohorts in the overall population of firms. It is shown in Sutton (1995b) that the size at stage t of a firm which entered the industry at stage τ can be described asymptotically as (1 + x) where x is a Poisson variable with parameter[17]

$$(1-p)\left\{\frac{1}{\tau} + \frac{1}{\tau+1} + \ldots + \frac{1}{t}\right\} \tag{10}$$

Estimating Concentration Ratios

We now turn to the implications of this for the relationship between the number of firms in the market, N, and a conventional measure of concentration, the k-

---

and Simon (1977)), which leads to a very simple calculation of (8) can be used only if it is *assumed* that $E(n_{i,t}/N_t)$ converges to a limiting value for each i. (See Appendix 2.)

[17]Were size distribution data available for separate age cohorts of businesses, this would provide a useful further test of the present model.

firm concentration ratio.

In what follows, we maintain the exponential approximation (9) throughout[18,19]. The properties of the k-firm concentration ratio, defined as the fraction of the t opportunities captured by the k largest firms among the N firms present, now follow from standard properties of the extreme value distribution for the exponential (Gumbel (1958) p. 116 ff.).

In Appendix 4, it is shown that for any integer k, as N increases the k-firm concentration ratio C_k tends towards

$$\frac{k}{N}\left(\gamma_k - ln\frac{k}{N}\right) \qquad (11)$$

where γ_k is a constant which depends on k. As k increases, γ_k approaches unity.

Given the range of values of k and N which are usually recorded in official statistics, the limiting formula

$$\frac{k}{N}\left(1 - ln\frac{k}{N}\right) \qquad (11)'$$

will often be adequate in empirical applications (see Appendix 4).

---

[18]The use of an exponential distribution, rather than a geometric, is unproblematic unless p is very close to unity. If $p = 1$, all opportunities are filled by new entrants and all firms are of size 1. For p close to, but not equal to 1, the size distribution is geometric, with $f(1) = p$.

[19]The smooth Lorenz curve generated by the exponential distribution approximates the piecewise linear schedule generated by the geometric distribution. This approximation is close enough for empirical purposes unless the top k firms include firms of size unity. This would happen, for a given k, if a sufficiently high proportion of firms were of size 1. What we require is that $k/N << 1-p$. Since we are normally concerned with the top tail of the distribution and since p normally lies in the range 0.1 - 0.2, this qualification is unimportant in practice.

Expressions (11) and (11)' define a Lorenz curve for the industry. The limiting form (11)' can be derived directly using elementary arguments, as follows: Let the size distribution be described by the exponential distribution (9), and consider the ratio between the proportion of opportunities accounted for by firms of size x or greater, and the number of firms in this size band.

From (9) we have

$$\int x\, f(x)\, dx = -\left(x + \frac{1}{p}\right) e^{-px}$$

$$\int f(x)\, dx = -e^{-px}$$

whence the proportion of products (or plants) accounted for by firms of size $\bar{x}$ or greater is

$$\int_{\bar{x}}^{\infty} x f(x)\, dx \;/\; \int_{-0}^{\infty} x f(x)\, dx = (1 + p\bar{x})\, e^{-p\bar{x}} \qquad (12)$$

and the proportion of firms in this size band is

$$\int_{\bar{x}}^{\infty} f(x)\, dx \;/\; \int_{0}^{\infty} f(x)\, dx = e^{-p\bar{x}}$$

Write $e^{-p\bar{x}}$ as z, whence $\bar{x} = -(1/p) ln z$. Substituting this in (12), we have: if firms are ranked in descending order of size, a proportion z of firms accounts for a proportion G(z) of sales, where

$$G(z) = z\,(1 - ln z) \qquad (11)''$$

This corresponds to equation (11)' above, with z = k/N.

The Lower Bound

In motivating Condition 1, it was argued that while large firms might suffer a relative disadvantage via 'strategic effects', they might in practice be able largely to evade any such disadvantage by a suitable choice of product or plant locations. On the cost side, it was argued that any disadvantage suffered by large firms could be avoided via a 'replication' strategy. However, in many industries, larger firms may enjoy some cost advantage over smaller rivals by way of economies of scale or scope. For this reason, inter alia, Condition 1 was stated in terms of an inequality constraint.

Expression (11) has been derived by replacing the inequality constraint of Condition 1 by the independence postulate of Condition 1', in order to derive a characterization of the least skew distribution of firm size consistent with Condition 1. In Appendix 2, it is shown that for any process satisfying the inequality constraint of Condition 1, the Lorenz curve of the corresponding limiting distribution lies further from the diagonal than the limiting curve given by (11)'. Hence, expression (11)' provides an asymptotic lower bound to concentration as a function of the number of firms:

Proposition 1: For any fixed ratio k/N, an asymptotic lower bound to the k-firm concentration ratio is given by

$$\underline{C}_{k/N} \geq \frac{k}{N}\left(1 - ln\frac{k}{N}\right)$$

This result has two interesting features:

(i) The shape of the size distribution, and so the lower bound to concentration, is *independent* of the entry parameter p. 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, in which theory led to a family of size distributions of varying skewness, parameterised by p. In Simon's work, this parameter was linked to the level of the entry rate of new firms to the market. Other early models also led to a *family* of size distributions; in Hart and Prais (1956), for example, the variance of the lognormal distribution could be linked to the variance of the distribution of 'shocks' to firm size between successive periods. In the present setting, expression (11)' contains no free parameters whose measurement might be subject to error. Rather, Condition 1 leads to a quantitative prediction regarding the lower bound to concentration, conditional only on the assumed constancy of the entry rate (Condition 2).

(ii) Various countries publish data on k-firm concentration ratios for several different values of k. Proposition 1 implies that the various k-firm ratios are all bounded below by a curve which approximates (11)' In the next section, we take advantage of this in pooling data for various reported k-firm concentration ratios.

The lower bound given in Proposition 1 lies far above the minimal level $C_{k/N} = k/N$ corresponding to firms of equal size (See Figure 1 and Table 1). It is also well separated from the family of Lorenz curves derived from 'Gibrat's Law', using 'reasonable' parameter values.

k/N	$C_{k/N} = \frac{k}{N}\left(1 - ln\frac{k}{N}\right)$
.1	.33
.2	.52
.3	.66
.4	.77
.5	.85
.6	.91
.7	.95
.8	.98
.9	.99

Table 1. Predicted lower bounds for thek-firm concentration ratio in an N-firm industry.

The prediction of Proposition 1 is set out in Table 1 and can readily be tested using published data on concentration ratios and firm numbers. In practice, however, data on firm numbers is widely seen as being problematic, for reasons noted in the next section. It is of interest, therefore, to ask whether the theory can be tested in the absence of satisfactory data on firm numbers. So long as data is available for two or more concentration ratios, it is possible to proceed as follows: if the size distribution converges to some stationary distribution, and is generated by transition probabilities satisfying Conditions 1 and 2, then if we know the m-firm concentration ratio C_m, we can place a lower bound D_k on the k-firm concentration ratio for any $k < m$. This lower bound D_k will coincide with the true k-firm concentration ratio if the size distribution is exponential. In Appendix 2, we establish:

Proposition 2. Let N_m be defined implicitly by the equation:

$$C_m = \frac{m}{N_m}\left(1 - ln\frac{m}{N_m}\right)$$

Then a conditional lower bound to the k-firm concentration ratio is

$$D_k\ (C_m) \geq \frac{k}{N_m}\left(1 - ln\frac{k}{N_m}\right) \qquad (14)$$

This procedure allows us to compute a series of lower bounds to C_k conditional on C_m; a range of computed bounds is shown in Table 2, for m = 50 (The 50-firm concentration ratio is the highest normally reported).

C_{50}	D_{20}	D_8	D_4
.1	.05	.02	.01
.2	.10	.05	.03
.3	.15	.07	.04
.4	.21	.10	.06
.5	.27	.13	.08
.6	.33	.17	.10
.7	.40	.21	.12
.8	.48	.25	.15
.9	.57	.31	.19
.99	.71	.41	.25

Table 2. Predicted lower bounds D_k for the k-firm concentration ratio conditional on an observed value C_{50} for the 50-firm concentration ratio.

IV. ROBUSTNESS

Before turning to empirical tests, we first consider the robustness of the predicted lower bound, relative to a number of special features of the present model.

Industry Growth Patterns

The analysis has been couched in terms of a sequence of opportunities, and the limiting distribution relates to the situation in which the total number of opportunities becomes large. The results are independent of the rate at which these opportunities arise over time.

Size of Opportunities: Firm-specific Efficiency Differences

In seeking to describe a lower bound to concentration, the model eliminates all inessential sources of asymmetry. In particular, it assumes that all opportunities are of the same size. Insofar as opportunities are discrete, but differ in size, the size distribution will be *more* skew, and concentration will lie above the bound specified in Proposition 1.

If some firms have a higher probability of capturing opportunities than others, due for example to firm-specific efficiency differences, then the distribution will again be *more* skew than the benchmark case.

The Pattern of Entry

The ancillary Assumption 2 on the constancy of entry rates is arbitrary, but it defines a useful benchmark. If the rate of entry of new firms increases over time, then the size distribution becomes more skew, and the bound given in Proposition 1 remains valid. If the rate of entry falls, however, the size

distribution becomes less skew, and the predicted lower bound may be violated.

Here, we follow Simon and Ijiri in quantifying the size of the deviations from the predicted lower bound which would follow for empirically reasonable patterns of firm entry. But how much do entry patterns differ in practice from constancy? One way of checking the size of such deviations directly is to identify opportunities with new plants (establishments) and look at the relationship between the rate of entry of new plants versus the net rate of growth of firms in the industry over time. A useful benchmark is obtained by taking all U.S. 4-digit homogeneous goods industries[20] from 1947 to 1977. Figure 1 shows, on the vertical axis, the firm/establishment ratio for 1977 divided by the firm/establishment ratio for 1947; this is plotted against the fractional increase in the total number of establishments over the same period. It is clear that, for industries in which the number of establishments continued to grow substantially from 1947 onwards, the proportionate rate of growth in the number of establishments is very close to the proportionate rate of growth in the number of firms (observations cluster around unity to the right of the figure). On the other hand, those industries which attracted few new establishments after 1947 showed widely varying experience: the growth rate in the number of firms was at most about equal to that of establishments, but in three cases fell to less than 0.5.

This suggests the following interpretation: over the first half of an industry's growth phase, a constant value of p is a reasonable assumption. Over the latter half, p is likely to decline, and it may in some cases fall by a factor of 0.5 or so. With this in mind, two following robustness tests were tried. (i) A series of simulations of the process were carried out, in which p was replaced by p/2 over the latter half of the industry's history (i.e. after half the final number of products had been entered). (ii) The process was simulated with p falling linearly from its initial value to zero, over the course of the industry's history.

[20]The set of industries used here is that introduced in Section IV below.

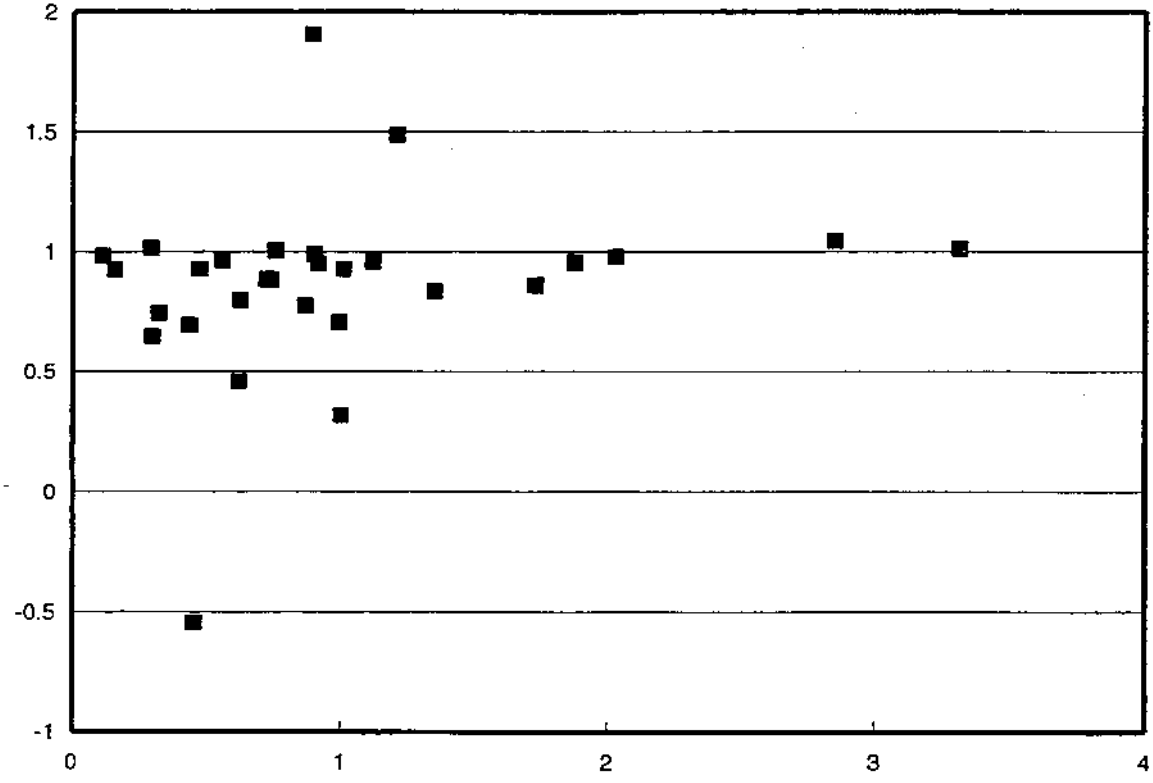

Figure 1. Growth in firm numbers and establishment numbers for the U.S., 1947-1977. The horizontal axis shows the fractional rate of increase in the number of establishments over its 1947 value. Industries for which this increase is less than 10% were excluded. The vertical axis shows the ratio between the fractional increase in the number of firms and the fractional increase in the number of establishments. The data relates to those industries listed in Section VI, and is confined to U.S. 4-digit industries whose SIC definitions were unchanged over this period.

In both cases, it was found that the shift in the Lorenz curve was quite small, (Figure 2). Indeed, its size is such that it might be difficult to detect in small samples[20].

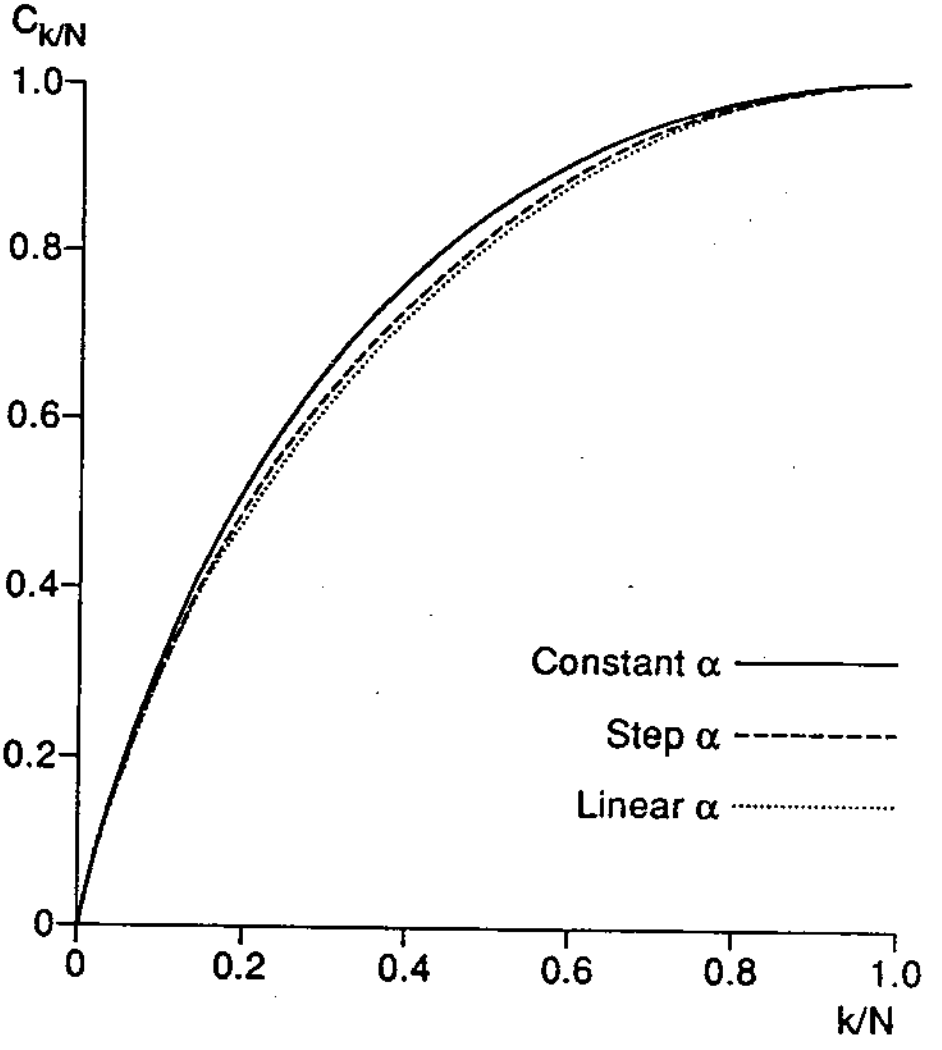

Figure 2. The solid curve shows the bound predicted by Condition 2, ('constant p') together with those curves generated by the robustness tests described in the text.

[20]To investigate this, a Lorenz curve was constructed for the industries shown in Figure 2, and the sample was split into two groups, according as p fell by more or less than the median amount, during the latter half of the period. The histograms of residuals between C_4 and its predicted lower bound were examined separately for each group; there was no detectable difference between the two groups. This is consistent with the above suggestion that empirically reasonable changes in p over time lead to differences in the lower bound which are so small as to be difficult to detect empirically.

Shakeout, Acquisition, and Decline

A number of related issues arise regarding the assumption that opportunities are permanent, and firms never decline.

- It is a common feature of industry histories that a substantial fraction of early entrants may exit the industry after some time ("Shakeout"; See Klepper and Graddy (1990)). The effect of a one-off shakeout of some point in the industry's history on the long run size distribution in the model depends on how the opportunities vacated by exiting firms are re-allocated among surviving firms. The descriptions of the shakeout process in the literature suggest that a natural model is one where the size of plant shifts upwards, and current industry production is re-allocated to a smaller number of 'surviving' plants. If 'surviving plants' are selected randomly, if all future plants are of the new 'larger' size, and if the fraction p of new plants introduced by entrants is unchanged, then the limiting size distribution retains its exponential form.

- Some 'disappearances' of firms occur via acquisition. If acquiring firms were drawn disproportionately from the low end of the size distribution, this could cause a violation of the proposed bound; in practice, however, acquiring firms tend to be drawn from the upper end of the distribution. (Acquisition activity is *one* of several mechanisms which will cause Lorenz curves for 'typical' industries to lie further from the diagonal than the proposed bound.)

- The model does not allow for any process of industry decline, in which market opportunities vanish as products are withdrawn or plants shut down. To do so would require an additional assumption, analogous to Condition 1, regarding the relationship between the size of an incumbent firm and the likelihood that this

firm will be the next firm to shut down a plant. We make no assumption here on this question, and so no constraint can be placed on the likely size of the effect which might be involved. It remains an empirical question whether the bound might be systematically violated in declining industries[21]

V. FROM THEORY TO MEASUREMENT

In order to test Propositions 1 and 2 empirically, we need first to address two serious problems which arise in relating the theoretical measures C_k and N to their empirical counterparts.

Aggregation

The model set out above relates to a single well defined market in which all active firms produce similar substitute goods. In implementing the model empirically, we are faced with data in which the SIC industry may encompass different subgroups of firms whose activities are focused on different product lines, or mixes of product lines, within the industry.[22] Such problems become

---

[21]A game-theoretic analysis suggests that, for *some* forms of technology, there will be a tendency for the sizes of the largest firms in the industry to converge as industry output falls (Ghemawat and Nalebuff (1990)). This could lead to a violation of the bound.

[22]This problem is quite distinct from another issue which arises in practice: that some firms may also be active in other industries. As noted earlier, we avoid this issue here in that all the data presented relates to firms' sales *within the industry* (i.e. to businesses). It is worth remarking that, since it is the larger firms which are more often diversified, the size distribution of firms in terms of total reported sales (which was widely studied in the Growth of Firms literature) should be more skew than the limiting distribution characterized by equation (11)'.

more serious as we move to higher levels of aggregation. For the U.S., concentration measures are reported both at the 5-digit 'product market' level and at the 4-digit 'industry' level. Most countries, however, report ratios only at the 4-digit level. The present model may reasonably be applied at the 5-digit level; it is less clear whether it is appropriate to apply it at higher levels of aggregation.

We begin by distinguishing two separate problems which arise with aggregate data. Firstly, there is the case in which the 'industry' encompasses two quite independent product markets A and B. The distinguishing feature of this case is that A-products are produced by one group of firms, and B-products by a different group. Each product market may be described by the present model, but they may differ in respect of the parameter p, and so in terms of average firm size. We refer to this as the case of 'Independent sub-industries'.

A separate problem arises, which we label 'Interdependent sub-industries'. Here, new products are again of 'type A' or 'type B', but a single group of firms produces both product types. A new product of either type may be introduced by any firm currently active in the industry, whether that firm is already producing A, or B, or both.

The following results are established in Appendix 3:

Proposition 3(a): ('Independence"): In the 'independent' sub-industry case, the lower bound to the measured values ($C_{k/N}$,N) lies at or above the lower bound specified by Proposition 1. If mean firm size in all sub-industries is equal, then the measured values *coincide* with those defined by Proposition 1; otherwise, they lie strictly within the bound. (The bound is valid, but is not tight.)

Proposition 3(b): ('Interdependence'): In the 'interdependent' sub-industries case, the lower bound to the measured values ($C_{k/N}$,N) coincides with the lower bound specified by Proposition 1.

The overall conclusion, then, is that the bound specified by Proposition 1 continues to be valid for aggregate data, but may not be tight.

Measuring Firm Numbers

A separate problem arises in measuring the number of firms in the industry. Here, the problem arises at the lower end of the distribution. We often find a fringe of very small firms allocated by the Census to a particular manufacturing industry, whose activities do not extend to the production of a standard line of core industry products, but are confined to small-scale ancillary activities. If such firms are included, the effect can be represented as an aggregation effect, in that it introduces an 'independent' subindustry whose mean firm size is small relative to that of the main industry. This will cause a bias of the form noted in Proposition 3(a): the bound is valid but not tight.

Statistical procedures in some countries deal with the problem of fringe firms by applying a standard cutoff level for firm size, including only firms with at least 20 employees, say, in the reported figures. This device is an imperfect one; it will reduce the bias associated with inappropriate inclusions at the expense of introducing a reverse bias associated with inappropriate exclusions.

The exclusion of firms causes the reported value of N to be lower than the true value and this will lead to an inappropriately high estimate of $\underline{C}_k$ using Proposition 1. Hence we might observe a violation of the proposed bound. Where a cutoff level is used, therefore, we cannot say anything *a priori* as to the net direction of bias in reported ($C_{k/N}$,N) values relative to Proposition 1. The direction of the bias is known when the reporting of N is complete, but is

indeterminate if a cutoff size is used.

These problems with the measurement of N are sufficiently serious to warrant placing particular emphasis on the conditional predictions provided by Proposition 2, which do not require the use of N values.[23]

VI. EMPIRICAL EVIDENCE I: PRODUCT LEVEL DATA

Since the theory relates to the size distribution of *businesses*, the appropriate context in which to test the predictions of Propositions 1 and 2 is that of individual product markets; in terms of U.S. data, this corresponds to the 5-digit SIC level. Data for 4, 8, 20 and 50 firm concentration ratios is available at this level, but no data is available for the number of firms active in each market. (Asset concentration ratios are unavailable at this level.) At the 4-digit or 'industry' level, data is available both for 4, 8, 20 and 50 firm concentration ratios and for the number of firms active in each market.

Few other countries report a wide range of sales-concentration ratios even at the 4-digit level. Figures are available for Germany however, at a level slightly more aggregated than the U.S. 4-digit level, for 3, 6, 10, 25 and 50 firm ratios. The number of firms is also reported, though subject to a rather high cutoff size (20 employees).

A comparison of U.S. and German experience is of particular interest in the

---

[23]It is also worth noting that the aggregation problem is also eased in this setting. To see this, imagine an industry consisting of two subindustries of widely differing mean firm size. Since Proposition 2 deals only with the upper tail of the distribution, and since the top k firms will, for sufficiently large k, be drawn predominantly from the sub-industry with the higher mean firm size, it follows that the (C_k,N) values will again (approximately) satisfy Proposition 1.

present context, since a central claim of this approach is that little can be said about 'average' or 'typical' size distributions, but that the lower bound relationship should be stable in spite of possibly wide fluctuations in 'average' experience. It is therefore of particular interest to compare experience in economies between which average experience differs widely. International comparisons of concentration levels regularly note the fact that U.S. levels are relatively high, and German levels relatively low.[24] In what follows, we begin with an examination of U.S. data at the 5-digit level (at which German data is unavailable), and then turn to a comparison of U.S. and German experience at the 4-digit level.

As explained earlier, we focus attention in what follows on those industries in which neither Advertising nor R&D play a major role. Since we wish to define a corresponding set of industries for the U.S. and Germany, which use different industry definitions at lower levels of aggregation, this is done by confining attention to those 2-digit industry groups in which advertising and R&D intensities are very low. In what follows, we report results for the following set of industry groups:

20	Food and Drink (low advertising industries only)[25]
22	Textiles
23	Clothing
24	Lumber
25	Furniture
26	Paper
27	Printing (excluding 2711, 2721, 2731)

[24]Some part of this difference may be attributed to differences in industry definition (levels of aggregation), but substantial differences remain present even in data sets consisting of closely matched pairs of industries (see Sutton (1991) for an example.)

[25]The Food and Drink sector divides into two sets of (4-digit) industries, in one of which advertising intensity is high; and one in which it is very low (advertising/sales < 1%; Sutton (1991), Chapter 5). Markets falling within the latter set of industries are included in the present dataset.

31 Leather
32 Stone, Clay and Glass (excluding 3211, 3229)
33 Primary Metals
34 Metal Products

The Conditional Prediction

We first examine the conditional prediction of Proposition 4 for U.S. 5-digit data. Proposition 4 predicts a lower bound to C_k, for k = 4, 8 and 20, as a function of C_{50}. Figure 3 shows a scatter diagram in which each point represents a single 5-digit industry in 1977. The value of C_{50} is plotted on the horizontal axis, and the value of C_4 on the vertical axis. The solid curve shows the lower bound $D_k(C_{50})$ predicted by Proposition 2. The lowest possible value for C_k, given C_{50}, is attained by a distribution in which all firms have the same size, and this corresponds to the ray from the origin $C_k = (k/50)C_{50}$ which lies below this curve. (At the other extreme, the highest possible value for C_k, given C_{50}, is $C_k = C_{50}$, corresponding to the diagonal). Figure 4 shows a histogram of the differences between the observed value C_4 and the predicted lower bound $D_k(C_{50})$. The model predicts that this histogram should lie wholly above zero, and if the bound is 'tight', we should see a sharp cutoff at zero. (At the upper end, we expect the histogram to fade out slowly, with no suggestion of any 'upper bound'.)

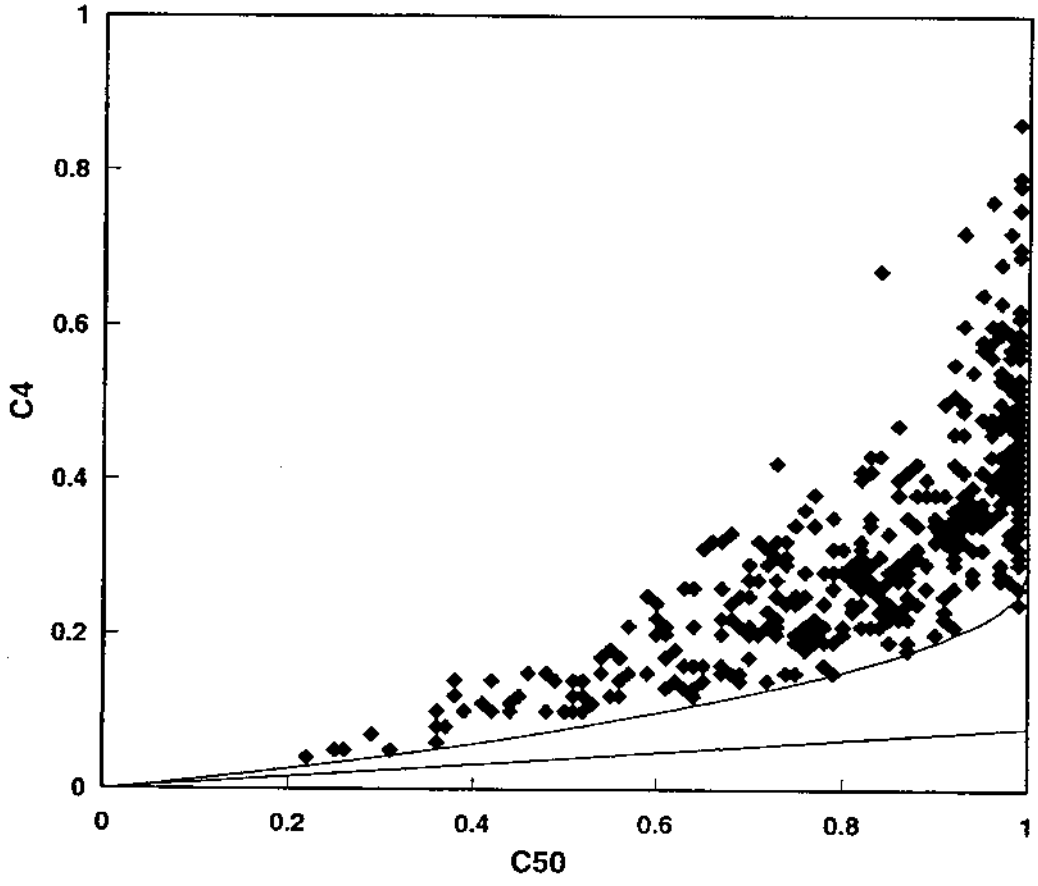

Figure 3. Testing Proposition 4 for U.S. data at the 5-digit level, 1977: a scatter diagram of C_4 versus C_{50}. The solid curve shows the lower bound $D_k(C_{50})$ predicted by Proposition 4. The ray shown below this curve corresponds to the symmetric equilibrium in which all firms are of equal size. (This data is not available for the 1987 Census.)

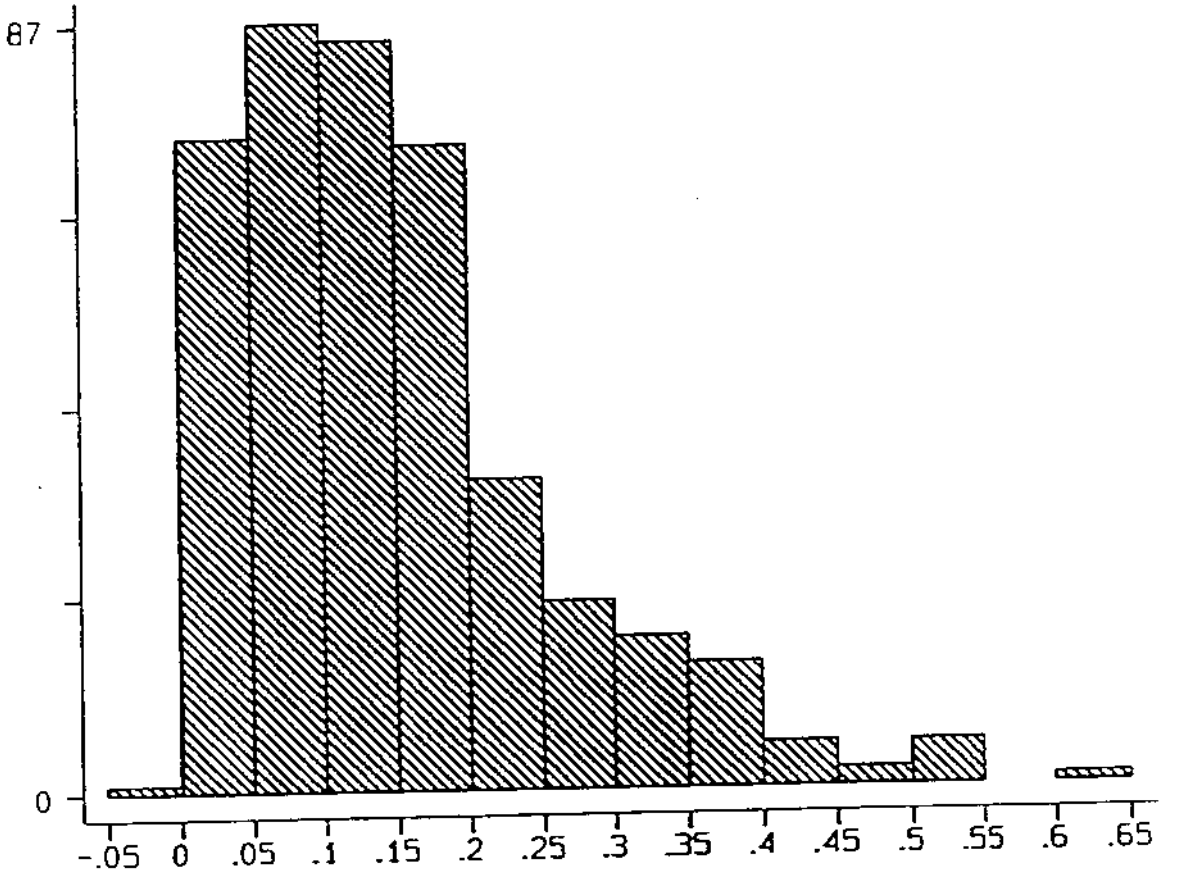

Figure 4. A histogram of differences between the actual concentration ratio C_k and the predicted lower bound $D_k(C_{50})$ for the data shown in Figure 3.

A summary measure of the performance of the prediction for various concentration ratios is shown in Table 3, which records the fraction of observations which violate the lower bound, and the fraction of observations lying above, but within 5 percentage points of, the bound. Table 3 also records the equivalent results for U.S. data at the 4-digit level, and for the corresponding German dataset.

It appears that the prediction of Proposition 1 performs reasonably well. While violations do occur, they are rare; moreover, the cutoff at the lower bound seems sharp. The form of the histogram in Figure 4 is of particular relevance. This is not suggestive of a more-or-less symmetric distribution of 'residuals' about some 'true model'. Rather, the strongly asymmetric shape of the histogram is suggestive of a bounds relation.

Dataset	k	No of Obs	Within 5%	Below
U.S. 5-digit 1977	4	420	0.176	0.002
	8	421	0.109	0.000
	20	420	0.112	0.000
U.S. 4-digit 1987	4	205	0.146	0.000
	8	207	0.082	0.000
	20	205	0.083	0.000
German (≅4 digit) 1990	3	80	0.350	0.063
	6	80	0.263	0.038
	10	80	0.238	0.025
	25	80	0.313	0.013

Table 3. The conditional prediction for U.S. 5-digit data (top panel). The column 'Within 5%' shows the fraction of C_k values which lie above, but within five percentage points of, the predicted lower bound $D_k(C_{50})$. The column 'Below' shows the fraction of data points which violate the lower bound. (U.S. 5-digit data was not reported in the 1987 Census.) The second and third panels show equivalent results for U.S. and Germany at the 4-digit level.

The Lorenz Curve

We now turn to an examination of the Lorenz Curve predicted by Proposition 1. Here, problems of aggregation as well as associated problems related to the measurement of N play an important role.

In comparing the U.S. and German data-sets in what follows, the related problems of aggregation and of measuring firm numbers will be evident[26]. Two points relating to the datasets are relevant. Firstly, the U.S. data includes all firms, while the German census adopts a high cutoff value, counting only firms with more than 20 employees. Following the results reported in Section V, we therefore expect 4-digit data for the U.S. to be biased upwards from the lower bound (i.e. to be less tight). As to the German data, however, the presence of the cutoff value in reported firm numbers means that the net direction of bias is indeterminate.

In Figure 5, we show the pooled data sets for the U.S. in 1987 and for Germany in 1990. Here, we take advantage of the implications of Proposition 1 to pool all available concentration ratios on the same figure. In Figure 5, each 4-digit industry is represented by a set of points, one for each available concentration ratio. For the U.S. data set, shown in the top panel, each industry that has over 50 firms is represented by three points (k = 4, 8 and 20). The horizontal axis shows the fraction k/N of firms to which the reported ratio corresponds, while the vertical axis shows the reported ratio C_k for 1987. The bottom panel shows

---

[26]The ideal test of the claims in Proposition 3 regarding aggregation problems would lie in comparing the scatter of observations (C_k,N) with the predicted Lorenz Curve for data collected for the same country at two different levels of aggregation. This direct test is, unfortunately, not possible for either the U.S. or Germany. While both 4 and 5-digit C_k data is published for the U.S., the number of firms active in each product market is not recorded at the 5-digit level; while for Germany, data is published only at one level of aggregation.

the corresponding scatter for Germany in 1990.

A comparison of these figures shows that, in spite of substantial differences in average concentration levels, both datasets appear to conform well to the predicted lower bound[27].

Overall, the bound specified in Proposition 1 appears to perform reasonably well, in that violations of the bound are infrequent, and the bound is fairly tight. The fact that the bound is tighter in the German dataset is consistent with the predicted consequences of aggregation and measurement problems. In spite of wide differences in average concentration levels between the two countries, the lower bound appears to be closely similar in both cases[28].

Beyond these remarks, it is difficult to make any precise claim regarding the goodness of fit of the bound. It might seem of interest to estimate confidence intervals around the predicted lower bound, and to compare observed deviations with this. However, such confidence intervals would relate to a population of firms *all* of which evolved according to the limiting process (Condition 1'). The present hypothesis is that different industries will evolve according to different processes each satisfying Condition 1; and the expected proportion of points lying below the bound is therefore unspecified by the theory. Confidence intervals calculated using Condition 1' could at best only provide an upper limit to the number of violations expected under the theory.

---

[27]Data for the U.S. is available over a long period, beginning in 1947. An examination of the corresponding data for each census year from 1947 - not shown here - indicates that, in spite of substantial changes in overall industrial structure, the lower bound has displayed considerable stability over time.

[28]The present tests have been extended to all 4-digit industries. The only substantial violation of the bound occurs for the Carbon Black industry in the U.S.. It is interesting to note that this industry has been steadily declining over the past 50 years (see footnote 21).

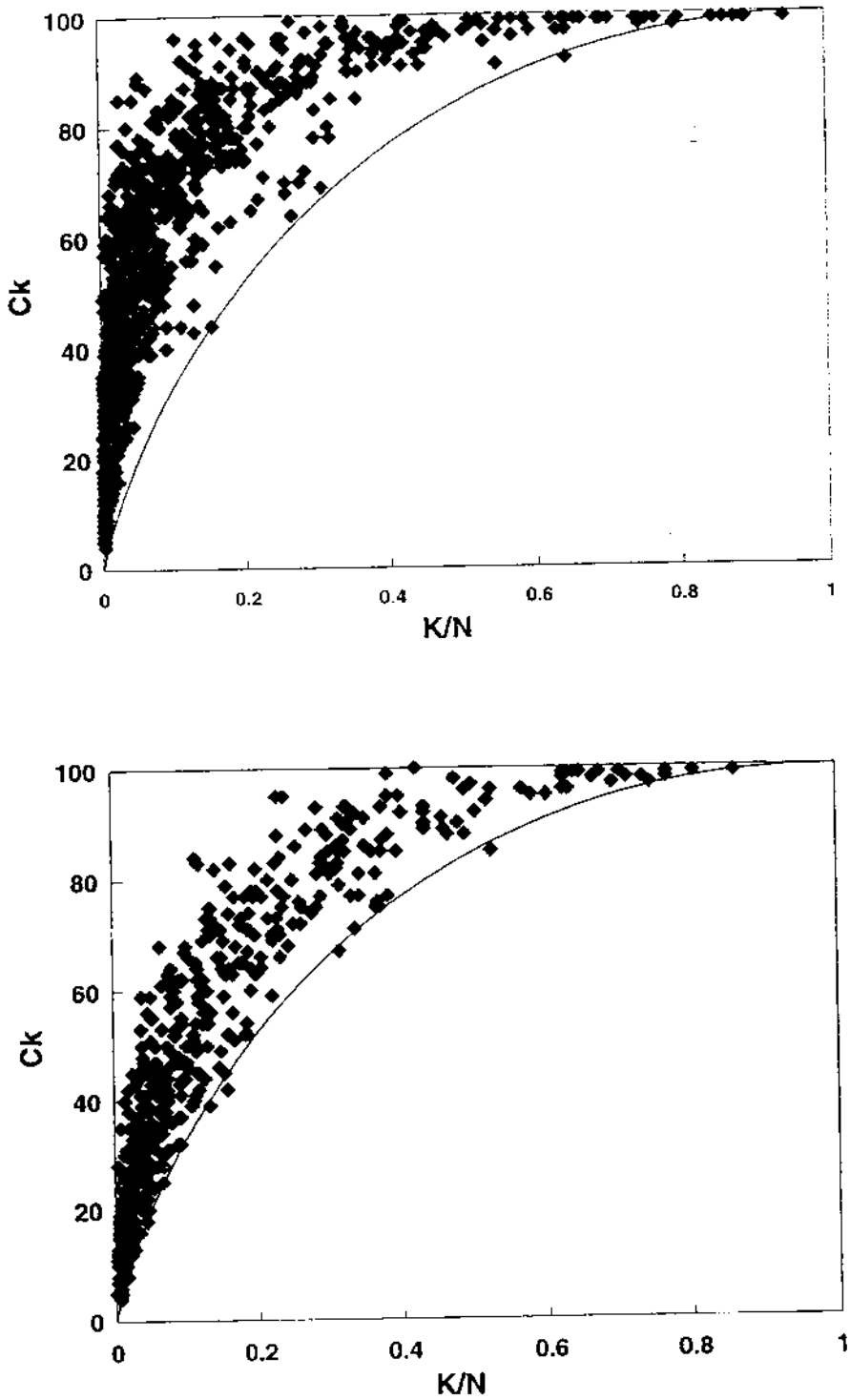

Figure 5. Testing Proposition 1. The top panel shows the scatter diagram of C_k against k/N for pooled data (k = 4, 8 and 20) for the U.S., 1987, at the 4-digit level. The Lorenz curve is the limiting curve defined by Proposition 1 (equation (11)'). The bottom panel shows data for Germany, 1990 (k = 3, 6, 10 and 25).

Rather then consider the goodness of fit of this bound, viewed in isolation, we prefer to take it as providing a potentially useful null hypothesis against which richer models, and in particular strategic models, can be tested.

VII. CONCLUDING REMARKS

The motivation for this paper lies in the claim that any adequate theory of market structure will need to encompass two aspects of the problem: the role of strategic factors, and the part played by 'independence effects'. One noteworthy feature of the existing literature is that it contains two traditions that focus on different aspects of market structure, and that use different mathematical approaches. The modern literature based on multistage games is directed towards the 'Bain' tradition which focuses on explaining cross-industry differences in concentration be reference to industry-specific factors. This literature emphasies strategic factors, but while it makes some testable claims as to concentration ratios, it has nothing of interest to say about the shape of the size distribution. If a market consists of a large number of independent submarkets, this kind of model just says that there will be many equilibria, corresponding to different size distributions. In some cases, these may include equilibria where all firms are of the same size. Such outcomes are rare, or nonexistant, in practice. The older Growth-of-Firms literature on the other hand, focused on the shape of the size distribution and gave a central role to stochastic factors that might contribute to skewness, but ignored strategic effects. Recent attempts to re-work this approach to encompass various 'economic' mechanisms offer one way of building a bridge between the two traditions. A recent survey of this literature concluded, however, that it appeared difficult to place any interesting restrictions on the form of the size distribution, once the model was enriched in this way (Schmalensee (1989)).

The present paper introduces a different way of building a bridge between the

two traditions. This approach begins by placing a bound on the degree of skewness that might result from 'independence effects' and cost considerations alone. With this as a benchmark, it may prove easier to develop testable predictions regarding the influence of strategic factors.

The usefulness of this benchmark will depend on whether we can identify an appropriate domain for the model, i.e. a set of industries in which the 'least skew distribution' defined by the model is sometimes attained, but rarely violated. In terms of the Lorenz curve representation, we need a limiting Lorenz curve that is often attained ('tight') but is seldom exceeded by a substantial margin. The importance of having both these features, is that some strategic factors ('externatlities') may cause a shift of the Lorenz curve towards the diagonal while others ('escalation') may cause it to shift away from the diagonal. The empirical results reported above suggest that the present bound may be satisfactory enough on both these counts to provide us with a usable benchmark.

In a companion paper, Sutton (1995b), we first re-cast the present results in a game-theoretic setting, and then extend the model to include strategic factors.

APPENDIX 1.

Mathematical Notes

1. Deriving Equation (4)

The state $N_{t+1} = N$ can be entered from state $N_t = N$ or from state $N_t = N-1$. The associated probabilities are:

N(t) = N & 'No Entry': $\binom{t-1}{N-1} p^{N-1} (1-p)^{t-N} \times (1-p)$

N(t) = N-1 & 'Entry': $\binom{t-1}{N-2} p^{N-2} (1-p)^{t-N+1} \times p$

Note that these two expressions sum to give the unconditional probability

$$\text{Prob}(N_{t+1} = N) = \binom{t}{N-1} p^{N-1} (1-p)^{t-N+1}$$

It follows that

$$\text{Prob}\ (N_t = N \mid N_{t+1} = N) = \frac{\binom{t-1}{N-1}}{\binom{t}{N-1}} = 1 - \frac{N-1}{t}$$

$$\text{Prob}\ (N_t = N-1 \mid N_{t+1} = N) = \frac{\binom{t-1}{N-2}}{\binom{t}{N-1}} = \frac{N-1}{t}$$

2. Checking equation (7a,7b)

We first consider the case i = 1. Here, (suppressing the time subscript on N_t to ease notation) equation (7) reduces to

$$\frac{1}{N} E(n_{1,t} \mid N) = \frac{N-1}{t-1} \qquad (7a)$$

We aim to show that this satisfies equation (5). Substituting (7a) on the r.h.s. of equation (5) yields:

$$\frac{N-1}{t}\left\{E(n_{1,t}|N-1)+1\right\}+\left(1-\frac{N-1}{t}\right)\left\{E(n_{1,t}|N)-\frac{E(n_{1,t}|N)}{N}\right\}$$

$$=\frac{N-1}{t}\left\{\frac{(N-1)(N-2)}{t-1}+1\right\}+\frac{t-(N-1)}{t}\left\{\frac{N-1}{N}\cdot\frac{N(N-1)}{t-1}\right\}$$

$$=\frac{N(N-1)}{t}=E(n_{1,t+1}|N)$$

We now turn to the case $i > 1$. From equation (7) we have:

$$E(n_{i,t}|N) = N\binom{t-i-1}{t-i+1-N}\Big/\binom{t-1}{N-1}$$

$$E(n_{i,t}|N-1) = (N-1)\binom{t-i-1}{t-i+2-N}\Big/\binom{t-1}{N-2}$$

$$E(n_{i-1,t}|N) = N\binom{t-i}{t-i+2-N}\Big/\binom{t-1}{N-1}$$

We aim to show that equation (6) is satisfied.

The r.h.s. of equation (6) is:

$$\frac{N-1}{t}\left\{E(n_{i,t}|N-1)\right\}+\left(1-\frac{N-1}{t}\right)\left\{E(n_{i,t}|N)-\frac{E(n_{i,t}|N)}{N}+\frac{E(n_{i-1,t}|N)}{N}\right\}$$

Inserting the above expressions, and extracting the common factor

$$\frac{(t-i-1)!\,(t-N)!}{(t-i+2-N)!\,(t-1)!}=x, \text{say}$$

from each term, this reduces to:

$$x\left\{\frac{(N-1)^2(N-2)}{t}(t-N+1)+\frac{t-N+1}{t}(N-1)^2(t-i+2-N)+\frac{t-N+1}{t}(N-1)(t-i)\right\}$$

which in turn reduces to:

$$x\left\{\frac{N(N-1)}{t}(t-N+1)(t-i)\right\}$$

Similarly, the l.h.s. of (6) reduces to:

$$E(n_{i,t+1}|N) = N\binom{t-i}{t-i+2-N}\Big/\binom{t}{N-1} = x\left\{\frac{N(N-1)}{t}(t-N+1)(t-i)\right\}$$

APPENDIX 2.

The General Process

In the text, the process was analysed for the special case of Condition 1', in which each active firm has an equal probability of capturing the next opportunity. Here, we examine the general case corresponding to Condition 1, in which this probability is nondecreasing in firm size. Depending on the relationship between firm size and the probability of capture, it may not be the case that $\frac{1}{t}E(n_{i,t})$ tends to a limiting value. Here, we do not discuss conditions for convergence, but simply develop a characterization result which specifies the relevant properties of a stationary distribution of firm size, if such a distribution exists.

We define the function,

$$g_i(t) = \frac{E(n_{i,t})}{1+p(t-1)}$$

where $E(n_{i,t})$ denotes the (unconditional) expectation of n_i at stage t.

Let $Prob_i(t)$ denote the probability that an opportunity which is captured by some incumbent at stage t is captured by a firm of size i.

We now characterise the properties of $g_i(t)$ at $t \to \infty$, on the assumption that there are two sequences of constants $\{\bar{g}_i\}$ and $\{\bar{\pi}_i\}$ such that

$$\begin{aligned} \lim_{t\to\infty} g_i(t) &= \bar{g}_i \\ \lim_{t\to\infty} Prob_i(t) &= \bar{\pi}_i \end{aligned} \qquad \text{(A1)}$$

Condition 1 implies that the ratio $\bar{\pi}_i/\bar{g}_i$ is nondecreasing in i. Write $\bar{\pi}_i/\bar{g}_i$ as ϕ_i, where ϕ_i is a nondecreasing sequence. Consider the behaviour of $n_i(t+1)$ for $i \geq 2$. This takes the value $n_i(t) + 1$ with probability $(1-p)Prob_i(t)$; the value $n_i(t) - 1$ with probability $(1-p)Prob_{i-1}(t)$; and the value $n_i(t)$ otherwise. It follows that for all $i \geq 2$ we have for any fixed t:

$$E(n_{i,t+1}) = E(n_{i,t}) - (1-p)Prob_i(t) + (1-p)Prob_{i-1}(t) \qquad \text{(A2)}$$

Now from the definition of $g_i(t)$, it follows that:

$$\begin{aligned} E(n_{i,t+1}) - E(n_{i,t}) &= (1+pt)g_i(t+1) - (1+p(t-1))g_i(t) \\ &= (1+pt)(g_i(t+1) - g_i(t)) + pg_i(t) \end{aligned}$$

whence from (A1) it follows that

$$\lim_{t\to\infty}[E(n_{i,t+1}) - E(n_{i,t})] = p\bar{g}_i$$

Similarly

$$\begin{aligned} &\lim_{t\to\infty}[Prob_i(t) - Prob_{i-1}(t)] \\ &= \bar{\pi}_i - \bar{\pi}_{i-1} = \phi_{i-1}\bar{g}_{i-1} - \phi_i\bar{g}_i \end{aligned}$$

Taking limits in equation (A2), then, we have

$$p\bar{g}_i = (1-p)(\phi_{i-1}\bar{g}_{i-1} - \phi_i\bar{g}_i)$$

or

$$\frac{\bar{g}_i}{\bar{g}_{i-1}} = \frac{(1-p)\phi_{i-1}}{p+(1-p)\phi_i} = \frac{(1-p)\phi_{i-1}}{p+(1-p)\phi_i} \quad , \quad i \geq 2 \qquad \text{(A3)}$$

Following a similar argument for i = 1, we have

$$En_1(t+1) = En_1(t) + p - (1-p)\,Prob_1(t)$$

whence by the same argument we obtain

$$p\bar{g}_1 = p - (1-p)\phi_1\bar{g}_1$$

or

$$\bar{g}_1 = \frac{p}{p+(1-p)\phi_1} \qquad \text{(A4)}$$

Equations (A3) and (A4) describe the limiting distribution. For the special case analysed in the text, we have $\phi_i = 1$ for all i and so (A3) and (A4) coincide with the geometric distribution specified by equation (8b) of the text. Denote this special (geometric) density as $\bar{f}_i$.

Since $\bar{g}_i$ and $\bar{f}_i$ are densities,

$$\sum_{i=1}^{\infty} i\bar{g}_i = \sum_{i=1}^{\infty} i\bar{f}_i = 1 \qquad \text{(A5)}$$

Moreover, since a new firm enters with probability p each period, it follows that the mean firm size equals 1/p, as in the basic process, whence

$$\sum_{i=1}^{\infty} \bar{g}_i = \sum_{i=1}^{\infty} \bar{f}_i = 1/p \qquad \text{(A6)}$$

A convenient way of comparing the general density $\bar{\bar{g}}_i$ with the geometric density $\bar{f}_i$ is as follows: the constants ϕ_i can be interpreted as a set of weights attached to firms of different sizes, which determine their relative probabilities of capturing the next opportunity. The process is obtained by introducing some nondecreasing sequence of weights ϕ_i to the equal probabilities assigned in the basic process. Any increasing sequence ϕ_i can be reached by successive multiplication of the $\{\bar{f}_i\}$ by a sequence of step functions ϕ_i^m, for $m \geq 2$, which increase the probability for firms of size greater than or equal to m, while lowering it for firms below that size, viz.

$$\phi_i^m = a < 1 \quad , \quad i < m$$
$$= b > 1 \quad , \quad i \geq m$$

It is clear from inspection of (A3) and (A4), and recalling (A5), (A6), that applying this step function to any $\bar{\bar{g}}_i$ generates a new distribution $\bar{\bar{g}}_i'$ with the following properties: $\bar{\bar{g}}_i' > \bar{\bar{g}}_i$ and $\bar{\bar{g}}_i'$ crosses $\bar{\bar{g}}_i$ at two points. In other words, the operator ϕ_i^m shifts weight to the tails of the distribution, and so moves the corresponding Lorenz curve further from the diagonal.

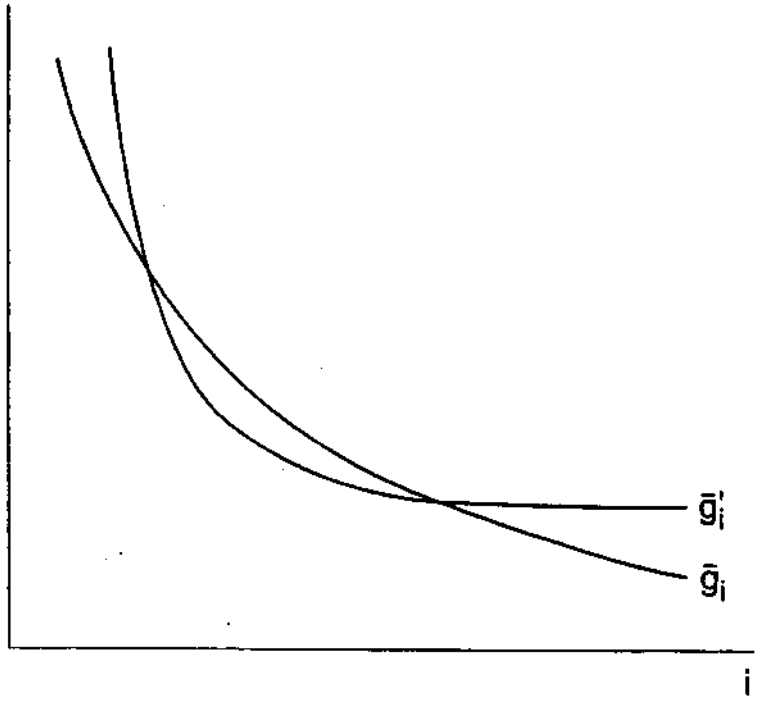

Figure A1. The effect of the operator ϕ_i^m on the density $\bar{\bar{g}}_i$.

This argument can be extended to the case of the conditional lower bound specified in Proposition 2 of the text. To see this, suppose the m-firm concentration ratio is known, but not the number of firms.

In Figure A2, the heavy curve L is a (rescaled) Lorenz curve, which shows the fraction of plants owned by the top N firms; the (absolute) number of firms N is shown on the horizontal axis. We denote by N_{true} the (unobserved) true number of firms. Say we have an observed value for the m-firm concentration ratio $C_m = y$. Then define $\hat{N}$ implicitly using

$$y = \frac{m}{\hat{N}}\left(1 - ln\frac{m}{\hat{N}}\right) \qquad (A7)$$

The value $\hat{N}$ thus defined is shown in Figure A2. Also shown is the Lorenz curve E for the corresponding exponential distribution; this defines C_k as a function of N for all k on $0 \leq N \leq \hat{N}$, viz.

$$C_k(N) = \frac{k}{\hat{N}}\left(1 - ln\frac{k}{\hat{N}}\right)$$

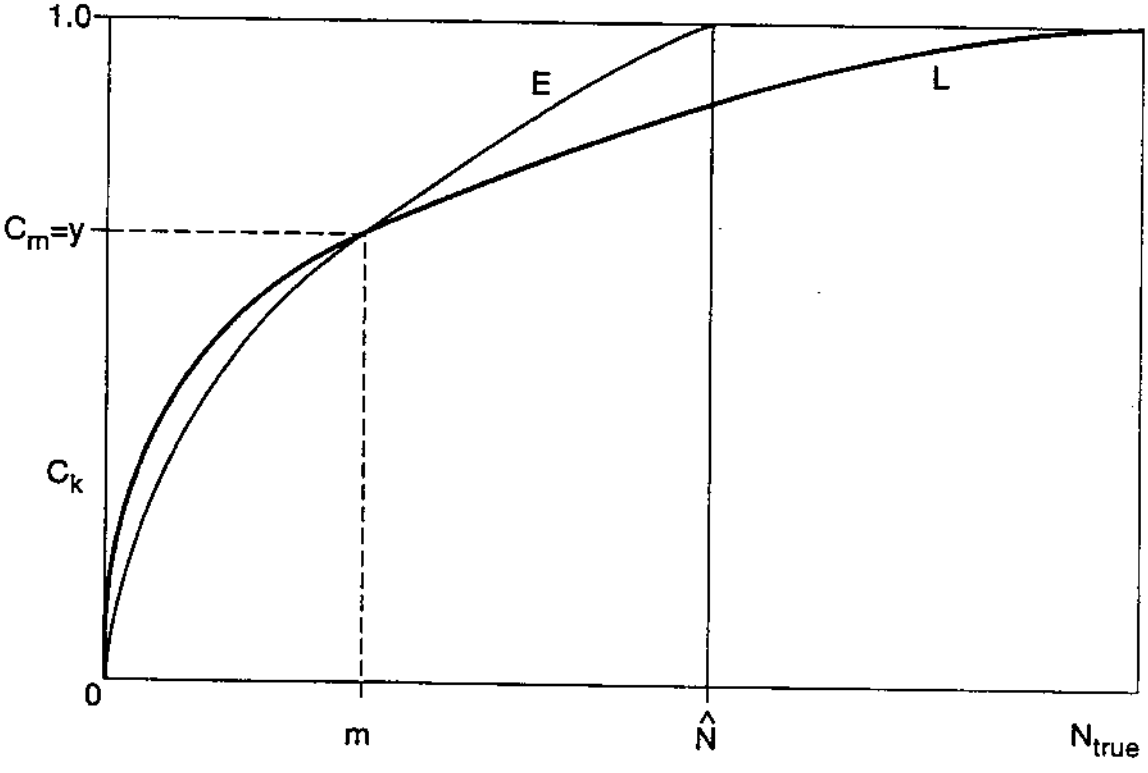

Figure A2. The Conditional Lower Bound.

This curve coincides with the true (rescaled) Lorenz curve at the origin and at the point (m,y). The size distribution to which it corresponds is an exponential density. Hence the crossing property (Figure A1) ensures that L lies wholly inside E within the box $0 \leq N \leq m$, $0 \leq C_k \leq y$. It follows that curve E defines a lower bound to C_k for all $k \leq m$.

APPENDIX 3.

Aggregation Effects

Independent Sub-industries

Consider two independent industries i = 1, 2, whose evolution is described by Conditions 1' and 2 of the text. Denote the total number of opportunities arising in industry i as T_i. Let industry i have entry parameter p_i. The size distribution of businesses in each sub-industry converges to an exponential $f_i(x)$ with mean $\mu_i = 1/p_i$.

The density function describing the size distribution for the 'industry' is now a weighted average of the exponential densities for each industry, viz.

$$f(x) \rightarrow \frac{\sum_i T_i f_i(x)}{\sum_i T_i}$$

If the parameter p_i is the same for all sub-industries, f(x) is exponential. Otherwise, it is a distribution of the type described in Appendix 2 (Figure A1), whose Lorenz curve lies further from the diagonal than that described by Proposition 1.

Interdependent Sub-industries

Consider an industry which comprises two product markets. The industry evolves over time as follows: we begin with one active firm which produces one product variety; label that product as x and the second product as y. At each date t = 1, 2, 3 ... one new product is entered. With probability φ, this is an X type product and with probability $(1-\varphi)$ it is a Y type product. This new product is entered by a new entrant to the industry with probability p. With probability (1-p) it is entered by an active firm.

We assume here that all firms active in the industry - whether or not they currently produce in product market X - have an equal probability of introducing the new X product or plant; and likewise for Y. ("Interdependent Industries") Hence if there are N active firms in the industry at time t, then each of these has an equal probability of introducing the next product introduced by an incumbent.

Now the measured size of a firm is described by the total number of products it offers, i.e. by (x+y). But the evolution of (x+y) is described exactly by the model set out in the text; so Proposition 1 applies.

APPENDIX 4.

The Lorenz Curve

It was noted in the text that when T, and so N, is large, the properties of the lower bound follow from the standard properties of the extreme value distribution for the exponential (Gumbel (1958), p.116ff.)

The mean of the m-th smallest value among N draws is given by

$$\bar{x}_m = \sum_{N-m+1}^{N} \frac{1}{i}$$

and so the expected value of the sum of the sizes of the k largest firms equals

$$\sum_{1}^{N} \frac{1}{i} + \sum_{2}^{N} \frac{1}{i} + \ldots + \sum_{k}^{N} \frac{1}{i} = k + k \sum_{k+1}^{N} \frac{1}{i}$$

Hence, for a given N, the expected value of the k-firm concentration ratio is

$$\frac{k}{N}\left(1 + \sum_{k+1}^{N} \frac{1}{i}\right)$$

Given the form of the expression in the summation sign, it is natural to express this in terms of *ln*(k/N). For k = 1, expression (10) is asymptotically equal to

$$C_{1/N} = \frac{1}{N}\left(\gamma + ln\frac{1}{N}\right)$$

where γ is Euler's constant (= .577 ...)(Gumbel (1958), p. 116).

For $k \neq 1$, it will be convenient to define γ_k implicitly by

$$\frac{k}{N}\left(1 + \sum_{k+1}^{N} \frac{1}{i}\right) = \frac{k}{N}\left(\gamma_k - ln\frac{k}{N}\right)$$

Computed values of γ_k are shown in Table A1 for those values of k that are commonly reported in official statistics. The asymptotic result of Proposition 1 corresponds to the case where k is large.

k	1	4	6	8	10	20	50
γ_k	.577	.880	.919	.939	.951	.975	.990

Table A1. Values of γ_k

REFERENCES

Bain, J., (1966), International Differences in Industrial Structure: Eight Nations in the 1950s, CT: Greenwood Press

Dunne, T., Roberts, M. and Samuelson, L., (1988), 'Patterns of Firm Entry and Exit in U.S. Manufacturing Industries', Rand Journal of Economics, XIX, pp.416-515.

Ericson, R. and Pakes, A., (1995), 'Markov-Perfect Industry Dynamics: A Framework for Empirical Work,' Review of Economic Studies, vol. 62, pp. 53-82.

Evans, D., (1987b), "The Relationship Between Firm Growth, Size and Age: Estimates for 100 Manufacturing Industries," Journal of Industrial Economics, vol. 35, pp. 567-581.

Ghemawat, P. and Nalebuff, B., (1990), 'The Devolution of Declining Industries, Quarterly Journal of Economics, 420, pp. 167-186.

Gibrat, R. (1931), Les Inégalités Économiques. Applications: Aux Inégalités des Richesses, a la Concentration des Entreprises, Aux Populations des Villes, Aux Statistiques des Familles, etc., d'une Loi nouvelle: La Loi de L'Effet Proportionnel, Paris: Librairie du Recueil Sirey.

Gumbel, E.J., (1958), Statistics of Extremes, Columbia University Press.

Ijiri, Y. and Simon, H., (1964), 'Business Firm Growth and Size', American Economic Review, 54, pp.77-89.

Ijiri, Y. and Simon, H., (1977), Skew Distributions and the Sizes of Business Firms, Amsterdam: North-Holland Publishing Co.

Jovanovich, B., (1982), 'Selection and Evolution of Industry', Econometrica, 50, p.649-70.

Lucas, R.E. (1978), 'On the Size Distribution of Business Firms,' Rand Journal of Economics, vol. 9, no. 2, pp. 508-523.

Klepper, S. and Graddy, E., (1990), 'The Evolution of New Industries and the Determinants of Market Structure', Rand Journal of Economics, 21, pp. 27-44.

Rao, C.R., (1973), Linear Statistical Inference and its Applications, Chichester: Wiley.

Scherer, F.M., (1980), Industrial Market Structure and Economic Performance, (2nd ed.), Chicago: Rand McNally.

Schmalensee, R., (1989), 'Inter-Industry Studies of Structure and Performance,' in R. Schmalensee and R. Willig (eds.), Handbook of Industrial Organisation, volume 2, Oxford: North Holland.

Shaked, A. and Sutton, J., (1987), 'Product Differentiation and Industrial Structure', Journal of Industrial Economics, XXXVI, pp. 131-146.

Simon, H. and Bonnini, C., (1958), 'The Size Distribution of Business Firms', American Economic Review, vol 48.

Sutton, J., (1991), Sunk Costs and Market Structure, MIT Press.

Sutton, J., (1995a), "Gibrat's Legacy," London School of Economics, unpublished.

Sutton, J., (1995b), 'The Size Distribution of Businesses: Part II: A Game-Theoretic Model', STICERD Discussion Paper No. EI/10, Economics of Industry Group, STICERD, London School of Economics.