ECONOMIC RESEARCH REPORTS

N-PRODUCT NATURAL MONOPOLY AS "NATURAL CARTEL" -- ON SCALE ECONOMIES UNDER CAPITAL RATIONING

by
William J. Baumol
and
Ralph E. Gomory

RR # 95-18 **July 1995**

C. V. STARR CENTER FOR APPLIED ECONOMICS

NEW YORK UNIVERSITY
FACULTY OF ARTS AND SCIENCE
DEPARTMENT OF ECONOMICS
WASHINGTON SQUARE
NEW YORK, N.Y. 10003

N-PRODUCT NATURAL MONOPOLY AS "NATURAL CARTEL" -- ON SCALE ECONOMIES UNDER CAPITAL RATIONING[1]

William J. Baumol and Ralph E. Gomory*

*Director, C.V. Starr Center for Applied Economics, New York University and President, The Alfred P. Sloan Foundation, respectively. The authors are very grateful for the support of the Sloan Foundation and the C.V. Starr Center for their support of the research reported here.

It is a great delight to take advantage of this opportunity to dedicate this paper to Richard Quandt. Besides being a scholar of very great creativity he is also a very good friend. For the rest of us, his retirement comes much too soon, and we trust it is largely a formality.

The paper that follows seems appropriate for the occasion in two respects. It follows up on a subject that many years ago one of the current authors had explored together with Richard Quandt (Baumol and Quandt, 1965), and it is an offshoot of work in a different area of economics to which the two current authors have dedicated themselves as their primary analytic activity for the past several years.

I. The Issue and Implications of the Analysis.

The model we will describe here deals with an extreme case, but we believe that if the premises are weakened somewhat to bring them closer to reality the implications either remain largely unchanged or are easily modified where appropriate. We deal with an n-product industry characterized by strong scale economies that persist throughout the relevant output ranges. The industry consists of two firms, both of whose activities are circumscribed by one effective constraint: the limited capital resources available to each enterprise. We will later indicate why the capital rationing constraint is not as artificial as it may appear at first glance and why, in many cases, it seems to be a reasonable approximation to reality.

It is well known that a world of scale economies is one in which industries are natural monopolies for reasons that hardly need review. But this does not mean that such an industry, loosely defined, will necessarily consist of a single firm. If the scale economies are not accompanied by economies of scope so that there is no cost disadvantage in specialization in one or a limited number of products, the industry can

[1] This article is a first application to an issue in industrial organization of a method of analysis originally formulated by Ralph Gomory and previously applied by the present authors to issues in the theory of international trade. See, e.g., Gomory (1994), Gomory and Baumol (1994) and Baumol and Gomory (1994)

be composed of a number of firms, each the sole producer of some subset of the industry products. That is, each individual product of the industry will be supplied by a monopolist, but in equilibrium firms can coexist if no two of them undertake to produce the same commodity. Thus, the bottom line is that market forces will push the industry into the form of a cartel in which each firm is assigned its exclusive (product) territories. We see that, in addition to such recognized forms of industry structure as "natural monopoly" and "natural oligopoly" the market is capable of creating a "natural exclusive-territory cartel," whose attributes will be explored here.

If each monopolistically-supplied product offers incremental profits, then each firm in the industry will have an incentive to supply as many of the products as possible. However, effective capital rationing changes this. The firm can take on production of another commodity only by withdrawing resources from somewhere else. Either it must reduce its output of some of its other products, or it must abandon one or more products altogether. This means that if the firm were the only enterprise in the industry, profit-maximization (or loss minimization) would require it to pick and choose among the alternative goods to which it can devote itself, seeking to determine that product mix that yields the highest profits to the enterprise.

In a multi-firm industry, however, a firm, call it firm J, is not free to choose its product lines. For some other firm may have beaten it to the production of some commodity, I, that J would like to supply. It is then apt to be very difficult for J to invade the field. For scale economies give the incumbent monopolist a substantial cost advantage over any entrant, unless the entrant is prepared from its date of entry to come close to matching the scale of output of the incumbent. In addition, there is no reason to assume that the firms in the industry have identical production functions and so it is at least possible that firm J will have another handicap in its pursuit of product I.

We will find that in fact any way of dividing up the set of n products between the two firms, giving each firm a monopoly of any good in its output set is a prospective equilibrium, and that there are altogether 2^n-2 such prospective equilibria altogether, a number that obviously grows far faster than the number of products in the industry. Further, we will find that a large proportion of these prospective equilibria are locally stable, that is, a small deviation from any one of them will elicit pressures making for a return to that initial equilibrium. We will also provide a graphic representation of any equilibrium in what we can call "profit-product-share space." This representation will be used to show that the set of prospective equilibria has an orderly structure. Indeed, it will be suggested that the points representing the prospective equilibria for either one of the firms fall into a well-defined region that tends to fill up with equilibrium points as n increases. This region of prospective equilibria has a characteristic shape that is robust, and has a simple explanation as well as substantial economic implications.

It may be helpful to interpret our discussion in terms of a concrete example. As will become clearer as we proceed, air passenger transportation is an apt illustration. Of course, it is not a duopoly, but the number of domestic airlines of substantial size it small -- something on the order of eight carriers. The n products offered by the airlines are the different routes along which they fly, each origin-destination pair constituting such a product. While there are economies of scale, beyond some point they are probably exhausted, and result in average cost curves that are closer to being flat bottomed rather than steadily descending. This, as we will see, is entirely compatible with an extended form of our analysis and accounts for the absence of perfect specialization -- the fact that on many routes service is provided by several airlines rather than only one. Finally, while it is not quite true that capital is rigidly rationed to the carriers, their persistent losses undoubtedly make them an industry unattractive for funding by investors. The persistence and near ubiquity of these losses, it will be suggested, are not a fortuitous phenomenon but can to a considerable degree be accounted for by the scale economies that underlie our model.

The airline example suggests that our analysis may be able to cast light on a number of other industries as well, notably many of the industries undergoing partial or virtually total deregulation in the U.S. and many of those being privatized in other countries. For a considerable proportion of these owe their earlier regulation or nationalization to the economies of scale inherent in their technology that was previously deemed to undermine the prospects for effective competition, and all of them can be presumed to be multi-product firms. While some of these industries will violate other assumptions of our model, there is reason to suspect that a number of them will resemble the model sufficiently to render the analysis pertinent to an understanding of their workings and to formulation of appropriate public policy toward them.

II. The Model.

We employ the following assumptions:

1) Each firm has a well defined profit function $\pi_j(y_j)$ of the vector of outputs $y_{i,j}$ of commodity I by firm J. Each firm adopts output levels of each of the goods it produces at which the marginal profits are equal.

2) The values of the output variables are nonnegative, and any I for which this variable is zero is, of course, one of the items that is not produced by firm J.

3) Labor and other inputs, with the exception of capital, are obtainable in the market at fixed prices in any desired quantity. Thus, the cost of the inputs other than capital are taken into account simply as deductions within the profit functions, and the corresponding input quantities need not be taken into account explicitly in the production functions.

4) There is a production function, $f_j(y_j,k_j) = 0$, for each firm, J, where, as before, y_j is firm J's output vector and k_j is its vector of capital inputs, $k_{i,j}$, the quantities of capital it assigns to its output of good I.

5) The capital constraint is not effective for firm J= 1,2 in any assignment in which firm J produces less than some number, N_j* of commodities, but it is effective if it produces more than N_j** commodities ($N_j^* < N_j^{**}$) so that, in the latter case,

(2.1) $\Sigma_i k_{i,j} = k_j^* =$ the total quantity of capital available to firm J.

6) Production is carried out on a monopoly basis, that is,

(2.2) $y_{i,1}y_{i,2} = 0$, for all I.

7) Finally, we assume that there is some output level of each commodity that yields a positive incremental economic profit.

We use the term monopoly-production assignment to describe any set of the ys that are set equal to zero and that satisfies (2.2), for any such set of these ys defines which firm is the non producer of each good. Obviously, if every good is produced by someone, it also tells us, residually, which firm is the producer of every good..

Since we assume that the firms are profit maximizers, for every monopoly assignment the output of each good by each firm and the corresponding use of capital is obtained by maximization for each firm of its profit function subject to a set of constraints. These constraints are (a) the set of zero-valued variables specified by the assignment, (b) the firm's production function, (c) the capital-availability constraint (2.1) and (d) the requirement that every variable receive a value that is non negative. This is, clearly, a nonlinear programming problem for each of the two firms.

We also obtain from the solution of these two nonlinear programs the maximum profit obtainable by each of the firms from the given monopoly-production assignment.

Though, as we will see later, there generally is a substantial subset of these assignments that turn out not to be equilibria, a very considerable proportion of those assignments are equilibria, and all of the latter are locally stable. The crucial difference entails the difference between acquisition and abandonment of a product. In some such monopoly-production assignment a firm may find that it can increase its profit by dropping one of the goods from its product line and moving the capital that thereby becomes available into expansion of the outputs of some or all of its other products. If the sunk costs are low this change can presumably be carried out with relatively little difficulty. If, on the other hand, the sunk obligations are high, the change may be infeasible and the equilibrium will then be stable. We may refer to the low sunk cost case as a profitable-abandonment assignment.

In contrast, we will speak of a profitable acquisition assignment as one in which an increase in profit from that offered by the initial assignment requires the firm to add to its product line some product currently supplied by its rival, a step that can be far more difficult than abandonment to carry out. Scale economies mean that success requires that firm's new product to emerge like Athena, full grown from the head of Zeus. In our world of scale economies, entry on a modest scale is entry foredoomed to failure. There is little point here in reviewing the costs and risks such instantaneously full-scale entry entails In any event, it is easy to construct a dynamic model of the equilibration process in which small departures from a profitable acquisition assignment set into play forces that move the variables of the model back into the neighborhood of the initial assignment. For this reason, all profitable-acquisition assignments are locally stable and each of them is a profit-maximizing equilibrium for its given assignment.

We have no general formula indicating the number of such locally-stable equilibria. However, we do know how many monopoly-production assignments there are altogether. Later, after we have constructed our basic graph, we will see that the number of profitable acquisition assignments (our locally-stable equilibria) as well as the number of profitable-abandonment assignments, is apt to be substantial when the number of commodities produced by the industry is large.

For the moment, however, we note that with two firms producing n commodities, the number of monopoly-production assignments will be exactly 2^n-2. For each of the commodities can be assigned to either of the two firms, meaning that there are 2^n possible assignments altogether. If we exclude the two extreme cases in which either of the firms is left with no commodities to produce, so that the industry is no longer a duopoly, then the expression just given clearly represents the number of monopoly-production assignments. It is clear that this number increases rapidly with n. For example, in a 20-product industry the number of monopoly-production assignments is somewhat greater than one million.

III. On Exhaustion of Scale Economies when Outputs Grow Sufficiently Large.

Empirical evidence indicates that scale economies are a widespread phenomenon in modern industry (see, e.g., Nadiri, 1994). Nevertheless, studies of individual firms also suggest that something close to a "flat-bottomed average-cost curve" is common in practice.[2] This is a case where the firm enjoys scale economies up to some combination of its input quantities, but that beyond this output vector further increases in the outputs of the enterprise raise total costs proportionately to the increase in outputs, at least over some considerable range of production quantities (beyond which average costs presumably begin to rise).

For our purposes this observation is important because it leads to a modification in our conclusions that brings them closer to reality. We have suggested that in a world of universal scale economies that extend throughout the relevant output range there will be complete specialization, with no product supplied by more than one producer. In terms of our airline example, in this case no origin-destination pair will be served by more than a single carrier. Reality is patently very different from this. The most obvious explanation, one supported by some empirical evidence, is that in the airlines, beyond some scale of operation along any given route the scale economies enjoyed by a carrier begin to erode. This means that if two airlines operate along such a route,

[2] As has been pointed out elsewhere, where the outputs of a multi-product firm are subject to economies or diseconomies of scope, the concept of "average total cost" of any one product or any set of the firm's products is not even definable because there is no defensible way to attribute the fixed and common costs, which are characteristically substantial. In our discussion it will be convenient to think of the firm's different products as being substantially independent in production, so that each has its own well defined cost, whose magnitude is not affected by the outputs of any other of the products of the firm.

growth in the volume of passengers served by one of them will give it no cost advantage over the other. Thus, multi-firm operation looses the instability by which it would be characterized if scale economies were to persist at all relevant output levels.[3]

Nevertheless such a route retains something of the property that we have elsewhere called "retainability." That is, a small firm whose volume deprives it of the cost savings permitted by scale economies will have a hard time seeking to take the market away from one of the route's larger incumbents. Even a large airline will find its entry into a new route handicapped by the necessity of sinking a substantial amount into the entry process.[4] Thus, the industry retains its character as a "natural cartel," but its individual products now can be supplied by an oligopoly rather than a monopoly.

Though nonspecialized equilibria, then, clearly add to the realism of the model, it remains true, as we have proven elsewhere (Gomory 1994) that the graphic analysis to which we will turn next can be carried out without distortion and much more easily by focussing on the perfectly specialized assignments. Accordingly, we will employ that simplifying device in the remainder of the discussion.

IV. The Basic Graph

We are now in a position to show the orderly relationship among our set of monopoly-production assignments with the aid of a graph that has already been mentioned several times. For this purpose it is convenient to define one other variable, share of products, $Z_1 = N_1/(N_1+N_2)$ where N_j is the number of products produced by firm J. Using firm 1 as the basis for the representation, we will then plot, in turn, for each monopoly assignment, the absolute profit of each firm on the vertical axis of the graph, while on the horizontal axis we will plot the share of products supplied by firm one. The use of only firm one's share of product on the horizontal axis permits the equilibrium points for both countries to be plotted on a single graph. For, obviously, $Z_2 = 1 - Z_1$, so that as we move from left to right in the graph the share of the world's goods produced by firm 1 rises from zero to unity, while as we go from right to left, the same is true for the share of products supplied by firm 2.

Now, for each monopoly assignment and each value of relative profit on the

[3] For an analysis of the number of firms that can survive in an industry with flat-bottomed cost curves see Baumol and Fischer (1978) and Baumol, Panzar and Willig (1988, Chapter 2).

[4] This brings up the much-debated issue of the degree of contestability that characterizes air-passenger transportation. The evidence seems to indicate that, as some participants in the deregulation discussion observed, the fact that airplanes represent "capital on wings" makes the markets quite contestable. Nevertheless, the cost of establishment of a new route, including advertising, planning, acquisition of slots and so forth, constitutes a significant deviation from perfect contestability and imparts some degree of retainability to the position of major carrier along some route.

horizontal axis we obtain two data points, one representing the absolute profit of firm 1 as a function of its product share, and the other representing the absolute profit of firm 2, also as a function of firm 1's product share. In other words, each monopoly assignment is represented in the graph by two points, one directly above the other (unless they happen to coincide). For the moment we will ignore the point corresponding to the absolute profit of firm 2 and will therefore focus exclusively on the point representing the absolute and product share of firm 1. Figure 1 shows two of the points in such a graph. Figure 2 shows the graph with the points all actually calculated for a specific eleven-product model. Here, the dots obviously represent the profits yielded to firm 1 by the different assignments. It will be seen that the dots fall into a well-defined region. We have also drawn in upper and lower regional boundaries, to which we will refer respectively as firm 1's upper and lower profit frontiers . Later, we will show how those frontiers are calculated, given the relationships constituting a particular model. (See Appendix A).

The shape of both the upper frontier and of the region as a whole should be noted. Both frontiers can be taken to start off at the origin at their leftward ends. Then the region moves upward as we proceed toward the right, and finally it begins to descend toward a point on the right-hand vertical axis. Thus, the region has, roughly, the shape of a crescent, with the upper profit frontier being hill shaped. These shapes are typical, but not universal. In particular, if the quantity of capital available to a firm is sufficiently small the region of monopoly assignments will have no descending portion.

It can be shown (see Gomory 1994) that as n, the number of products in the model, increases the number of dots representing firm 1's profits tends to "fill up" the monopoly assignment region for firm 1, meaning that if one selects arbitrarily any point P in the region and any distance ϵ, however small, then there is a value n* such that for any n>n* an assignment point will lie within a distance from P that is less than ϵ.

So far, then, we have shown that the number of monopoly assignments and, hence, the number of candidate equilibria that are locally stable, grows very rapidly as the number of commodities in the model increases. We have also indicated that there is a pattern constituted by these assignments whose orderliness is perhaps surprising. Finally, we have indicated that the points representing the assignments define a region that tends to fill up with these points. For further implications we must next try to explain why the region has the shape just described and what that shape implies.

V. Economics of the Shape of the Region of Monopoly Assignments.

This section provides an intuitive discussion of the shape of the assignment region and of its economic interpretation. The easiest starting point for an understanding of the relationships is to think of a move from left to right in the diagram as the capture of an increasing share of the industry's products by firm 1 from firm 2. The left-hand endpoint of the region of firm 1 assignment point lies on the left-hand vertical axis (at the origin) and corresponds to the unique assignment in which firm 2 produces all of the industry's goods and none is produced by firm 1. The reverse is patently true at the right-hand end point on the right-hand vertical axis where the duopoly has degenerated into a firm 1 monopoly, with the latter supplying every industry product exclusively by itself. In between, the upper profit frontier will rise. This is an immediate implication of assumption 7, that tells us that if firm J produces any commodity it can earn a positive incremental profit. It is obvious that both of the end point assignments are unique, so that there can be only a single assignment point on either vertical axis. This is why the upper and lower profit frontiers that bound the region of assignment points must meet at the two vertical axes.

We see immediately why, starting from the left vertical axis, as one moves toward the right in the graph the upper frontier must have a positive slope. For a rightward move means that firm 1's share of industry products rises, and implies that its absolute profit will rise.

Until firm 1 is assigned the production of more than n_j* commodities the capital constraint is assumed not to be effective, so that its total profit will increase monotonically at least up to the value of product share, Z_1, corresponding to the first such assignment. However, once the capital constraint becomes effective for firm 1, the addition of further commodities to that firm's product line requires the reassignment of some of its limited capital stock to those goods and the consequent withdrawal of some of its capital from the goods it had already been producing. Beyond some point this process is likely to go too far, that is, the firm's capital-constrained profits will have been reduced by the assignment to it of an excessive number of products. Then, there will be some intermediate set of products in firm 1's product line that will maximize its profits, and where n, the number of products supplied by the industry, is very large the corresponding assignment point in the graph will lie approximately at the highest point on firm 1's upper profit frontier.

In principle one can determine the profit maximizing set of goods in the firm's product line through a linear programming approximation to the integer programming problem that is posed by the choice between inclusion and exclusion of each product in turn. The inclusion of any additional products in the firm's product line must then reduce the firm's total profits so long as it supplies nonzero quantities of those items. Thus, a move to the right in our graph toward assignments that entail the production by firm 1 of more products than this profit-maximizing assignment must result in a downward movement in the upper profit frontier and a decline in the height of the corresponding portion of the region.

VI. The Relation of the Monopoly Assignment Regions for the Two Firms.

To get to the economic implication of the graph it is necessary next to turn to the region of monopoly assignment points for firm 2. and then to bring the two regions together into a single diagram. To do this we recall that the firm 1 region is generated by the points (π_1, Z_1) for each such assignment, but that each assignment also yields the corresponding point, (π_2, Z_1), for firm 2. It is this second set of points that provides us with firm 2's region of monopoly assignments in exactly the same way as we generated that for firm 1--with one technicality as an exception. For convenience in plotting the two regions in a single graph, we represent firm 2's absolute profits not as a function of its own relative profits, but as a function of the relative profits of firm 1 (see Figure 3). The resulting region can now be drawn into our basic figure, superimposing it, in Figure 4, upon firm 1's region as reproduced from Figure 2. The shape of firm 2's region will be very similar qualitatively to that of firm 1, except that the former will be shaped roughly as the mirror image of the latter. This is because, by definition, $Z_1 + Z_2 = 1$, that is, the two firms between them produce 100 percent of the industry's products. As firm 2's product share increases from zero to unity the product share of firm 1 will decrease correspondingly from unity to zero. In other words, while one moves, as usual, from left to right in interpreting the firm 1 assignment region, in interpreting the corresponding region for firm 2 one moves from right to left, from the right-hand vertical axis toward the vertical axis on the left. Consequently, the height of firm 2's region can well be zero at its right-hand end, it will then rise as one moves leftward from that point, and then, normally, the region will descend toward a point on the left-hand axis, the point of degenerate duopoly in which firm 2 has captured every industry product.

Focusing next on the upper profit frontiers of the two firms, we note that point Max I, the maximum point for firm 1, lies to the right of Max II, the corresponding point for firm 2. This is generally true, and it simply means that the profit maximizing assignment(s) for firm 1 must entail its having a larger share of total industry products than it produces in an assignment that maximizes the profit of firm 2.

The fact that maximum point Max II of firm 2's upper profit frontier always lies to the left of point Max I, the corresponding point for firm 1, enables us to divide the graph into three zones. The first is the zone to the left of Max II, in which the upper frontiers of both firms on average have a positive slope, the intermediate zone between Max II and Max I, where firm 1's frontier slopes upward while that of firm 2 slopes downward on average, and the third zone, to the right of Max I, where both upper frontiers have an average-negative slope.

This last observation lends itself to a direct economic interpretation. In the rightmost of the three zones an assignment point near firm 1's upper frontier has that firm as the manufacturer of more products than serve its own best interest. A reduction in its product line can then generally increase its profits. However, if firm 1 succeeds in divesting itself of some set of goods and thereby enhances its own profits, the absorption of that set of goods into firm 2's product line is also apt to enhance the profits of this second enterprise. In other words, a move from an assignment point in this rightmost zone can offer mutual benefits to the two firms. A corresponding observation clearly applies to the zone to the left of Max II, where the upper frontiers of both firms generally slope upward. However, in the central zone, that between points Max II and Max I, there is generally conflict in the interests of the two enterprises. Any move that benefits one is likely to be detrimental to the other. This central zone, then, can be interpreted as the zone of inherent rivalry. Such rivalry, can, of course, constitute a temptation to collude, or can even introduce the urge to merge. These possibilities will, however, not be examined further here.

VII. Equilibrium and Non Equilibrium Monopoly-Production Assignments.

As was discussed earlier, even in the case where scale economies are ubiquitous and do not evaporate as outputs increase, not not all of the monopoly-production assignment points are equilibria, or, at least, not stable equilibria. Clearly, the vast bulk of them are not profit-maximizing assignments for either firm, for most of them do not lie at either Max I or Max II. Given the assignment, however, that is, given which products are supplied by which of the two firms, each assignment point entails maximization of profit under the constraint constituted by the production assignment in question. Yet, from almost any such point a firm can hope to increase its profit by moving to another assignment. The problem is that with strong scale economies, if the move requires the capture of an industry from the firm's rival, the move to a more profitable assignment is costly and discontinuous. As a result of scale economies, a small move in the direction of the preferred assignment is bound to fail.[5]

Matters are quite different, however, where the higher profit assignment that the firm is seeking to attain entails the elimination of some good or goods from its product line. While closing down production of a commodity may be more costly than is generally recognized, that cost is nevertheless apt to be modest in comparison with that of acquisition of a product from a rival supplier. Consequently, in those cases where abandonment does not leave the firm worse off because of continuing sunk obligations, assignments in which either firm can increase its profit by abandonment of products are not equilibria or, if one prefers to consider them as equilibria of a degenerate variety, they are certainly unstable. Hence, such assignments are not members of the set of (locally) stable equilibria.

Which of the points in our graph remain in the set of stable equilibria and which do not? Since a movement from left to right in the graph generally corresponds to the gain of products by firm 1 and the loss of those products by firm 2, it follows that to the right of Max I any assignment is likely to entail for firm 1 a product line that contains more than the profit-maximizing set of products. The same, evidently, is true for firm 2, for any assignment point to the left of Max II. Consequently, in the absence of sunk obligations, assignment points to the left of Max II or to the right of Max I will generally

[5] It should be obvious that the same observation applies more generally, even in the absence of scale economies, if embarkation on the production of a commodity not formerly in the firm's product line entails heavy startup costs, even if those costs are fixed and once-and-for-all outlays. Elsewhere, we have referred to such a product as a retainable commodity (Gomory and Baumol 1994), that is, once a producer succeeds as supplier of the commodity in question it is very costly and hence difficult for an intended rival to take that business away from the incumbent producer. Thus, the product is relatively retainable against attempted acquisition by others.

not be equilibria, whereas most assignment points between the two maxima of the upper profit frontiers will be locally-stable equilibria.

The conclusion from all this is that when there are no sunk obligations most equilibrium points can be expected to lie within the zone of rivalry, where the interests of the two firms are opposed because each can gain only at the other's expense by taking away one of the other's products. That is not easy, because of the heavy cost of invasion of the territory of a firm by its rival, and so each may persist in its product line even though a modification of that line of outputs, if successful, promises an increase in profits. Still, an attempt to capture a lucrative product will be tempting to both enterprises, and so attempts at invasion are to be expected from time to time.

VIII. On Weak Capital Constraints, Small n and Other Possible Modifications of the Model.

Before delving further into the implications of the model, a few observations should be offered on the consequences of weakening of some of its premises in order to broaden its applicability.

The fundamental premise of the analysis is the rationing of capital to the firms. In the model we have assumed that the capital resources available to the firm are represented by an immutably fixed number. The firm has a given quantity of funds available to it on prespecified terms and it cannot obtain a penny more. In reality, many if not most firms are indeed circumscribed by a capital constraint, and the management of a typical firm is apt to feel itself constantly beset by a shortage of investable resources. Yet, in reality, the capital constraint is usually considerably more flexible than has been assumed here. Normally, the firm can expand its financial resources, but it can do so only at a rising cost. Moreover, that rising cost will sometimes be prohibitive, so that during a given period, having raised some intended funding, the firm will seek to avoid any additional recourse to the capital market unless some emergency forces it to do so. We can regard our premise as a special (extreme) case of a rising cost of capital, the case in which the capital-supply curve is L shaped (the L lying on its back), with additional funds priced at infinity. Slight curvature of the supply curve of funds to the firm, rather than a sharp corner, is undoubtedly more realistic. We conjecture, however, that if the curved portion is sufficiently narrow and rises sufficiently rapidly, no change in our analysis is required. Firms will behave as our model describes, with some small degree of flexibility in their acquisition of funds, a possibility that will serve to reduce somewhat the risk entailed in its acquisition of capital in an amount near to the limit that is available to the enterprise, but with little else affected. A more gradual rise in the cost of capital must undoubtedly complicate the calculations that underlie our diagram because then it is no longer safe to assume that a fixed stock of capital represents a harmless approximation to the facts and entails no basic distortion of the firm's decision process. Still, the basic points of the analysis remain valid. With substantial scale economies in every product stable equilibria will entail monopoly production assignments, the number of such assignments will remain exactly as large as in our model, and the points representing these assignments in our graph will fall into the same sort of regions as described by our graphs. Moreover, it will in general remain true with a materially rising cost of capital as the funding of the firm increases, that there will come a point beyond which further additions to its product line will reduce profits, just as in our model.

It was also observed earlier that the upper profit frontiers need not be hill shaped. It is easy to see that the firm 1 frontier can rise monotonically from the origin, or that of firm 2 can fall monotonically from left to right. This will obviously occur if capital is not really scarce, because then, as has been shown, the more products that a firm can capture, the more profit it will earn under our assumption that every additional commodity offered by the industry can yield profit if sold by an increment monopoly. In effect, in this case the profit-maximization problem entails a knapsack model in which the knapsack has more than enough capacity to hold its prospective contents.

Since the obverse of scarcity of capital is a large number of products each requiring a substantial amount of investment, it is clear that monotonicity of the upper profit frontier can also occur if, with a given amount of capital available to the firms, the number of products supplied by the industry is relatively small. For in that case, the larger the number of the limited set of products offered by a given firm, the larger its profits will be Thus we conclude that for the full set of possibilities described by our model to be pertinent, the amount of capital that it is practical for the firm to acquire must be substantially limited relative to the number of products offered by the industry.

Finally, we offer a few additional remarks on the consequences of relaxation of our other fundamental premise--that scale economies are entailed in the production of each commodity supplied in the industry and that they hold throughout the relevant range of outputs. We have already noted that where diminishing returns to scale are present in the relevant range, production of a given commodity can be carried out by both firms, and if the industry contains more than two enterprises, even more can participate in production of that good. We have also shown elsewhere (see Gomory and Baumol 1994) that as the share of the industry's goods subject to diseconomies of scale increases, the region of monopoly production assignments shrinks away from both vertical axes because it becomes impossible for either firm to capture all of the products and hence all of the profits of the industry. In the limit, when all of the industry's products are of the diminishing-returns variety, the region collapses to a single point. In an industry with many firms, that point will presumably correspond to an assignment that yields maximum profit to every firm, but that maximum profit will, of course, be zero.

IX. Scale Economies, Propensity to Incur Loses and Capital Rationing.

Firms in reality never find themselves subjected to absolute rationing of their capital, as we have just observed. Even in very unfavorable circumstances most firms are able to acquire additional funding at a price. But if that price is sufficiently high, the premise that the amount of capital available to the firm is absolutely fixed is not too bad an approximation to reality. Clearly, the conditions in when the firm is likely to find itself subject to capital rationing entail either a period of extreme stringency in the capital market or a set of circumstances that make that firm particularly unattractive as a place for investors to place their funds. Persistent and widespread losses, like those experienced by the airlines in recent years, seem to constitute a good example. Thus, the capital-rationing premise underlying our model seems not to make that model irrelevant to reality.[6]

Nevertheless, one may well raise the question whether this poor financial performance of the airlines is largely fortuitous, so far as our analysis is concerned, and so cannot be counted upon as a robust feature of the appropriate analysis, or whether it is more deeply embedded in the structure of the airlines and, perhaps, some other industries. Here the old doctrine that scale economies industries are subject to "destructive competition" appears to be pertinent. Imagine an industry characterized by flat-bottomed cost curves so that survival of competition in any particular product is not ruled out, but in which firms vary in size, with a number of the smaller firms operating on a scale that puts them in the declining portions of their average cost curves.[7]

The problem besetting the industry in this case is that these smaller firms can be expected to be subject to two pressures that lead them to adopt prices that are uncompensatory to themselves and that force their larger rivals to adopt prices that also impose losses upon the latter. The two pressures that beset the smaller carriers are, first, the higher costs they incur as a result of the fact that their small volumes have not

[6] The fact that the airlines have been incurring substantial losses need not conflict with our model's assumption that additional products generally yield positive incremental profits. A money-losing firm can clearly lose even more if deprived of relatively remunerative products by its competitors. Thus, the airlines seem to constitute a variation on the old adage--they earn a dollar on every route, but lose it all in their aggregation.

[7] There are, obviously, some economies of scale in airlines. An airplane clearly operates with a per-passenger cost that is lower the closer to a capacity load of passengers it carries. In addition, a large aircraft operating at capacity incurs a smaller cost per passenger than a smaller one. Yet, the advantage of an airline that is larger in terms of number of passengers flown over a given route is not primarily a matter of scale economies, conventionally defined. Such a much-used airline will typically offer more flights and more frequent departures than its smaller rivals. That makes it easier for the former to attract additional passengers, as does the fact that its frequent-flyer bonuses are more attractive because they offer a greater variety of places to which they provide bonus trips. The result is that the larger airlines tend to have a larger proportion of their seats filled on a typical flight, and this, together with lower marketing costs offers them a cost advantage over smaller airlines. We are indebted to Professor Elizabeth Bailey and her extensive knowledge of the industry for these observations.

brought them to the scale of output that minimizes their average costs, and second, because their outputs fall in the range of declining average costs, the marginal cost of any of their products will be below the average cost. In those circumstances, a smaller firm is likely to experience losses at the prices offered by their larger, lower-cost rivals for the homogeneous product in question. However, the smaller firm will be able to reduce its losses by additional sales that can be attracted by any price above marginal cost, even if that price is below the supplier's average costs and, very likely, below the average costs of its competitors.[8] With products homogeneous, the larger rivals will be forced to match those uncompensatory prices and widespread losses will be the predictable result. Moreover, with each product homogeneous among different suppliers no firm will be able to raise its price to a compensatory level unless every other supplier does so. The applicability of this parable to our airline example should be obvious.

The implication is that in an industry sharing the production characteristics of air transportation, protracted and widespread losses are to be expected so long as the medium sized firms survive or if they are replaced by entrants whose investors hope to succeed even though so many others have failed before them (note the many tiny airlines that continue to open for business). As a result, a firm in the arena can be expected to find it difficult to get significant amounts of additional capital except on very onerous terms. In other words, in such an industry something close to capital rationing can be expected to be characteristic.

X. Some Welfare Implications: Inefficient Equilibria and Violation of Comparative Advantage.

One noteworthy implication of our analysis is that a cartel form can occur in an industry not only as the result of conspiracy; but as a consequence of the market forces that thrust the cartel structure upon the industry. Where scale economies extend throughout the relevant range, the cartel will automatically tend to provide exclusive product territories to each of the firms that constitute the industry. Where the economies of scale in the individual industry products are exhausted after a relatively large but economically viable level of output, the individual products will be supplied by an oligopoly rather than a monopoly, but entry into that oligopoly will still not be easy.

Another basic implication of the analysis so far is the tendency of the market mechanism to ensure conflict in the interests of the capital-constrained duopolists. In a sense this entails some degree of efficiency in the workings of the market even here,

[8] If this is a correct explanation of the airlines' dilemma it follows that for the larger airlines, with their comparatively low costs, additional routes must be incrementally profitable, up to a point, as the previous footnote suggested. This must be true despite the low fares imposed on them by their smaller competitors. Of course, this story is not one of long-run equilibrium, since eventually the smaller airlines in this scenario must prove unviable and drop out of the industry.

because what we have seen is the tendency to avoid assignments that entail unused opportunities for mutual gains to the two firms. In this case, moreover, the mutual gains to the duopolists are not, in general, obtained at the expense of consumers, for what is avoided is an assignment in which the allocation of commodities among the two firms leaves room for improvement. In other words, there seem to be forces making for the assignment of products to those firms that can do the best job of producing them. We will see next, however, that even though there may be a modicum of substance to this argument, the circumstances described in our model leave considerable scope for inefficiency. Indeed, we will see that a considerable portion of the stable equilibria may fail to attain efficiency--meaning, as usual, that they offer unused opportunities to increase some output quantities without any increased use of the scarce input. This result would appear to conflict with the standard presumption that the interests of a profit-seeking monopoly or an oligopolistic firm are best served by efficiency in its operations. That presumption retains its validity in our analysis, but there nevertheless remains a critical source of inefficiency--the possibility of inefficient assignment of product line to the different specialized firms in the industry.

First, it should be noted that the nature of the premise on capital rationing--that each firm is endowed with a fixed amount of capital, with the implication that this capital is not transferable from one firm to the other--introduces here the relationship between efficiency and *comparative advantage*. It should be clear intuitively from the usual textbook parables that with fixed input quantities in the hands of two economic agents, if each of them takes on tasks in which it is *comparatively* the more efficient in the use of the scarce resource, then efficiency is assured, because then the transfer of a task from one of the agents to the other must reduce some output. This can, in fact, be proved rigorously for our model. However, it can also be demonstrated that in a scale economies world the rules of comparative advantage, while still sufficient, are no longer necessary for efficiency (see Baumol and Gomory [1994] for full discussion of these matters). Thus, in our model there are apt to be equilibria that are efficient, even though they violate comparative advantage. The reason, of course, is that under scale economies, specialization, *in itself* is a source of efficiency. For specialization is an opportunity for the sole producer of a good to produce larger amounts of that item than anyone would if production of the item were divided among several suppliers. The large outputs of the goods then yield the efficiency benefits of scale economies, even if the particular assignment violates the rule of comparative advantage.

An example will show that in this model the market does not prevent the occurrence of inefficient equilibria. Consider an equilibrium assignment in which firm 1 is exclusive producer of good A, with output $Y_{a,1}$ of good A, and firm 2 has a similar position $Y_{b,2}$ in good B, and suppose the industry produces only those two goods. Then a *ceteris paribus* interchange of the two goods between the two firms obviously yields another monopoly production assignment $[Y_{b,1}, Y_{a,2}]$. Moreover, it entails the *acquisition* of a product by each firm, so that the new assignment, too, can be a locally stable equilibrium. One can then construct a production frontier for the two products for each of the firms.

Now consider the case in which the two frontiers intersect. In that case, if $Y_{a,1} > Y_{a,2}$ it must be true that $Y_{b,1} < Y_{b,2}$. Then, the equilibrium $[Y_{a,1}, Y_{b,2}]$ must clearly dominate the other equilibrium, with both using the same capital quantities. Hence, the second equilibrium must be inefficient.

It follows from all this that the combination of scale economies and capital rationing can yield serious efficiency problems, because once an assignment becomes an equilibrium, its local stability makes departure from it difficult. One may well suspect that the modifications of the model suggested earlier to bring the construct closer to reality do not eliminate those difficulties. There is, indeed, reason to believe that such path-dependent efficiency traps are not altogether rare, as Brian Arthur and his associates have been suggesting (see, e.g., Arthur, 1990). Then it may require something like a substantial innovation or a major strategic misjudgment by one of the participants to dislodge the economy from such a trap once it has fallen into it.

Bibliography

Arthur, Brian, "Positive Feedbacks in the Economy, Scientific American, February 1990, 92-99.

Baumol, W.J. and Dietrich Fischer, "Cost-Minimizing Number of Firms and Determination of Industry Structure," Quarterly Journal of Economics, 92, August 1978, 439-467.

Baumol, W.J, J.C. Panzar and R.D.Willig, Contestable Markets and the Theory of Industry Structure, San Diego: Harcourt, Brace, Jovanovich, Revised Edition, 1988.

Baumol, W.J. and R.E. Gomory, "On Efficiency and Comparative Advantage in Trade Equilibria Under Scale Economies," New York: C.V. Starr Center for Applied Economics, Economic Research Report RR 94-13, April 1994.

Baumol, W.J. and R.E. Quandt, "Investment and Discount Rates Under Capital Rationing--A Programming Approach," Economic Journal, 75, June, 1965,

Gomory, R. E., "A Ricardo Model With Economies of Scale," Journal of Economic Theory, 62, April 1994, 394-419.

Gomory, R.E. and W.J. Baumol, "Shares of World Output, Economies of Scale and Regions Filled With Equilibria," New York: C.V. Starr Center for Applied Economics, Economic Research Report RR 94-29, October 1994.

Nadiri, M. Ishaq and Theofanis P. Mamuneas, "Infrastructure and Public R&D Investments, and the Growth of Factor Productivity in US Manufacturing Industries," C.V. Starr Center for Applied Economics, Economic Research Report RR 94-26, August 1994.

Appendix A

Let $\pi_{i,j}$ be the profit for firm j making product i. In any assignment of the capital of firm i in the amounts $k_{i,j}$ to the products we must have

$$\Sigma_i \, k_{i,j} = K_j$$

Where K_i is the total capital available to firm i. In any locally profit maximizing assignment we must also have

$$\frac{dp_{i,j}}{dk}(k) = \frac{dp_{q,j}}{dk}(k)$$

for all i and q. Otherwise it would pay to shift capital marginally between the i^th^ and q^th^ products.

We will assume that the profit functions $\pi_{i,j}$ (k) have the general appearance shown in
Figure A1. We will discuss $k_{i,j}$ (μ) the capital associated with the slope μ, or equivalently, the marginal return on the amount of capital μ.

If the profit functions have the form shown in Figure A1 we can see that a firm dividing its capital among many products will tend to have a large marginal return μ and one that spends its capital on a few will tend to have a small μ,

We will consider all possible μ. Note that once μ is chosen then $k_{i,j}$ (μ) is known as is $\pi_{i,j}$ $(k_{i,j}$ $(\mu))$.

Consider the maximization problem

$$\text{(A.1)} \qquad \begin{gathered} Max \; \Sigma_i \, x_{i,1} \; \pi_{i,1} \; (k_{i,1}(\mu)) \\ \Sigma_i \, x_{i,1} \; k_{i,1}(\mu) \; = \; K_1 \qquad 0 \leq x_{i,j} \leq 1 \end{gathered}$$

In (A.1) the $x_{i,j}$ are the choices of capital assignment. If $x_{i,j} = 1$, the capital is assigned, if $x_{i,j}=0$ it is not assigned. For any choice of μ (A.1) will be a linear programming problem with one equation - a "knapsack problem". In the solution to (A.1) there will be at most one value of the $x_{i,j}$ that is neither 0 nor 1. The value of the objective function in (A.1) will give us an upper bound on the profit that can be obtained by firm 1 in any specialized assignment of products to firms having marginal return μ.

By repeating this calculation for all μ we can obtain the upper profitability frontier. Figure 2 in the text.

Figure 1.

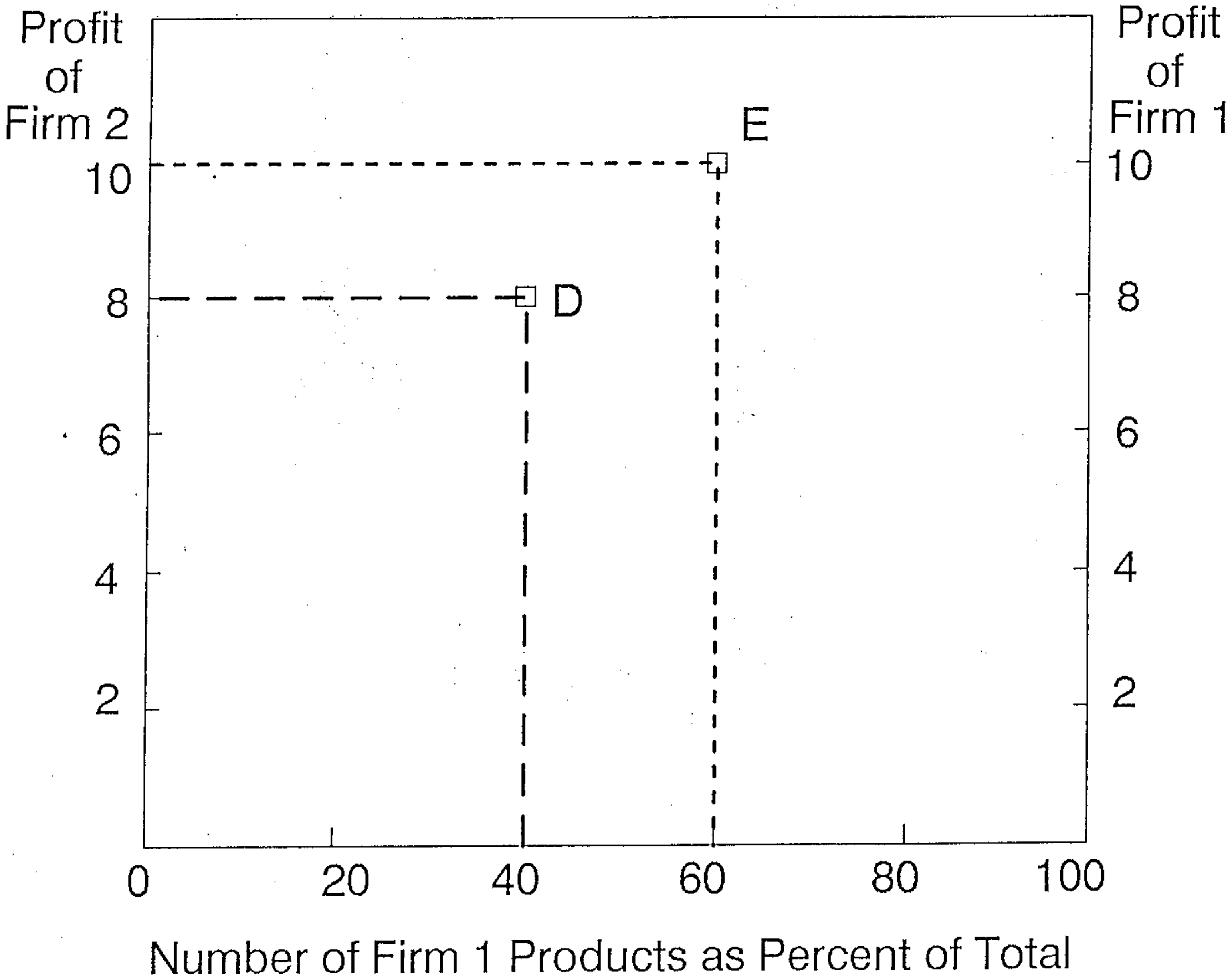

Profit
of
Firm 2
10
8
6
4
2
Profit
of
Firm 1
10
8
6
4
2
E
D
0
20
40
60
80
100
Number of Firm 1 Products as Percent of Total

Figure 2.

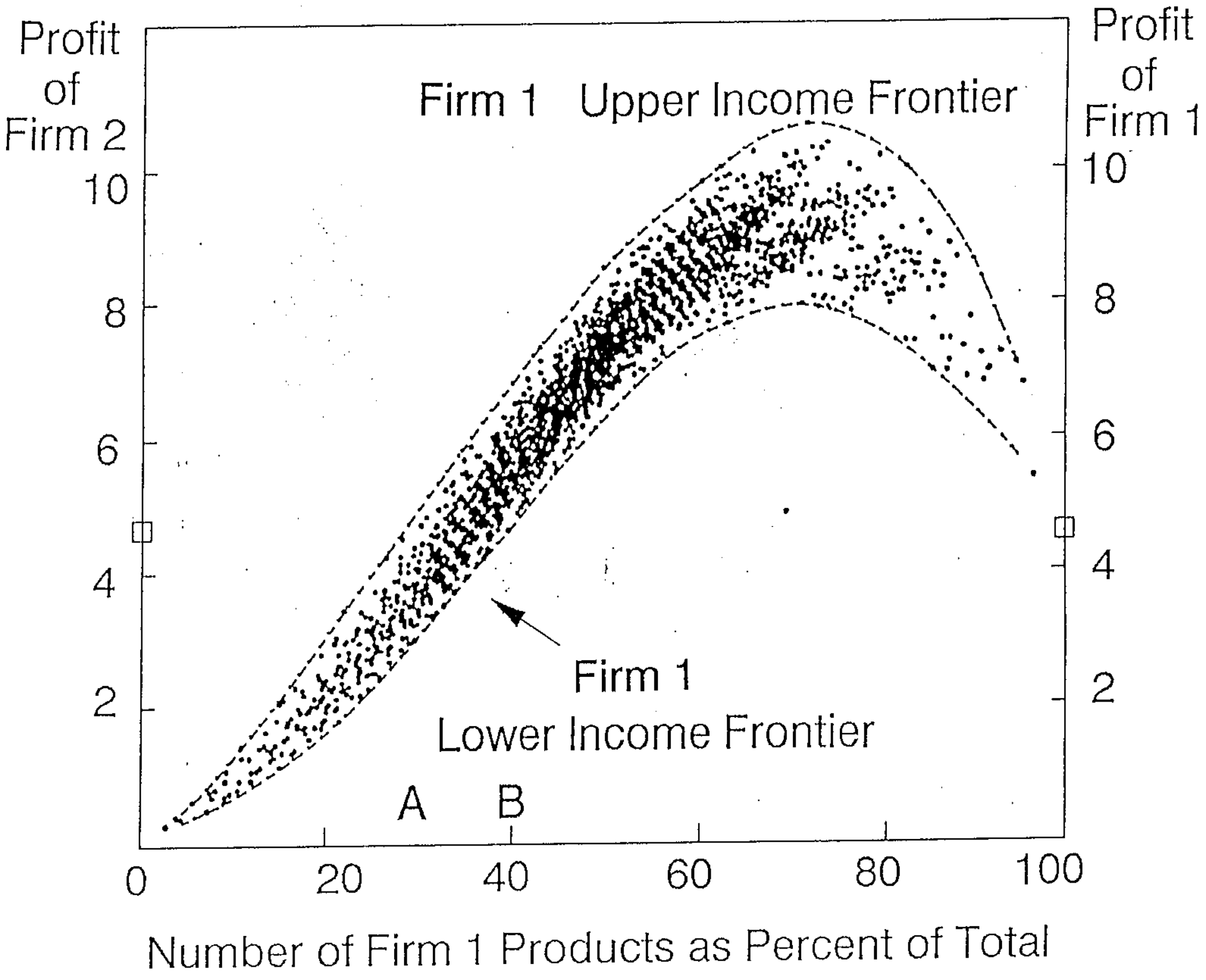

Profit
of
Firm 2
10
8
6
4
2
Profit
of
Firm 1
10
8
6
4
2
Firm 1 Upper Income Frontier
Firm 1
Lower Income Frontier
A
B
0
20
40
60
80
100
Number of Firm 1 Products as Percent of Total

Figure 3.

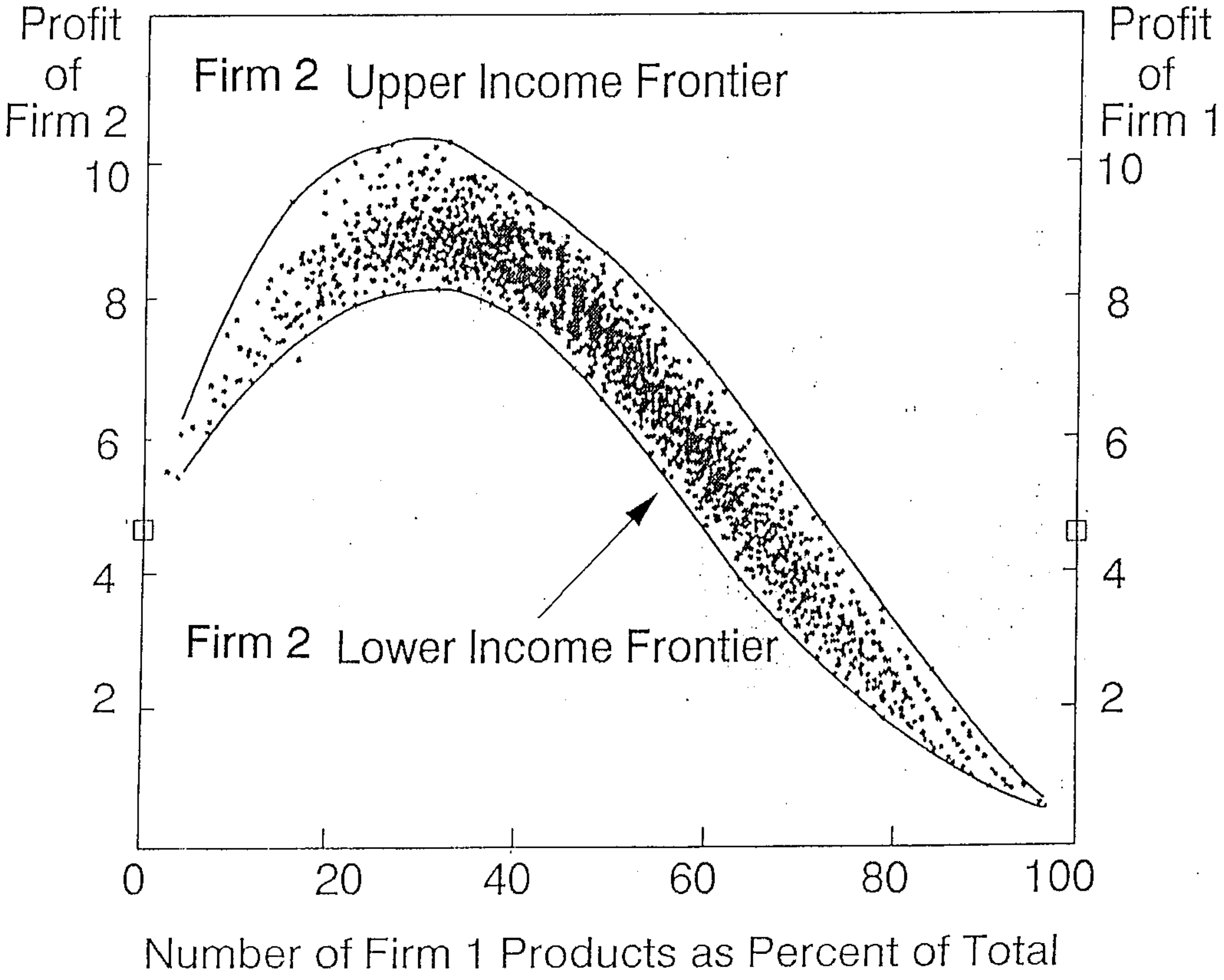

Profit of Firm 2
Profit of Firm 1
Firm 2 Upper Income Frontier
Firm 2 Lower Income Frontier
10
8
6
4
2
0
20
40
60
80
100
Number of Firm 1 Products as Percent of Total

Figure 4.

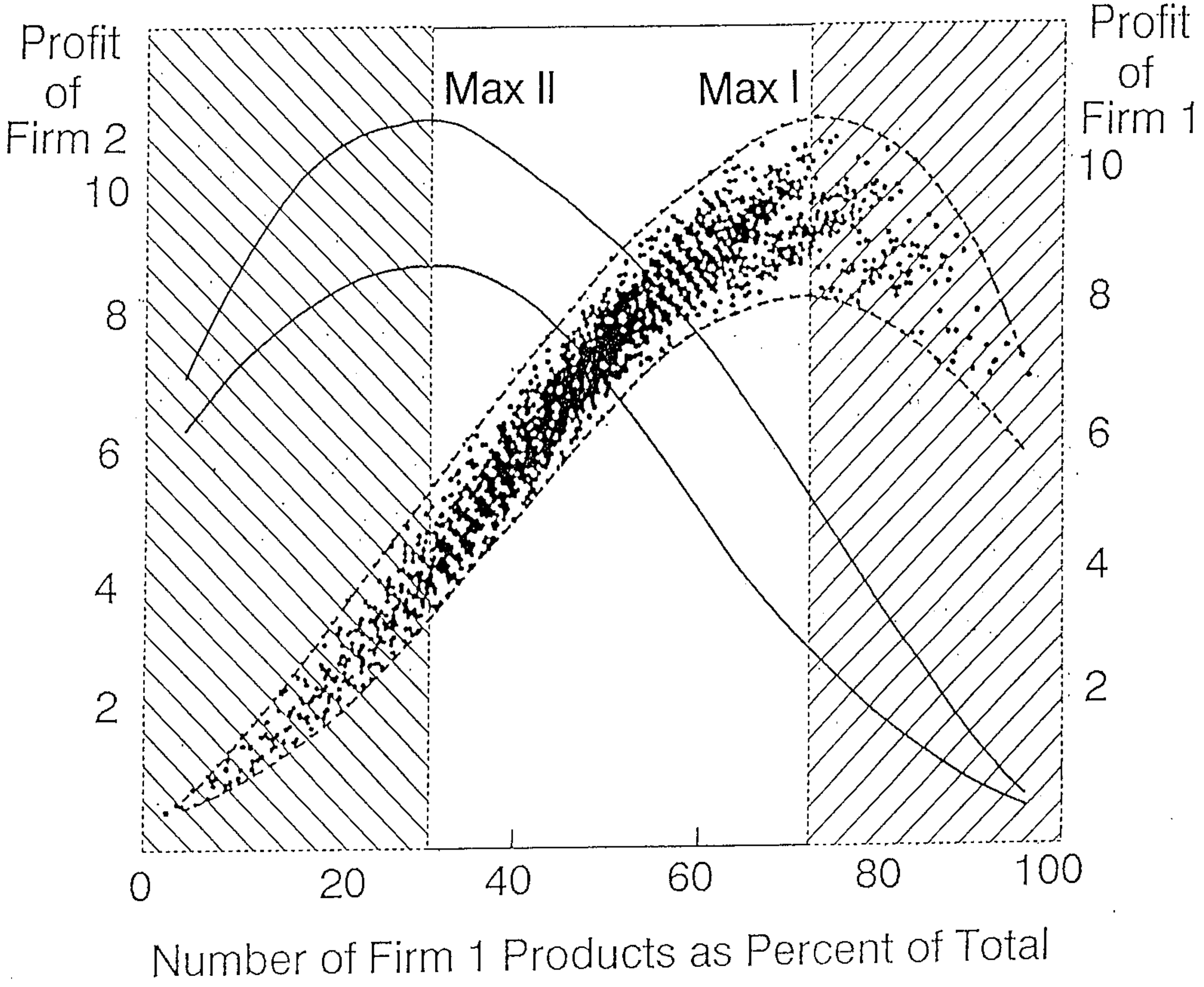
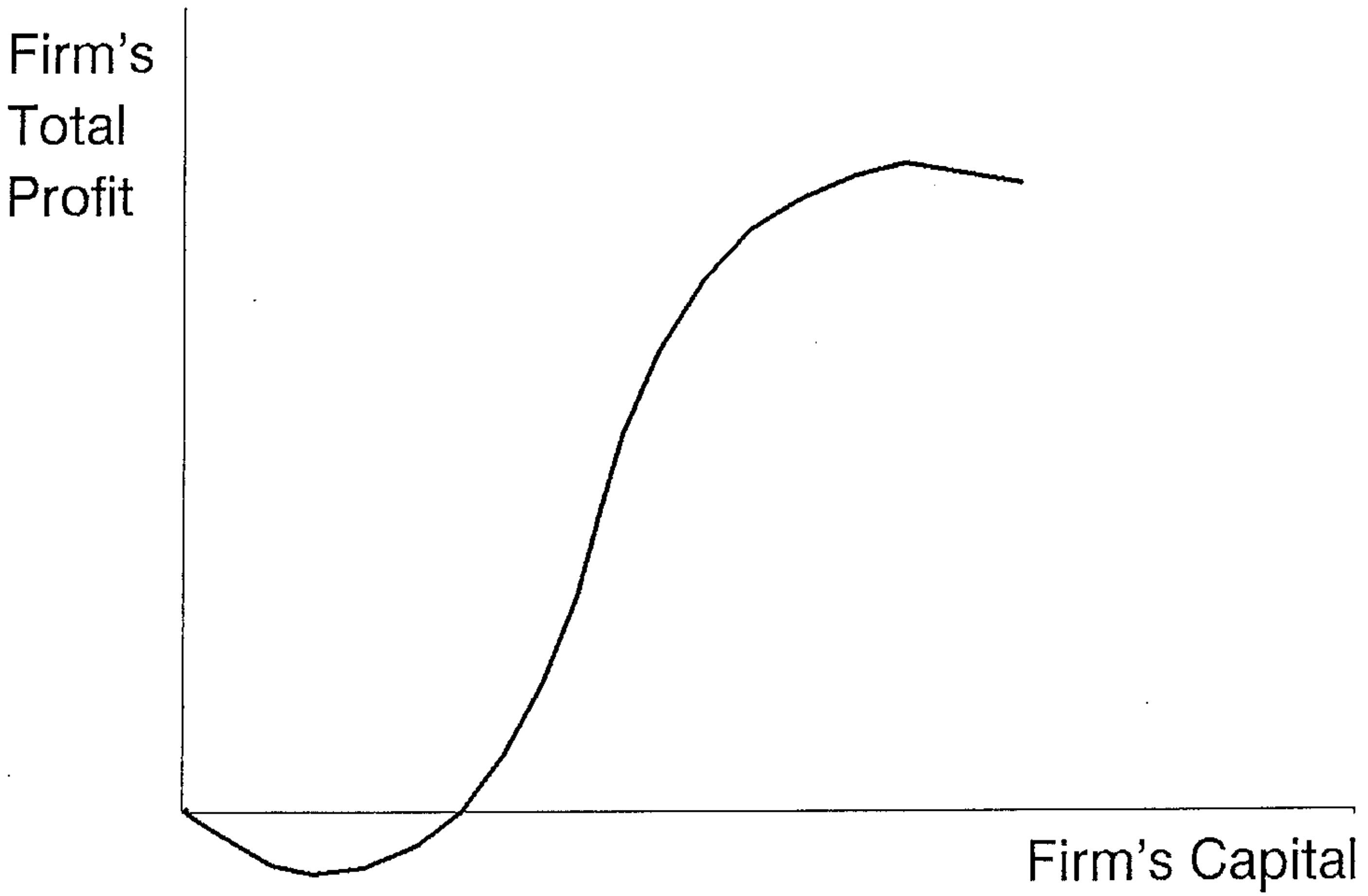

Profit
of
Firm 2
10
8
6
4
2
Max II
Max I
Profit
of
Firm 1
10
8
6
4
2
0
20
40
60
80
100
Number of Firm 1 Products as Percent of Total

Figure A1



File:duopgoa1 12 May 95