The Intuition Behind Sutton's Theory of Endogenous Sunk Costs

Daniel R. Shiman[1]
1/15/08

Abstract

To explain why many manufacturing industries have remained highly concentrated, John Sutton has developed a new theory of endogenous sunk costs which predicts that industries with significant endogenous sunk costs will have a lower bound on concentration even as industry demand increases. Endogenous sunk costs are fixed costs that firms can choose to invest in, which affect the price-cost margin of a firm. Any optional fixed investments in quality, advertising, and cost-reducing plant, that allow the firm to raise its price or lower its variable costs, qualify as endogenous sunk costs. This paper cuts through the mathematical complexity of the model and examines the intuition behind the theory, to provide a better understanding of how and when the theory of endogenous sunk costs can be applied, and its limitations. How the incentive for firms to invest in endogenous sunk costs increases as the market expands, and how this can limit the number of firms that can profitably remain in the market, is discussed. The paper examines the key conditions required for this effect, and shows how the same basic results can be obtained even if some of Sutton's assumptions are relaxed. It also examines the reasons for the model's unusual result that in some circumstances, a market expansion can actually induce a reduction in the number of firms, contrary to the usual expectation in industrial organization theory. A proof is provided of the proposition that concentration is bounded below for the Cournot model. The paper then discusses how endogenous sunk costs are likely to have a significant impact on industrial development in ways that Sutton did not discuss, for example when firms have a choice of variable cost-reducing investments.

Keywords: quality, endogenous sunk costs, vertical product differentiation

JEL codes: L11, L15

I. Introduction

John Sutton has tackled a puzzle in the field of industrial organization – why so many manufacturing industries become dominated by a few firms as they grow to a large size. His new theory of endogenous sunk costs, which builds on and extends existing theory, appears to provide the explanation. In his book *Sunk Costs and Market Structure* (Sutton 1991; *see also* Shaked & Sutton 1983; Sutton 1986; Shaked & Sutton 1987) Sutton has made a persuasive case, based on extensive theoretical analysis covering a broad range of models, and detailed empirical evidence including numerous case studies and econometric analysis, that industries that have significant endogenous sunk costs, such as advertising and R&D, will develop differently from other industries as their markets grow, with less entry and greater concentration. His theory of

[1] E-mail: dshiman@starpower.net. Author is an economist at the Federal Communications Commission. The opinions expressed in this article are those of the author and do not necessarily represent the views of the Federal Communications Commission or its commissioners.

endogenous sunk costs thus helps explain why many large industries tend to remain highly concentrated even as demand for their products expand, which is contrary to the usual expectations of economic theory. As a result, his work has attracted attention in the economics literature (Carlton 2005; Matraves & Rondi 2005; Carlton & Perloff 2004). It has been applied to the packaged goods industries (Bronnenberg, Dhar, & Dubé 2006), supermarkets (Ellickson 2006), banking (Dick 2007), restaurants and newspapers (Berry & Waldfogel 2003), as well as more broadly (Robinson & Chiang 1996).

This paper examines the intuition behind how endogenous sunk costs affect market concentration. While Sutton has focused in his work on providing theoretical and empirical support for his basic proposition that concentration remains high in industries with endogenous sunk costs, this paper interprets the mathematics that underlie the model, using the familiar tools used by economists. Because of the complex nature of his model, with a three stage game and three choice variables, which yields very complex equilibrium equations that cannot be analytically solved for (for his basic Cournot model), the model could prove quite difficult to use for further theoretical and empirical work. The discussion provided in this paper of the basic principles involved should assist researchers interested in either further developing the theory, or in empirically applying the theory of endogenous sunk costs to their analysis of particular industries.[2] Other papers that have discussed some of the theoretical underpinnings of Sutton's analysis, but not as completely as here, are Bresnahan (1992) and Schmalensee (1992). This paper also discusses the key assumptions that underlie the model, and demonstrates how the theory's conclusions can be obtained with a relaxed set of assumptions. Proofs of some basic results are provided in the appendices, along with a full solution to the Cournot model of endogenous sunk costs. The paper then shows how endogenous sunk costs encompass a wide range of firm investments, including investments in advertising, R&D, product innovation, process (cost-reducing) innovations, plants of different sizes, and whenever firms have a choice of fixed sunk investments that affect the price-cost margin.

The theory of endogenous sunk costs has its origins in the advertising and vertical product differentiation literatures, in which firms can raise consumers' willingness to pay for their products by spending money on advertising and R&D (e.g., Shaked & Sutton 1987; Dixit & Norman 1978; Carlton & Perloff 2004, ch. 14). Vertical product differentiation models have been explored in a number of papers. The primary focus has been on the impact of market structure on the quality of products provided, and whether firms produce too much or too little quality and too many or too few products of differing quality, relative to the social optimum (Spence 1975; Swan 1970; Sheshinski 1976; Levhari & Peles 1973; Tirole 1988). Another strand of the literature, pioneered by Akerlof, examines the impact of consumers' imperfect information about product quality on the performance of markets (Akerlof 1970; Tirole 1988). There is also a substantial literature concerning firms' investment in advertising and its impact on consumers' willingness to pay (Dorfman & Steiner 1954; Nerlove & Arrow 1962; Dixit & Norman 1978).

[2] The focus of this paper is on Sutton's theoretical arguments and model concerning the impact of endogenous sunk costs on industry structure, as outlined in chapters 2 and 3 of his 1991 book, and not on his extensive empirical analysis.

The problem that particularly interests Sutton is the high concentration that persists in many manufacturing industries, even though in many of these industries demand has grown and output has substantially increased. In traditional models of entry the existence of fixed costs can create an upper bound on the number of firms that can profitably enter the market. Entry can only be profitable if the short-run profits that firms make post-entry are sufficiently large to cover the fixed costs of entry. The larger these fixed costs, the less likely that a given number of firms will each earn sufficient profits to make entry profitable. How many firms can profitably exist also depends on the price level for the industry, which depends on, according to Sutton, the "toughness of price competition" (Sutton 1991, pp. 9, 33-35, 131-34, 157-60). The tougher is price competition between firms, the lower the price-cost margin for the industry, which leads to lower profits and fewer firms in the industry.[3]

Sutton's innovation is to distinguish between exogenous sunk costs, which are fixed costs incurred upon entry that are necessary to participate in the market, and endogenous sunk costs, which are sunk costs of varying sizes that firms can choose to invest in (such as advertising and R&D) to increase their price-cost margin, but whose size does not depend on the level of production. For industries with just fixed entry costs (i.e., exogenous sunk costs), most economic models predict that an increase in the size of the market would, in the long run, allow more firms to profitably enter, thus reducing industry concentration. This follows quite simply from the tendency for increases in demand to increase individual firm profits, either from a rise in prices, quantities, or both. As the size of the market increases, it becomes more likely that a potential entrant would gain sufficient additional short-run profits after entry to cover the sunk costs needed for entry. For a wide class of homogenous goods models, including Cournot and joint profit-maximization, Sutton in his book shows that as market size grows there is no lower limit to market concentration. Thus as demand increases, the number of firms entering the market can be expected to grow without limit, and, as Sutton describes it, "the industry converges to a fragmented structure" (Sutton 1991, p. 35).

The existence of endogenous sunk costs, however, can significantly change the relationship between market size and concentration. Endogenous sunk costs are fixed investments a firm can make in those things that can either increase the value of the firm's product to customers, such as advertising or improvements in product quality from product innovations, or lower its marginal cost of producing each unit, such as process innovations. Thus they reflect investment in advertising, marketing, and R&D. The theory applies to all sunk investments that increase the price-cost margin on each unit of a good sold. Unlike most models of vertical product differentiation, the theory of endogenous sunk costs is a model of long-run equilibrium behavior by an oligopoly of firms competing in both price and quality. Within the endogenous sunk costs model firms can choose the size of the investment they want to make,

[3] Note that Sutton's "toughness of price competition" does not refer to a particular price level, but is a function that expresses the relationship between concentration and prices. Sutton 1991, pp. 9, 33-35. This factor that Sutton introduces is important in both his theoretical and empirical analysis. In his theoretical work it allows Sutton to generalize some of his results, to show that they do not depend on any particular economic model and its associated price-cost margin. In his empirical analysis changes in the toughness of price competition are a key factor (along with endogenous sunk costs and certain other factors) in his analysis of why concentration changed over time, in the case studies he provides for a number of manufacturing industries. Sutton's introduction of this factor has been commented on favorably in the literature (Bresnahan 1992, p. 141; Carlton 2005, p. 5).

with larger investments having a greater impact on the price-cost margin.[4] The cost of the investment, however, does not depend on the quantity produced.[5] The theory makes the bold prediction that for industries with significant endogenous sunk costs, as the size of the market increases there is a lower bound to industry concentration, and the increase in sales might even cause the number of firms to shrink (Sutton 1991, ch. 3). This is caused by a rise in the sunk costs necessary to enter and participate in the market, which may equal or even exceed the additional revenues generated by the growth of demand for firms' goods. Sutton shows that this effect occurs under many conditions, using a series of models and proofs.

The key to the model is that a firm can choose to invest in endogenous sunk costs, and because an investment of a given size and cost increases the profitability of each unit sold (i.e., the price-cost margin) by a fixed amount, the benefit from investing in endogenous sunk costs increases with the quantity sold. Therefore as the market size and firms' output increases, their incentive to invest in endogenous sunk costs also increases. However, because most of the benefits of this investment derive from the perceived quality advantage obtained relative to other firms in the market, if all firms invest equally in endogenous sunk costs, the investment produces little or no increased industry profits in the long run.[6] The competitive advantage gained by each firm's investment in advertising and quality is largely negated when other firms also make this investment. The result is an "arms race" of investment in advertising, product improvements and cost reductions by all firms, fueled by an expanding market. Indeed, firm profitability can actually fall in the long run in an expanding market if the additional profit gained from the increased volume of sales is exceeded by the increase in investment in exogenous sunk costs. And any firm that tries to avoid making this collectively-unprofitable investment will be driven out of the market, because it will lose enough sales to competitors that made the investment to make it unprofitable to stay in business. Thus competitive pressures drive firms to make investments in quality and advertising which in the long run mostly serve to raise the cost of participating in the market.

II. The Cournot Model of Endogenous Sunk Costs

While the theory of endogenous sunk costs applies across a variety of models, for his detailed analysis and simulation results Sutton relies heavily on a Cournot model, modified appropriately. The Cournot model is summarized here in order to demonstrate how the basic framework of the model works. Examination of this model also helps reveal the crucial assumptions that drives some of the theory's results.

It is useful first to examine how an expanding market affects concentration in a Cournot model in which firms have only exogenous sunk costs, to see how industries without endogenous

[4] This paper assumes that endogenous sunk costs are fixed costs that affect the price-cost margin, but a more general definition could be used for endogenous sunk costs, that would likely yield similar results to Sutton's framework. This definition would include three requirements: (1) Firms have a choice of investment opportunities to invest in, each with its own sunk cost; (2) The cost of a particular choice does not depend on the quantity of output; (3) The revenue gained does depend on the quantity of output. Thus brand proliferation, which Sutton discusses in reference to the ready-to-eat cereal market (Sutton 1991, ch. 10), could be considered an endogenous sunk cost under this broader definition.

[5] Because the cost of each investment is fixed while the impact on profits depends on the quantity of output, there is a kind of scale economies produced by endogenous sunk costs. This is explored later in section IV of this paper.

[6] As discussed later, firms may gain benefits in the short run from temporary advantages in quality and cost. Sutton's focus is on the long run, and thus ignores these quasi-rents from short-term competitive advantage.

sunk costs will behave (Sutton 1991, ch. 2). These industries have only a fixed cost of entry, and firms cannot make additional investments to improve their price-cost margin. Modifying Sutton's notation a little, there is a fixed sunk cost of entry σ, and a constant marginal cost c. Let each firm i produce q_i units of output, the industry output be $Q = \Sigma q_i$, and the industry price be P. Sutton assumes that industry demand is determined by $Q = S / P$, so $P = S / Q$, with S representing the size of the market. The market is modeled as a two stage game. In the first stage, firms decide whether to enter and incur the fixed cost σ, which is an exogenous sunk cost. In the second stage, firms in the market choose the level of output q to maximize their profit. Total profits for each firm i are $\pi_i = (P-c) q_i - \sigma$. The short-run profits, which are the second stage profits from the firm's choice of q_i and exclude fixed costs, will in this paper be represented by $z_i = (P-c) q_i$ and called "gross profits."[7] Since decisions made in the first stage depend on the likely profits to be made in the second stage, to find the equilibrium result we first solve for firms' choice of output in the second stage, assuming a fixed number of firms N, which yields their expected gross profit. Solving the first-order conditions yields the equilibrium gross profit for each firm in the second stage of $z_i^* = S/N^2$, so the long-run total profit is $\pi_i = S/N^2 - \sigma$.[8] Firms enter in the first stage if they can anticipate making sufficient gross profit in the second stage to at least cover the cost of entry, i.e., if $z_i \geq \sigma$. Therefore in the first stage firms enter until $\pi_i = 0$ (or as close as possible, keeping $\pi \geq 0$), so in equilibrium $S/N^2 - \sigma \approx 0$ or, denoting the number of firms in equilibrium as N* and letting it vary continuously for simplicity, $N^* \approx \sqrt{S/\sigma}$. Thus as the market size S rises, the maximum number of firms N* rises without limit, and industry concentration, defined as 1/N*, converges to zero, yielding a fragmented market structure.[9]

To analyze the impact of endogenous sunk costs Sutton employs a three stage game, adding the decision to invest in endogenous sunk costs as a separate stage (Sutton 1991, ch. 3). We will call firms' expenditures in endogenous sunk costs (represented by A(u)) an investment in "quality", and this will cover investments in advertising, R&D, and any other fixed cost that raises consumer's willingness to pay. The key assumptions concerning quality that Sutton makes for his model are: (1) the investment in quality is a sunk cost, the cost of which does not vary with the quantity sold; (2) consumers are willing to pay more for a firm's good to the extent that it is higher in quality than other firms' goods (i.e., if one firm raises its quality u above the level of its competitors' quality, call their quality level $\overline{u}$, then consumers will pay a higher price by a factor of $u/\overline{u}$); (3) if a firm raises its quality relative to other firms, its output is adjusted by a factor of $u/\overline{u}$ when calculating total industry output, for the purpose of determining the market price (i.e., $Q = \Sigma q_k + (u/\overline{u})\, q$, if the other firms have output q_k, for calculating $P = S/Q$); (4) if all firms raise their quality to the same level there is no change in the price consumers are willing to pay and in the quantity demanded (i.e., consumer expenditures PQ = S are constant for a market

---

7 Sutton calls the second stage profits "π." In this paper "z" will refer to the firm's (short-run) profits exclusive of fixed costs, i.e., its gross profits, and "π" will refer to the firm's total (long-run) profits. The name "gross profits" is used because it is an accounting term that is closest (but not necessarily the same) to our definition of z.

8 Assuming that $z_i = P q_i - c q_i$, then the first order conditions are $\partial z_i/\partial q = P + \partial P/\partial q\, q_i - c = 0$ for N firms. The equilibrium solution is $Q^* = S (N-1) / c N$, $P^* = cN / (N-1)$, and hence $z^* = S/N^2$.

9 We get essentially the same result if a linear demand function is used for the model with exogenous sunk costs, such as $P = g - bQ$ (so $Q = (g-P)/b$). In this case the second stage equilibrium profits are $z_i^* = [(g - c)/(N+1)]^2 /b$, and the first stage equilibrium number of firms is $N^* = [(g-c) / \sqrt{(\sigma b)}] - 1$. If the market expands, with an increase in g or decrease in b, then N* also increases without limit.

of a given size S, and Q is not quality adjusted for a change in the overall level of quality).[10] The first two assumptions are essential to the analysis of endogenous sunk costs, while the last two can be relaxed, as discussed later. The last assumption is likely most appropriate concerning investments in advertising, and its ability to provide firms with an advantage relative to their competitors by changing the "perceived quality" of the good, as Sutton calls it, which might not affect the inherent quality of the product.

Firms choose whether to enter and incur fixed cost σ in the first stage. In the second stage they choose the level of quality u, at a cost of A(u), in the second stage. A(u) is a fixed cost relative to the quantity sold, but varies with the level of quality chosen. In the third stage they choose output q_i, which yields gross profit (profit exclusive of fixed costs) z_i. Assuming that the market price is P, and that all of the firms begin with the same quality $\bar{u}$, then if one firm raises its quality to u, it will be able to charge $(u/\bar{u})\,P$ for each unit sold, and gain gross profit

$$\mathbf{z_i = ((u/\bar{u})\,P - c)\,q_i} \qquad \textbf{(1)}$$

The total fixed cost for the firm (fixed relative to the quantity sold) is the sum of the costs of entry and of raising quality, so $F=\sigma+A(u)$. The long-run profit for the firm is $\pi_i = z_i(q_i,u_i) - F$, or

$$\boldsymbol{\pi_i = ((u/\bar{u})\,P - c)\,q_i - A(u) - \sigma} \qquad \textbf{(2)}$$

Again, to find the equilibrium outcome, it is necessary to solve for the last stage first, and work backwards through the stages. So the first calculation is to determine the optimal profit-maximizing level of output q_i for every firm in the third stage, while assuming that the number of firms and their choices of quality are fixed, and that one firm has raised its quality above the others. The assumption that one firm has raised its quality above others in the analysis of the third stage is necessary to determine the firm's incentive to raise quality in the second stage.

The second stage equilibrium choice of u for all firms is determined using the Nash assumption that every firm assumes that the other firms hold their quality constant. Sutton assumes that the cost of raising quality is $A(u)=a/\gamma\,(u^{\gamma} - 1)$, and that there is a floor to the level of quality u, so $u \geq 1$. This functional form for A(u) is a general function that allows the cost of increasing quality to grow quickly or slowly, depending on the size of the parameter γ, and thus permits us to investigate behavior under a broad range of cost conditions for an industry. Solving the first order conditions shows that under certain conditions firms have an incentive to invest in higher quality, and that in equilibrium all firms will raise their quality to the same level in the second stage.

In the first stage firms enter until $\pi = 0$, yielding an equilibrium number of firms N^*, assuming for convenience that N is continuous. The full solution to the Cournot model of endogenous sunk costs is provided in Appendix A to this paper. Appendix A also calculates sufficient conditions for an interior solution in quality that are looser than what Sutton specified.

[10] The industry output Q used to calculate the price in P = S/Q thus reflects only changes in relative quality between firms, and not the overall quality of the product.

The equations for the equilibrium number of firms in the first stage are very complex and cannot be solved analytically. Sutton asserts, however, that under a broad range of conditions, the equilibrium number of firms N* is bounded above, and as the market size increases industry concentration tends to an asymptotic value greater than zero (Sutton 1991). This paper provides a proof of this proposition in Appendix B, along with estimates of the upper bound and the asymptotic value. Indeed, if the entry cost σ is low enough such that $\sigma < a/\gamma$, then the relationship between S and N* is not "monotonic", such that as the market expands the number of firms in the market rises, reaches a peak, and then falls. Thus an increase in market size may actually induce a reduction in the number of firms in the industry.[11]

III. The Intuition Behind Endogenous Sunk Costs Theory

To understand the endogenous sunk costs model it is important to note two key effects at work here. First, each firm in an industry with endogenous sunk costs has an incentive to invest in quality, but the firms collectively gain little or no benefit if all do so.[12] Just as in the traditional Cournot model firms' individual output in equilibrium exceeds the joint profit-maximizing level, firms here invest in more quality in equilibrium than would be collectively desirable. There is, in effect, a negative externality of individual firms' decision to invest in quality, impacting on the profits of the other firms. Sutton makes the rather strong assumption for his model that market demand and the market price are completely unaffected if all firms equally raise their quality, and thus any investment by firms in endogenous sunk costs to raise quality is wasted money (from a long-run profit-maximizing point of view). Since this assumption is not a necessary feature of the theory,[13] Sutton's Cournot model will sometimes be treated as a special case of the general theory of endogenous sunk costs in the following discussion.

The second key effect is that firms' individual incentive to raising their quality increases with market size. The reason for this lies in how investments in exogenous sunk costs affect the

[11] For his theoretical analysis concentration is defined as 1/N*. Therefore for concentration to have a lower bound, as Sutton asserts exists for industries with endogenous sunk costs, there has to be an upper bound to the number of firms N*. Note that in his empirical work he focuses on the 4-firm concentration ratio, since he allows for numerous small fringe firms to exist in a market. He asserts that these small firms are, in effect, serving a different market segment or niche, compared to the larger firms. Thus the impact of endogenous sunk costs in a growing market is observed empirically not by a reduction in the number of firms, but in an increase in size of the largest firms, and a squeezing out of the mid-sized firms. Sutton 1991, pp. 174-75.

[12] Firms' failure to benefit from an increase in quality is a long-run equilibrium result. In the real world first-moving firms are likely to gain a temporary advantage from their investments, thus allowing them to gain quasi-rents. In some circumstances this advantage can last a long time or be permanent if there are first-mover advantages to be gained, such as from establishing a brand name or gaining a patent. Meanwhile, followers may benefit from their ability to copy the leaders, if that is possible, which gives them a lower cost of raising quality. Sutton's endogenous sunk costs model does not take these two effects into consideration. If incorporated into the model, the presence of first-mover advantages would likely accelerate investment in quality, while potential benefits from being a follower would probably slow it down.

[13] If instead consumers are assumed to react to an increase in overall increase in quality, i.e., an increase in utility gained per unit purchased, with an increase in purchases (e.g., $Q = S\ r(u) / P$, with $dr/du>0$), then an industry investment in quality would shift the demand curve to the right, raising all firms' gross profits in the short run. Industries with endogenous sunk costs will still have a lower bound on concentration if firms have a sufficiently strong incentive to invest in endogenous sunk costs that their equilibrium choice of quality exceeds the joint profit-maximizing level.

firm's costs and revenues. The cost of raising quality is a fixed cost, which does not vary with the quantity produced. The benefit from a particular increase in quality, which raises by a fixed amount the price-cost margin of the good, rises with an increase in quantity sold. The firm will only invest a dollar in raising quality if it gains that back in added gross profits. With higher sales a given increase in the price-cost margin yields greater incremental revenues, which makes an investment in quality more desirable.[14] Therefore as the market grows in size, firms will likely find it profitable to invest in higher quality. The size of this investment depends on the cost of investing in quality, as reflected in the shape of the function A(u).

The result of these two effects is that an expansion of the market leads to increased investment in endogenous sunk costs and hinders entry. Because this investment yields little or no additional revenues in the long run, since all firms match it, it eats into profits and reduces the profitability of entry. Thus for a potential entrant, the attraction of serving a larger market is largely offset by the higher cost of entry. Incumbent firms' profits could even fall as a result of an increase in market size, before entry or exit occurs, likely leading to firms exiting and raising concentration. This counter-intuitive result will occur if an increase in market size induces an investment in quality that exceeds the increase in gross profits caused by the increase in sales. While this can happen under certain circumstances, the model more generally predicts that industries with significant endogenous sunk costs will always have a lower bound to concentration.

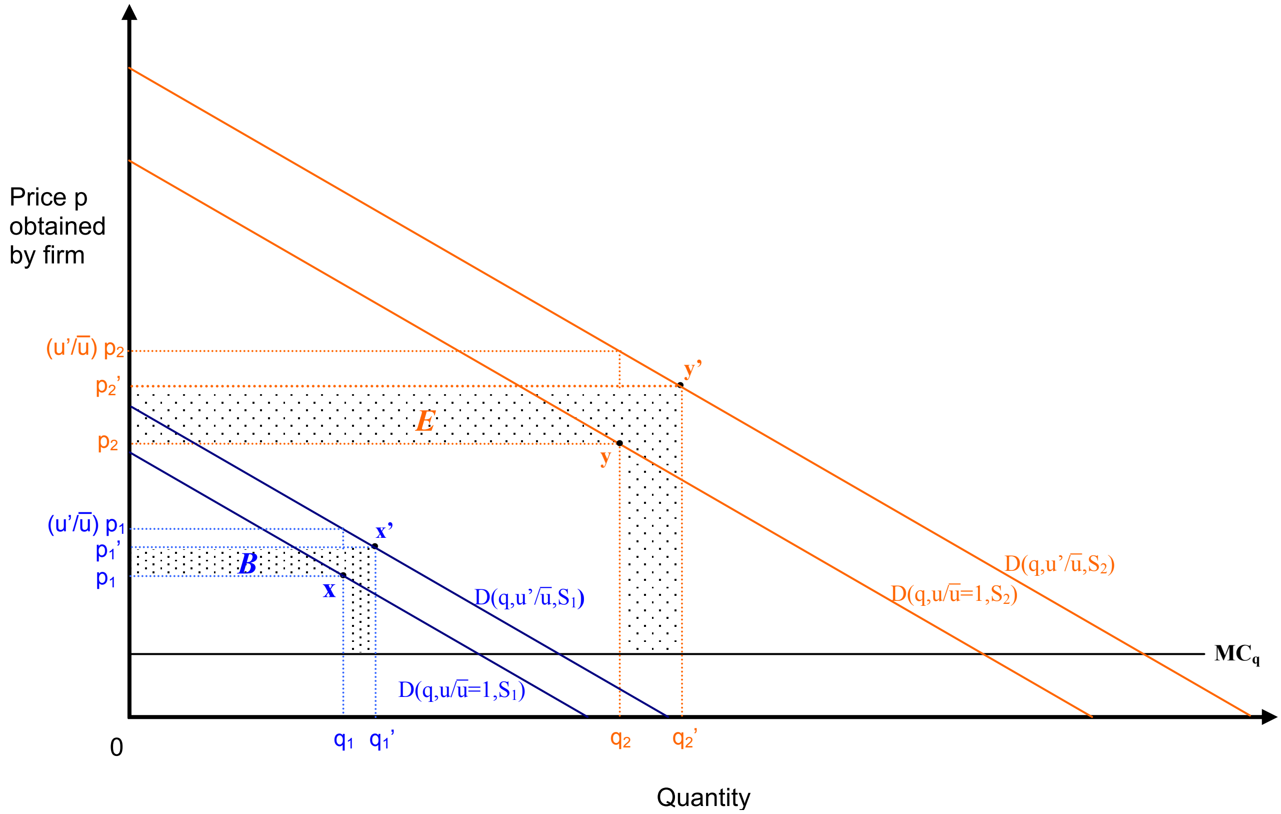
The basic intuition underlying the theory of endogenous sunk costs is illustrated in the following diagrams. Some of the analysis here is based on the economics literature on advertising (Dixit & Norman 1978; Carlton & Perloff 4th ed., ch. 14). Note that most of the key effects discussed here do not depend on the use of the Cournot model, supporting Sutton's contention that his results will apply broadly to a variety of industries that have substantial endogenous sunk costs.

The impact of an increase in market size on firms' incentive to invest in quality is illustrated in Figure 1. This diagram is adapted from the usual diagram employed for examining the incentive to invest in advertising (Carlton & Perloff 4th ed., p. 458; Dixit & Norman 1978, p. 4). We examine one firm's incentive to deviate from the average market quality $\bar{u}$. The firm has constant marginal cost with respect to output of MC_q. As represented by the firm's demand curve $D(q,u/\bar{u},S)$, the price the firm receives depends on the firm's choice of output q, its quality u relative to the industry average quality $\bar{u}$, and the size of the market S. The firm is assumed to charge a price directly proportional to its quality advantage over the industry average, such that $p_d = u/\bar{u}\ p$, where p_d and u are the deviant firm's price and quality. We ignore for the time being the likely response of other firms to changes in quality and output by the deviant firm.

[14] Thus if a firm's investment in quality raises the price it can charge by $(u/\bar{u}-1)P$, then the firm will gain increased revenue of $(u/\bar{u}-1)P\ q_i$. Since the investment in quality A(u) does not depend on the firm's sales, the profitability of the investment in higher quality, which is equal to $(u/\bar{u}-1)P\ q_i - A(u)$, rises with q_i. For example, if the rise in price due to the quality improvement is $2 per unit, and the firm sells 10 units, then it would be profitable to invest in this quality improvement if the improvement cost less than $20. If the firm instead sells 20 units, then the firm would be willing to pay up to $40 for the improvement. See Figure 1 for a diagrammatic depiction of this.

Figure 1

The Profitability of Increasing Quality u as Demand Increases for the Firm's Output



The market size is initially S_1, and assuming the firm produces at the same quality as everyone else, so $u=\bar{u}$, the firm has demand curve $D(q,u/\bar{u}=1,S_1)$. The firm's profit-maximizing choice of output and price is assumed to be at x, selling q_1 units at price p_1. The firm's gross profit is $(p_1-MC_q)\, q_1$, which is the area between the $p=p_1$ and MC_q lines, from $q=0$ to $q=q_1$. Suppose that the firm has the option of investing to raise its quality to u' from the industry average $\bar{u}$. If it invests in u', it will raise its demand curve to $D(q,u'/\bar{u},S_1)$, which yields a price of $(u'/\bar{u})\, p_1$ at output q_1. The firm increases its output to q_1', and thus moves to x' with price p_1' on the diagram.[15] This increases its gross profits by the area *B* on the diagram. It will only make this investment if this increase in gross profit exceeds the rise in fixed cost A(u) of raising u to u'. Assuming for simplicity that $\bar{u}=1$ and so $A(\bar{u})=0$, the firm gains additional profit from raising its quality to u' of *B* – A(u'), and it will not make the investment if this is less than zero.

Assume next that the market expands to S_2, raising the firm's demand curve to $D(q,u/\bar{u},S_2)$, and that the firm's profit-maximizing choice of output and price is now q_2 and p_2, before any investment in quality. The impact of investing in raising quality to u' from u is to shift the demand curve to $D(q,u'/\bar{u},S_2)$, causing the firm to move from y to y', where it charges price p_2' and output q_2'. The additional gross profit from investing in u is *E*. Note that the investment in quality u' causes the same proportionate shift upwards in the demand curve D in

[15] With an increase in quality, the firm will not just be able to charge more for its product, but it will likely want to change the quantity it sells. For example in the Cournot model the firm's optimal choice of output is $q^* = q_k\,[(N-1) - (N-2) / (u/\bar{u})]$, which depends on the firm's quality u as well as the output of the other firms q_k.

the larger market as in the smaller market, with the firm's price rising by a factor of $u'/\bar{u}$ for a given quantity of output. Under most circumstances, however, *E* will be larger than *B*, since this proportionate increase in demand has a greater impact on firm revenues in the larger market, with its higher equilibrium price and/or quantity sold. In Sutton's basic model, in which the equilibrium market price is invariant to the market size ($p_2 = p_1$), the benefit to the firm from raising quality (i.e., the added gross profit) is always greater in larger markets, assuming marginal costs are constant. This is because the increase in the price-cost margin obtained by a firm from an investment in quality is applied to a larger quantity of output in the larger market, yielding a higher gross profit. In other models with less restrictive assumptions, the increase in market sales could cause a rise in the equilibrium price, as depicted in the figure, such that a rise in quality would cause a greater increase in price in absolute terms in the larger market than in the smaller market. I.e., since the change in price p_t in market t is $\Delta p_t = (u'/\bar{u})\, p_t - p_t = (u'/\bar{u} - 1)\, p_t$, if $p_2 > p_1$ then $\Delta p_2 > \Delta p_1$, and the firm will gain a higher price-cost margin for each unit sold in market 2.

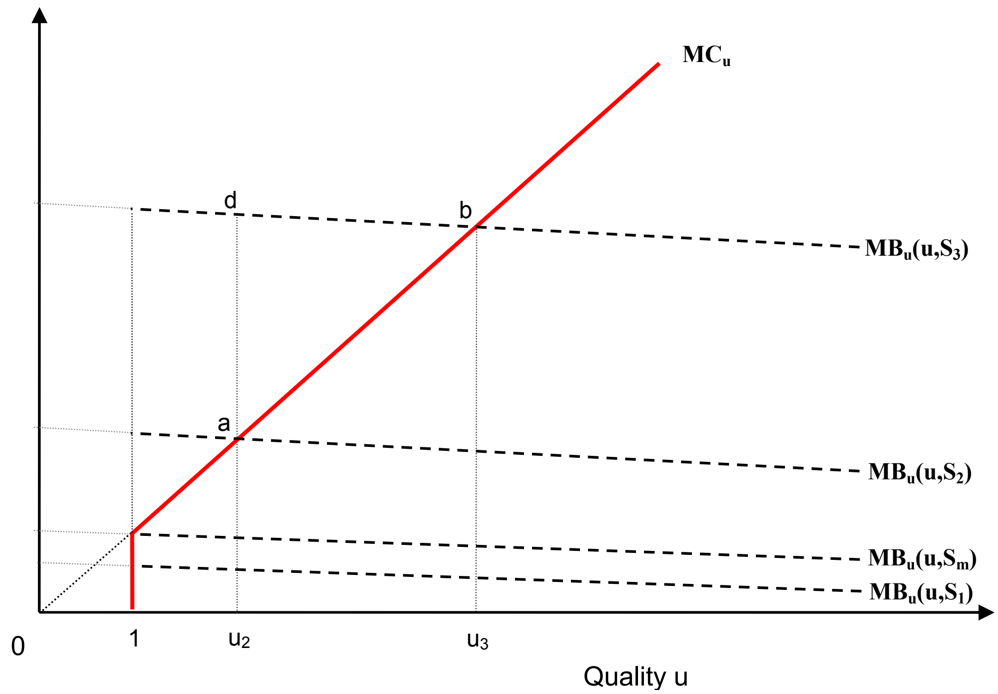
Since the cost of the investment A(u) is invariant with respect to output, the potential gain from such an investment in this larger market is *E* – A(u), which is larger than in the smaller market (*B*−A(u)). If the cost of the investment is between *B* and *E*, i.e. *B* < A(u) < *E*, then the firm will make the investment A(u) only in the larger market. Thus a rise in market demand increases the firm's incentive to invest in higher quality. The key assumption that drives this result is that investments in endogenous sunk costs are fixed with respect to their cost, but the benefit gained depends on the size of the market. It was also assumed for this analysis that there is some level of industry demand for which it is profitable for a firm to invest in endogenous sunk costs.

In Figure 2 marginal analysis is used to show how the firm's profit-maximizing choice of quality increases with market size. The MC_u curve represents the marginal cost of raising quality, and is equal to dA(u)/du. The MB_u curves represent the marginal benefit gained from raising quality for various market sizes, and are equal to dz/du, which is the increase in gross profits z caused by raising quality. We assume again that $\bar{u} = 1$. Figure 2 is related to Figure 1 in that the marginal benefit of raising u for different market sizes correspond to the values *B* and *E* in Figure 1 (the marginal cost of u cannot be observed in Figure 1, however). The marginal cost of quality curve slopes up in Figure 2 because it is assumed that raising consumers' willingness to pay by a given proportion becomes more costly for higher levels of u. Thus the cost of raising consumers' willingness to pay by 20 percent through an investment in quality is more than double the cost of raising it by 10 percent.[16] Because these are assumed to be fixed costs, this curve is invariant to the firm's output and the market size. The marginal benefit of increasing quality, on the other hand, does depend on the market size, as shown in the figure, with larger market sizes represented by higher subscripts ($S_3 > S_2 > S_1$). Consistent with our previous discussion of Figure 1 and how the benefit to the firm from investing in quality will usually be

[16] Using Sutton's functional form of $A(u) = a/\gamma\, (u^{\gamma} - 1)$, with a and γ being fixed parameters, then the marginal cost of increasing u is $MC_u = dA(u)/du = a\, u^{\gamma-1}$, and since Sutton assumes a>0 and γ >1, its slope is $d^2A(u)/du^2 = a\, (\gamma - 1)\, u^{\gamma-2} > 0$. In addition, if fixed costs $F = \sigma + A(u) = \sigma + a/\gamma\, (u^{\gamma} - 1)$, in order to satisfy the second order conditions on a firm's choice of quality u and obtain a well-behaved solution, Sutton assumes that $\gamma > \max\{1, 2a/3\sigma\}$. See Appendix A below, and Sutton, 1991, pp. 53-54 and Appendix 3.1, pp. 369-71. As noted in Appendix A below, this restriction could be relaxed a little to $\gamma > \max\{1, a/2\sigma\}$ and still satisfy the second order conditions.

larger in larger markets (so $E > B$), we assume that growth in the market causes the marginal benefit curve to shift up. As noted before, in Sutton's model this is always the case.[17]

Figure 2
The Profit-Maximizing Choice of Quality for the Single Firm as the Market Size Increases



A firm will want to invest in raising quality if the investment increases gross profits z more than it increases fixed costs A(u). The firm's profit-maximizing level of quality is therefore where the marginal cost and marginal benefit curves intersect, i.e., the level of u where $dz/du = dA(u)/du$. At any other level of quality the firm can increase profits by moving to this profit-maximizing level of quality.

Assume that the initial market size is S_2. Then the firm's profit-maximizing level of quality is u_2. As the market expands, and the market size increases to S_3, the firm's profit-maximizing choice of u rises from u_2 to u_3, since the marginal benefit of increasing u has risen. The additional profits the firm expects to gain by increasing u in response to the expansion of the market are represented by the area of the triangle abd. Note that the other firms in the market will also raise u, and thus in equilibrium no one actually earns more gross profits in Sutton's model. The market's "supply" curve of quality for every level of demand is the solid line, which coincides with the marginal cost of quality curve for $S \geq S_m$.

At small levels of output there may be no investment in endogenous sunk costs. This can be observed in Figure 2 for market size S_1. At this level of output, since the marginal benefit of increasing u is below the marginal cost, the firm will not want to increase its quality above its

[17] In Sutton's model in equilibrium, using equation (1) and assuming $\bar{u} = 1$, $MB_u = \partial z_d/\partial u = P\, q_d^* + u\, \partial P/\partial u\, q_d^* + (uP - c)\, \partial q_d^*/\partial u = 2S\ [u - 1 + 1/(N-1)] / [u + 1/(N-1)]^3$. Thus MB_u rises linearly in S.

initial level of u=1. Only as the market size exceeds S_m (e.g., at $S=S_2$) does it become profitable for the firm to invest in raising its quality. S_m is the market size at which the marginal benefit and marginal cost curves intersect at u=1, the minimum quality permitted in Sutton's model. Therefore S_m is the minimum level of market size at which it becomes profitable for firms to invest in endogenous sunk costs, which Sutton calls the switching point.[18] Since all firms produce the same quantity of output in equilibrium, there is a minimum firm output $q_m = S_m / N$ for which the firm begins to invest in endogenous sunk costs.

It should be obvious from Figure 2 that the existence of a minimum market demand at which firms begin to invest in raising quality is due in large part to Sutton's restriction that firms cannot lower their quality below its initial level of u=1. In Sutton's framework at u=1 firms have zero investment in advertising or R&D (since A(u=1)=0). If this restriction were eliminated, allowing u to fall below one,[19] there would be no discontinuity in investment at $S=S_m$, although S_m would still represent the minimum market demand at which firms would want to increase quality beyond its starting level. For example, if the starting market size were S_1 in Figure 2, firms would want to *lower* their quality below one, which Sutton's model does not permit. Not until the market size grew to exceed S_m would they begin to raise quality above its initial value.

The existence of this minimum market demand S_m as a discontinuity in firms' investment functions plays a significant role in Sutton's analysis. In particular, Sutton's result that the relationship between S and the equilibrium number of firms N* is non-monotonic under certain conditions (see Figure 5, Figure 6, and Figure 7 below) is the result of the lower bound placed on quality, with the number of firms peaking at S_m. As discussed later, without this restriction the relationship would always be monotonic.

As with any marginal analysis, whether and how much an increase in market size induces an increased investment in endogenous sunk costs depends on how this investment affects revenues and costs. The curvatures of the marginal benefit and marginal cost curves in Figure 2 will have a significant impact on how a firm's incentive to invest in quality is affected by a change in market size. For example, if the marginal cost curve is steeper than shown, reflecting a high cost to raising quality, then a rise in market size from S_2 to S_3 will cause a smaller increase in quality u.

Figure 3 illustrates how the firm's preferred monetary investment in quality rises with its output, and how this is affected by the shape of A(u), which determines the marginal cost of

---

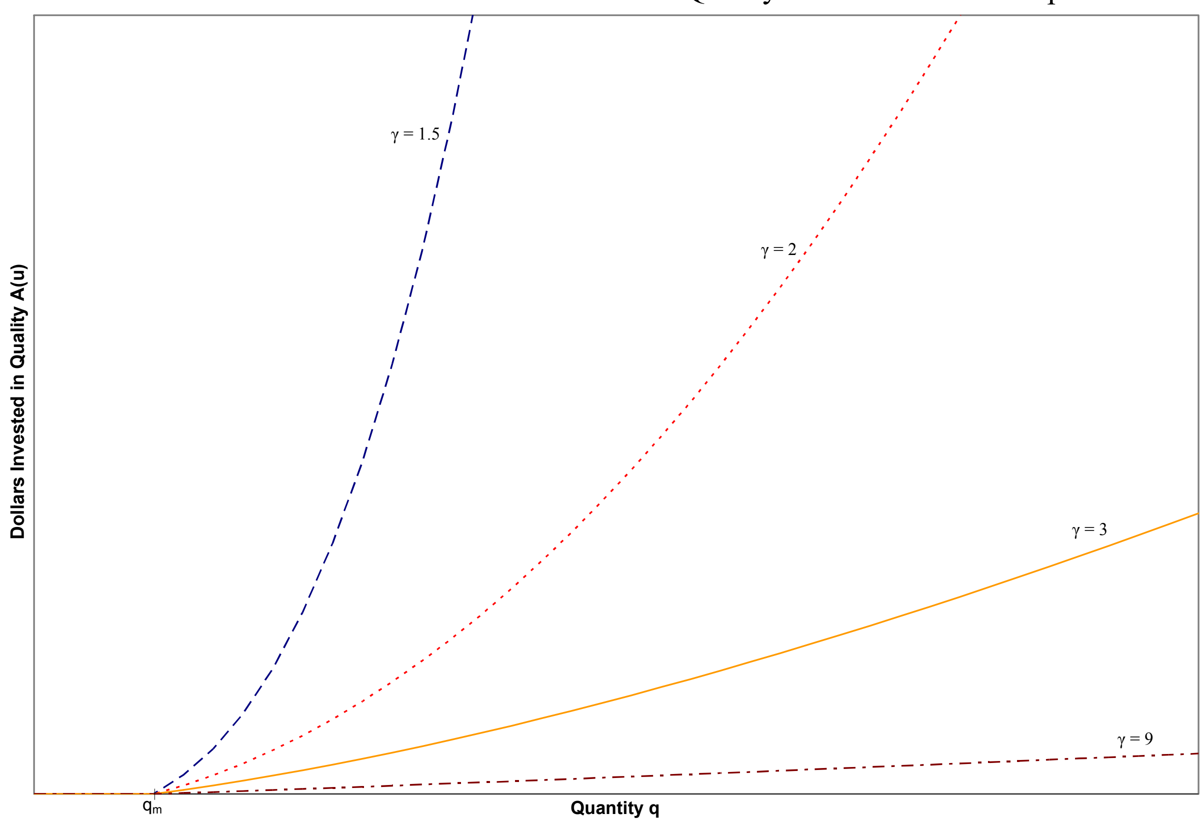
[18] For some industries, S_m is quite high. These industries may have only exogenous sunk costs, because the returns to advertising and R&D and other fixed investments in quality-enhancements and cost-reductions are too low.

[19] While the idea of firms lowering quality below its initial level (so consumers willingness to pay actually falls) may sound nonsensical, for example if it refers to the design of the product (why would anyone switch to an inferior design?) or to the money that is spent on advertising (how can anyone spend negative dollars on advertising?) or R&D (the desirability of the product should not decline in the absence of R&D expenditures), in certain contexts it might be a reasonable possibility. For example, a firm's initial level of advertising and marketing may not be zero, especially if it employs a staff for the functions of marketing and design. If a firm fails to continue to invest in advertising and marketing, its product's sales may decline, especially if it is the kind of product that is subject to the whims of fashion (e.g. clothing). With some products the need for advertising may decline once the product has achieved consumer acceptance and retail distribution (hence the rise of generic and "No-Ad" goods). In some industries particular equipment may be needed to achieve higher quality (e.g., movie industry) and lower costs (e.g., chemical plants). In small markets firms might decide to not purchase this equipment and allow their quality to fall.

quality.[20] In the model the parameter γ determines how quickly the marginal cost of quality rises with u, with higher γ associated with a faster increase in the cost of raising quality (i.e., a steeper MC_u curve in Figure 2). From Figure 3 it is evident that the firm's preferred investment rises more quickly for lower values of γ, when the marginal cost of increasing u is low. Thus the flatter the MC_u curve (low γ), the faster spending on endogenous sunk costs rises with an increase in the firm's output. q_m is the minimum output level for the firm at which it becomes profitable to invest in raising quality.

Figure 3

Firm's Preferred Choice of Investment in Quality as a Function of Output



The analysis so far has shown that an expansion in the market provides each firm with an incentive to invest in higher quality, and that the size of this investment depends on various structural factors. The resulting Nash equilibrium in the Cournot model for the second stage, with the number of firms fixed, has every firm investing equally in higher quality, as shown in Appendix 1.[21] This investment in quality turns out to be collectively harmful to industry profits

---

[20] Because of the difficulty in solving for the firm's preferred choice of quality for the second stage in the Cournot model (assuming $\bar{u}$=1, the preferred quality u* is implicitly defined in $a\,u^{\gamma-1}\,[u+1/(N-1)]^3 / [u-1+1/(N-1)] = 2S$), which incorporates other firms' changes in output (but not quality) in reaction to the deviating firm's new choice of quality, a simplified calculation is used for Figure 3 which assumes that other firms' output is fixed. Holding other firms' choice of output and quality constant, the firm's preferred quality is $u^* = (pq / a\bar{u})^{1/(\gamma-1)}$, and its investment in endogenous sunk costs is $A(u^*) = a/\gamma\,((pq / a\bar{u})^{\gamma/(\gamma-1)} - 1)$, provided that $q>q_m$. These results should not differ much from the Cournot model's predictions for the 2nd stage. The diagram assumes that a=5, $\bar{u}$ =1, and p=1, and q=[1,40].

[21] A typical method of demonstrating the (Cournot) Nash Equilibrium for this kind of problem involves determining the best-response (reaction) function for each firm, which is each firm's preferred response to other firms' choice of

in the long run, however, because competition between the firms will tend to drive the price back down to cost, thus negating much of the benefit of raising consumers' willingness-to-pay.

Thus as the market expands the increase in revenues will be accompanied by an increase in fixed costs. The rise in profits is then reduced by the growing investment in endogenous sunk costs, and this will reduce firms' incentive to enter the market. Thus increased investment in endogenous sunk costs prevents concentration from falling as quickly as would be expected from the increase in market size, compared to markets that lack endogenous sunk costs. Under certain circumstances, indeed, an expansion in market size can induce a large enough investment in endogenous sunk costs to turn profits negative for existing firms, causing some firms to exit. In this case firms' induced investment in endogenous sunk costs exceeds the additional revenues gained from supplying a larger market, such that the larger market size has actually caused a decline in existing firms' long-run profitability and a rise in market concentration.

Nor can any firm avoid investing in endogenous sunk costs, if the market expands and other firms raise their quality. Firms that attempt to keep their quality the same as before in the face of a market expansion and a resulting increase in average market quality, will find that customers will only purchase their good at a discount to the market price. Thus as the market expands resistors to the hike in quality will see their price-cost margin shrink and profits fall, and eventually will lose money and have to exit the market.[22]

The impact of expanding industry sales on individual firms' investment in endogenous sunk costs, and on their profits, in equilibrium is illustrated in Figure 4. This simulates the Cournot model's second stage equilibrium, in which the number of firms is fixed and the firms choose output and quality.[23] Note that, unlike in the analysis discussed above, the results plotted in this figure incorporate firms' reactions to each other's choices of output and quality, and the resulting Nash equilibrium. The simulation results in the figures following rely on the Cournot model. Figure 4 shows how rising sales can affect gross and long-run profits, fixed costs, and equilibrium quality. Initially, when sales are low ($S<S_m$), firms invest in no endogenous sunk costs and quality stays at $u^*=1$. As the market grows beyond the threshold S_m firms begin to invest in quality. This increases fixed costs $F(u^*)$ and lowers profits. Within this particular simulation firms' fixed costs $F(u^*)$ eventually exceed their gross profit z^* at sales levels of $S>S'$,

---

quality, and finding the Nash equilibrium. Sutton used assumptions, however, that were sufficient for determining the equilibrium without using best-response functions. In fact best-response functions cannot be calculated in Sutton's model, since Sutton defined industry output assuming that either all firms have the same quality, or just one firm deviates from the industry average ($Q=\Sigma q_k + (u/\bar{u})q_d$). To calculate a best-response function requires reformulating the basic setup, by defining industry output for situations when two or more firms change their quality. One possibility that is compatible with his assumptions is to let the lowest quality be a floor: $Q=\Sigma(u_k/u_m)q_k$, with $u_m=\min(u_1,u_2,\ldots)$.

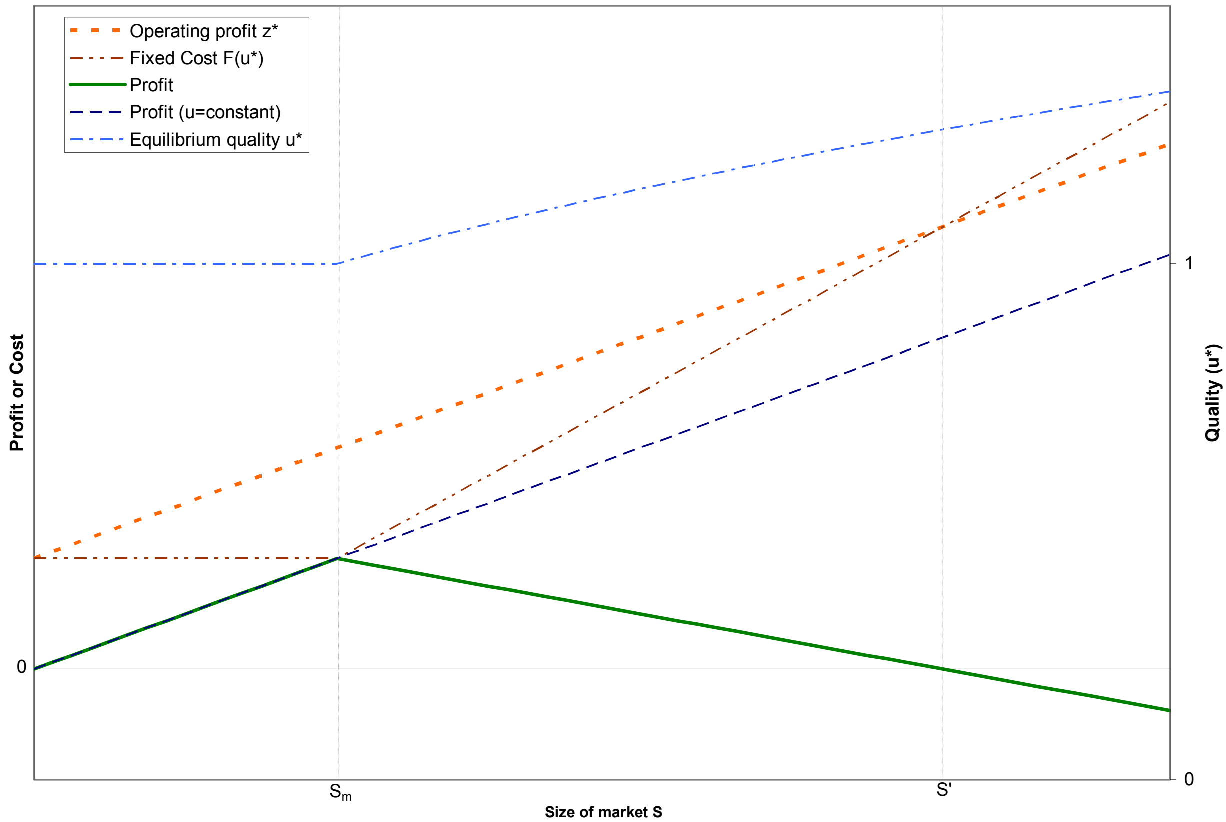
[22] Not raising quality allows the firm to avoid having to pay A(u) in the long run, but reduces the firm's gross profits, because it gets a lower price, equal to $u/\bar{u}\ p$. The reverse effect from that shown in Figure 1 is now obtained. Remember that in the long-run (1st stage) equilibrium, profits are zero, even if all firms have invested in a higher quality. Examining Figure 1, at output q_1 the firm gives up *B* by not investing in A(u), and thus has a change in profits of A(u)-*B*. If output rises to q_2, the foregone revenue rises to *E* while the savings remains at A(u). Eventually the firm loses money and exits, if it does not increase quality to match the other firms.

[23] The equations are provided in Appendix A for the 2nd stage equilibrium. The parameter values for Figure 4 are N=4, σ=20, γ=3, a=180, and S=[320, 1520]. All quantities shown are for each firm in the market, in equilibrium in the second stage. The equilibrium quality u* is graphed according to the right vertical axis.

so for S>S' total profits fall below zero.[24] If firms did not invest in endogenous sunk costs, on the other hand, so quality stayed constant, profits would continue to rise with S for $S>S_m$, as shown in the figure with the heavy dashed line.[25]

Figure 4

Impact of Change in Market Size on Second Stage Equilibrium, with Fixed Number of Firms



This simulation illustrates the basic argument that as the market expands, the increase in endogenous sunk costs will be significant enough to make further entry unprofitable, and can even cause firms to leave the market, thus increasing market concentration.[26] Entry is likely to occur when profits are greater than zero. The rise in endogenous sunk costs eats into profits, however, which dampens entry. When S>S' profits are negative and some firms will likely exit the market. Contrast this with what would happen if this industry had no endogenous sunk costs, and quality was constant. In this case profits would continue to rise, as shown in the figure by the constant-quality profit line, attracting further entry. The growing gap between the constant quality and variable quality profit lines shows how differently industries dominated by

---

[24] There is another scenario, with a different choice of parameters σ, γ, and a, in which profits for a fixed number of firms approach but never fall below zero.

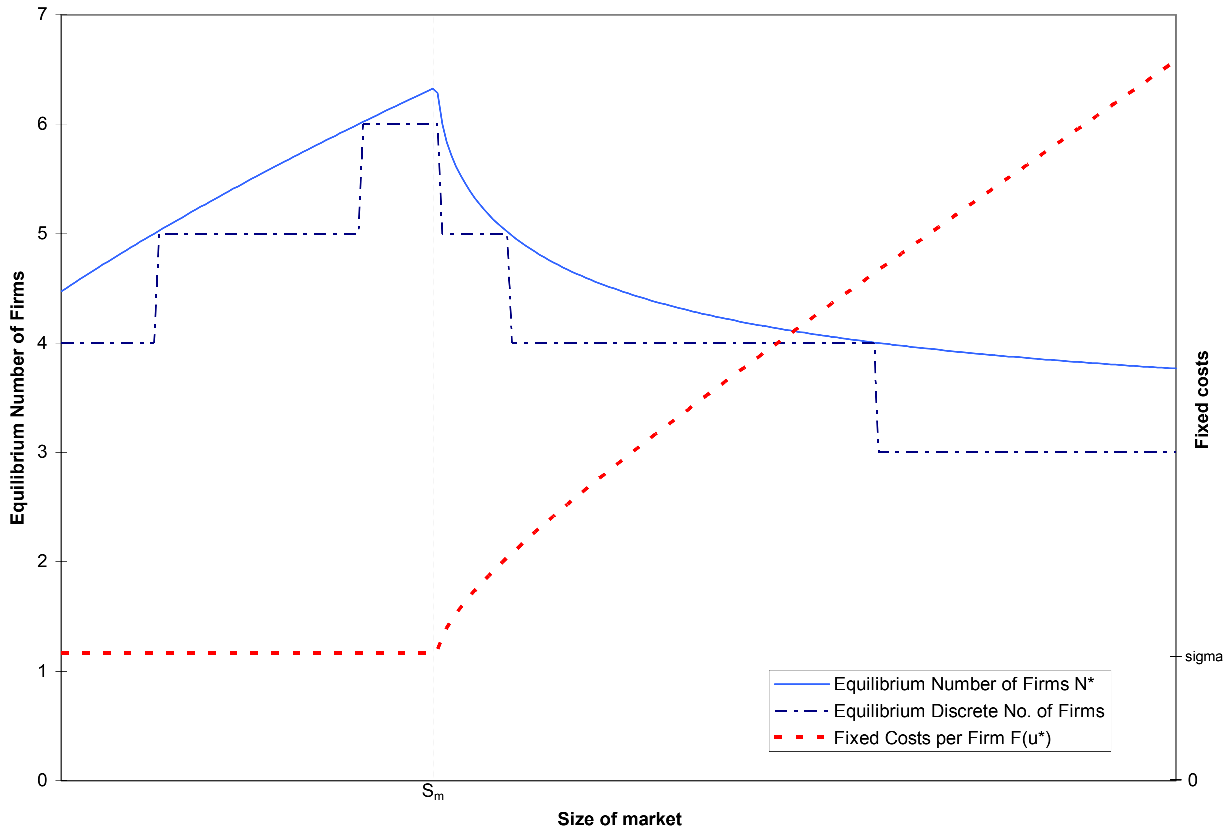
[25] The line "Profit (u=constant)" shows how profits continue to rise as S increases when firms make no investments in endogenous sunk costs. In this case $\pi = S/N^2 - \sigma$, since u=1 and A(u)=0.

[26] While Sutton concluded only that there is just an upper bound on the number of firms in the market, we focus for the moment on the possibility, which he said was of "central importance" (Sutton 1991, p. 59), that increases in market size will cause firms to exit the market and increase concentration.

exogenous sunk costs will evolve as market sales increase, compared to industries with substantial endogenous sunk costs.

The full first stage equilibrium endogenously determines the quantity and quality of output and the number of firms in the market, for a given market size. Firms enter if they can profitably do so, and incumbents exit if they are unprofitable. Thus the equilibrium condition for the first stage of the model, in which firms decide whether to enter or exit the market, is simply that expected firm profits should be greater than or equal to zero. The impact of an increase in market size is illustrated in the simulation results presented in Figure 5. The equilibrium number of firms N* is plotted, assuming that it can vary continuously to ensure that long-run profits are zero, as well as the discrete integer number of firms which can participate in the market with non-negative profits.[27] Fixed costs are plotted according to the right axis, and are calculated per firm assuming the equilibrium number of firms is continuous.

Figure 5
Impact of Change in Market Size on Equilibrium Number of Firms



[27] The equilibrium equations are the same as for Figure 4, with N* now allowed to vary continuously to satisfy the equilibrium condition $\pi = z^* - F(u^*) = 0$. N* is then defined implicitly by equation (7), which cannot be solved analytically, so numerical methods must be used. Note that gross profits equal fixed costs ($z^*=F(u^*)$). For Figure 5 the same parameter values were used as for Figure 4, except N* can vary and S=[400, 1600]. S_m is larger here than in Figure 4 because its size depends on the number of firms in the market (*see* Appendix A).

For small market sizes ($S<S_m$) fixed costs are constant, since it is unprofitable for firms to invest in endogenous sunk costs. As the market size increases profits rise and firms enter the market, so long as S remains below S_m. The rise in the equilibrium number of firms is what we would expect to occur for an industry with just exogenous sunk costs. Once S exceeds the threshold S_m, however, firms have an incentive to invest in endogenous sunk costs, and fixed costs begin to rise. The rise in fixed costs is sufficient to cause profits to fall, driving some firms out of the market. The expansion of the market has thus caused concentration to rise, despite the rise in gross profits.[28]

While Sutton's focus, and ours so far, has been on a possible rise in market concentration in equilibrium, Sutton does not predict that an expansion of the market in an industry with endogenous sunk costs will necessarily raise concentration, but just that concentration has a lower limit. Sutton argues that for a broad range of conditions, as the market expands fixed costs will rise sufficiently to limit the growth of the equilibrium number of firms N*. In other words, N* may continue to grow, but will always be bounded above. In fact, in his model N* asymptotically approaches a limit, call it N_a. Sutton's proof that N* is bounded above in his Cournot model is incomplete, however. A complete proof is provided in Appendix B to this paper.

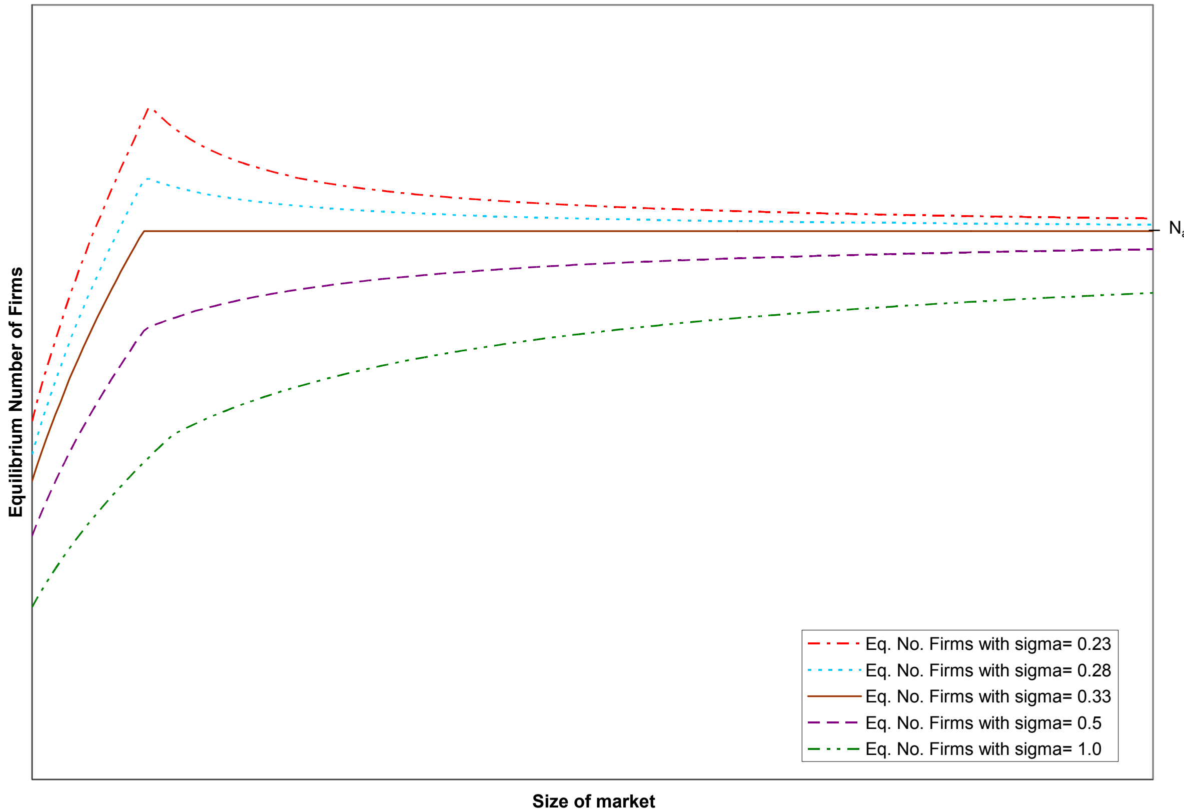
These results are demonstrated in the simulation results shown in Figure 6.[29] As the market size grows, the equilibrium number of firms N* initially increases, then tends toward the asymptotic value N_a. How N* behaves as S rises depends on certain parameters, and on whether S is greater than S_m. The threshold for firms to invest in endogenous sunk costs, S_m, is that point on each curve where it suddenly changes slope (S_m is different for each curve). For small market sizes, with $S<S_m$, N* rises as S increases, since $N^*= \sqrt{(S_m/\sigma)}$, and reaches its maximum at $S=S_m$. For larger market sizes ($S>S_m$), how N* behaves depends on the relationship of σ to a/γ, as shown in the figure. Because the scenario of $\sigma=0.33$ has $\sigma=a/\gamma$, the equilibrium number of firms is constant at $N^*=N_a$ for $S>S_m$. The scenarios $\sigma=0.23$ and $\sigma=0.28$ have $\sigma<a/\gamma$, and they peak at S_m and then fall. The scenarios $\sigma=0.5$ and $\sigma=1.0$ have $\sigma>a/\gamma$, and they steadily rise to approach, but never exceed, N_a. N* is bounded above in all cases.[30]

[28] There are a number of ways that the decline in profits can cause the number of firms to shrink. Those firms that are too small to make the necessary investment, or do not recognize the importance of those investments, or are unsuccessful in making the investments, will find that the price they obtain is insufficient to cover their costs, and thus may decide to leave the market. Alternatively, in the face of rigorous competition in quality, some firms may choose to merge together, to gain the size needed to profitably make the necessary investments in quality.

[29] The simulation used these values in the Cournot model: $\gamma = 3$; $a = 1$; $\sigma = 0.23, 0.28, 0.33, 0.5, 1.0$; with S varying from 1 to 13. N* is defined by equation (7) in the text.

[30] In effect, N* is bounded above by $2+½ \max\{\gamma,a/\sigma\}$, as shown in Appendix B.

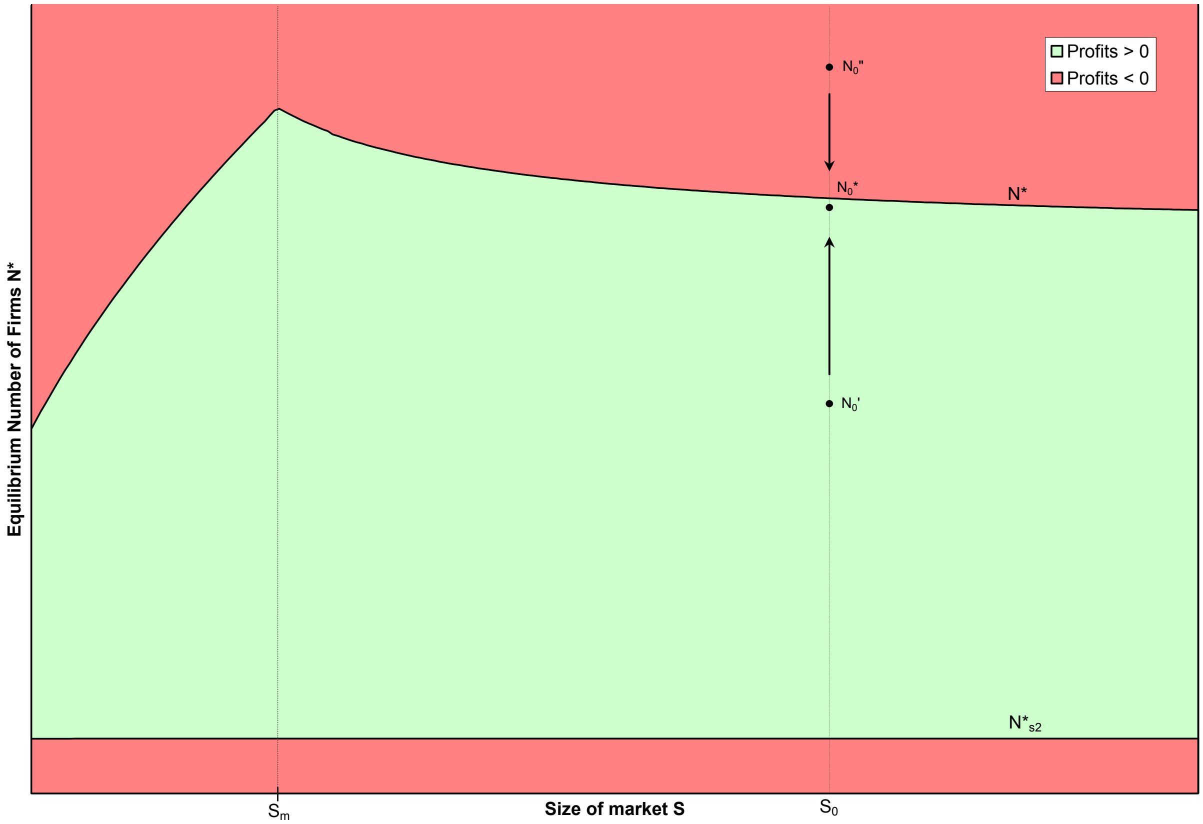
Figure 6
Equilibrium Number of Firms as Market Size Grows, for Various Values of σ



When entry and exit will occur is illustrated in Figure 7 below, assuming that $\sigma < a/\gamma$. For any given S and N, we can calculate the second stage equilibrium profits π^*, according to equation (5) provided in Appendix A. N^* and N^*_{s2} are the solutions to $\pi^* = 0$, but as explained in Appendix A, only N^* is the equilibrium outcome for the model. The area between N^*_{s2} and N^* is where firms are making positive profits. Firms are losing money ($\pi^* < 0$) when $N > N^*$ or $N < N^*_{s2}$. Referring to the figure, if for market size S_0 there are N_0' firms in the market, then firms are making positive profits, and firms enter until $N=N_0^*$. If there are N_0" firms, then firms are losing money, and some firms exit until $N=N_0^*$. Thus the number of firms always converges on N^*.

Figure 7

First Stage Equilibrium Analysis – Areas of Profitability (assuming $\sigma<a/\gamma$)



It was mentioned before that Sutton assumed that firm quality could not be less than one, such that firms could only raise, and not lower, their initial quality. If we relax this assumption, and just assume that u>0 and $A(u) = (a/\gamma)\, u^{\gamma}$,[31] we end up with similar results as seen in Figure 6 for the scenarios with high σ, in which $\sigma>a/\gamma$. N* rises with S, and is bounded above by N_a.[32] The non-monotonicity in growth of N* discussed by Sutton, as seen in Figure 6 for low values of σ, is the result of the assumption that firm quality is bounded below.

IV. Cost-Reducing Investments

Endogenous sunk costs represent quantity-invariant sunk costs that affect the price-cost margin. As discussed earlier, these can involve investments in advertising and quality, which raise the price consumers are willing to pay for the good (call these "value-enhancing"

---

[31] This formulation of A(u) is proposed in order to avoid negative costs being incurred if u<1. Changing the formulation for A(u) is not a significant change to the model, since it is equivalent to letting $\sigma - a/\gamma = \sigma'$ in Sutton's model, with σ' representing just a shifted value of exogenous sunk costs.

[32] The equilibrium results with this modification are similar to Sutton's model, except that N* behaves according to equation (7) (Sutton's equation for N*) for low sales ($S<S_m$) as well as high sales. For low sales firms here reduce their quality below 1, just as for high sales they increase their quality above 1. Equation (7) can still be used if $\sigma - a/\gamma$ is replaced with σ'. This is because Sutton's formulation of $F(u) = (a/\gamma)\,(u^{\gamma} - 1) + \sigma = (a/\gamma)\, u^{\gamma} + \sigma - a/\gamma$ changes to the modified formulation of $F(u) = (a/\gamma)\, u^{\gamma} + \sigma'$. Note that there is a steady decline in the number of firms (the low σ scenarios in Figure 6 for $S>S_m$) only if $\sigma' = \sigma - a/\gamma < 0$, implying a negative cost of entry when u is close to zero.

investments, regardless of whether consumers are better off from this investment). They can, however, also involve investments in equipment and R&D that lower the marginal cost of production (call these "cost-reducing" investments). This section examines how cost-reducing investments can be analyzed within the endogenous sunk costs framework. Either type of endogenous sunk cost can be modeled to show how investments in endogenous sunk costs affect industry concentration. Sutton in his model uses the value-enhancing formulation, with an increase in quality u relative to other firms' quality $\overline{u}$ allowing the firm to charge a proportionally higher price. We could, instead, use the cost-reducing formulation if an increase in quality[33] is assumed to cause the marginal cost to fall.[34] If a rise in quality u is allowed to have the same proportional effect on the marginal cost as it did on the price in Sutton's analysis, then the long-run profit equation is obtained, assuming again that fixed costs = A(u) + σ,

$$\boldsymbol{\pi_i(u,q) = P\, q_i - (c/u)\, q_i - A(u) - \sigma} \qquad \textbf{(3)}$$

It turns out that the two formulations yield virtually identical results within the Cournot model in terms of second stage and first stage equilibrium quality u, gross and long-run profits, and number of firms N*.[35] A rise in industry quality $\overline{u}$ now has the effect of lowering firms' marginal cost and, in equilibrium, there is a proportionate drop in price, leaving gross profits the same. Thus competitive pressures force all firms to invest in raising u and lower their marginal cost, which in the long run leads to lower prices, no increase in gross profits, and lower long-run profits.

Sutton uses the value-enhancing version in his models, mostly because of his focus on advertising as an endogenous sunk cost. In many industries, however, firms have a choice of technologies and plant sizes they can research and use. These investments generally take longer to implement compared to changes in advertising, but they can have a powerful impact on costs and ultimately on concentration, and thus should not be ignored.

Examination of cost-reducing investments allows us to relate endogenous sunk cost theory to the traditional literature on the impact of the shape of firm cost curves and scale economies on industry structure. Parsing equation (3) yields the total cost equation of TC = (c/u)q + σ + A(u). Once the firm has chosen a particular u, call it u_0, the firm has short-run cost curve TC = $(c/u_0)q + \sigma + A(u_0)$ and the average total cost is ATC = $c/u_0 + (\sigma+A(u_0)) / q$. The choice of u therefore determines which short-run cost curve the firm will have (*see* Bresnahan 1992, p. 143). The firm's goal in the second stage is then to choose u to minimize this total cost, given the quantity of output it expects to produce. As expected q rises, the cost-minimizing

---

[33] The discussion in this paper retains its generic use of the term "quality" as a proxy for advertising, R&D, improvements in service, plant investments, and product and process innovations requiring a fixed investment that affect the price-cost margin for every unit sold through either raising the price obtained or lowering the variable cost of producing each unit.

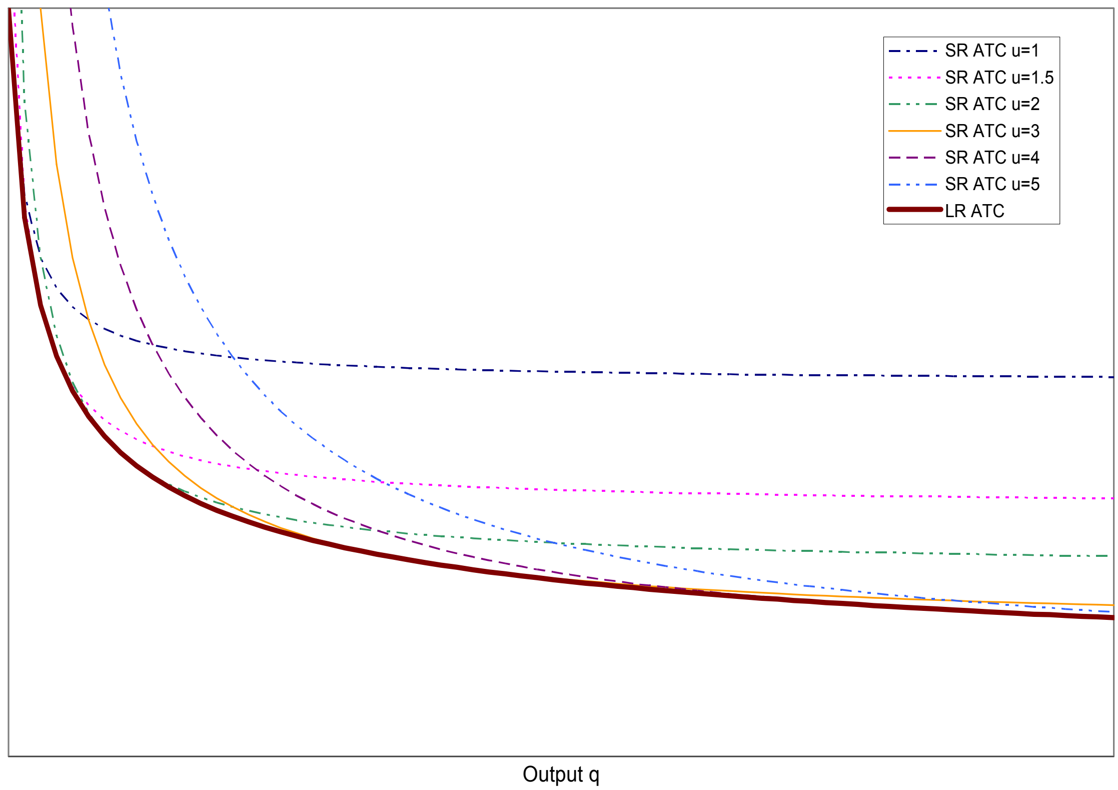
[34] Thus the marginal cost is constant with respect to output, but not with respect to quality.

[35] Using the Cournot model with equation (3) replacing equation (2), the third stage equilibrium price is $P = (c/\overline{u})(1+ 1/[(u/\overline{u})(N-1)])$, and output is $Q = S\, u\, (N-1) / [c\, ((u/\overline{u})\, (N-1) + 1)]$. The resulting gross and long-run profit equations are identical to the value-enhancing Cournot model solutions in Appendix A, and we obtain the same equilibrium quality and number of firms for the second stage and first stage equilibria. This is not surprising given that in the value-enhancing Cournot model equilibrium profits are not affected by the marginal cost nor by firms' quality if all firms choose the same quality ($u=\overline{u}$).

choice of u, call it u*, should also rise.[36] This yields a long-run average cost (LRAC) curve that is the lower envelope of all possible short-run average total cost curves.[37] The relationship between the short-run ATC and the LRAC is illustrated in Figure 8, with several short-run ATC curves plotted, each corresponding to a different value of u.[38] Note that if firms can't vary the "quality" of their production costs, the firm's long-run average cost curve is one of these short-run ATC curves.

Figure 8

Short-run and Long-run Average Cost Curves, Allowing Quality u to Vary



Note the similarity of this diagram to the standard diagrams in many microeconomic textbooks that show the relationship between the short-run ATC curves and the LRAC.[39] The curves here look a little different because the short-run cost curves are not U-shaped, but instead exhibit scale economies for all levels of output, because of the high fixed costs and constant marginal costs of this model.

The implication here is that market size can have an impact on firm plant size. This means that the minimum efficient scale (MES) is endogenously determined, and that investments in plant and equipment are not just exogenous sunk costs, but also contain an element of

[36] From $A(u) = a/\gamma\ (u^{\gamma}-1)$ we get $u^* = (cq/a)^{1/(\gamma+1)}$ as the choice of quality u that minimizes the total cost function $TC(q,u) = (c/u)q + \sigma + A(u)$. This yields total cost curve $TC(q;u^*) = (cq)^{\gamma/(\gamma+1)}\ a^{1/(\gamma+1)}\ (1+1/\gamma) + \sigma - a/\gamma$.

[37] This point was also made in Bresnahan 1992, pp. 142-43.

[38] The example assumes that σ=5, c=5, a=5, and γ=2, and plots the values for q=1 to q=70.

[39] E.g., see Carlton & Perloff p. 34.

endogenous sunk costs. There is a danger, then, that plant costs could be misidentified as being solely exogenous sunk costs, with the consequence that the measured σ could rise as market size increases. This form of endogenous sunk costs may have played a role in the history of the salt and sugar industries, which Sutton in his book describes as industries with homogenous goods that only had exogenous sunk costs, since they had high plant costs but no advertising costs. Sutton ascribed the rise in concentration in these industries mostly to an attempt by firms to reduce the toughness of price competition in order to maintain profit margins, but he also noted there was an increase in exogenous sunk costs in some of the industries (i.e., a rise in the minimum efficient scale) (Sutton 1991, ch. 6). He could be right that the observed increase in plant and firm sizes could be due to exogenous factors such as changes in technology increasing the minimum efficient scale, but it is also possible that they were caused by investments in endogenous sunk costs (including investments in integrated operations in the US sugar industry that he mentions were important for efficiency (Sutton 1991, p. 148)), that were made profitable by the increase in industry sales, which in turn led to lower prices in the long run. The change in plant costs due to a rise in MES should be considered endogenous if the more efficient plant were available earlier, but were not profitable investments until the market grew sufficiently in size. Thus increased market size may have contributed to an eventual increase in market concentration because of induced investments in endogenous cost-reducing sunk costs.

V. Implications and Extensions

The theory of endogenous sunk costs is fairly versatile. It can be used to examine a wide range of industries, and can incorporate a variety of kinds of investments that qualify as endogenous sunk costs. This section provides a brief discussion of some additional areas in which the theory can be applied, and how it can be expanded to examine a different set of problems other than the market size-concentration relationship.

First, careful examination of the basic principles of the framework reveals that many of the predicted effects will likely be observed even if some of the technical assumptions are relaxed. For example, Sutton's assumptions concerning the second order conditions ($\gamma > \max\{1, 2a/3\sigma\}$), the constant total consumer expenditures regardless of the level of market quality, and the restriction that quality cannot be less than one, can all be loosened to some extent, as discussed above.

More fundamentally the endogenous sunk costs model will be useful for any situation where firms have a choice of investment opportunities for which the benefit, but not the cost, depends on the size of output. The choices the firms make will then depend on the size of their market. In his initial empirical work Sutton focussed on the role of advertising as an endogenous sunk cost, such that industries with significant spending on advertising have tended to become concentrated as they grew in size (Sutton 1991). His later work has examined how R&D can also be an endogenous sunk cost that affects product quality and hence industrial development and market structure (Sutton 1998; Sutton 2000). The endogenous sunk cost model is a useful tool for examining many kinds of investments in sunk costs, both value-enhancing and cost-reducing. As discussed in the section on cost-reducing investments, firms in manufacturing and chemical industries may invest in larger plant as a result of growth of their market.

As should be apparent from the above discussion, the endogenous sunk costs model can be applied to a wide range of industries. Essentially any industry with a range of investment decisions that involve fixed costs, whose benefit depends on the quantity of output, can be analyzed using the endogenous sunk costs model. For example, the puzzle of the automobile industry's frequent style changes, which Kwoka (1993) claims to be a drag on profits, can be explained as investments in endogenous sunk costs.[40] Another example might be Andrew Carnegie's frequent investments in new plant during periods of rapid market expansion in the late nineteenth century, scrapping his old plant in order to lower his unit costs of producing steel, which drove out many of his competitors with low prices (Hughes 1987, p. 318).[41] Researchers have applied the model to many industries in the U.S., including packaged goods (Bronnenberg, Dhar, & Dubé 2006), supermarkets (Ellickson 2006), banking (Dick 2007), and restaurants and newspapers (Berry & Waldfogel 2003). Industries in the media and information goods sector also have significant endogenous sunk costs, because they are typically characterized by high fixed costs, low marginal costs, and consumer demand is sensitive to the quality of the product (as well as advertising for it). For example, the quality of movie making depends on the investment made by movie companies in actors, scripts, locations, production, and special effects. Software is another industry for which firms must decide on how much money to spend on improving their product and adding features. In many software categories (word processing, spreadsheets, browsers, games) an initial burst of entry has been followed by consolidation and eventually domination by a few big companies, or sometimes one company. As endogenous sunk costs theory predicts, the cost of producing a competitive product has substantially increased as these markets have expanded in size and firms have invested in raising the quality of their product.

As Sutton points out, many factors besides endogenous sunk costs must be taken into consideration when analyzing the development of a particular industry. One factor that figures prominently in his empirical analysis is the toughness of price competition. As Sutton makes clear in his theoretical and empirical analysis, the toughness of price competition is affected by various industry characteristics (e.g., whether the good is homogenous) and structural factors (whether firms are able to form a cartel or coordinate prices, government regulations on price and entry), and can vary significantly between industries. His case studies on a number of manufacturing industries show that changes in factors affecting the toughness of price competition, such as increased product differentiation, the formation or breakdown of a cartel, or changes in government regulation on pricing and entry, has had a significant impact on firm profitability and industry concentration (Sutton 1991, ch. 6). He also identifies other factors, many of them industry-specific, that have affected the development and concentration of these industries, such as first-mover advantages, segmentation of the market, brand proliferation by major producers, and responsiveness of demand to a particularly effective advertising campaign (Sutton 1991, chs. 7-11). To apply the model to other kinds of industries besides manufacturing

[40] Kwoka observed that the big US auto manufacturers invested in frequent style changes and heavy advertising because they increased a model's sales. But he also found that the net effect of the frequent style changes was to leave total sales in the market essentially unchanged, and asked "why the companies persisted in this very expensive, but ultimately self-cancelling, strategy." While he was uncertain as to its causes, Kwoka found that it drove out smaller firms and increased industry concentration. Kwoka 1993, pp. 66-68. Endogenous sunk costs theory appears to provide the answer.

[41] Carnegie reportedly said at a board meeting "Well, what shall we throw away this year?" Hughes 1987, p. 318.

additional factors would need to be considered, such as changes in the structure of distribution channels for the media and information goods industries.

VI. Conclusion

The theory of endogenous sunk costs is a useful tool for explaining the persistence of concentration in many industries. Industries that are characterized by significant endogenous sunk costs will, as their market expands, increase their investments in these sunk costs, which will limit entry. This is the result of the two key effects this paper has identified and illustrated, which are the individual incentive firms have to invest in endogenous sunk costs, even though these investments yield little additional profits when other firms do likewise, and the impact of a growth in the market on this incentive to invest. The consequence of this is that these industries will tend to remain or become more concentrated.

Further work is needed to determine how broadly Sutton's results apply. There are a number of circumstances for which his analysis does not apply, and thus we do not know if the conclusions would be the same. Within his Cournot model with endogenous sunk costs, for example, we want to know the likely equilibrium result for all likely parameter values. Thus it could be useful to know what happens if the second order conditions for quality u^* are not met, such that $\gamma < N + 3/N - 3$. Models allowing market demand to expand with increased quality would also be useful to examine.

This paper suggests that this theoretical framework could be fruitfully applied in a number of areas, to a wide range of industries with endogenous sunk costs. There are many industries that have substantial endogenous sunk costs, which have become quite concentrated. The growing information industries sector, including media and software, is a prime example. This paper also suggests that other kinds of fixed investments could be analyzed as endogenous sunk costs, including cost-reducing investments and R&D.

Bibliography

Adams, Walter and James W. Brock (1990), "The Automobile Industry", ch. 4 in Walter Adams, ed., The Structure of American Industry, 8th ed.

Akerlof, George (1970), "The Market for 'Lemons': Qualitative Uncertainty and the Market Mechanism", Quarterly Journal of Economics, 84, 488-500.

Berry S., and J. Waldfogel (2003), "Product Quality and Market Size," NBER working paper #9675.

Bresnahan, Timothy (1992), "Sutton's Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration: Review Article", RAND Journal of Economics. Spring, 23(1), pp. 137-52.

Bronnenberg, Bart J., Sanjay K. Dhar, and Jean-Pierre Dubé (2006), "Market Structure And The Geographic Distribution Of Brand Shares In Consumer Package Goods Industries", working paper, July 18, 2006.

Carlton, Dennis (2005), "Barriers to Entry", NBER Working Paper # 11645, Sept.

Carlton, Dennis and Jeffrey Perloff (2004), Modern Industrial Organization, 4th ed.

Dick, Astrid A. (2007) , "Market Size, Service Quality, and Competition in Banking" Journal of Money, Credit and Banking 39 (1), Feb., 49–81.

Dixit, Avinash and Victor Norman (1978), "Advertising and Welfare," The Bell Journal of Economics, 9 (1), Spring, pp. 1-17.

Dorfman, R, and Steiner, P. O., (1954). "Optimal advertising and optimal quality." American Economic Review, 44, 826-836.

Ellickson, Paul B. (2006), "Does Sutton Apply to Supermarkets?" Duke University, working paper.

Hughes, Jonathan R. T. (1987), American Economic History, 2nd ed.

Kwoka, John (1993), "Automobiles", in Larry L. Duetsch, Industry Studies, pp. 63-84.

Levhari, D. and Peles, Y., "Market Structure, Quality and Durability", The Bell Journal of Economics and Management Science, 4 (1), Spring 1973.

Matraves, Catherine and Laura Rondi (2005), "Product Differentiation, Turbulence and Sustainability of Competitive Advantage", working paper, March.

Nerlove, M., and Kenneth Arrow (1962), "Optimal Advertising Policy under Dynamic Conditions", Economica, 29, May, 129-42.

Robinson, W.T. and J. Chiang (1996), "Are Sutton's Predictions Robust?: Empirical Insights Into Advertising, R&D, and Concentration," Journal of Industrial Economics, 44, 389-408.

Schmalensee, Richard (1992), "Sunk Costs and Market Structure: A Review Article", The Journal of Industrial Economics, June, XL (2): pp. 125-33.

Shaked, Avner and John Sutton (1987), "Product Differentiation and Industrial Structure", The Journal of Industrial Economics, December, 36 (2): pp. 131-46

Shaked, Avner and John Sutton (1983), "Natural Oligopolies", Econometrica, 51 (5), Sept., pp. 1469-84.

Sheshinski, Eytan (1976), "Price, Quality and Quantity Regulation in Monopoly Situations", Economica, New Series, 43 (170) (May), pp. 127-137.

Spence, A. Michael (1975), "Monopoly, Quality, and Regulation", The Bell Journal of Economics, 6 (2), Autumn, pp. 417-29.

Sutton, John (1986), "Vertical Product Differentiation: Some Basic Themes", American Economic Review, 76 (2), May, pp. 393-98.

Sutton, John (1991), Sunk Costs and Market Structure: Price Competition, Advertising, and the Evolution of Concentration, MIT Press, Cambridge MA.

Swan, P.L., "Durability of Consumer Goods", American Economic Review, 60 (5), Dec. 1970.

Tirole, Jean (1988), The Theory of Industrial Organization, MIT Press, Cambridge MA.

Appendix A: Solution to Sutton's Cournot Model of Endogenous Sunk Costs

Let P = market price, c = marginal cost, and u = quality. The fixed costs are $FC = \sigma + A(u)$. Let $A(u) = a/\gamma\,(u^\gamma - 1)$, with $u \geq 1$. Thus $MC_u = dA/du = a\,u^{\gamma-1}$. For the demand side, Sutton assumes $Q = S / P$, so $P = S / Q = S / \Sigma\, q_j$. If one firm (subscript d) raises its quality to u, while other firms (subscript k) keep their quality fixed at $\bar{u}$, then the deviant firm's price is $P_d = (u/\bar{u}) * P$, and the market quality-adjusted output used in the demand equation is $Q = \Sigma\, q_k + (u/\bar{u})\, q_d$. Thus the output of the deviant firm is quality-adjusted in the calculation. The market price is $P = S / Q = S / (\Sigma\, q_k + (u/\bar{u})\, q_d)$, and so $\partial P/\partial q_k = -S / Q^2$ and $\partial P/\partial q_d = -(u/\bar{u})\, S / Q^2$.

Let $z_i(q_i,u_i)$ be firm i's gross (short-run) profit function, so $z_i = (u/\bar{u})\, P\, q_i - c\, q_i$. Therefore the long-run profits π_i are

$$\pi_i = z_i(q_i,u_i) - FC = (u/\bar{u})\, P\, q_i - c\, q_i - a/\gamma\,(u^\gamma - 1) - \sigma$$

3rd Stage Solution:

In the third (last) stage, we assume that one firm has deviated in quality (u) from the market level of quality ($\bar{u}$). Firms choose q_i, assuming the number of firms N and quality u_i is fixed. Then for the deviant firm the gross profit function is

$$z_d(u,q) = P_d\, q_d - c\, q_d = (u/\bar{u})\, S\, q_d / Q - c\, q_d$$

and for the other firms, whose quality is fixed at $\bar{u}$, it is

$$z_k(u,q) = P_k\, q_k - c\, q_k = S\, q_k / Q - c\, q_k$$

So the first order conditions (FOC) are

$$\partial z_d/\partial q_d = (u/\bar{u})\, S / Q - (u/\bar{u})^2\, S\, q_d / Q^2 - c = 0$$
$$\partial z_k/\partial q_k = S / Q - S\, q_k / Q^2 - c = 0$$

which yield solutions for q_d and q_k

$$q_d = Q / (u/\bar{u}) - (c / [S\,(u/\bar{u})^2])\, Q^2$$
$$q_k = Q - (c/S)\, Q^2$$

The solutions to the 3rd stage are then:

$$Q^* = S\,(u/\bar{u})\,(N-1) / [c\,((u/\bar{u})\,(N-1) + 1)]$$
$$q_k^* = S\,(u/\bar{u})\,(N-1) / \{c\,[((u/\bar{u})\,(N-1) + 1)]^2\}$$
$$q_d^* = S\,(N-1)\,[(u/\bar{u})\,(N-1) - (N-2)] / \{c\,[(u/\bar{u})\,(N-1) + 1]^2\} = q_k^*\,[(N-1) - (N-2) / (u/\bar{u})]$$
$$P^* = c\,(1 + 1/[(u/\bar{u})\,(N-1)])$$
$$P_d^* = (u/\bar{u})\,P^* = (u/\bar{u})\,c\,(1 + 1/[(u/\bar{u})\,(N-1)])$$
$$z_d(u;q^*) = S\,\{[1 - 1 / [(u/\bar{u}) + 1/(N-1)]\}^2$$
$$z_k(u;q^*) = S / [((u/\bar{u})\,(N-1) + 1)]^2$$

2nd Stage Solution:

For the 2nd stage equilibrium we assume that the number of firms is fixed at N, and that firms are choosing quality. The long-run profit functions for the deviant firm and for the other firms are

$$\pi_d(u;q^*) = S\,\{1 - 1 / [(u/\bar{u}) + 1/(N-1)]\}^2 - a/\gamma\,(u^\gamma - 1) - \sigma$$
$$\pi_k(u;q^*) = S / [(u/\bar{u})\,(N-1) + 1]^2 - a/\gamma\,(u^\gamma - 1) - \sigma$$

The FOC for the deviant firm's choice of u given choice of q* in 3rd stage, evaluated at $u = \bar{u}$ (all firms choose the same quality in equilibrium), assuming the Second Order Condition (SOC) for u* (discussed below) is met and that u*>1, is:

$$\partial\pi_d(u;q^*)/\partial u = (2S/\bar{u})\ \ \{1 - 1 / [(u/\bar{u}) + 1/(N-1)]\} / [(u/\bar{u}) + 1/(N-1)]^2 - a\, u^{\gamma-1}\,|_{u=\bar{u}}$$
$$= (2S/\bar{u})\ \ (N-1)^2 / N^3 - a\, \bar{u}^{\gamma-1} \qquad = 0$$

So

$$\bar{u} = u^* = [(2S/a)\ \ (N\text{-}1)^2 / N^3]^{1/\gamma} \qquad (4)$$

Because Sutton requires that quality cannot be lower than one (i.e., u≥1), A(u*) is greater than or equal to zero, so

$$A(u^*) = \max\{a/\gamma\ (u^{*\gamma} - 1),0\} = \max\{S\ (2/\gamma)\ \ (N\text{-}1)^2 / N^3 - a/\gamma,0\}$$

Therefore the fixed costs in equilibrium are:

$$F(u^*) \ = A(u^*) + \sigma\ = \max\{S\ (2/\gamma)\ \ (N\text{-}1)^2 / N^3 - a/\gamma,0\} + \sigma$$

Since $u=\bar{u}$ in equilibrium, we get from the 3rd stage solution to $z_k(u;q^*)$:

$$z_d(u^*,q^*) = z_k(u^*,q^*) = S/N^2$$

Long run profits in 2nd stage equilibrium are then:

$$\boldsymbol{\pi = S / N^2 - \max\{S\ (2/\gamma)\ \ (N-1)^2 / N^3 - a/\gamma,\ 0\} - \sigma} \qquad (5)$$

Note that for the 2nd stage equilibrium:

$$q^* \quad = S\ (N-1) / c\ N^2$$
$$Q^* \quad = S\ (N-1) / c\ N$$
$$P^* \quad = c\ N / (N-1)$$
$$u^* \quad = \max\{[(2S/a)\ \ (N-1)^2 / N^3]^{1/\gamma},1\}$$
$$= \max\{[(2cq^*/a)\ \ (N-1) / N]^{1/\gamma},1\}$$

1st Stage Solution:

Firms enter so long as the gross profits they earn cover their fixed costs. The equilibrium number of firms is then the maximum N for which z* ≥ F*. We assume that the number of firms N is a continuous variable, such that in equilibrium π = z*–F* = 0 or z*=F*, so we get

$$\boldsymbol{S / N^2 = \max\{S\ (2/\gamma)\ \ (N\text{-}1)^2 / N^3 - a/\gamma,0\} + \sigma} \qquad (6)$$

As shown below, if $S>S_m$, where S_m is the switching point, then u*>1. Assuming that $S>S_m$, and thus $A(u^*) = S\ (2/\gamma)\ \ (N\text{-}1)^2 / N^3 - a/\gamma > 0$, this can be rearranged to get Sutton's equation:

$$\boldsymbol{N^* + 1/N^* - 2 = \gamma/2\ [1 - N^{*2}\ (\sigma - a/\gamma) / S]} \qquad \text{- Sutton's equation} \qquad (7)$$

This is the equation that Sutton uses in his book. It implicitly defines the equilibrium number of firms N* as a function of the size of the market S (and of the parameters a, γ, σ), for $S>S_m$. Note that N* cannot be solved for analytically, and that we require that $S>S_m$.

Another way to write this is

$$[N^{*2} - (2+ \gamma/2)\ N^* + 1] / N^{*3} = (a - \sigma\ \gamma) / 2S$$

Note that the term in brackets is a quadratic equation, with roots N_{a1}, N_{a2}:
$N_{a1} = [4+\gamma + \sqrt{(8\gamma+\gamma^2)}] / 4$, $N_{a2} = [4+\gamma - \sqrt{(8\gamma+\gamma^2)}] / 4$

Thus N* is defined implicitly for $S>S_m$ by

$$\mathbf{(N^* - N_{a1})\,(N^* - N_{a2}) / N^{*3} = (a - \sigma\,\gamma) / 2S} \qquad (8)$$

Note that equation (8) can be rearranged to create a third degree polynomial equation in N*,[42] so there are potentially three solutions in N* that solve this equation for zero profits. From examination of equation (8) it can be seen that as S grows large, one of the solutions involves N* approaching N_{a1}, another has N* approaching N_{a2}, and if $a > \sigma\,\gamma$ (so the RHS is positive) a third has N* approaching ∞ as S rises. Call these possible solutions N^*_{s1}, N^*_{s2}, and N^*_{s3}, respectively, with $N^*_{s2} < N^*_{s1} < N^*_{s3}$. First we observe that the third solution N^*_{s3}, in which N* grows without limit as S rises (assuming that $a>\sigma\gamma$), is not a feasible solution because it requires that $u^*<1$, which violates the restriction that $u^*\geq 1$.[43]

From equation (5) it can be shown that the first and second solutions of N^*_{s1} and N^*_{s2} frame the region in {S,N} space where firm profits are positive. In other words, for any given S where $S>S_m$, profits are less than zero for $N>N^*_{s1}$ and for $N<N^*_{s2}$, and profits are greater than zero for $N^*_1 > N > N^*_2$. Thus if the number of firms is greater than N^*_{s1}, firms are losing money and some firms exit, so N falls, until profits are zero, which occurs at $N=N^*_{s1}$. If the number of firms N is between N^*_{s1} and N^*_{s2}, firms are making money, and entry occurs, increasing N, again until profits are zero, at $N=N^*_{s1}$. This effect is demonstrated in Figure 7 above. Thus N^*_{s1}, with N* approaching N_{a1} as S rises, is the equilibrium condition for the industry under Sutton's framework.[44] Our analysis of the equilibrium will refer to this solution, and our discussion of N* will be referring to N^*_{s1}.

Switching point S_m
Sutton requires that $u\geq 1$, so firms only invest in quality when the optimal quality $u^*>1$. From equation (4) it is evident that there is a minimum market size S for which firms just start to invest in quality, call it S_m, and thus $u^*>1$ only for $S>S_m$. The second stage equilibrium value (with N fixed) for S_m can be solved for by setting $u^*=1$ in equation (4) to yield

$$S_m = aN^3 / [2\,(N-1)^2] \qquad (9)$$

---

[42] Rearranging equation (8) yields the polynomial $(\sigma\gamma - a)\,N^{*3} + 2S\,N^2 - S\,(4+\gamma)\,N + 2S = 0$.
[43] We can show this through the use of a close approximation for this solution. Using the assumption that $(N^*-N_{a1})(N^*-N_{a2}) / N^{*2} \approx 1$, which holds for large N*, from equation 1)a)i)(8) we get $N^*_{s3} \approx 2S / (a-\sigma\gamma)$. This solution violates, however, Sutton's restriction that $u\geq 1$, since for this solution, $u^*<1$ for all N*. This can be seen from the equilibrium solution for $u^* = [(2S/a)\ (N^*-1)^2 / N^{*3}]^{1/\gamma}$. Using our large N* assumptions that $(N^*-1)^2/N^{*2} \approx 1$ and $N^* \approx 2S / (a - \sigma\gamma)$, our equilibrium solution for u* is $u^* = [(2S/a)\,(N^*-1)^2 / N^{*3}]^{1/\gamma} \approx [2S / aN^*]^{1/\gamma} \approx [2S / a\,(2S/(a-\sigma\gamma))]^{1/\gamma} = [(a-\sigma\gamma) / a]^{1/\gamma} < 1$. Thus this large N* solution requires that firms lower their quality below 1, which is not permitted in Sutton's framework. Profits for $N=N^*_{s3}$ are therefore negative, and thus N^*_{s3} is not a feasible solution here.
[44] If $N=N^*_{s2}$, then since profits rise with higher N, an entering firm could make positive profits. Thus N^*_{s2} is not an equilibrium.

For $S \leq S_m$, firms keep their quality constant at $u=\bar{u}=1$, so $F(u^*=1) = \sigma$, and $\pi^* = S/N^2 - \sigma$ in the second stage equilibrium. In the first stage equilibrium firms enter until $\pi^* = S/N^{*2} - \sigma = 0$, and therefore the first stage equilibrium solution for N* is, for $S \leq S_m$,

$$\mathbf{N^* = \sqrt{(S/\sigma)}} \qquad (10)$$

with the equilibrium number of firms rising with S, until $S=S_m$.

We can calculate the first-stage equilibrium value of the switching point S_m, and the equilibrium number of firms N* at $S=S_m$, call it N_m. From equation (10) we know $N_m = \sqrt{(S_m/\sigma)}$, and thus we have from equation (9): $N_m = [2+a/2\sigma \pm \sqrt{((2+a/2\sigma)^2 - 4)}]/2$. There are two solutions here, resulting from the ± operator, but only the larger solution, in which N*>1, is an equilibrium.[45] Thus the first stage equilibrium number of firms at the switching point is

$$N_m = [2+a/2\sigma + \sqrt{((2+a/2\sigma)^2 - 4)}] / 2 \qquad (11)$$

Thus given a and σ, firms begin to invest in quality in the long run when S exceeds the first stage equilibrium value for S_m of

$$S_m = [(2+a/2\sigma)^2 - 2 + (2+a/2\sigma) \sqrt{((2 + a/2\sigma)^2 - 4)}] / 2 \qquad (12)$$

Second Order Conditions for u

To determine whether u* calculated in the 2nd stage maximizes profits, we need to examine the second order conditions (SOC), by determining whether $\partial^2\pi_d(u;q^*)/\partial u^2 < 0$ at $\bar{u} = u^* = [(2S/a)\,(N-1)^2 / N^3]^{1/\gamma}$. This depends on whether

$$\partial^2\pi_d(u;q^*)/\partial u^2 = (2S/\bar{u}^2)\,[(N-1)^3 / N^3]\,(1-3/N) - a\,(\gamma-1)\,\bar{u}^{\gamma-2} < 0$$

or

$$(2S)\,[(N-1)^3 / N^3]\,(1-3/N) < a\,(\gamma-1)\,\bar{u}^{\gamma}$$

Replacing $\bar{u}$ with our equilibrium solution for u* and simplifying yields our SOC of

$$\boldsymbol{\gamma > N + 3/N - 3} \qquad \mathbf{(13)}$$

Using the zero profit equation for N*, this means we require $\gamma > N^* + 3/N^* - 3$. Unfortunately this formulation of the SOC requires knowledge of N*, which we cannot analytically solve for. Sutton points out (pp. 369-71) that imposing the restriction on γ of

$$\boldsymbol{\gamma > \max\{1, 2a/3\sigma\}} \qquad \mathbf{(14)}$$

is sufficient to guarantee that the equation (13) is met, and that this ensures that there is a "well-behaved" equilibrium solution for u* in pure strategies. This is because:

[45] As discussed earlier, the smaller solution for N* is when profits start to become positive, so firms will continue to enter until the larger N* is reached.

1. If $1 < N^* < 3$ equation (13) always holds, for $\gamma > 1$.
2. If $N^*>3$, and $\gamma > a/\sigma$, then since $N^* < 2 + \gamma/2$ (this is an upper bound to N^* if $\gamma > a/\sigma$, as shown in Appendix B), from equation (13) and the fact that the RHS of (13) is increasing in N for $N > \sqrt{3}$, it is sufficient to require $\gamma > \gamma/2 + 3/(2+\gamma/2) - 1$, which implies $\gamma > -3 + \sqrt{52} / 2 = 0.61$. Requiring $\gamma > 1$ will satisfy this.
3. If $N^*>3$, and $\gamma < a/\sigma$, then since $N^* < 2 + a/2\sigma$ (this is an upper bound to N^* if $\gamma < a/\sigma$, as shown in Appendix B), from equation (13) it is sufficient to require $\gamma > a/2\sigma + 3/(2+a/2\sigma) - 1$. Requiring $\gamma > \max\{1,a/2\sigma\}$ guarantees this is met, because
   a. if $a/2\sigma > 1$ then $\gamma > a/2\sigma > a/2\sigma + 3/(2+ a/2\sigma) - 1$ (since $3/(2+ a/2\sigma) - 1 < 0$);
   b. if $a/2\sigma \leq 1$ then $\gamma > 1 \geq a/2\sigma + 3/(2+ a/2\sigma) - 1$.[46]

   Sutton's requirement that $\gamma > \max\{1,2a/3\sigma\}$ ensures that $\gamma > \max\{1,a/2\sigma\}$.

From the foregoing discussion it should be clear that equation (14) is a sufficient condition, but the restriction on γ is greater than is necessary, partly to keep the calculations manageable. Indeed, from point #3 it is clear that we could have instead relied on the looser restriction:

$\gamma > \max\{1,a/2\sigma\}$ **(15)**

We assume that γ meets Sutton's restriction in equation (14) in this paper.

[46] Letting $y=a/2\sigma$, then $y \leq 1 \Rightarrow y^2 \leq 1 \Rightarrow 3 \leq 4-y^2 = (2-y)(2+y) \Rightarrow 3/(2+y) \leq 2-y \Rightarrow y+3/(2+y)-1 \leq 1$.

Appendix B: Proof that N* is Bounded Above for the Cournot model of Endogenous Sunk Costs

The proof of Sutton's proposition that N* is bounded above for his Cournot model is complicated. Sutton did not provide an explicit proof in his book. First note that for low levels of demand, firms do not invest in endogenous sunk costs. Thus for $S \leq S_m$, firms keep their quality constant at $u=\bar{u}=1$, so $F(u^*=1) = \sigma$, and $\pi^* = S/N^2 - \sigma$ in the second stage equilibrium. In the first stage equilibrium firms enter until $\pi^* = S/N^{*2} - \sigma = 0$, and therefore $N^* = \sqrt{(S/\sigma)}$, with the equilibrium number of firms rising with S, until $S=S_m$. Thus N* is maximized over $S=[0,S_m]$ at the point when sales $S=S_m$, with the value $N^*=N_m=\sqrt{(S_m/\sigma)}$.

The equilibrium number of firms N* cannot be solved for analytically from the Cournot model of endogenous sunk costs, but examination of the formula implicitly defining N* can help us understand how N* is bounded above, and asymptotically approaches a limit. In equilibrium in the first stage, firms enter until profits are zero, and therefore $\pi^* = 0$. Therefore, as shown in Appendix A above, for $S>S_m$ (such that $\bar{u} > 1$), the equilibrium number of firms N* is implicitly defined by

$$\mathbf{(N^* - N_{a1})\ (N^* - N_{a2})\ /\ N^{*3} = (a - \sigma\ \gamma)\ /\ 2S} \qquad (16)$$

where N_{a1} and N_{a2} are the roots of a quadratic equation, defined as

$$N_{a1} = [4+\gamma + \sqrt{(8\gamma+\gamma^2)}]\ /\ 4$$

and

$$N_{a2} = [4+\gamma - \sqrt{(8\gamma+\gamma^2)}]\ /\ 4.$$

Since $\gamma \geq 1$, we know that $N_{a1} > 1 > N_{a2} > 0$ always. We wish to determine whether the equilibrium number of firms N* is bounded above or not. It turns out that N* is always bounded above, and that N_{a1} is generally the asymptotic limit that the number of firms tends toward as the market size increases.[47] How N* approaches N_{a1} depends on certain parameters. Note that N_{a1} is somewhat less than $2+\gamma/2$. So N* is likely to be small, depending on the size of γ.

For high levels of demand, with $S>S_m$, we have to determine whether the investment in endogenous sunk costs is sufficient to prevent an unlimited increase in the number of firms in the market. The analysis is provided for each of three cases, depending on whether a is greater than, equal to, or less than $\sigma\gamma$.[48]

If $a = \sigma\gamma$, to satisfy equation (16), N* has to be equal to one of the roots, so for $S>S_m$, $N^* = N_{a1} = [4+\gamma + \sqrt{(8\gamma+\gamma^2)}] / 4$, which is a constant and does not depend on S. Thus the number of firms is constant when $a = \sigma\gamma$, for $S>S_m$.

If $a < \sigma\gamma$, then N* has to be less than N_{a1}, such that N_{a1} is the upper bound for the number of firms. This is evident from equation (16), where the right hand side is negative. Since $N^*>0$,

[47] *See* Sutton 1991, pp. 62-63.
[48] Sutton frames this as whether σ is greater than, equal to, or less than a/γ. Sutton 1991, pp. 56-60.

$S>0$, and $N_{a1}>N_{a2}$, the left hand side is negative only if $N^*< N_{a1}$ while $N^*> N_{a2}$. Thus N* can never exceed N_{a1} in equilibrium, and N_{a1} is an upper bound for N*.

It can be also shown that, for $a < \sigma\gamma$ and $S>S_m$, N* is always growing as market size increases, i.e. $dN^*/dS > 0$, and that N* stays below N_{a1}. Taking the derivative with respect to S of equation (7), which is Sutton's formulation for N*, yields $dN^*/dS - 1/N^{*2}\, dN^*/dS = -\gamma N^*\, dN^*/dS\, (\sigma - a/\gamma)/S + \gamma N^{*2} (\sigma - a/\gamma) / 2S^2$, and thus

$$dN^*/dS = \{ N^{*2} (\sigma\gamma - a)\} / \{2S^2 [1 - 1/N^{*2} + (\sigma\gamma - a) N^* / S]\} \qquad (17)$$

Since we assumed that $a < \sigma\gamma$, the numerator is always positive, and since $N^*\geq 1$, the denominator is also positive. Therefore $dN^*/dS>0$, so N* rises as S increases. Indeed, as $S\to\infty$, since N* is bounded above, $dN^*/dS\to 0$, as we would expect with a convergent series.[49]

If the cost of entry σ is low, so $a > \sigma\gamma$, then the right hand side of equation (16) is positive. The equilibrium involves N* approaching N_{a1}, with $N^* > N_{a1}$ to ensure that the LHS is positive. In the Cournot model the upper bound on N* is the value obtained when $S=S_m$, when firms begin investing in quality. The equation for dN*/dS from equation (17) can be restated, using equation (16), as $dN^*/dS = -\gamma (\sigma - a/\gamma) N^{*2} / \{2S^2 [1 - (4+\gamma)/N^* + 3/N^{*2}]\}$. The roots of the quadratic equation in the denominator can be calculated, to get:

$$dN^*/dS = (a-\sigma\gamma) N^{*4} / \{2S^2 [(N^*-N_{t1}) (N^*-N_{t2})]\} \qquad (18)$$

where $N_{t1}(\gamma) = [4+\gamma + \sqrt{((4+\gamma)^2-12)}]/2$ and $N_{t2}(\gamma) = [4+\gamma - \sqrt{((4+\gamma)^2-12)}]/2$. Thus if $a>\sigma\gamma$, then dN*/dS is less than zero if N* satisfies the key condition that $N_{t1}>N^*>N_{t2}$. Note that $4+\gamma>N_{t1}>1>N_{t2}>0$, and N_{t1} and N_{t2} are invariant to S. Since we assume that $N_{t2}<1$, and that $N^*>1$, we are only concerned whether N* is less than N_{t1}.

We can show that at $S = S_m$, $N^* = N_m< N_{t1}$, and therefore according to equation (18), $dN^*/dS < 0$, so that as S rises, N* has to fall.[50] We know that if $\gamma > 0$, then $N_{t1} = [4+\gamma + \sqrt{((4+\gamma)^2-12)}]/2 > 3+\gamma$, and that $N_m = [2+a/2\sigma + \sqrt{((2+a/2\sigma)^2 - 4)}]/2 < 2+a/2\sigma < 3+a/2\sigma$. Thus $N_{t1}>N_m$ holds if $3+\gamma > 3+a/2\sigma$, or $\gamma > (1/2)\, a/\sigma$. However, as discussed in Appendix A above, Sutton assumed that $\gamma > \max\{1,(2/3)\, a/\sigma\}$ to ensure an equilibrium solution to the choice of u*, such that $\gamma > (2/3)\, a/\sigma$. Therefore $N_{t1} > N_m$, and thus according to equation (18), $dN^*/dS <0$ at $S=S_m$. Since N* is falling as S rises for $S>S_m$, N* must have reached its maximum at $S=S_m$.

Therefore by following how N* changes in response to a rise in S, we can show that N* is bounded above, for $a>\sigma\gamma$. When demand is low, such that $S\leq S_m$, as market size S rises, then N* rises (since $N^*=\sqrt{(S/\sigma)}$), to reach a peak at $N^*=N_m$, where $S=S_m$. For larger levels of demand, when $S>S_m$, N* falls steadily, and converges on N_{a1}. N_m is therefore an upper bound on the value of N*. Thus we get the non-monotonicity discussed by Sutton, with N* rising and then falling with growing market size, if $a > \sigma\gamma$.

---

[49] We have not, however, proven that N* converges on N_{a1}, but that N_{a1} is an upper bound when $a < \sigma\gamma$.

[50] In fact, if N* is less than N_{t1} for some S, then, with $\sigma < a/\gamma$, it must be less than N_{t1} for all $S>S_m$. This follows from equation (18), in which N* can never begin (at $S=S_m$) below N_{t1} and then, as S rises, increase in value to exceed N_{t1}, since $dN^*/dS<0$. Nor can N* begin at a level above N_{t1} and for higher S fall below N_{t1}, since $dN^*/dS>0$. Note that it is impossible for N* to equal N_{t1} and still solve the equation.

Thus we can determine the upper bound to N^* for the three scenarios. If $a = \sigma\gamma$ then $N^* = N_{a1} = [4+\gamma + \sqrt{(8\gamma+\gamma^2)}] / 4$, while if $a < \sigma\gamma$ then $N^* < N_{a1}$, and if $a > \sigma\gamma$, N^* peaks at $N_m = [2+a/2\sigma + \sqrt{((2+a/2\sigma)^2 - 4)}]/2$. Since $N_{a1} < 2+\gamma/2$ and $N_m < 2+a/2\sigma$, the general statement holds that N^* is bounded above by $2 + ½ \max\{\gamma,a/\sigma\}$.