COASE, THE NATURE OF THE FIRM, AND THE PRINCIPLES OF MARGINAL ANALYSIS

Neil Kay
Economics Department,
University of Strathclyde,
100 Cathedral St,
Glasgow,
G4 OLN

0141-548-3867
neilkay@aol.com

First draft, January 6th 2005

Abstract

Coase's introduction of transaction costs into economic analysis was recognized by the award of the Nobel Prize for Economics in 1991. The announcement of the award by the Royal Swedish Academy of Sciences was accompanied by its statement that Coase for the first time had produced "robust and valid" explanations based in economic theory for two major questions, that is why do firms exist, and why is each firm a certain size?

In this paper we argue that Coase's analysis has structural flaws that raises legitimate questions as to whether it can indeed be regarded as providing a robust and valid approach to these questions. We argue that his approach to the boundaries of the firm is at best incomplete with indeterminate outcomes, at worst wrong and misleading, and that these problems are based on a misreading of the principles of marginal analysis. We explore these issues by drawing on the same principles of marginal analysis available to Coase at the time of writing his 1937 paper.

COASE, THE NATURE OF THE FIRM, AND THE PRINCIPLES OF MARGINAL ANALYSIS

Ronald Coase was awarded the Nobel Prize for Economics in 1991 for his work on transaction costs. The announcement from the Swedish Academy of the award of the prize cited his contribution as essentially composed of two stages represented by two papers: (1) Coase (1937) where he analyzed the nature of the firm in terms of transactions costs, and; (2) Coase (1960) where he looked at the relationship between property rights and transaction costs.

In discussing his first stage contribution looking at the nature of the firm (Coase, 1937), the Academy stated that in his first major study "The Nature of the Firm" Coase provided robust and valid solutions to two questions which had seldom been subjected to strict economic analysis, that is why do firms exist and why is each firm a certain size?:

> "Coase introduced transaction costs and illustrated their crucial importance. Alongside production costs, there are costs for preparing, entering into and monitoring the execution of all kinds of contracts, as well as costs for implementing allocative measures within firms in a corresponding way. If these circumstances are taken into account, it may be concluded that a firm originates when allocative measures are carried out at lower total production, contract and administrative costs within the firm than by means of purchases and sales on the market. Similarly, a firm expands to the point where an additional allocative measure costs more internally than it would through a contract on markets." (Royal Swedish Academy of Sciences, 1991) [1]

Coase's contribution has been immense and has stimulated, informed and enriched many areas of economics over the last several decades. The point that there may be costs of market exchange, and that these costs can underpin the creation and maintenance of the institutional structure and functioning of the economy is as profound as it is simple.

However, this paper will argue that his basis for establishing the boundaries of the firm (and summarized above by the Swedish Academy) is at best incomplete with indeterminate outcomes, at worst wrong and misleading. What is surprising is that these difficulties appear to have gone unrecognized, a problem perhaps being that while Coase (1937) is much cited, he may be less read. There has been insufficient attention to what he actually said, and analysis has instead tended to centre around second or later generation research which often differs radically from Coase's original formulation of the problem.

The arguments here have implications for the theory of the firm and associated research agendas and we discuss some of the issues below. We shall do so by drawing on the same principles of marginal analysis available to Coase at the time of writing his 1937 paper. While there has been substantial work in recent years on the nature of the firm, much of it building on the foundations laid down by Coase, it is necessary to look at these issues from the perspective of what Coase was trying to achieve with the tools and principles at his disposal in the 1930's. That means examining his arguments in the light of what Coase described in his paper as "two of the most powerful principles of economic analysis developed by Marshall, the idea of the margin and that of substitution" (Coase, 1937, p. 386). Coase clamed that he had explained the nature of the firm by applying the principles of marginal analysis developed in other contexts by Marshall and others, and we shall judge him by these standards in this paper[2].

We start in Section 1 by noting a simple problem for Coase's analysis; if the size of the firm can be determined by considering marginal costs of alternative modes of governance, then why can demand side changes have dramatic effects on the boundaries of the firm, even in the apparent absence of any changes in these same marginal costs? In Section 2 we explore the foundations of Coase's analysis of the limits to the size of the firm more fully, and then examine his application of marginal analysis in Section 3. In Section 4 we examine the sources of the problems encountered with Coase's analysis with the help of an analogy drawn from analysis of multiplant operations. The role and relevance of transaction benefits or gains is considered in Section 5, and we finish with a concluding Section 6.

1. A Problem

Coase (1937) argues that analysis of the firm may be illuminated by the "principle of marginalism" (p.404) and relates this to the problems of the setting of the boundaries of the firm. This is summarized by Coase in a simple rule;

> "A firm will tend to expand until the costs of organising an extra transaction within the firm become equal to the costs of carrying out the same transaction by means of an exchange on the open market or the costs of organising in another firm" (Coase, 1937, p.395).

Over fifty years later, Coase (1988) reiterated this as stating that "the limit to the size of the firm is set when the scope of its operations had expanded to the point at which the costs of organizing additional transactions within the firm exceeded the costs of carrying out the same transactions through the market or in another firm. This statement has been called a 'tautology'. It is the criticism people make of a position which is clearly right" (1888, p.19).

However, Coase's statement does not in itself give a sufficient foundation for analyzing the extent of the firm and the setting of its boundaries. For example, in December 2000 GM announced that its boundaries would contract with the closure of its Vauxhall Vectra plant at Luton UK, and with the loss of 2,000 jobs. The stimulus for the closure was a fall in demand for Vectras, in five years its sales had fallen from 2.6mill units to 2.1 mill units due to a trend to smaller more fuel efficient cars (English et al 2000). This was a trend which echoed what had happened in even more dramatic fashion through the 1980's when the after effects of the 1970's oil crises led to foreign (especially Japanese) car manufacturers expanding the boundaries of their firms by making steep inroads into the US market at the expense of domestic manufacturers such as GM. To some extent the expansion was reflected in increased exports to the US, while to some extent it represented the development of Japanese "transplants" or multinational expansion into the US (Singleton, 1992).

These expansions and contractions in firm boundaries were clearly stimulated by demand side considerations[3]. We can push this point further with a mental experiment involving two scenarios, one in which almost all consumers prefer Japanese cars because they associate them with fuel efficiency, and another in which a new Vauxhall Vectra is a status symbol for most consumers that overrides mundane considerations such as cost. The important thing about both scenarios is that they are conceivable, and indeed it is possible to find individual consumers in the real world whose demand characteristics are consistent with either extreme. Some car buyers may always prefer to buy Japanese because of the economy image, while there is also a Vauxhall Vectra Owners Club for enthusiasts. We are just imagining alternative scenarios in which one or other of these behavioural traits are highly frequent or even dominant in the population at large. We also assume that the costs of organising transactions within the firm and the costs of market exchange (however measured) are the same in both scenarios.

The boundaries of GM would look very different in these respective scenarios. In the "prefer Japanese" scenario, the boundaries of GM would shrivel towards nothingness, the extent of the shriveling depending on the strength of the "prefer Japanese" trait in the population at large. In the "Vectra status symbol" scenario, we would expect the boundaries of GM to push outwards with a concomitant expansion in domestic and foreign investment in Vectra, and associated plants and subsidiaries, the limits to this expansion being dominated by the degree to which Vectra-mania infected the global population. But none of this is captured by Coase's dictum that the expansion of the firm (and by implication its contraction) continues to the point where the costs of organising within the firm becomes equal to the costs of market exchange. In our two scenarios we have only varied the demand side, and Coases "tautology" above does not give any obvious guidance as to why in one scenario GM heads towards the dustbins of history, while in the other it moves towards ruling the automotive world.

Clearly there would seem to be, at best, some incompleteness regarding the ability of the Coasian framework to help delineate the boundaries of the firm. In the next section we shall explore this point further by examining Coase's analysis of the limits to the size of the firm.

2. Coase and the Limits to the Size of the Firm

When Coase deals with the setting of the boundaries of the firm, he focuses on the cost side. Coase asks "Why is not all production carried on by one big firm?" (1937, p.394). He concludes that there would appear to be certain possible explanations:

> "First, as a firm gets larger, there may be decreasing returns to the entrepreneur function, that is, the costs of organizing additional transactions within the firm may

> rise. Naturally, a point must be reached where the costs of organizing an extra transaction within the firm are equal to the costs involved in carrying out the transaction in the open market, or, to the costs of organizing by another entrepreneur. Secondly, it may be that as the transactions which are organized increase, the entrepreneur fails to place the factors of production in the uses where their value is greatest, that is, fails to make the best use of the factors of production. Again, a point must be reached where the loss through the waste of resources is equal to the marketing costs of the exchange transaction in the open market or to the loss if the transaction was organized by another entrepreneur. Finally, the supply price of one or more of the factors of production may rise, because the 'other advantages' of a small firm are greater than those of a large firm. Of course, the actual point where the expansion of the firm ceases might be determined by a combination of the factors mentioned above. The first two reasons given most probably correspond to the economists' phrase of 'diminishing returns to management.'" (Coase, 1937, pp.394-95).

Interestingly, one of the clearest and succinct summaries of what was to become known as diminishing returns to management had been set out earlier by Marshall (1920);

> "The small employer has advantages of his own. The master's eye is everywhere; there is no shirking by his foremen or workmen, no divided responsibility, no sending half-understood messages backwards and forwards from one department to another. He saves much of the book-keeping, and nearly all of the cumbrous system of checks that are necessary in the business of a large firm" (p.284).

Marshall is recounting what would later be described as principal-agent problems, control loss and other costs of bureaucracy that might be associated with large firms. However, the discussion is set in the context of the generally superior managerial advantages of large firm operation, and not pushed by Marshall to the point where diminishing returns to management might be a dominant and pervasive feature limiting firm expansion.

Coase then argues that:

> "Other things being equal, therefore, a firm will tend to be larger:
>
> a. the less the costs of organizing and the slower these costs rise with an increase in the transactions organized.
>
> b. the less likely the entrepreneur is to make mistakes and the smaller the increase in mistakes with an increase in the transactions organized.
>
> c. the greater the lowering (or the less the rise) in the supply price of factors of production to firms of larger size." (1937, pp 396-97)

However, for completeness we would expect to see a corresponding discussion of what happens to the transaction costs (Coase's "marketing costs") as the level of that activity rises. Do transaction costs of market exchange experience diminishing returns just as do Coase's costs of organising within the firm, or are they subject to constant returns or even continuously increasing returns? As we shall see, the answer to this question is crucial if marginal analysis is to be applied to the question of the boundaries of the firm as Coase argues. And to explore this question we have to establish what are the sources of Coase's "marketing" costs, or costs of market exchange.

Coase argued (1937, p. 391) that a series of market contracts may be substituted by one contract between the entrepreneur and the owner of a relevant factor of production, in turn reducing the costs of making that transaction. And what were these costs of using the market mechanism?

> "The most obvious cost of "organizing" production through the price mechanism is that of discovering what the relevant prices are. This cost may be reduced but it will not be eliminated by the emergence of specialists who will sell this information. The costs of negotiating and concluding a separate contract for each exchange transaction which takes place on a market must also be taken into account" (1937, pp.390-91)[4].

Coase also noted that other costs of using the market derived from risk and uncertainty and the difficulties of forecasting impeding the formation of long term contracts.

At this point, a "modern" approach to the question of what constitutes the costs of market exchange would introduce notions of opportunism and asset specificity (Williamson, 1985, 1998). However, we are still endeavoring to pursue the analysis on the terms set by Coase in his original framework, and indeed we are reinforced in setting aside issues of opportunism by Coase's own subsequent rejection of both fraud and opportunism as significant sources of transaction costs:

> "…opportunistic behavior of the type we are discussing would … normally be unprofitable and this argument has added force since a firm acting in this way will certainly be identified … the propensity for opportunistic behavior is usually effectively checked by the need to take account of the effect of the firm's actions on future business. But, of course, there are also contractual arrangements which reduce the profitability of opportunistic behavior and therefore make it more unlikely" (1988b, p.44)

Coase also raises doubts about the relevance of asset specificity (1988b, pp. 42-44) but as Williamson would agree, these reservations are redundant because if opportunism is not a serious problem in exchange transactions, then neither is asset specificity (Williamson, 1985, p.31)[1].

We note in passing that if Coase is correct, he has effectively undermined almost all of the enormous body of work which has fashioned models and approaches on notions of opportunism and asset specificity on the transaction cost foundations laid by his 1937 article. But that is a wider problem than the issue we are concerned with here. At this

[1] See also Love (2005) and Love and Roper (2005) for discussion of Coase's effective rejection of Williamson's version of transaction cost economics.

point we are interested in a more limited problem. In what form are the costs of market exchange identified by Coase likely to be encountered in practice?

Dahlman (1979) explored how Coasian transaction costs may be embodied and expressed in market exchange. In his analysis, Dahlman is referring primarily to Coase (1960). However, since Coase (1988b) argues that his two articles (1937 and 1960) are just different applications of "the concept of transaction costs" (p.35), it is reasonable to regard Dahlman's elaboration of Coasian transaction costs as applicable to the 1937 analysis also[5].

> "In order for an exchange between two parties to be set up it is necessary that the two search each other out, which is costly in terms of time and resources. If the search is successful and the parties make contact they must inform each other of the exchange opportunity that may be present, and the conveying of such information will again require resources. If there are several economic agents on either side of the potential bargain to be struck, some costs of decision making will be incurred before the terms of trade can be decided on. Often such agreeable terms can only be determined after costly bargaining between the parties involved. After the trade has been decided on, there will be the costs of policing and monitoring the other party to see that his obligations are carried out as determined by the terms of the contract, and of enforcing the agreement reached. These, then, represent the first approximation to a workable concept of transaction costs: search and information costs, bargaining and decision costs, policing and enforcement costs." (Dahlman, 1979, pp.147-48).

So the categories of Coasian transaction costs elaborated by Dahlman may be summarized as search, information, bargaining, decision, policing and enforcement costs. Dahlman argues that these classes of cost all have in common that they represent "resource losses due to lack of information", and that, "it is really necessary to talk only about one type of transaction cost: resource losses incurred due to imperfect information" (p. 148).

But if these are all "resource costs" of market transactions, what kinds of resources are we likely to be talking about, and where would they be found? If we have two firms making an exchange in the market they might use intermediaries (such as other firms, in which case it implies a further layer or level of transactions to make this transaction), but otherwise the resources they could be expected to utilize and deploy in pursuing search, information, bargaining, decision, policing and enforcement activities would be drawn, inter alia, from their own planning, purchasing, sales, legal and marketing departments. Depending on the nature and significance of the transaction, the firm might also draw upon the resources of the general/senior management of either/both firms. These activities would appear to be the most obvious source of the resource costs that Dahlman argues constitutes Coasian transaction costs.

However, trying to pin down the nature of these "resource costs" merely raises a further problem. Since these resources involve costs associated with internal management functions, it is difficult at first sight to see any obvious difference between these resource costs of market exchange and costs of internal organization. Demsetz (1988) makes a similar point:

> "One person phones another and directs him to purchase specific assets by a certain time if they can be acquired for less than a stipulated price. Is this activity transacting or managing? Knowing the answer would allow us to determine if an increase in the cost of this activity is expected to lead to the substitution of one firm for two or two for one. Since the call might be from an owner/manager of a firm to his employee in the purchasing department or from a customer/investor to the brokerage house whose services he purchases, it is hard to know whether we are dealing with a transaction or management cost until we already know whether we are discussing a firm or a market … p.149)

In short, in such circumstances there would seem to be no qualitative difference between transaction costs of market exchange and costs of internal organization. But we should not be surprised to discover this. The key to unraveling these problems is provided by Fourie (1993) when he looked at the distinction between firms and markets: "a market, unlike a firm, cannot produce. Therefore market relations can only link firms (producing units)". We could add that a market, unlike a firm, does not make decisions[6], it can only provide part of the environment for the making of decisions. Even when conducting market exchanges, management and (associated resource costs) do not float around in some disembodied ether called a market, management (and associated resource costs) are to be found where managements' offices, wages and employers are to be found, that is inside firms.

It would seem that pursuing the Dahlman route of trying to pin down where the "resource costs" of market mediated transactions would be located and embodied leads us to one rather clear conclusion; if two firms conduct a market exchange, then the transaction costs associated with that exchange would be reflected in resource costs (costs associated with internal management functions) incurred by one or (more likely) both trading partners. In other words, costs of internal organization and transactions costs of market exchange are both embodied as resource costs of internal management functions[7], the essential difference between the two reducing to the simple point that in the former case they are located in a single firm, in the latter case they are more probably shared by both firms.

But if costs of both internal organisation and market exchange are ultimately reducible to the same wellspring in the form of the coordinating ability of the management function, a further implication would seem to follow. The arguments made by Coase (above) that the internal expansion of the firm "must" be characterized by diminishing returns to management would seem to be equally applicable to the firm expanding its activities in the form of market exchange agreements - or, at the very least, our default position should be that in the absence of any evidence to the contrary, if there are forces that "must" lead to diminishing returns from expanding the firm, we would expect that these same forces must also lead to diminishing returns from firms expanding activity in the form of market exchange agreements. So we shall assume for the moment that it is reasonable to assume in a Coasian framework that market exchange activity carried out by the firm is subject to diminishing returns, though we shall also acknowledge other possibilities below.

3. Coase and the Application of Marginal Analysis

Coase introduces his paper by arguing that it represents the extension of marginal analysis into exploring and illuminating the nature of the firm;

> "It is hoped to show in the following paper that a definition of a firm may be obtained which is not only realistic in that it corresponds to what is meant by a firm in the real world, but is tractable by two of the most powerful instruments of economic analysis developed by Marshall, the idea of the margin and that of substitution, together giving the idea of substitution at the margin" (1937, p. 386-87)

Loasby (1971) argues that, "In explaining the allocation of economic decision-making between the market and a directing authority by applying marginal analysis to the cost of each kind of decision-process, Coase significantly reinforced the marginalist paradigm" (p.881) and Demsetz, (1988, p.145) notes that, "this comparison of transaction and management costs has become the focusing conceptualization of the transaction cost theory in all applications to the theory of the firm of which I am aware"

We can explore how Coase's application of marginalism works in practice. As we noted above, Coase argues that a firm expands until the costs of organising an extra transaction within the firm become equal to the costs of carrying out the same transaction by means of a market exchange (or the costs of organising in another firm).

We can start by summarising Coase's decision rule for the firm as follows:

$$\text{Expand as long as } MCt < MC_m$$

Coase also implies there will be an optimal size of firm where:

$$MCt = MC_m$$

Where MCt is the marginal cost of organizing an extra transaction within the firm and MC_m is the marginal cost of organizing that transaction on the open market or in another firm. Coase argues that as long as MCt is less than MC_m, the firm will expand and get larger, the process stopping at the optimal size of firm where $MCt = MC_m$.

The problem is, that is not how the "principle of marginalism" works in economics. In general, marginalism works by relating marginal benefit to marginal cost (however defined) in different contexts. Blaug summaries the general principle[8] coming out of the emergent marginalist revolution that Coase was attempting to extend;

> "The principle at issue is that of equalizing marginal values: in dividing a fixed quantity of anything among a number of competing uses, 'efficient' allocation implies that each unit of the dividend is apportioned in such a way that the gain of transferring it to one use will just equal the loss involved in withdrawing it from another. Whether we refer to allocating a fixed income among a number of consumer goods, or a fixed outlay among a number of productive factors, or a given amount of time between work and leisure, the principle always remains the same. Moreover, in each case the allocation problem has a maximum solution if and only if the process of transferring a unit of the dividend to a single use among all the possible uses is subject to diminishing returns … The whole of neoclassical economics is nothing more than the spelling out of this principle in ever wider contexts." (Blaug, 1983, p 312).

However, we have an immediate problem if we try to interpret Coase's expansion rule in the light of the principles set out by Blaug. Coase's expansion rule equates marginal cost to marginal cost, there is no accounting for gain or benefit from expansion, so there is no way of telling whether such expansion is efficient or not from the point of view of the firm.

In the next section we shall explore this point further with the help of an analogy using the economics of multiplant operations.

4. An Analogy: the Economics of Multiplant Operations

We can explore the proper application of marginal analysis to the problem of the size of the firm by drawing an analogy with the determination of output in the case of multiplant operations. The economics of multiplant operation was in principle tractable with the tools available to Coase at the time[9], and the analogy is perhaps more appropriate than might be first thought since both the textbook economics of multiplant operations and Coase's analysis are about the application of marginal analysis in a world of decreasing or diminishing returns. Both help determine the scale of the firm's activity, and the allocation of that activity between alternative contexts (plants in one case, the market and the firm in the other) such that marginal costs are equalized. Indeed, Coase's analysis could reasonably be described as the economics of multimode operations, the alternative modes being the firm or the market. The most obvious differences between the conventional treatment of multiplant operations and Coase's analysis of multimode operations is, of course, that we are dealing with internal production costs in the former case, and costs of both internal and external governance in the latter case.

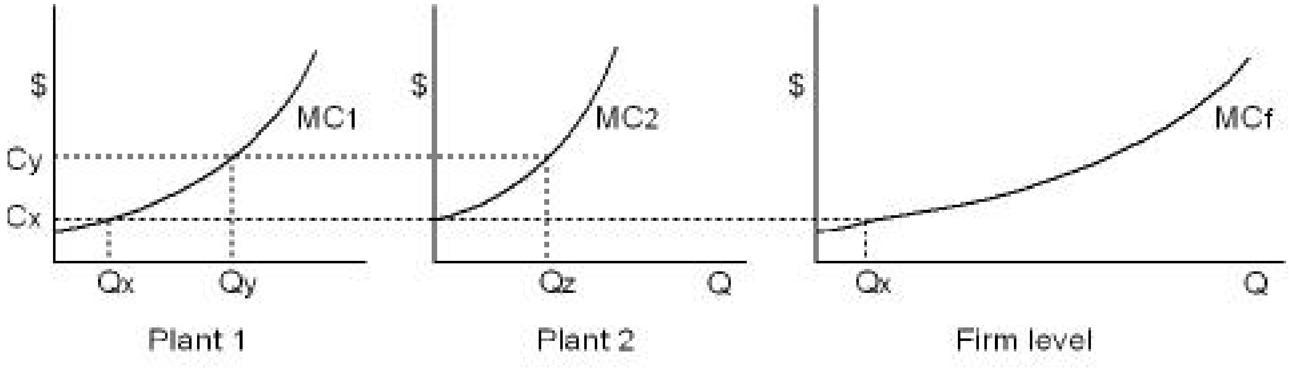

Figure 1: Marginal costs in multiplant operation

Bearing these caveats in mind, Figure 1 shows the marginal costs of production for a two plant firm, the marginal costs of production for plants 1 and 2 being MC_1 and MC_2 respectively (reflecting diminishing returns), while MC_f is marginal cost of production at the level of the firm. As can be seen, MC_1 lies below MC_2 up to output Q_x, and for any level of production up to Q_x the firm would only operate Plant 1. At Q_x, we have $MC_1 = MC_2 = C_x$.

Now, suppose the firm was operating at some output level Q_n, such that:

$$0 < Q_n < Q_x$$

In those circumstances, the firm would be operating just Plant 1, since $MC_1 < MC_2$ at Qn.

We now want to apply marginal analysis to the question of what the size of firm would be (here measured in terms of scale of output). If we were to apply Coase's version of marginal analysis above, we would instruct management to expand output of Plant 1 as long as $MC_1 < MC_2$, and cease expansion at the point where the marginal costs of operating Plant 1 have increased to the point where $MC_1 = MC_2$, which is at Q_x with a marginal cost of C_x. That would give us a Coasian version of equilibrium size of the firm (measured in terms of output rather than transactions) of Q_x represented by Plant 1 output.

The problem is that there is essential missing information if we wish to establish the optimal size of the firm (whether we measure it in terms in terms of output or transactions). Q_x is not the only output level that satisfies the condition $MC_1 = MC_2$, for example in Figure 1, $MC_1 = MC_2$ with Plant 1 output of Q_y, Plant 2 output of Q_z, and a marginal cost of C_y. Indeed, there are an indefinite number of combinations of Plant 1 and Plant 2 levels of output that would satisfy the condition $MC_1 = MC_2$, the number only limited by the divisibility of output and any relevant capacity constraints.

So what is the source of the problems in applying Coasian marginal analysis to the problem of multiplant operations? The answer is, of course, that it entirely neglects the importance and role of the demand side in marginal analysis.

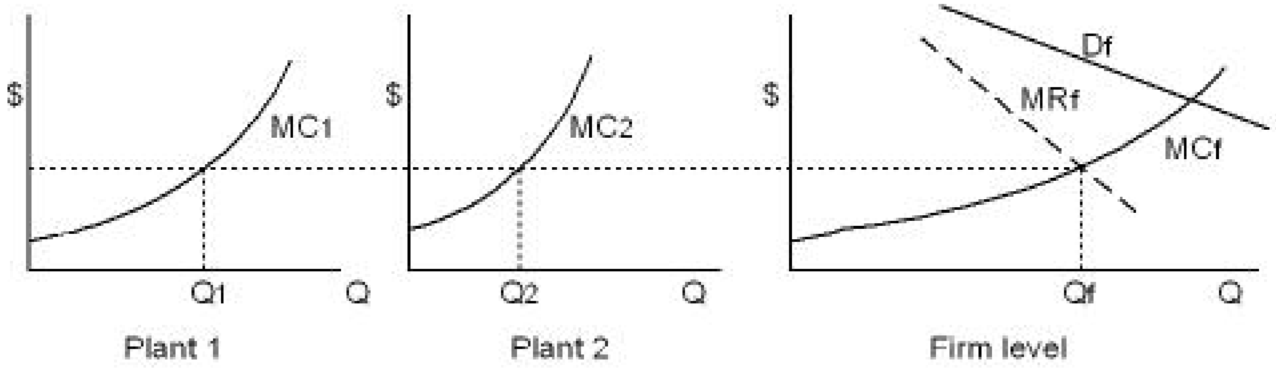

Figure 2: Allocation of output in multiplant operation

Figure 2 has augmented the earlier figure with demand at the level of the firm, D_f, and a corresponding marginal revenue curve, MR_f (we are assuming in this case that the firm is a monopolist). The standard decision rule here for optimal multiplant operation is:

Expand as long as MR > MC

And the firm will be maximizing profits and in equilibrium where:

$$MR_f = MC_f = MC_1 = MC_2$$

If we wanted to measure the size or scale of the firm here in terms of output, it would be $Q_f = Q_1 + Q_2$.

Now, it is the case that $MC_1 = MC_2$ in the multiplant case, but not as a direct principle of expansion itself, rather it is the indirect result or outcome of the firm expanding according to the decision rule "expand as long as MR > MC" rule. It is this that determines the scale of the firm's activity, $MC_1 = MC_2$ is the part that describes how that activity is distributed or allocated between the different plants of the firm as a consequence of applying that rule.

The problems with Coase's expansion rule of setting $MC_f = MC_m$ is that while this would be a necessary condition to be fulfilled in application of marginal analysis, it is not sufficient to give coherent rules for deciding the optimal size of firm. As our GM example suggests, and the multiplant analogy confirms, without explicit consideration of the demand conditions we cannot say anything about how far the firm would expand its boundaries. Ironically, while Coase (1937, acknowledges (Austin) Robinson (1934) and Kaldor (1934) and even draws on their analysis to justify diminishing returns to management, he does not heed Robinson's warning that "it is impossible, that is, in any case to regard the optimum firm as one would wish to regard it, as a size of firm determined independently of demand and independently of the environment in which the firm would work" (Robinson, 1934, p. 256-7). Robinson points that the size of firm is dependent on the rate of growth of the market and probable variations of demand as well as cost conditions (p. 256). Kaldor also points out "the limitation upon the size of firm … is sufficiently accounted for by the supply and demand curves with which it is confronted (1934, p. 73, italics added). In terms of the marginalist revolution that Coase claims to be following, both Robinson and Kaldor are absolutely correct.

We can illustrate this if we assume the firm's boundaries encompasses activities A, B, C, D, all are profitable, and the marginal costs of coordinating transactions associated with each activity inhouse are just equal to the marginal cost of alternatives involving market transactions or in other firms (satisfying Coase's condition or "tautology" for the limit to the size of the firm). Now suppose technological innovations render obsolete activities C and D but A and B would still be profitable even in the absence of C and D. If the firm is to maximise profits, it will close down activities C and D and contract its boundaries to just encompass A and B, even though there has been no change in the marginal cost of organizing transactions inside or outside the firm. Just as a fall in price (and marginal revenue) may lead to cutbacks in the multiplant case even if marginal costs do not change, so a fall in revenue from activities may lead to firms scaling back their boundaries even if transaction costs of organizing activities in alternative modes do not change.

The logic of expansion may be seen more clearly in the multiplant case if we take a limit case and assume that Plant 2 is obsolete and inefficient and that $MC_2 > MR$ for all levels of output. In that case the problem boils down to the simple case of setting output where $MR = MC_1$

An analogous situation would arise in the Coasian perspective if MC_m (marginal cost of market exchange) were such for all transactions that it was simply not worth while organising any transactions through market exchange or in another firm. In that case, the only marginal cost that would matter would be MC_t (marginal cost of internal organisation). What would be the appropriate marginalist rule then as far as the optimal size of the firm is concerned? Just as in the multiplant case it would boil down to comparing the marginal benefits of expansion with the marginal costs of expansion and expanding as long as marginal benefits exceed the marginal costs[10]. Here, if we started with a small firm and if marginal benefits exceeded these marginal costs for a large number of transactions we would finish up with a large firm. If marginal benefits exceeded these marginal costs for only a small number of transactions we would still finish up with a small firm. If marginal benefits did not exceed these marginal costs for any new transactions, the firm would stay the same size.

There is a further problem with the Coasian application of marginal analysis to the size of the firm which our multiplant analogy can help signpost. Even if the demand side is introduced into transaction cost analysis, this would not be sufficient to guarantee that there would be an equilibrium solution to the problem of the size of the firm. As Blaug notes above, the application of the equi-marginal principle to the allocation problem by the theorists that Coase was drawing on have maximum solutions if and only if there are diminishing returns. Indeed, the prior authorities on the role and existence of diminishing returns to management cited by Coase (Kaldor, 1934, Robinson, 1934) both identified this issue as central to helping the establish the equilibrium size of the firm.

To some extent, we can draw some reassurance from our arguments above. If there is a case to be made that there are diminishing returns from the internal expansion of the firm as Coase argued, then these same arguments should equally apply to the situations in which firms expand their activities in the form of market transactions.

However, this is not sufficient to solve the problem, and we can see why by again drawing on the multiplant analogy above where expansion in both plants is subject to diminishing returns. Suppose Plant 1 has followed the Coase expansion rule and expanded to Q_X in Figure 1 where the marginal costs of Plant 1 are now equal to those associated with Plant 2. However, if activity switches to Plant 2 at this point, then diminishing returns in Plant 2 means that any expansion of Plant 2 will result in the marginal cost of Plant 2 rising once more above that of Plant1. But what will happen then? If we follow the Coase rule of expanding in our original mode if the marginal cost of that mode is less than the marginal cost of the other mode, this will encourage a switch back to Plant 1, and further expansion there, which in turn means marginal cost of Plant 1 rises again… and so on. In the absence of any "off-switch" for production (which in Figure 2 is where MR = MC), the Coase criteria of expansion where MC of one mode is less than MC of the other will result in continuing expansion along both modes, activity switching back and forward depending on which mode has suffered the vicissitudes of diminishing returns the least at any point in time. Again, no equilibrium is in sight because we have no "off switch" due to the fact that the marginal analysis has not been properly applied.

The implications for the size and scale of the firm using Coase's expansion rule is equally clear. Applying the rule would lead to unrestricted expansion of the firm, or more accurately no basis on which the expansion of the firm could be restricted, the only effect of the rule that the marginal cost of internal organization and market transactions should be equal being to allocate the mix of these activities between the two modes of firm and market governance[11], rather than to limit the scale of these activities[12].

5. The Role of Transaction Benefits

The notion that gain or revenue aspects have to be added to transaction cost is not novel and has already been strongly argued by Dietrich (1993). He argues that transaction cost economics is based on partial reasoning and that analysis of choice of governance structure has to be extended to include differential abilities to generate revenue rather than just transaction cost efficiencies (p. 166).

Although Dietrich does not refer to Coase and concerns himself instead with Williamson's version of transaction cost economics, from our earlier discussion we can argue that his criticism is equally valid as far as the earlier foundations laid down by Coase is concerned. Indeed, if we are correct, the source of any subsequent problems with analysis such as that developed by Williamson (1975 and 1985) may stem in large part from the directions signposted earlier by Coase.

Further, our analysis so far should not be misinterpreted to suggest that there has been little or no exploration of demand or benefit aspects in analysis of alternative forms of governance. On the contrary, there has been significant coverage in recent years of such issues, and much analysis that has involved the notions of benefits, gains and value of transactions. A problem is that much analysis of the role of benefits or gains from transactions tends to be limited to a specified transaction or set of transactions considered in isolation. For example, the incomplete contracts literature is primarily concerned with the nature of ownership and financial structure of the firm. In their exploration of the foundations of the incomplete contracts literature, Hart and Moore (1999) develop a model of contracting in which parties attempt to maximise expected surplus, and trades involve prices as well as costs. In these respects they may be seen as introducing essential demand side considerations missing from Coase's analysis.

However, while looking at the value of one transaction in isolation under alternative governance regimes may help generate insights as to why that transaction is carried out inside or outside the firm, that research agenda does not directly address the Coasian question of the determination of the optimal size of firm. Holmstrom and Roberts (1993) look at how the question of the boundaries of the firm has been investigated in economics, following Coase's 1937 article and its insights, and also building on Williamson (1975, 1985). They conclude; "it seems to us that the theory of the firm, and especially work on what determines the boundaries of the firm, has become too narrowly focused on the holdup problem and the role of asset specificity … It is also questionable whether it makes sense to consider one transaction at a time when one tries to understand how the new boundaries are drawn. In market networks, interdependencies are more than bilateral, and how one organizes one set of transactions depends on how the other transactions are set up" (91-92)[13].

We could add that it may be in these "interdependencies" of internally organized tasks and market transactions that the sources of Coasian diminishing returns to management are to be found, if they are to be found at all.

However, even if we were to allow for benefit/value considerations to be added to the Coasian agenda, there would still remain serious problems. Williamson (1998) argues that; "declaring that the transaction is the basic unit of analysis usefully moves economics in the direction of being a science of contract, as against a science of choice" (p.36), but there are problems when we pursue the original Coase agenda of combining this unit of analysis with marginal analysis to identify the scale and scope of the firm. For example, work looking at the issue of the size and scope of the firm has traditionally drawn on notions based on homogenous or standardardised units of output or employment (e.g. Kaldor 1934, Robinson, 1934, and Dietrich, 1993). It is more difficult to how the same type of analysis of the size or scope of the firm can be undertaken using transactions as the basic unit. Yet if we are accept Coase's argument that his paper is simply the application of marginal analysis as developed by Marshall, then an obvious corollary of his conclusion that the limit to the size of firm is given by the marginal transaction is that the scale or scope of firm can be measured in terms of the sum of all transactions as the firm expands towards that limit.

The problem with using "transaction" as our measure for size of firm (rather than say, output, employment, or value) is, as Coase himself notes, transactions are highly diverse[14]. Suppose, for example, that the management of a firm is hosting a meeting with their trading partner. This management has to decide two matters: (1) whether to propose merger with the other firm or continue the present trading arrangements; and (2) whether to have lunch sent up from the company canteen or to send out for sandwiches from the delicatessen round the corner. Both represent choices between organising the activity inhouse or through market exchange, but they clearly raise issues of measurability, comparability and commensurability of units in dealing with problems such as size and scope of firm if we were to regard them as constituting supposed "units" of transactions.

In practice, the size and scope of firm is likely to be highly path dependent and influenced by the interplay between resources and opportunities facing the firm at any given time, with costs and benefits from specific forms of governance only a part of these considerations. Further, as writers from Marshall (1920) through Kaldor (1934) to Penrose (1959) have pointed out, and as Coase himself also recognized, dynamic factors can mean that if there is any equilibrium size of the firm, it may be a moving one (Coase, 1937, pp. 404-05).

6. Conclusion

Suppose that someone today was to argue that they had applied the marginalist principles developed by Marshall to the economics of multiplant operations and now claimed that optimal output for a two-plant firm was to be found by expanding the output of one plant until its marginal cost became equal to the marginal cost of a second plant. We can be reasonably confident that it would not take long before they were advised that this was a misapplication of marginalist principles. Any attempt to establish the optimal scale of output (and by implication, one potential measure of the size of firm) using that criterion would almost certainly be flawed, misleading and inaccurate. If it produced the correct measure of optimal size of firm it would only be by accident.

Since Coase clamed to be following these same marginalist principles, Coase's rule for the optimal size of firm (expand until the costs of internal organisation equal costs of market exchange) may also be regarded as a misapplication of marginalist principles and equally flawed, misleading and inaccurate. If the optimal size of firm is found using this approach, it would only be by accident.

Coase was right, it is possible to analyse the boundaries of the firm using marginal analysis, but it has to be by comparing marginal benefits with marginal costs, not simply comparing one set of marginal costs with another set of marginal costs. The latter route simply does not provide a sound analytical base for analysing the scale and scope of the firm. On the assumption that the arguments made here are correct, it is puzzling that this basic point has apparently been overlooked for so many decades. It can only be presumed that the reason for this is that Coase's principle here has been accepted as obvious and even tautologous, and not really subjected to sufficiently hard examination as to its meaning and implications. In 1972, Coase described his 1937 article as "much cited and little used" (Coase, 1972, p. 62). While it has certainly been even more frequently cited and used since 1972, there is perhaps still less evidence that it has been read properly and its analysis and implications followed to logical conclusions. If this had been done, it is difficult to imagine the problems identified here being effectively overlooked for so many decades.

We have argued here that Coase's basis for deciding the boundaries of the firm and the related question of its optimal size was at best incomplete with indeterminate outcomes, at worst wrong and misleading. Where the balance falls between these two judgments is left open here. The basic problem is that Coase treated what is essentially a necessary condition to be satisfied to generate an optimal outcome (marginal costs must be equalized across all activities) as a rule to be followed that would also be sufficient to generate this optimal outcome when, as we have seen, it is not.

The implications of this are profound. Coase's analysis is the foundations of transaction cost economics. If transaction cost economics can be regarded as a paradigm in one of the many senses[15] used by Kuhn (1970), then the determination of the boundaries of the firm may be regarded as its paradigm problem. Coase's paradigm problem should have been expressed and explored in transaction value terms and not transaction cost terms if, as he claimed, it was to have been seen as consistent with the marginalist revolution pioneered by Marshall and others. The foundation paper of transaction cost economics (Coase 1937) fails to deal with its foundation problem, and is at best seriously incomplete with indeterminate outcomes, at worse simply wrong. This has been overlooked since the paper was published, quite possibly because the field of transaction cost that followed Williamson (1975)'s interpretation of Coase has been largely concerned with transactions expressed in the form of discrete choice problems (e.g. the make or buy decision) rather than organisational choices involving the transaction expressed as a continuous variable, which is what was implied by Coase's arguing that the boundaries of the firm could be determined using marginal analysis with the transaction as the basic unit for that marginal analysis. Had those who followed Coase actually tried to apply his analysis to his paradigm problem, it is probable that the inconsistencies and difficulties identified here would have been uncovered at an earlier date.

It is acknowledged that analysis of individual transactions can help illuminate important aspects of firm behaviour and that much valuable work and been carried out in this context. However, in contrast to the arguments of Coase and Williamson, it has been argued separately (Kay, 1997, 1999) that the transaction is not the most appropriate currency for exploring the nature of the firm. Instead, it can be argued that the nature of the firm and associated problems such as the determination of its boundaries may be best explored by analyzing the nature of the decisions (and the relative advantages and disadvantages possessed by alternative modes in handing different types of decisions) in conjunction with analysis of questions of resources, competences and capabilities. It has been argued here that the mishandling of value considerations is a structural flaw running through transaction cost economics and that the source of the problem can be traced right back to Coase's original article. But it is also important to note that problems with transaction cost economics may not be resolved by simply adding transaction benefits on top of transactions costs. The nature of the firm is almost certainly more complex and interesting than could be revealed from any attempt to decipher the runes of transactions, even if marginalist principles are deployed correctly.

REFERENCES

Arrow, K, (1994) Foreword, in Arthur (1994), pp.ix-x.

Arthur, W. B. (1994) Increasing Returns and Path Dependence in the Economy, Ann Arbor, University of Michigan Press.

Blaug, M. (1983) Economic Theory in Retrospect, Cambridge, Cambridge University Press

Campbell, D. and Klaes, M. (2005) The principle of institutional direction: Coase's regulatory critique of intervention, Cambridge Journal of Economics, 29, 263-88.

Coase, R. H. (1937) The Nature of the Firm, Economica, 4, 386-405.

Coase, R. H. (1960) The Problem of Social Cost, Journal of Law and Economics, 3, 1-44.

Coase, R. H. (1972) Industrial Organization: A Proposal for Research, in V. R. Fuchs (ed.) Policy Issues and Research Opportunities in Industrial Organization. New York, National Bureau of Economic Research, 59-73

Coase, R. H. (1988a) The nature of the firm: meaning, Journal of Law, Economics, and Organization, 4, 19-32.

Coase, R. H. (1988b) The nature of the firm: influence, Journal of Law, Economics, and Organization, 4, 33-47.

Dahlman, C. J. (1979) The problem of externality, Journal of Law and Economics, 22, 141-62.

Dietrich, M. (1993) Transaction costs … and revenues, in, Pitelis, C. (1993) (ed.) 167-83.

Demsetz, H. (1988) The theory of the firm revisited, Journal of Law, Economics and Organization, 4, 141-61.

English, E. G. Jones and A. McSmith (2000) Vauxhall to axe 2,000 car workers, News Telegraph, London, The Telegraph Group, 13th December.

Fourie, F C. v. N. (1993) In the beginning there were markets? In C. Pitelis (ed.) (1993) 41-65.

Hart, O. and J. Moore (1999) Foundations of incomplete contracts, Review of Economic Studies, 66, 115-38.

Holmstrom B. and J. Roberts (1998) The boundaries of the firm revisited, Journal of Economic Perspectives, 12, 73- 94.

Kaldor, N. The Equilibrium of the Firm. Economic Journal, 19, 60-76.

Kay, N. M. (1997) Searching for the firm: the role of decision in the economics of organizations, Industrial and Corporate Change, 9, 683-707.

Kay, N. M. (1999) Loasby and decisions: a non-Coasian perspective on the nature of the firm, in, S. C. Dow and P. E. Earl (eds.) Contingency, Complexity and the Theory of the Firm: Essays in Honour of Brian J. Loasby, Cheltenham. Edward Elgar, Vol. 2, 67-91.

Kay, N. M. (2000) Pattern in Corporate Evolution, Oxford, Oxford University Press.

Kuhn, T. S. (1970) The Structure of Scientific Revolutions, Chicago, University of Chicago Press.

Loasby. B. J. (1971) Hypothesis and Paradigm in the Theory of the Firm, Economic Journal, 81, 863-885.

Love, J. H. (2005) On the opportunism-independent theory of the firm, Cambridge Journal of Economics, 29, 381-97.

Love, J. H. and Roper, S. (2005) Economists’ perceptions versus managers decisions: an experiment in transaction-cost analysis, Cambridge Journal of Economics, 29, 19-36

Marshall, A. (1920) Principles of Economics, London, Macmillan.

Masterman M. (1970) in I. Lakatos and A Musgrave, (eds), Criticism and the Growth of Knowledge, Cambridge, Cambridge University Press, 59-89.

Patinken, D. (1947) Multiple-plant firms, cartels and imperfect competition, Quarterly Journal of Economics, 61, 173-205.

Penrose, E. T. (1959) The Theory of the Growth of the Firm, Oxford, Blackwell

Pitelis C. (ed.) (1993) Transaction Costs, Markets and Hierarchies, Oxford, Blackwell

Robinson, A. (1934) The problem of management and the size of the firm, Economic Journal, 19, 241-56.

Royal Swedish Academy of Sciences (1991) Press Release: The Sveriges Riksbank (Bank of Sweden) Prize in Economic Sciences in Memory of Alfred Nobel for 1991, Stockholm.

Singleton, C. J. (1992) Auto industry jobs in the 1980's: a decade of transition, Monthly Labor Review, February, 18-27.

Wang, N. (2003) Coase on the nature of economics, Cambridge Journal of Economics, 27, 807-829.

Williamson, O. E. (1975) Market and Hierarchies: Analysis and Antitrust Implications, New York, Free Press.

Williamson, O. E. (1985) The Economic Institutions of Capitalism, New York, Free Press.

Williamson, O. E. (1998) Transaction cost economics: how it works; where it is headed, De Economist, 146, 23-58.

Zajac, E. J. and C.P. Olsen (1993) From Transaction Cost to Transactional Value Analysis: Implications for the Study of Interorganizational Strategies, Journal of Management Studies, 30, 131-145.

ENDNOTES

[1] See also Campbell and Klaes (2005) p.269.

[2] But see also Wang's (2003) assessment of Coase as a heterodox economist for an alternative perspective.

[3] See Kay (1997, p.38) and Kay (2000, p.687) for earlier arguments about the importance of demand side considerations in setting the boundaries of the firm.

[4] See also Love (2005) and Love and Roper (2005) for discussion of Coase's effective rejection of Williamson's version of transaction cost economics.

[5] It is also worth noting that Coase was editor of the Journal of Law and Economics at the time that Dahlman published his article there. It is reasonable to presume that if Coase felt that Dahlman had not represented or developed Coase's notion of transaction costs reasonably, that this would have been picked up in the normal course of editorial review.

[6] It could be argued that firms themselves do not make decision, decisions are made by individuals (singly or collectively) in a firm and that the proper role of the firm in this context is to provide an environment for decision-making, just as the market provides an environment for decision-making. This point could be accepted while noting that it does not affect the central point being made here, that the costs of market exchange and internal organisation are both essentially reducible to resource costs of management

[7] That does not necessarily mean that the same management team always deals with both internal and external relationships. For example, in Demsetz's example, the same management function and indeed management may be involved in both internal and market exchange activities, but in other cases it may be that a particular management function and managers are more likely to be associated with a specific form of governance, e.g. legal specialists with market contracts.

[8] Here Blaug is also making a distinction between "marginal analysis as such" and "maximizing behavoir subject to constraints", a distinction which he notes only represents a formal rather than a substantive difficulty in the analysis.

[9] However, to the best of my knowledge, the first systematic approach to the problem using these tools was attempted by Patinken (1947)

[10] All relevant costs and benefits would have to be taken into consideration, for example marginal benefits might be measured as the marginal revenue from the incremental transaction net of production and distribution costs.

[11] There may be added complications. To the extent that both forms of governance draw upon the same set of managerial resources, the respective cost functions may not be fully separable – for example expending more general management resources on the coordination of market transactions could add to the coordination problems if these resources are also further applied to internal activities.

[12] We have only considered the possibility of diminishing returns to both modes of governance, which we have argued is a reasonable interpretation consistent with Coase's analysis. But it is worth noting that increasing returns have been the source of interest and study in economics from Adam Smith onwards (Arrow, 1994) and in recent years it has become the subject of intense research in the discipline (Arthur, 1994, p. xx). However, As Blaug implies above, the marginalist revolution which Coase was celebrating and attempting to extend generally saw increasing returns as destructive of the notion of equilibrium.

[13] The managerial literature also tends to look at transactions or ventures in isolation or as one-offs when they introduce demand side considerations (e.g Zajac and Olsen 1993).

[14] "Nothing could be more diverse than the actual transactions which take place in our modern world" (Coase, 1937, p.396).

[15] See Masterman (1970) for a discussion of the many different ways that the term "paradigm" was used by Kuhn.