1982

The Firm: A Coordinator of Contracts

Yoram Barzel

7065

ECONOMICS AND LAW WORKSHOP

THE FIRM: A COORDINATOR OF CONTRACTS

Tuesday April 20, 1982

82-11

Time: 4:00 p.m. Room: 4032 SSC

Major funding for the Centre for Economic Analysis of Property Rights has been provided by The Bureau of Corporate Affairs, Consumer and Corporate Affairs, and by the Academic Development Fund, The University of Western Ontario. The views expressed by individuals associated with the Centre do not reflect official views of the Centre, The Bureau of Corporate Affairs, or The University of Western Ontario.

Subscriptions to the Workshop papers and the Working Paper Series are $40 per year for institutions and $25 per year for individuals. Individual copies, if available, may be purchased for $3 each. Address all correspondence to John Palmer, Centre for Economic Analysis of Property Rights, The University of Western Ontario, London, Ontario, CANADA N6A 5C2

Preliminary Draft
Comments Welcome

THE FIRM: A COORDINATOR OF CONTRACTS*

Yoram Barzel
February 1982

Any exchange requires resources for its execution and the terms of exchange depend on how the transaction is performed. Part of the gain from exchange is lost, however, as each of the parties spends resources to divert to himself a larger portion of the gains. Constraining this action can reduce the shrinking of the pie. One method to effect and police such constraints is for the parties to engage a monitor and reward him according to the reduction in the cost of transacting he brings about. The reward will be more effective if it takes the form of residual payment, and if the transactors become employees, so they can more easily be constrained. This arrangement constitutes a "firm." The firm is run by the entrepreneur who tailors and coordinates the employment contracts of the exchange partners inducing them to enhance their productive effort and to restrain the dissipation when dealing with each other.

The firm, then, is perceived as an organization capable of executing certain transactions at a cost lower than that of other forms of organization.[1] The greater the cost of effecting an exchange, the less likely it will be conducted in the market. That cost is higher

*I am greatly indebted to A. Alchian, C. Hall, J. Hause, and G.J. Stigler for their comments. Also L. DeAngelo and W. Oi.

the greater are the difficulties in identifying or measuring the physical properties of exchanged goods such as the ammounts of ingredients in fruits and the "reliability" of appliances and the greater the difficulties in determining the values of the goods since it is most difficult then to allocate the gains from trade among transactors.

If the rationale for the firm offered here is correct, then it follows that (1) the boundaries between firms will be found in junctures in the production process where the product is easiest to measure; (2) a product exchanged between firms will be more fully specified than when it is used internally by a firm; and (3) of the workers using large scale pieces of equipment, the fraction employed by the firm owning the equipment will be smaller as the individual workers' net effect on the equipment becomes easier to discern. Additional implications are offered below.

When a worker uses equipment jointly with other workers, the costliness of determining accurately his net contribution to output may lead to resource expenditures to claim that output. The firm may be effective in reducing this cost. The relationship between the ease of measuring the individual contributions and equipment scale, however, seems tenuous. Thus firm's size may be associated with the scale of equipment, but the fundamental force determining size is the ease or difficulty in measuring commodities as they pass from one person to another. When these difficulties affect a sequence of operation, it may be most economical to carry all of them within a single "integrated" firm.

I. Introduction

Market transactions have, as a rule, many attributes and some of these are provided at no marginal charge. Stock brokers do not charge marginally for information, movie theaters do not charge more for good than for poor seats, woodbins in lumber yards contain pieces of varying qualities all priced equally, and restaurants do not lower their charges to speedy eaters.[2] Individuals consume these costly-to-produce attributes, or spend resources to acquire them to the point where their net marginal values are zero. People then come "too early" to the theater to get superior seats, spend "too many" resources trying to get the best wood pieces, and eat "too slowly" in restaurants. It is hypothesized that within the firm, exchange partners' contracts are coordinated so as to limit the exploitation or the "over use" of unpriced attributes.

The within firm flow of intermediate commodities from one worker to another constitutes exchange no less than when the commodities are sold. Even team production can be viewed as an exchange, though it is almost impossible to measure the exchanged values. It is shown below that market exchanges between two independent contractors that would have been exceptionally costly may be more efficiently monitored if the exchange is within the firm, where both parties are employees of the same employer. This leads naturally to a definition of the type of business organization of concern here called the "firm": A "firm" is said to exist when one person, called entrepreneur, contracts for services from two (or more) other people, paying them by their inputs (rather than output), and sells the combined output to others. Two persons working on successive production processes and both selling

their labor services to the same employer constitute (part of) a firm.[3]

Employment contracts are a central feature of the firm, and these contracts are also characterized by unpriced attributes. In particular, when pay is by the hour, the employee's effort may, at the margin become an unpriced attribute. Why do firms pay their employees mainly on the basis of time inputs, when in fact labor's net contribution depends on its effect on the value of output?[4] The contribution of a worker (or of any other factor) is seldom known with total accuracy. Still, some hourly workers' output is easier to observe and to measure than others'. Why are employees whose output is relatively easy to identify not paid by output? And, how are employees whose output is difficult to identify and who are paid on the basis of time input induced to produce?

The question could be turned asking: Had it been relatively inexpensive to identify a worker's net contribution to output, would he become a firm's employee or would he be self-employed? It will be shown below that if the employer were always able to know the precise value of an employee's net contribution to output, the distinction between employment in a firm and self-employment becomes inconsequential. When pay exactly matches the net contribution, inefficiencies such as that associated with shirking will be avoided. If a worker's net contribution to output is easy to measure, however, he could as easily operate as an independent contractor. As a distinct form of organization, then, the firm will <u>not</u> pay its workers their net marginal contributions. If, on the other hand, a firm would employ such a worker and pay him by time (rather than by output) it could not sur-

vive since it has to bear the cost of shirking. If contract coordination is an explanation of the firm, then the costliness of measuring labor contribution is a necessary condition for its existence.

Perfectly accurate measurement of a worker's performance is expensive, perhaps prohibitively so. Performance, then, will be measured with error providing an opportunity for shirking or the transfer of wealth. Its extent, however, is a function of the contractual constraints. When exchange is conducted within the firm, both parties are employed by the same employer. The firm is able to curb the exploitation of inaccurate measurements in internal exchanges by contracting with its employees such that their gain from dissipating activities is lessened. "Hierarchy" is only incidental; contract coordination is the firm's crucial feature.

The presence of discrepancies between individuals' maximization and joint maximization induce efforts to effect the distribution of income. The proposition that the divergence between individual and joint maximization is eliminated when both parties are employed by the same firm requires reexamination. Employing a polluter and a "pollutee" within the same firm by itself does not eliminate the pollution problem -- the externality is not automatically internalized. A self-employed polluter who is rewarded according to the value of his measured net output which takes no account of the adverse effect of the pollution. An employed polluter who is rewarded for the difference between his measured output and his measured input will pollute exactly as when he is self-employed. Only when he is rewarded on some other basis such as his time input will he behave differently and only then can the externality be internalized.

Similar considerations apply to all exchanges. The delineation of the physical properties of the exchanged commodity and the difficulty in pricing it are not confined to the exchange among firms. Every problem that is present in the market is likewise present within the firm and the incentive to exploit a free attribute is not automatically eliminated merely by making the exchange an inter-firm operation.[5] Moreover, coordination within the firm is costly to effect and generates its own, albeit different, losses. An exchange contract will be allocated to the firm only when the reduction in losses brought about by coordination exceeds the cost of effecting it. When the cost of one rises whereas that of the other stays constant, it is expected that the latter will become more prevalent.[6]

II. The Cost of Exchange, the Performance of Labor and the Firm

Although people exchange because of the expected gains, the exchange itself produces the opportunity for one party to gain at the other's expense. This opportunity occurs because the physical properties and the values of traded items are not costlessly known. The cost of knowing and attaining exactly the appropriate marginal equalities with respect to every attribute would be prohibitive. To lower these costs, traded commodities lump together numerous attributes, and the quantities of some of the attributes are varied but not marginally priced. Though the legal ownership of these commodities may be clear, because of the costliness of metering the attributes, they may not be fully priced. For example, apples are not identical; nevertheless they are often sold at the same price per pound. The seller, in effect, relinquishes the rights to the differential value between the

more and the less valuable ones; part of the effort by buyers is in competing to appropriate that value.[7,8]

Output measured with error is less valuable than when it is more accurately measured, but more accurate measurement is more expensive. A person performing by himself a two-stage production process values the information gained from measuring the outcome of the intermediate step. Presumably, he will bring the accuracy of measurement to the point where its marginal value is equal to the marginal cost. Since private and joint maximizing coincide here, and abstracting from differences from specializing, the measurements undertaken by such a person will be optimal.[9]

In contrast, consider the two stage process that is split between two self-employed workers. Suppose that the output of the first person is offered for sale to the second at some unit price; that the seller measures the units he offers for sale to the optimal level of accuracy (i.e., the accuracy a single person performing both operations will attain); that the buyer is free to inspect the product and to buy or refuse any unit; and that the buyer's measurement cost is the same as the seller's.

By assertion, a seller's last dollar devoted to measurement will increase the value of the product by one dollar. The first dollar the buyer would spend on measuring will increase the (social) value of the good by not quite one dollar. By selecting the best units, however, he is able to appropriate some of the value that otherwise would have accrued to the seller. The units rejected, obviously, are estimated to be least valuable among those selling at the going price. The buyer gains from his own measurement by getting better information on the

product he is about to use; he also gains by retaining only units of above average value. To the buyer, the differential in value among units selling at a given price is a free attribute which he can appropriate by the expenditure of resources in the form of his own measurement cost. Thus the buyer's return from extra measuring is greater than his cost and he will engage in it. This measurement by the buyer, however, is carried beyond the jointly maximizing level.

In anticipation of the buyer's action, the seller has the incentive measure and meter his product more accurately. In this way he will retain some of the value otherwise appropriated by the buyer. These added measurements, however, similarly go beyond the jointly maximizing level.[10] Thus, the buyer, the seller, or both have incentives to over-measure relative to the joint maximizing level.

Two independent workers engaged in exchange, then, can be made better off if some way is found to reduce the over-measurement. The larger the "optimal" measurement error, the more resources are likely to be spent on "excess" measurement, and the stronger the joint incentive for the exchange parties to agree to constrain themselves. The constraint may take the form of the buyer ceding (part of) his right to choose. In the case of intermediate products the transactors have another option; they may agree to follow the instructions of a third party — the employer. As will be suggested below, a firm employing both workers can facilitate the imposition of the appropriate constraint. The purchase of labor services, however, entails measurement problems also, and these must first be discussed.

Consider the effects of the costs of measuring the way labor is used. Had all relevant measurements been costless, an employee could

be paid the exact value of his net contribution to output. The employer, effortlessly, would have calculated the gross value of that output from which he would have deducted the costs of raw materials and space, including costs such as those due to wear and tear of equipment and of the effect of the particular worker on the productivity of others. Since the employee would fully bear the consequences of any change in his own behavior, there would be no reason to deny him complete freedom in choosing his hours, pace of work, and so on. But then the employer and the firm become superfluous. There would be no difference in behavior between such an employed worker and a self-employed one who would choose to perform precisely the same functions. He would buy the same amounts of materials; occupy the same space; use the same equipment at the same pace and care; interact with other workers as before and obtain the same net income from selling his output.[11]

Suppose now that because of measurement problems, the employee is paid by the hour. Some stipulations are necessary to get the employee to exert himself and to produce any output. So long, however, as these stipulations fall short of attaining precisely the same outcome as that obtained when pay is on the basis of output value, the work package would not be worth as much to the employee as the former one.[12] Indeed under specialized production the costliness of measuring the work performance necessarily introduces free attributes and the associated problem.

It may seem that the employer possesses the means to police the performance of the employee, and that the right incentives can be offered to attain the optimal, or at least nearly optimal, employee pro-

ductiveness. After all, the worker will gain from performing well because otherwise he may be fired, because he would like to be promoted, etc.[13] What does it mean, however, that an employee performs well when the truly valued output is too costly to precisely measure? The employer may observe the sweat on the employee's brow; the noise around his workbench; the amount of some raw material used and so forth. These, however, are inputs or output-proxies and are unlikely to be perfectly correlated with true output. If output is measured with error, so is the worker's compensation. Had it been possible to costlessly observe an employee's net contribution to output, resource allocation would be more efficient. It would have been advantageous, then, to compensate him directly by his output. The fact that pay is tied to input implies that output is too costly to measure.

The input measure by which the employee is paid must be correlated with output. Otherwise, as a result of the worker's minimization of effort for a given pay, no useful output will be forthcoming. Even within the firm, then, the output of every worker has to be measured somehow. Can't whatever output-measure being used in assessing workers' performance also be used to exchange the output across firms, thereby dispensing with the employer-employee relationship? These measures, however, are subject to error that could be exploited. It is now shown that the employee's incentive to shirk can be turned around to combat dissipation in exchange.

An employed buyer such as a hired maid or a restaurant worker will necessarily spend less resources on selecting commodities for the employer than the latter would when buying for himself. The contract of an employed buyer must stipulate performance standards such as that he

completes a minimal number of purchases within an hour, and that the goods acquired meet certain minimum specifications. The delegation of the purchasing function to the employee implies that the employer will not enforce quality standards to the exact degree as when he himself is buying. Otherwise, the employer would have to spend as many resources on enforcement as he himself would have spent on selection, rendering the delegation pointless.[14] The employed buyer will make the least effort yielding a given pay. Thus to the extent that the quality of purchases is not enforced, the extra effort in selecting better items will not take place.[15]

More generally, selling apples by the pound and permitting buyers to choose implies that along the "quality" dimension the marginal charge is zero.[16] The amount of resources a person will spend to obtain this attribute depends on the net (perceived) reward. The employed buyer has a competing attribute to exploit -- his effort level. He is not fully penalized for shirking along this margin, and thus he will work less strenuously in, among others, obtaining the free attribute. The incentive to spend resources on obtaining more of the zero-priced attribute, therefore, is restrained by the employed worker's incentive to minimize effort. The two effects tend to cancel each other and it is expected that people will look for ways to take advantage of this feature.

The consumer buying for himself will fully exploit the free attribute in the purchase of apples while an employed buyer will partly neglect it to the detriment of his employer. The employer may purchase by himself rather than delegate the function.[17] If a principal delegates the buying function to an employee-agent, he may insist that the

other party to the exchange -- the seller does likewise, so that both the buyer and the seller will sort less. However, each of the two principals then requires information on the contractual terms constraining the other's agents, and the cost of such information seems high. Similarly, a formal contract to the same effect between buyer and seller seems excessively costly to monitor. It is asserted that the cost of coordinating contracts will be lowered if a single firm employs the two principals. Thus, the greater the cost saving by such coordination, the more likely the transaction will be carried within the firm.

Within the firm, as shown above, employees are never paid precisely on the basis of their marginal productivity. Any employee, being a maximizer, will shirk in any margin where his effort is not remunerated. When the firm employs both parties to an exchange, part of the shirking will occur in their dealings with each other. Employment contracts are expected to be designed so that shirking will be especially directed towards dissipating activities. It is the entrepreneur's ability to contract and relatively cheaply to mesh the constraints on pairs of employees that provides the inexpensive coordination of their incentives. As a result of the coordination, less effort is spent in competing away the value of the free attributes. Thus shirking is harnessed to perform a useful function.

Approaching the problem from a different angle, it is argued that one way to reduce the (joint) waste from the competition for the free attributes is to engage a monitor to supervise or police the exchange. If the reduction in dissipation costs in the exchange exceeds the sum of the supervisor's pay and of the cost of transacting with him, it is

worth while doing that. To induce the supervisor to maximize the net value of the exchange, his reward will be tied to that gain. On the other hand, to lower the costly competition for the unpriced margins in the exchange, the two exchange parties' rewards will be divorced, in part, from the value of the exchange. The different roles played as well as the sharing of the gains from this arrangement are reflected in the contracts of the three parties. The supervisor or the entrepreneur is the residual claimant -- the claimant to the gross value of the exchange less input payments, whereas the two exchange parties are his employees whose pay is not strictly a function of the value of the exchange and thus their dissipating incentive is lowered. The three constitute (part of) a "firm."

The construction industry is a convenient subject for the eventual testing of the hypothesis that costly to measure exchanges will be conducted within the firm. There are two major relevant sources of variability in the way this industry operates. One is the consequence of the diversity of structures ranging from single family homes to large office buildings. The other results from the substantial regional differences in the cost of materials. The variability in construction technology generates variability in the ability to evaluate the different output components. It is hypothesized that the more difficult it is to assess the output of a particular worker, the more likely he is to become a construction firm employee. Conversely, the easier it is to measure a worker's output, the more likely he will become an independent contractor.[18]

Pursuing the argument further, consider a home owner who wishes to repair his house. Would he do it himself, or would he use a contrac-

tor? Holding constant such factors as the contractor's travel cost and the how common his specialized knowledge is, it is predicted that the more likely it is for a contractor to employ the specialist for the type of job at hand, the more likely is the home owner to repair his home himself. If the employed specialist's output is difficult to measure, it is also relatively costly for the home owner to transact with him. On the other hand, if the specialist regularly works as a sub-contractor, his output is revealed to be easy to measure and thus he is also likely to be retained by the home owner.

Still another implication is as follows. Consider a function that is sometimes carried out within a firm and at other times is performed by an independent contractor. It is expected that the specifications, or measurements, will be more detailed and more rigid in the latter case. Within the firm it is already cheaper to avoid the costs associated with excess measurement and thus it will not be carried as far.[19]

In the next section the function of the firm is discussed in conjunction with transacting for the use of equipment. Particular attention is given to the problems that arise because a third party is introduced; a party which in turn may gain from the transfer to wealth.

III. The Role of the Firm in Monitoring Equipment

The market skills of a person are best complemented by some particular amount of capital goods. How can he obtain use of the goods when his wealth is not adequate for acquiring them?[20] Consider the use of capital equipment that requires a single operator, postponing

for later the problems of several workers sharing equipment. The person can get properly equipped by (1) borrowing money to purchase the equipment; (2) leasing the equipment; or (3) becoming the employee of an employer who provides the equipment. If the person takes an unsecured loan to purchase the equipment, default becomes a free attribute. If the lender uses the equipment as a collateral, the borrower can appropriate some of the value of the loan by running the equipment too hard and taking the "profit" out, and default on the loan when the value of the equipment falls to zero. A similar problem arises with equipment lease. In both cases the equipment will be used harder and with less care than if the operator of the equipment fully owned it.[21]

The lender or the lessor can lower the loss by constraining the equipment's use (e.g., restricting rental cars to paved roads), or by switching to equipment which is less amenable to abuse. An owner-operator will not constrain himself in the same way. Thus these capital market exchanges lead to losses similar to those associated with commodity exchange.

The capitalist could moderate equipment abuse by employing the operator on an hourly basis. The operator's incentive to push the equipment too hard is curtailed because his remuneration is only partly based on measured output.[22,23] This arrangement, however, is subject to a severe drawback. When the worker leases the equipment by the hour, he will tend to overuse it. Symmetry requires that the capitalist that "leases" the worker by the hour, will tend to "overuse" him. The employer could, for example, speed up the equipment (overloading people), or when the output level is already stipulated, he may provide inputs of mean quality lower than expected. In short,

he will employ whatever practices that are not in violation of the hourly labor contract and which maximize the net value of the operation to him.

Whether the worker leases the equipment or whether the equipment owner hires the worker, the resource cost of production is higher than when the owner of the equipment is also its (self-employed) user.[24] At the same time, to compensate the "exploited" parties, the hourly rates of leasing or of wages will be higher than when overuse is avoided. The intensity margin, however, is still priced at zero, and thus the excessive use is not eliminated. Additional stipulations may be adopted to prevent the exploitation. The owner of a leased machine may install a governor to prevent speeding. When the owner employs the operator, the latter may stipulate the use of a governor. Such stipulations are costly to enforce and thus the attempt by one party to transfer wealth from the other will persist.

Within the firm the incentives of both parties (as employees of one employer) can be structured so that when they interact, fewer resources are spent on acquiring free attributes. When an entrepreneur employs the worker largely on an hourly basis and rents the capital on a fixed rental basis, the difference between the cost of inputs and the value of output accrues to the entrepreneur. In reducing the two resource owners' expenditures on acquiring free attributes when they interact, the employer's income increases correspondingly.

The entrepreneur, however, will also gain by exploiting the two in essentially the same ways they might have exploited each other. If, however, the entrepreneur's function is turned to a manager who is also rewarded partly on the basis of his performance and partly for

his time, the severity of the problem is reduced. The owners supply the capital, but employ a manager rather than run the firm themselves. This "separation of ownership and control" serves to lower their reward (and therefore the incentive) from engaging in efforts to transfer wealth.[25] The owners may still gain from bankruptcy. Lenders can curb this incentive further if they stipulate that equity will constitute a significant portion of total capital and that the rate at which profits can be withdrawn is constrained.

IV. Scale Economies and the Size of the Firm

It is commonly claimed that the size of the firm depends on the efficient scale of equipment. Couldn't a single machine or structure be utilized efficiently by several firms? It is obviously advantageous to use the most efficient equipment, which may require numerous people to work with it. But there does not seem to be a technological reason for all of them to be employed by a single firm. A connection between the argument here and equipment scale is as follows. As the scale of a piece of equipment gets larger and more operators work with it, measuring the net output of the individual worker may become more difficult. Employing all these workers within a single firm may reduce the cost arising from overusing the equipment. The difficulty of measuring workers' output, however, is not inherent to large scale equipment, and no clear necessary relationship between firm size and the size of equipment seems to emerge.[26]

Suppose that the efficient scale of some equipment would require several people to work alongside with it. As the equipment is used by a worker, its value will fall by an amount which depends on the exact

way he handles it. The reduction in value of the equipment one of these workers causes must be netted out to obtain his true output. It is hypothesized that when all the users' effects on the equipment are easy to discern, or to measure, they will buy or lease shares in it. When their effects are difficult to measure, a single employer is expected to employ all its operators on a time basis.

Computers appear to be a prime example of the former. A large computer will accomodate numerous users. The damage a user inflicts on the computer is confined to the preemption of its use by others. It is easy to tell how long a user occupies the computer to determine the cost he imposes, and it is not difficult to charge for it. In other words, unpriced attributes are of little importance in this case. Thus even though the physical scale and the value of some computers are large, they can be effectively utilized by many independent users. The number of people working with the computer does not determine the number of workers employed by the firm owning the computer.[27]

Similarly, the occupants of large office buildings need not all be employed by the same firm. Even when a user damages the building and lowers its value the effect is easy to assign and to measure, and thus the transaction for the use of space does not involve significant unpriced margins. Transacting in the market then seems preferable to transacting within a firm.

In both examples, even though the scale of the physical capital and its value are large, the firms owning them may be modest size employers since the users of their capital are not their employees, but rather renters paying fixed fees. Indeed, the very same factors

also facilitate the division of ownership of the physical capital among several firms thus permitting severing the relationship between firm size and the unit size of the physical capital.

In other and seemingly more common cases any one worker's effect on the value of a large piece of equipment may be more difficult to gauge. If shares in such equipment are leased to several jointly using self-employed users, each is expected to take advantage of the free attributes such as the pace at which the equipment is run, the expense and effort given to lubrication and so on. Thus some of the value of the equipment will be dissipated. If, on the other hand, all the common users of the equipment are employed by the owner of the equipment, the employment contracts could be so formed and so coordinated that the abuse will be lowered.[28]

Firm size may sometimes be determined entirely independently of equipment scale. Consider a local industry such as construction. Within a particular market area the degree of specializing depends on the size of the market. It is asserted that the exact scope of each specialty is determined, in part, by the ease of exchange. One worker's task will end and another will commence at a juncture where the product is easy to measure. The smaller the market is, the smaller the degree of specializing, and the greater the choice in selecting easy to measure junctures to separate among the different specialties. As market size, and with it specializing increase, the greater will the difficulty be in measuring the product in the transition between specialists. Given the hypothesized edge of the firm in that situation, it is predicted that the larger the market size, the larger the fraction of local industry employment within firms and the smaller the

fraction of the self-employed. Moreover, as the market size grows larger, the larger the chance that the difficulty in measuring will occur in successive junctures and thus average firm size is also expected to be larger.

V. Theft as a Free Attribute

With theft, as is obvious, individual and joint maximization diverge and wealth is transferred at the cost of resources. This is a feature, then, that theft has in common with the costs of transacting considered here. It will now be shown that organizing exchange within the firm may lower the loss associated with some theft. This ability derives from the firm's low cost of contracting with its employees and modifying their perception of the gains from theft.

In many hardware stores, customers count (or weigh) items such as nuts, bolts, nails and washers and mark their prices. The cashiers routinely accept customers' statements regarding what they should be charged. Whereas the procedure invites theft through deliberate understatement, the store's savings in personnel cost evidently are even greater.[29] The theft opportunity is equivalent to the presence of a free attribute; some valued items can be acquired at prices below their costs. Resources then will be spent on acquiring them and their consumption will be excessive.[30]

Consider now a repairman whose task requires a large variety but small quantities of the items marked by the customer. When the repairman is self-employed, the entire gain from theft accrues to him. On the other hand, if he is employed and rewarded largely by the hour rather than by output value, the gain from theft, and therefore the

incentive for it are lowered. The employer of such repairmen may purchase these items on a wholesale basis avoiding the theft premium that must be paid by honest customers in the retail store.[31] By providing the items to his workers free, he will also save, as the store does, on the cost of transacting.[32] Whereas the store provides the free attribute through the theft opportunity, the employer provides the free attribute legally, and because of the repairmen's hourly wage contract, the incentive for excessive use is curbed. The firm, then, is able to lower the cost of some form of theft in a way comparable to that used to lower the cost of legal free attributes.

Theft among neighbors may also be constrained by consolidating the holdings into a firm. A rancher located next to a wheat farmer will gain if the cattle ate some of the wheat and thus will not restrain the cattle as much as if he owned the farm. The farmer may erect an extensive fence to lower the loss caused by the cattle.[33] Suppose, alternatively, that the two units were owned by a single person who hired the two operators and paid them by the hour. The gain to the ranch operator from the cattle eating the wheat is lowered, and similarly lowered is the gain to the farm operator from erecting the extensive fence. Thus the firm using the appropriate employment contracts avoids theft losses, though at the costs arising from employees' shirking.

Is it appropriate to set plain theft on an equal footing with the voluntary provision of free attributes such as the restaurant waiter explicitly offering another free cup of coffee or of the grocer permitting selection from his apple bin? Perhaps not, but the distinction is not always sharp. How would one classify the behavior of the

customer asking to taste the cheese he does not intend to buy, or the winery visit just to savor the samples? And isn't a polluter stealing from his neighbors, using their property as a dump without paying rent? The ambiguity is underscored by the term "moral hazard" describing the entirely legal overconsumption of insured services. The legality of overconsumption is probably a reflection of the difficulty, or high cost, of enforcing a law prohibiting the practice. This might also be true in other instances where attributes are provided free because legal action to constrain consumption that would have been then termed "theft" is prohibitively costly.

VI. Final Comments

Any exchange, because of the cost of transacting, contains some free attributes. Transactors will spend resources to acquire these attributes and will consume too much of them. When a worker is hired by the hour to execute a transaction, the dissipation is lowered because he will shirk in obtaining free attributes. The sale of labor services by the hour, however, itself offers some free attributes that the buyer can exploit. This is where the firm attains its special position. The employer of the first exchange party can hire the second too so that both become employees of the same firm. The firm's organizational advantage is its low cost of coordinating its employees' contracts to curtail their dissipating activities.

It was suggested above that one may view the entrepreneur as being engaged by the exchange parties to supervise their exchange behavior. Wouldn't they rather engage an arbitrator who will rule on how the gains from exchange in each transaction should be divided? To do an

effective job, however, the arbitrator must undertake the same elaborate measurement that his action is supposed to supercede. The success of the entrepreneuer derives not from after-the-fact arbitration but rather from changing the rules by which exchange is conducted. Because of the unavoidable "side effects" that employees' contracts entail, the entrepreneur is also given the power of command over his employees. The fundamental feature of the firm, however, is that the entrepreneur contracts with his employees in such a way such that their effort is channeled productively and their tendencies to shirk are exploited, being directed towards the otherwise dissipating activities. This coordination of behavior seems to be the entrepreneur's central function.

A final comment regarding the scope of the firm. As a rule a worker exchanges with two or more other workers. He obtains materials at one end and delivers a product at the other. To the extent that measurement is costly at both ends the firm employing such a worker may employ his exchange partners at both ends. It may also employ the partners of the partners. The breaks in the chain and thus the separation between firms will occur at the junctures where free attributes play a relatively minor role. The boundaries between firms will be found where the exchanged products are measured relatively accurately and cheaply.

FOOTNOTES

[1]In "The Nature of the Firm" Coase pointed out that the use of the price mechanism is costly because of the need of "discovering what the relevant prices are," and of "negotiating and concluding a separate contract for each exchange transaction." He stated that the firm will lower these costs since there the "direction of resources is dependent on an entrepreneur." He did not, however, explain how the entrepreneur directs resources. Neither did he show how the firm can function without separate and presumably elaborate contracts with each of its employees.

Given the approach here, Coase was correct in explaining the role of the firm in terms of the cost of transacting not because the firm's transaction costs are lower than those in the market but because the firm has <u>different</u> transaction costs. This paper, then, follows Coase in arguing that the firm is a transaction cost phenomenon, but provides more specific prediction on when the firm will supplant the market.

[2]A stylized example further illustrates the variety of free attributes and hints at their ubiquity. Compare two bookstore patrons. One walks in on a Wednesday morning, gets an expensive and slow selling book off the shelf, pays cash and departs. The other parks in the store's lot on Christmas Eve, turns over numerous books, obtains extensive help from a saleslady, buys a paper-back, gets it wrapped, pays with a check, and then returns the book. Each of the differences constitutes a freely provided attribute, the costs of which are lumped into the books' prices.

[3]Alchian and Demsetz' "team production" problem is that of measuring the individuals' output. They argue that the problem is resolved by organizing production within a firm where the residual claimant monitor is "the central party common to all contracts with inputs." (783) Here too the difficulty is in measuring, but it pertains to any exchange. Alchian and Demsetz' entrepreneur is the monitor who reduces shirking whereas here he is a contract coordinator who induces shirking towards the otherwise dissipating activities.

[4]Cohen, who also recognizes this problem, argues that this mode of payment is partly the consequence of the difficulty in measuring the product and partly due to the insurance motive. Lloyd R. Cohen, "The Firm: A Revised Definition," S.E.J., October 1979, 46: 580-590.

[5]Klein, <u>et al</u>., assert that the "hold-up" and the "appropriable quasi rent" problems (which can be viewed as competition for free attributes) are resolved by integration without, however, showing how this is accomplished.

[6]The complexity and variability of market transactions, as is perhaps already evident, is a major feature underlying this paper. Whereas the recognition of the complexity may be sufficient for the theoretical discussion, the ability to derive testable implication

requires considerable particular knowledge. Most of the implications offered below come from industries such as construction with which most laymen have some rudimentary acquaintance. Thus the restricted set of activities brought up in conjunction with testing reflects the lack of (at least this) economist's knowledge of the characteristics of other industries. It seems, however, that ultimately similar implications could be derived for other lines of activity.

[7]The measurement problem of whatever is used on the other side of the exchange is abstracted from.

[8]Williamson's reason for the costliness of exchange resembles that propounded here. For instance, he argues that haggling arises because of the difficulty of measuring commodity attributes. He says that "Although this haggling is jointly (and socially) unproductive, it constitutes a source of private pecuniary gain."(115) He states that the firm can reduce such losses since their "integration harmonizes interests..." He does not demonstrate, however, how this is attained.
Williamson, Oliver E., "The Vertical Integration of Production: Market Failure Considerations," American Economic Review, Papers and Proceedings, Vol. 51, May 1961, 112-123.

[9]In his pioneering work on information, Stigler considers information as another good so that (under competition) its social marginal value equals its social marginal cost.

[10]For a more detailed demonstration of this point see Barzel, "Measurement Costs and the Organization of Markets."

[11]A risk averse employee would wish to insure against income fluctuations. It is often claimed that employers provide such insurance. When all relevant measurements of an employee's performance are costless, as assumed above, insurance could be provided as easily by insurers as by the employer, since moral hazard can be costlessly controlled. The firm's edge in insuring its workers may arise from its superior position of controlling the moral hazard when measurement costs are positive. But then again the firm becomes a transaction or measurement cost phenomenon.

[12]The worker will gain from working less strenously, but the corresponding reduction in pay will exceed the value of the better working conditions.

[13]For such incentives to be effective, long-term employment prospects must be offered, including situations when otherwise long-term employment would not have been practiced.

[14]This assumes that the employer and the employee are equally adept at sorting. If purchasing is delegated to an employee with comparative advantage in sorting it is even less likely that shirking can be prevented.

[15]The use of random sampling, itself resource consuming, is capable of partly, but never fully, resolving this problem. Moreover,

the verification that a sample is indeed random is costly, and the motives of the sample taker are always suspect.

[16]A similar argument applies to price. Since it takes resources to determine the equilibrium price, buyers and sellers do not know exactly what the right price is. To the extent that bargaining takes place it may be partly to determine whether the highest price the buyer will pay exceeds the lowest price the seller will accept. It also serves to acquire the right to the indeterminate price. This, then, is another free attribute of a transaction and real resources will be spent to obtain it.

[17]Supermarket managers, whose pay is more closely linked to the store's profitability than other employees', usually are personally involved in inspecting and counting the shipments from wholesalers.

[18]More specific predictions must await better knowledge of construction. A hypothetical example, however, may illustrate the point. Suppose that of two varieties of lumber used for the same purpose, one's rate of deterioration is more varied than the other's. It is predicted that workers using the more varied lumber are more likely to be employed by a contractor whereas those working with the more uniform material are more likely to be self-employed.

[19]By the same token when a firm buys or sells an item which it also produces, the specifications for the outside transactions are expected to be more comprehensive than when the product is produced and used internally.

[20]Smith's discussion of accumulation is at the firm level suggesting that the successful entrepreneur will reinvest his firm's profits till his firm reaches (what in current terminology is) the optimal size.

Jensen and Meckling's starting point is the cost of transacting (with its implicit free attributes) between the owner and his lender. They also take as given that there is an optimal firm size.

[21]If, as asserted, policing costs of loans are high, the larger a person's net worth is, the more likely he is to be self-employed. Had each person's wealth been just right to get him equipped for his (potentially) best skill, the problem discussed here would disappear. The ability of a worker to finance the capital equipment with which he works gives a new meaning to the concept of "capital-labor ratio."

[22]McManus, addressing the problem in a similar spirit, states: "If the owner of the dump truck and the driver choose to co-ordinate their actions within a firm, one of them, say the owner, will direct the behavior of the other within limits that are mutually agreed upon. The driver's income will become less sensitive to his rate of output and he will therefore have less pecuniary incentive to depreciate the value of the truck in his use of it." John C. McManus, "The Costs of Alternative Economic Organizations," Canadian Journal of Economics, August 1975, 8:334-50.

[23]Shirking may be directed also to maintenance. To lower the associated loss, the employment contract may stipulate some maintenance work, some related materials such as lubricating oil may be "generously" provided, or the owner may take care of such problems in some other way.

[24]These losses, sometimes called "residual losses" are similar to those described by Jensen and Meckling.

[25]The entrepreneur, of course, could have acted in the same way he is trying to persuade his manager to act; people, however, would be leery transacting with him since they know that in any occasion he can gain at their expense more than the employed manager can.

[26]Surely there is no large scale equipment to explain the size of the large law firm. Neither can equipment scale explain "vertical integration" even if it could account for the size of any of the "horizontal" components.

[27]Computer's use poses an interesting problem in that the information processed by a computer is potentially subject to theft. A user who wants to protect his information may choose to operate the computer he is using rather than rent time on another's. In this case monitoring the use of the computer is a serious problem. The scale of the firm, then, may be related to that of the computer, but because of transacting problems rather than because of sheer size.

[28]Had the effect of a worker on the value of the equipment been easy to measure it is expected that much of what is now observed as personalized equipment would have been shared and indeed that larger scale units would have been more common.

[29]The cost saving must be large enough to also cover the expected increase of theft by employees. The expected theft by customers makes theft by employees easier. The owner's control of inventory is already problematic making employees' theft harder to detect.

[30]On average, of course, the price charged for the merchandise must cover its cost. The theft opportunity, however, implies that some customers some of the time spend resources to exploit the opportunity, and then "overuse" the good. Additionally, non-thieves "underutilize" the good because they pay the high real price. Only when the (proportionate) understatement across all customers is uniform can the discrepancy between price and marginal cost disappear.

[31]The employer, too, could try to steal the merchandise from the retail store, but a large scale theft is probably easy to detect.

[32]The employees, as well as those of the store, must be constrained from engaging in theft by directly selling these items and pocketing the money received.

[33]The easier the assessment of the damage (i.e., the lower the cost of measurement) the easier it is for the parties to contract (di-

rectly, or through court action) to restrain themselves from the wasteful action.

APPENDICES

AI. Free Attributes and the Demand Elasticity

To facilitate the discovery of additional implications the free attribute problem is examined from another angle. The losses associated with free attributes can, at least some of the time, be constrained. Without constraint, a restaurant could not for long dispense salt free of charge; all other salt users, including the highway department, would get it right there. A constraint that would induce consumers to obtain that quantity of the free attribute they would have obtained if (marginal) price equalled marginal cost will eliminate excessive use. The constraint, however, would not be very useful if it is costly to impose, or if it burdens consumers with other resource costs. The restaurant could, for instance, provide salt in cumbersome dispensers. Salt use would indeed drop, but the restaurant's net revenue will suffer because customers' willingness to pay will also fall. On the other hand, if dispensers are smooth-operating but small, patrons' cost of using the salt for their meals is minimal, but high for other uses.

Whenever a costly-to-produce commodity is offered free of charge, its consumption will be "excessive." The higher the elasticity of demand for the commodity, the larger the increase in consumption when price is reduced to zero, and the larger the associated loss. The demand for a restaurant's salt is made less elastic if patrons are discouraged from taking quantities useful for melting snow in their driveways; the demand for the quality of apples by employed buyers is less elastic if their reward for "excess" quality is lowered; and the

demand for a store's parking is made less elastic if neighboring stores are required to provide parking at the same price. In each of these cases, the constraint raises the cost of, or lowers the return from undesired substitution.

An attribute will be offered free of charge only if demanders can be sufficiently constrained so that the loss from excess use is less than the cost of separately pricing the attribute. The constraints may be direct, as in the small container of salt example. Mostly, however, they are indirect, making the user perceive the benefits from excessive consumption as low. If the exchange is between employees of a single firm, the entrepreneur is in a position to write the employment contracts with that objective in mind. These constraints are advantageous regardless of whether these people are employed by the same firm. It appears, however, that the cost of monitoring and of policing the terms of such agreements are less if a central party takes charge of that task. The firm, or rather the entrepreneur, constitutes such a central party.

The constraint may take still another form. The demand elasticity facing a seller depends on whether the prices of substitute commodities change simultaneously with the price of his good or whether the prices of the substitutes remain constant.[1] The excess use of an attribute offered free by one seller will be curtailed if other sellers can be induced to also supply the attribute free. This seems to be one of the functions of two major organizing devices -- joint ownership and fair trade.

Consider first joint ownership. When an attribute is offered free, the door is opened for third parties to take a ride. Here too

the appropriate contract coordination may restrain the excessive consumption. The provision of parking constitutes a straightforward, relatively simple instance where the free rider (more appropriately, the free parker) problem may occur. In downtown locations rent is high, and the return from explicit metering and pricing parking spaces is sufficient to cover the cost. In suburban areas, however, the return from explicit pricing is lower and store owners find it profitable to provide their customers with free parking, covering that cost by charging higher prices for their merchandise.

Few people will use the lot of a seller located at some distance from others' while shopping elsewhere. Suppose, however, that the best location for a store is near other sellers. These can reduce costs by letting their customers park free at their neighbor's lot. The same applies to the neighbors too, and too little parking will be furnished. If parking is offered as a separately priced service, both the merchandise and the parking will be priced at marginal cost. Nevertheless, given the prior assertion, this is not necessarily preferable. In the aggregate, these sellers would have done better by coordinating their pricing methods. Such an attempt at joint maximization, however, may be difficult to bring about because each seller would do best if only he stayed out.

In shopping centers the problem is resolved differently. These centers are characterized by single owners who simultaneously rent out space to the various sellers and provide common free parking. Through the centralized management of a shopping center the contracts between each store and its customers are coordinated with those of its neighbors. The total rent an owner is able to charge is presumably higher

than when he either provides priced parking or lets each of his tenants take care of his own parking arrangement.[2]

Shopping center firms are expected to be more common in new than in equally low-rent old areas whose development preceded the dominance of the automobile. Old areas are subject to the difficulty of land assembly and of the hold-out problem.[3] Moreover, a new shopping center can be deliberately located where it is expensive to abuse the free parking privilege.

Behavior here is coordinated by a single firm -- the owner of the shopping center, who gets all tenants to indirectly provide parking free of charge. Each of the tenants alone would have abstained from providing the free attribute. When all of them offer it simultaneously, the excess consumption of each is restrained by the similarly low price charged by the others. Thus the loss from failing to equate marginal cost to price by a seller is restrained by the coordinated violation of a similar condition by others. In this way an explanation is provided for the existence of the shopping-center-landlord-firm and for its particular contractual relationship with its tenants.

Consider now fair trade as another form of contract coordination. Some two decades ago Telser hypothesized that fair trade is designed to prevent a retailer from taking a ride on the free information supplied by another.[4] This explanation is a special case of the more general phenomenon considered here. Retailers often supply customers with information at no charge which implies that the cost of explicitly pricing that product-information exceeds the return from the finer pricing practice. Each retailer, however, would gain if others will bear the cost of the information. But then too little information

would be supplied.

The manufacturer who imposes a minimum retail price for his product actually coordinates the behavior of his retailers. He forces each of them, though indirectly, to supply free services alongside with the product.[5] Thus the opportunity for one seller to take a ride on another is reduced. It is expected, then, that commodities demanded along with a significant amount of "sales" services would be fair traded.[6] The incentive for fair trade extends to the entire industry when the information is applicable to substitutes produced by competing manufacturers. In this regard, then, cartels might be efficient.

AII. Road Services

The provision of free services is a familiar feature of government operations. Does the logic of the argument here also apply to such public sector service? It is hypothesized that the coordination of behavior with regard to the use of this service is attained by its exclusive provision by the state at no direct charges.

The supply of road services is taken for granted as a proper function of government. Yet it is evident that roads are not "public goods"; congestion is a constant and severe problem associated with their use. The reason they are supplied by government seems to be that the cost of collecting fees for these highly valued services is excessive. It was suggested above, however, that the costliness of pricing is a pervasive feature in private markets, and thus cannot alone explain government supply. The answer seems to be in the difficulty private operators encounter in excluding free riders.

turnpikes which were favorably located...found profitable operation extremely difficult... . Collection of tolls entailed burdensome operating expenses, and ensuring honest and efficient performance by tollhouse keepers was so difficult that the right to operate tollgates was often sold for a fixed sum... . The traveling public showed considerable reluctance to part with money for tolls. Shunpikes -- roads around the tollgates -- appeared widely despite the best efforts of the turnpike companies, and teamsters waited until after sundown in order to pass free when no collector remained on duty. Perhaps most damaging of all, wherever public roads offered fair passage, as they often did especially during the more favorable seasons of the year, the teamsters demonstrated great interest in choosing more roundabout routes if tolls could be avoided."[7]

Accepting the assertion of the pricing difficulty, why weren't roads provided privately at no charge as a component of another transaction? A mill owner, for instance, might have covered such a cost by raising the milling fee. The road, however, would attract other free riders including the customers of a competing mill, and exclusion, as evidenced by the above quote was costly. Thus except when the demand for the road is confined to an exclusive use, "too few" roads would be provided. The incentive for public provision is evident.

But then once <u>some</u> roads are provided publicly at no charge, the profitability of nearby private roads, whether explicitly priced or not, would decline. This, in turn, calls for the extension of the public road system into areas in which the absence of <u>any</u> public roads, private roads would have been provided.

A similar problem arises within the public sector. The federal

government is in charge of the interstate road system. Given the purpose of such roads, it seems reasonable that they would be paid for from federal funds rather than from directly taxing the state through which such a road passes. But then the states may free ride on the federal roads. The "matching funds" that the federal government grants to states for road construction seem to be in response to that problem. The costs of such roads, as perceived by the states, is lowered, and the incentive to overuse the federal roads is also lowered.[8]

NOTES

[1]This notion in elaborated on in Barzel's Tying Arrangements.

[2]A similar argument applies to other shopping center services. For instance, advertisements for merchandise sold in a center's store are likely to induce added purchases from other sellers in the center. The shopping center management is expected to stipulate in contracts with tenants some minimal advertising expenditures.

[3]A city ordinance requiring each store to provide and maintain some minimal parking space is a possible solution here. Indeed, affected businesses are expected to push for such a measure.

[4]Lester G. Telser, "Why Should Manufacturers Want Fair Trade?" JLE, October 1960, Vol. III: 86-105. Another given explanation, not mutually exclusive with the above, is the reduction in consumers' search cost that may arise with price variability.

[5]Competition among retailers will force them to offer some extra free service if the fair trade price is maintained. It is not clear why that free service will be that of information, since each seller still gains when his customers obtained the information from his competitors.

[6]It is not clear, however, how one can classify commodities by these criteria.

[7]"The Transportation Revolution, 1815-1960," by George R. Taylor, Harper, New York, 1978, (reprinted from 1951 ed.), pp. 29-30. In England, turnpike companies obtained the rights to "erect bars against byelanes, close up ancient highways, divert others at their pleasure and compel every one to travel by the new road they had constructed." Sidney and Beatrice Webb, The Story of the King's Highway, p. 120, Frank Cass, London, 1963 reprint, first published by Longmans Green, 1913.

[8]If the state is not constrained in locating its roads, expected is a free ride on the federal roads, taking the form of constructing the bulk of the state roads at "too great" a distance from the federal roads and "too close" to each other.