CHAPTER 12

The boundaries of the firm revisited

BENGT HOLMSTROM AND JOHN ROBERTS

Bengt Holmstrom was born in 1949 in Helsinki, Finland. He earned his Ph.D. in economics from the Graduate School of Business, Stanford University, in 1978. When this article was published, he was Edwin J. Beinecke Professor of Management Studies, Yale School of Management. He is presently the Paul A. Samuelson Professor of Economics at Massachusetts Institute of Technology.

John Roberts is the John H. and Irene S. Scully Professor of Economics, Strategic Management and International Business in the Graduate School of Business, Stanford University. See also the chapter by Roberts and Milgrom.

Why do firms exist? What is their function, and what determines their scope? These remain the central questions in the economics of organization. They are also central questions for business executives and corporate strategists. The worldwide volume of corporate mergers and acquisitions exceeded $1.6 trillion in 1997. It is hard to imagine that so much time, effort and investment bankers' fees would be spent on adjusting firm boundaries unless there was some underlying economic gain. Indeed, the exceptional levels of merger and acquisition activity over the past two decades are a strong indication that economically significant forces do determine organizational boundaries.

The study of firm boundaries originated with the famous essay by Coase (1937), who raised the question of why we observe so much economic activity inside formal organizations if, as economists commonly argue, markets are such powerful and effective mechanisms for allocating scarce resources. Coase's answer was in terms of the costs of transacting in a world of imperfect information. When the transaction costs of market exchange are high, it may be less costly to coordinate production through a formal organization than through a market.

In large part thanks to the work of Williamson (1975, 1985), recent decades have seen a resurgence of interest in Coase's fundamental insight that firm

Reprinted in abridged form from Bengt Holmstrom and John Roberts, "The Boundaries of the Firm Revisited," *Journal of Economic Perspectives,* 12, 4(1998): 73–94. Reprinted with the permission of the American Economic Association.

boundaries can be explained by efficiency considerations. Our understanding of firm boundaries has been sharpened by identifying more precisely the nature and sources of transaction costs in different circumstances. In the process, the focus of attention has shifted away from the coordination problems originally emphasized by Coase and towards the role of firm boundaries in providing incentives. In particular, the most influential work during the last two decades on why firms exist, and what determines their boundaries, has been centered on what has come to be known as the "hold-up problem."

. . .

The next section of the paper will review the two strains of work that have dominated the research on the boundaries of the firm: transaction cost economics and property rights theory. Both theories, while quite different in their empirical implications, focus on the role of ownership in supporting relationship-specific investments in a world of incomplete contracting and potential hold-ups. There is much to be learned from this work.

In this essay, however, we argue for taking a much broader view of the firm and the determination of its boundaries. Firms are complex mechanisms for coordinating and motivating individuals' activities. They have to deal with a much richer variety of problems than simply the provision of investment incentives and the resolution of hold-ups. Ownership patterns are not determined solely by the need to provide investment incentives, and incentives for investment are provided by a variety of means, of which ownership is but one. Thus, approaches that focus on one incentive problem that is solved by the use of a single instrument give much too limited a view of the nature of the firm, and one that is potentially misleading.

. . .

Investment incentives are not provided by ownership alone

There is no doubt that hold-up problems are of central concern to business people. In negotiating joint venture agreements, venture capital contracts, or any of a number of other business deals, much time is spent on building in protections against hold-ups. At the same time, such contracts are *prima facie* evidence that hold-up problems do not get resolved solely by integration of buyer and seller into a single party – the firm. Indeed, there seems to be something of a trend today toward disintegration, outsourcing, contracting out, and dealing through the market rather than bringing everything under the umbrella of the organization. This trend has seen the emergence of alternative, often ingenious solutions to hold-up problems.

Japanese subcontracting

The pattern of relations between Japanese manufacturing firms and their suppliers offers a prominent instance where the make-buy dichotomy and related theorizing have been less than satisfactory. Although the basic patterns apply in a number of industries (including, for example, electronics), the practices in the automobile industry are best documented (Asanuma, 1989, 1992). These patterns have spread from Japan to the auto industry in the United States and elsewhere, and from autos to many other areas of manufacturing. These practices feature long-term, close relations with a limited number of independent suppliers that seem to mix elements of market and hierarchy. Apparently, these long-term relations substitute for ownership in protecting specific assets.

Two points of contrast in the treatment of specific investments between traditional U.S. practice and the Japanese model present particular problems for the received theory. The first concerns investments in designing specialized parts and components. Traditional U.S. practice featured either internal procurement or arm's length, short-term contracting. Design-intensive products were very often procured internally (Monteverde and Teece, 1982).[1] When products were outsourced, the design was typically done by the automaker, with the drawings being provided to the suppliers. This pattern is what hold-up stories would predict, for the investment in design is highly specific and probably cannot be protected fully by contracts; thus, external suppliers will not make such relationship-specific investments, for fear that they will be held up by buyers after their investments are in place. In stark contrast, it is normal practice for Japanese auto firms to rely on their suppliers to do the actual design of the products supplied. The design costs are then to be recovered through the sale price of the part, with the understanding that this price will be adjusted in light of realized volumes.

A second contrast: traditional U.S. practice has been that physical assets specific to an automaker's needs are owned by the automaker. This clearly applies in the case of internally procured items, but it also holds in cases where the assets are used by the external supplier in its own factory. For example, the dies used in making a particular car part will belong to the automaker, even though they are used in the supplier's plant on the supplier's presses. Again, this accords well with the transaction cost story of potential hold-up by the automaker.[2] In Japan, in contrast, these specific investments are made by the supplier, who

[1] However, this pattern did not become standard until decades after the founding of the industry. Earlier, something akin to the practices associated now with the Japanese was the norm. See Helper (1991).

[2] An alternative story is more in the line of Williamson's earlier discussions emphasizing inefficiencies in ex post bargaining. The useful life of a die far exceeds the one-year contracting period. If the supplier owned the die, changing suppliers would require negotiating the sale of the die to the new supplier, and this could be costly and inefficient.

retains ownership of the dies. This would seem to present the automaker with temptations to appropriate the returns on these assets, once the supplier has made the relationship-specific investment. Moreover, because the Japanese auto manufacturers typically have a very small number of suppliers of any part, component or system, the supplier would also seem to be in a position to attempt opportunistic renegotiation by threatening to withhold supply for which there are few good, timely substitutes.

The Japanese pattern is directly at odds with transaction cost theory. Meanwhile, the divergence in ownership of the dies between the two countries presents problems for attempts to explain ownership allocation solely in terms of providing incentives for investment.[3]

In Japanese practice, explicit contracting is not used to overcome the incentive problems involved in outsourced design and ownership of specific assets. In fact, the contracts between the Japanese automakers and their suppliers are short and remarkably imprecise, essentially committing the parties only to work together to resolve difficulties as they emerge. Indeed, they do not even specify prices, which instead are renegotiated on a regular basis. From the hold-up perspective, the prospect of frequent renegotiations over the prices of parts that are not yet even designed would certainly seem problematic.

The key to making this system work is obviously the long-term, repeated nature of the interaction.[4] Although supply contracts are nominally year-by-year, the shared understanding is that the chosen supplier will have the business until the model is redesigned, which lasts typically four or five years. Moreover, the expectation is that the firms will continue to do business together indefinitely. There has been very little turnover of Japanese auto parts suppliers: over a recent eleven-year period, only three firms out of roughly 150 ceased to be members of *kyohokai*, the association of first-level Toyota suppliers (Asanuma, 1989).

The familiar logic of repeated games, that future rewards and punishments motivate current behavior, supports the on-going dealings.[5] An attempted hold-up would presumably bring severe future penalties. As importantly, the amount of future business awarded to a supplier is linked to ratings of supplier performance. The auto companies carefully monitor supplier behavior – including cost reductions, quality levels and improvements, general cooperativeness, and so on – and frequent redesigns allow them to punish and reward performance on an on-going basis. In this sense, supplier relationships in Japan are potentially *less*, not more, locked in than in the traditional U.S. model, where at the

[3] Interestingly, Toyota followed U.S. practice in supplying the dies used by at least some of the suppliers to its Kentucky assembly plant (Milgrom and Roberts, 1993).

[4] Taylor and Wiggins (1997) argue that these long-term relations are also the means used in the Japanese system to solve moral hazard problems with respect to quality.

[5] Baker et al. (1997) [Ed.: published as Baker, Gibbons and Murphy 2002] present a formal analysis of the choice between external and internal procurement, taking into account the important fact that long-term relational contracts can be maintained both within a firm as well as across firms.

corresponding point in the value chain, the supplier is typically an in-house division or department.

Having a small number of suppliers is crucial to the Japanese system. It reduces the costs of monitoring and increases the frequency of transacting, both of which strengthen the force of reputation. Also, the rents that are generated in the production process do not have to be shared too widely, providing the source for significant future rewards. This logic underlies the normal "two-supplier system" used at Toyota. There is more than one supplier to permit comparative performance evaluation, to allow shifting of business as a reward or punishment, to provide insurance against mishaps, and perhaps to limit the hold-up power of each supplier, but the number is not chosen to minimize hold-ups.

The relationship is marked by rich information sharing, including both schedules of production plans necessary for just-in-time inventory management and also details of technology, operations and costs. The automakers also assist the suppliers in improving productivity and lowering costs: technical support engineers are a major part of the automakers' purchasing staff, and they spend significant amounts of time at the suppliers' facilities. All this in turn means that potential information asymmetries are reduced, which presumably facilitates both performance evaluation and the pricing negotiations.[6]

Perhaps the major problem in the system may be that the automakers are inherently too powerful and thus face too great a temptation to misbehave opportunistically. Indeed, many Japanese observers of the system have interpreted it in terms of the automakers' exploitation of their power. One counterbalance to this power asymmetry is the supplier association, which facilitates communication among the suppliers and ensures that if the auto company exploits its power over one, all will know and its reputation will be damaged generally. This raises the cost of misbehavior. In this regard, the fact that Toyota itself organized an association of the leading suppliers for its Kentucky assembly plant is noteworthy (Milgrom and Roberts, 1993).

An alternative solution to this imbalance would be for the automaker to own the dies, as in the United States. Here a property rights explanation may be useful: under this arrangement, the supplier would not have the same incentives to maintain the dies, since it must be very hard to contract over the amount of wear and tear and its prevention.[7]

[6] Strikingly, as automobile electronics have become more sophisticated and a greater part of the cost of a car, Toyota has ceased to rely exclusively on its former sole supplier, Denso, and has developed its own in-house capabilities in this area. Arguably, this was to overcome information asymmetries and their associated costs (Ahmadjian and Lincoln, 1997). In contrast, see the discussion of the effects of Ford's complete reliance on Lear for developing seats for the redesigned 1997 Taurus (Walton, 1997).

[7] See Segal and Whinston (1997) for a model in the property rights spirit that is relevant to these issues.

Mini-mills, exclusive sourcing and inside contracting

Another significant shift in the organization of production is illustrated by Nucor, the most successful steel maker in the United States over the past 20 years. Nucor operates mini-mills, which use scrap (mainly car bodies) as raw material for steel production. After an initial technological breakthrough, Nucor started to expand aggressively (Ghemawat, 1995). The strategy required much capital, and to save on capital outlays, Nucor decided to outsource its entire procurement of steel scrap. Traditionally, mini-mills had integrated backwards, partly to secure an adequate supply of raw material and partly because sourcing entails substantial know-how and so was considered "strategically critical." Chaparral Steel, another big mini-mill operator, continues to be integrated backwards, for instance.

In a break with the tradition, Nucor decided to make a single firm, the David J. Joseph Company (DJJ), its sole supplier of scrap. Total dependence on a single supplier would seem to carry significant hold-up risks, but for more than a decade, this relationship has been working smoothly and successfully. Unlike in the Japanese subcontracting system, there are certain contractual supports. Prices are determined by a cost-plus formula to reflect market conditions, and an "evergreen" contract specifies that the parties have to give advance warning (about half a year in advance) if they intend to terminate the relationship. Even so, there is plenty of room for opportunism. Despite transparent cost accounting (essentially, open books), DJJ can misbehave, since realized costs need not be the same as potential costs. Asset specificity remains significant even with the six-month warning period, since a return to traditional sourcing and selling methods would be quite disruptive and expensive for both sides. Indeed, one reason why the partnership has been working so well may be the high degree of mutual dependence: Nucor's share of DJJ's scrap business is estimated to be over 50 percent.

The success of Nucor's organizational model has led other mini-mills to emulate and refine it. In England, Co Steel has gone as far as relying on its sole supplier to make ready-to-use "charges," the final assemblage of materials to go into the steel-making ovens. The production technology for charges is quite complicated: about 20–30 potential ingredients go into each mixture, with the mix depending on the desired properties of the final product, and big cost savings can be had by optimizing the use of the different inputs. This activity entails much know-how and requires extensive information exchange with the steel plant to match inputs with final product demand. The charges must be prepared by the supplier on Co Steel's premises, both for logistical reasons and to facilitate information sharing. In transaction cost economics, such a cheek-by-jowl situation would be an obvious candidate for integration. Yet, the industry is moving in the direction of disintegration in the belief that specialization will

save on costs by eliminating duplicate assets, streamlining the supply chain, and providing better incentives for the supplier through improved accountability.

Related experiments of "inside contracting" include Volkswagen's new car manufacturing plant in Brazil, where the majority of the production workers in the factory are employees, not of Volkswagen, but of subcontractors that provide and install components and systems on the cars as they move along the line. It is too early to tell whether other firms will return to inside contracting, which used to be quite common in the United States up to World War I (Buttrick, 1952), and whether such a move will be successful. But evidently, even potentially large hold-up problems have not deterred recent experimentation.

Airline alliances

Another illustration of close coordination without ownership is provided by airline alliances, which have proliferated in recent years. Coordinating flight schedules to take advantage of economies of scope requires the parties to resolve an intricate set of issues, particularly ones related to complex "yield management" decisions on how to allocate seats across different price categories and how to shift prices as the flight date approaches. Information and contracting problems abound, and it is hardly surprising that tensions occasionally surface. For instance, KLM and Northwest Airlines recently ran into a dispute that had to be resolved by dismantling their cross-ownership structure. But interestingly, this did not prevent KLM and Northwest from deepening their commitment to their North Atlantic alliance by agreeing to eliminate, over a period of years, all duplicate support operations in the United States and Europe. With the completion of this deal, KLM and Northwest have made themselves extraordinarily interdependent in one of the most profitable segments of their business. A 13-year exclusive contract, with an "evergreen" provision requiring a three-year warning before pull-out, is the main formal protection against various forms of opportunism, but undoubtedly the real safeguard comes from the sizable future rents that can be reaped by continued good behavior.

Why don't the two airlines instead integrate? Regulations limiting foreign ownership and potential government antitrust objections are a factor, as may be tax considerations. However, an explanation we have been given is that airline cultures (and labor unions) are very strong and merging them is extremely difficult. Pilot seniority is a particularly touchy issue.

Contractual assets and network influence

In property rights theory, the boundaries of the firm are identified with the ownership of assets, but in the real world, control over assets is a more subtle

matter. "Contractual assets" can often be created rather inexpensively to serve some of the same purposes that the theory normally assigns to ownership: to provide levers that give bargaining power and thereby enhance investment incentives. What we have in mind here are contracts that allocate decision rights much like ownership; for instance, exclusive dealing contracts such as Nucor's, or licensing agreements of various kinds. Such "governance contracts" are powerful vehicles for regulating market relationships. With increased disintegration, governance contracts seem to have become more nuanced and sophisticated. They place firms at the center of a network of relationships, rather than as owners of a clearly defined set of capital assets.

. . .

Microsoft and the web of inter-firm relations centered around it provide another illustration. The stock market values Microsoft at around $250 billion, which is more than $10 million per employee. Surely very little of this is attributable to its ownership of physical assets. Instead, by leveraging its control over software standards, using an extensive network of contracts and agreements that are informal as well as formal and that include firms from small start-ups to Intel, Sony and General Electric, Microsoft has gained enormous influence in the computer industry and beyond. We are not experts on Microsoft's huge network of relationships, but it seems clear that the traditional hold-up logic does poorly in explaining how the network has developed and what role it serves. If one were to measure asset specificity simply in terms of separation costs, the estimates for breaking up some of the relationships – say, separating Intel from Microsoft – would likely be large. Yet these potential losses do not seem to cause any moves in the direction of ownership integration.

A similar pattern can be observed in the biotechnology industry (Powell, 1996). As in the computer industry, the activities of the different parties are highly inter-related, with different firms playing specialized roles in the development and marketing of different products. Most firms are engaged in a large number of partnerships; for instance, in 1996 Genentech was reported to have 10 marketing partnerships, 20 licensing arrangements, and more than 15 formal research collaborations (Powell, p. 205). Significant relationship-specific investments are made by many parties, and potential conflicts must surely arise after these investments are in place. Yet the system works, thanks to creative contractual assets – patents and licensing arrangements being the oldest and most ingenious – but also to the force of reputation in a market that is rather transparent, because of the close professional relationships among the researchers.

Firm boundaries are responsive to more than investment incentives

The examples above make clear that there are many alternatives to integration when one tries to solve hold-up problems. The examples also suggest that ownership may be responsive to problems other than underinvestment in specific assets. Speaking broadly, the problems relate to contractual externalities of various kinds, of which hold-ups are just one.

Resolving agency problems

An example of how agency issues can affect the boundaries of an organization is whether a firm employs its sales force directly, or whether it uses outside sales agents. The best-known example here involves electronic parts companies, some of which hire their own sales agents while others sell through separate supply companies (Andersen, 1985; Andersen and Schmittlein, 1984). Originally, Andersen (1985) appears to have expected that the observed variation in this choice would relate to the degree of asset specificity; for example, the extent to which investment by sales people with knowledge about products was specific to a particular company. Instead, measurement costs and agency concerns turned out to be central. An employee sales force is used when individual performance is difficult to measure and when non-selling activities (like giving customer support or gathering information about customers' needs) are important to the firm; otherwise, outside companies are used.

Holmstrom and Milgrom (1991a, 1994) rationalize this pattern with a model of multi-task agency, in which sales people carry out three tasks: making current sales, cultivating long-term customer satisfaction, and gathering and relaying information on customer needs. If the latter two activities are important and if the three activities compete for the agent's time, then the marginal rewards to improved performance on each must be comparable in strength; otherwise, the ill-paid activities will be slighted. Because performance in non-selling activities is arguably hard to measure, it may be best to provide balanced, necessarily lower-powered incentives for all three activities.

Offering weak incentives to an outside sales agent can be problematic, however, because the agent may then divert all effort to selling other firms' products that come with stronger rewards for sales. With an employee, this problem can be handled with a salary and a low commission rate, because the employee's outside activities are more easily constrained and promotion and other broader incentives can be used within the firm to influence the agent's behavior.[8] This

[8] For a further discussion of the idea that low-powered incentives are a major virtue of firm organization and can help explain firm boundaries, see Holmstrom (1996).

logic also explains why outside agents commonly receive higher commission rates than does an inside sales force.

A less familiar illustration of how ownership responds to agency concerns comes from multi-unit retail businesses. Some of these businesses are predominantly organized though traditional franchise arrangements, in which a manufacturer contracts with another party to sell its products in a dedicated facility, as in gasoline retailing. Others, including fast-food restaurants, hotels and pest-control services, are organized in what is called "business concept" franchising. The franchiser provides a brand name and usually other services like advertising, formulae and recipes, managerial training and quality control inspections, collecting a fee from the franchisee in return, but the physical assets and production are owned and managed by the franchisee. Sometimes franchisers (like McDonald's) own and operate a number of outlets themselves. Finally, other businesses are commonly organized with a single company owning all the multiple outlets and hiring the outlet managers as employees. Examples are grocery supermarkets and department stores. What accounts for such differences?

It is hard to see how the specificity of the assets – real estate, cash registers, kitchens and inventories – differs between supermarkets and restaurants in such a way that transactions cost arguments would lead to the observed pattern. Indeed, the assets involved are often not very specific at all. Alternatively, applying the Hart-Moore property rights model here would involve identifying non-contractible investments that are unavailable to the other party if the franchise agreement is terminated or the store manager's employment should end. Noncontractible investments by the center in building the brand might qualify on the one hand, but in many cases, it is hard to see what the investments of the operator might be. For example, a fast-food restaurant manager might invest in training the workers and building a clientele, but these investments would presumably still be effective even if the manager were replaced by another. Further, these should also be investments that vary across cases in such a way that it is more important to provide the strong incentives of ownership to the manager of the outlet in one case and to the central party in the other.[9]

An alternative approach based on the need to offer incentives for effort has been proposed by Maness (1996). This approach begins by noting that any elements of the retail outlet's financial costs that are sufficiently difficult to measure must accrue to the owner of the outlet as residual claimant, because they cannot be passed by contract to another party. Suppose then that all costs are non-contractible in this sense; that is, since the level or appropriateness of various costs cannot be well-monitored from outside, such costs cannot be part of an agreed-upon contract. Then, the only possibility for payments from the owner to the other party is on the basis of revenues. Indeed, actual franchise

[9] See Lutz (1995) for a formal model of franchising along these lines.

fees are almost always based on revenues and not on costs (Maness, p. 102) and incentive pay for employee managers is also often based on sales. In such a structure, the employee-manager has no direct incentive to control costs under central ownership, while the franchiser has no incentives for cost-reduction under local ownership. Because the efforts of either party might affect costs, this creates a potential inefficiency. The solution is to lodge ownership with the party to whom it is most important to give incentives for cost control. Maness then argues that cost control in a fast-food operation is more influenced by the local manager's efforts at staffing, training, controlling waste and the like, while costs in supermarkets are most influenced by the inventory and warehousing system, which can be centrally managed. Thus, an explanation emerges for the observed patterns: ownership is assigned to give appropriate incentives for cost control.

A more complex example involves gasoline retailing in the United States and Canada, which has been studied by Shepard (1993) and Slade (1996), respectively. They document a variety of contractual arrangements that are used in each country between the gasoline refining company and the station operator and, in the United States, significant variation in ownership of the station. While the physical assets used in gasoline retailing are quite specific to that use, a station can be switched from one brand to another with a little paint and new signs. Consequently, neither study attempted to explain the variation in contractual and ownership arrangements in terms of specific assets and hold-up.

Both studies find that the observed patterns are consistent with the arrangements being chosen to deal with problems of inducing effort and its allocation among tasks. These arrangements differ over the strengths of the incentives given to sell gasoline and other, ancillary services like repairs, car washes, or convenience store items. In turn, these ancillary services differ in the ease and accuracy of performance measurement. The observed patterns were generally consistent with their being selected to provide appropriately balanced incentives. For example, Shepard's (1993) work notes that in repair services, effort is hard to measure and, more importantly, monitoring the realized costs and revenues by the refiner may be tricky. This should make it less likely that the refinery will own the station and employ the operator and more likely that an arrangement will be adopted where the operator is residual claimant on sales of all sorts, either owning the station outright or leasing it from the refiner on a long-term basis. This is what the data show. In Slade's (1996) data, the presence of repair did not affect the ownership of the station (essentially all the stations were refiner-owned). It did, however, favor leasing arrangements, where the operator is residual claimant on all sales, and diminished the likelihood of commission arrangements, which would offer unbalanced incentives because the operator is residual claimant on non-gasoline business but is paid only a small commission on gasoline sales. The presence of full service rather

than just self-serve gasoline sales also favors moving away from the company-owned model, since it matches the returns to relationship-building with the costs, which are borne by the local operator.[10] However, adding a convenience store actually increases the likelihood of using company-owned and -operated stations in the U.S. data, which goes against this logic unless one assumes that monitoring of such sales is relatively easy.

Considering a broad variety of retailing businesses more generally, LaFontaine and Slade (1997) document that the contractual and ownership arrangements that are used are responsive to agency considerations.

. . .

Knowledge transfers and common assets

Information and knowledge are at the heart of organizational design, because they result in contractual and incentive problems that challenge both markets and firms. Indeed, information and knowledge have long been understood to be different from goods and assets commonly traded in markets. In light of this, it is surprising that the leading economic theories of firm boundaries have paid almost no attention to the role of organizational knowledge.[11] The subject certainly deserves more scrutiny.

One of the few economic theory papers to discuss knowledge and firm boundaries is Arrow (1975), who argued that information transmission between upstream and downstream firms may be facilitated by vertical integration. As we saw in the examples of Nucor and the case of Japanese subcontracting, however, this type of information transfer may actually work fairly well even without vertical integration. More significant problems are likely to emerge when a firm comes up with a better product or production technology. Sharing this knowledge with actual or potential competitors would be socially efficient and could in principle enrich both parties, but the dilemma is how to pay for the trade. Until the new ideas have been shown to work, the potential buyer is unlikely to want to pay a lot. Establishing the ideas' value, however, may require giving away most of the relevant information for free. Again, repeated

[10] A hold-up story is consistent with the fact that the presence of repair services favors dealer ownership over leasing arrangements in the U.S. data: a lessee who invests in building a clientele for repair work might worry that the refining company will raise the lease payments to appropriate the returns from this investment. This argument, however, does not do much to explain the pattern in the Canadian data, where the refiners own all the stations. One might also attempt to apply this logic to the choice between company-owned and leased stations by arguing that if the company owns the station it cannot motivate the employee-manager to invest in building a clientele because it will appropriate all the returns. However, this argument is not compelling without explaining how firms in other industries succeed in motivating their employees to undertake similar investments.

[11] In contrast, researchers outside economic theory have made much of the role of knowledge. See, for instance, Teece et al. (1994).

interactions can help here; in fact, even competing firms engage in continuous information exchange on a much larger scale than commonly realized. A common example is the extensive use of benchmarking, in which the costs of particular processes and operations are compared between firms. But when big leaps in knowledge occur, or when the nature of the knowledge transfer will involve ongoing investments or engagements, the issues become more complex. A natural option in that case is to integrate. Any claims about the value of knowledge are then backed up by the financial responsibility that comes with pairing cash flow and control rights.[12]

We think that knowledge transfers are a very common driver of mergers and acquisitions and of horizontal expansion of firms generally, particularly at times when new technologies are developing or when learning about new markets, technologies or management systems is taking place. Given the current level of merger and acquisition activity, and the amount of horizontal rather than vertical integration, it seems likely that many industries are experiencing such a period of change. The trend towards globalization of businesses has put a special premium on the acquisition and sharing of knowledge in geographically dispersed firms.

. . .

The problem with knowledge transfers can be viewed as part of the more general problem of free-riding when independent parties share a common asset. If bargaining is costly, the situation is most easily solved by making a single party responsible for the benefits as well as the costs of using the asset. Brand-names are another example of common assets that typically need to be controlled by a single entity.

Concluding remarks

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 hold-up problem and the role of asset specificity. Think of arraying the set of coordination and motivation problems that the firm solves along one dimension of a matrix, and the set of instruments it has available along the other. Put the provision of investment incentives in column one and ownership-defined boundaries in row one. Let an element of the matrix be positive if the corresponding instrument is used to solve the corresponding problem, and zero

[12] Stuckey (1983), in his extraordinary study of the aluminum industry, reports that knowledge transfer was an important driver of joint ventures.

otherwise. So there is certainly a positive entry in row one, column one: ownership does affect incentives for investment. We have argued, however, that both the first column and the first row have many other positive elements; ownership boundaries serve many purposes and investment incentives are provided in many ways.

. . .

The property rights approach, with its emphasis on incentives driven by ownership, may be a good starting point. . . . But this approach also needs to expand its horizon and recognize that power derives from other sources than asset ownership and that other incentive instruments than ownership are available to deal with the joint problems of motivation and coordination. We do not believe that a theory of the firm that ignores contracts and other substitutes for ownership will prove useful for empirical studies. The world is replete with alternative instruments and, as always, the economically interesting action is at the margin of these substitutions.