RSM 392H1: Competitive Strategy

University of Toronto Rotman School of Management Undergraduate Seminar
Professor: Kevin A. Bryan

Class Notes: Theories of the Firm

Modern economies by and large consist of firms. Nearly all firms *make* some of what they use, and *buy* the rest. For instance, a restaurant will generally cook the food, and have their employees serve it on plates owned by the restaurant, but will buy electricity and ingredients, and rent their storefront. In larger firms, the distinction between making and buying is more nebulous: large conglomerates like Hyundai and General Electric are vertically integrated in some cases – the firm owns and controls means of production from basic resource extraction to retail – and highly disintegrated in other cases, performing only a small portion of the total production of some good. There are large agricultural firms like ADM, performing many diverse activities in broad areas related to farming with a large full-time staff, as well as large farms who simply hire workers on a day-to-day basis and sell all of their output to a middleman as soon as it is picked. Owners have become rich from giant conglomerates like Carnegie Steel and from tiny "virtual" operations like John Jacob Astor's American Fur Company.

Even very similar firms treat the decision of making a good within the firm or buying it from outside in very different ways. For instance, both Apple and Intel rely on production facilities located outside of the United States. Intel owns an assembly plant in Costa Rica which, as of 2006, was responsible for five percent of the GDP of the entire country. Apple, on the other hand, contracts huge portions of his production to a Taiwanese supplier, Foxconn, who assembles iPhones in China on behalf of Apple. Why does Intel own their overseas production plant and Apple contract with an outside firm for assembly?[1]

Stepping back from the optimal *size* of a firm, the very *existence* of firms is something of a puzzle. Every first year economics student learns the First Fundamental Theorem of Welfare Economics: every Walrasian equilibrium, under certain assumptions, is Pareto efficient. This means that *completely* decentralized trade between agents generates an efficient outcome. Even if no single person or firm knows all of the steps involved in making a pencil, from extracting lead to efficiently cutting wood to figuring out demand for number 2 pencils in Saskatoon in August of 2016, shifts in the relative price of various resources is sufficient, under perfect competition, for every aspect

[1] On global supply chains and the theory of the firm, see Atalay, Hortascu and Syverson, American Economic Review (2014), and Ramondo, Rappaport and Ruhl, International Econonic Review (2014).

of the production process to be performed efficiently and independently by different tiny entities.

A lumberjack need only know the price of maple logs versus oak logs, knowing nothing about the use to which the logs will eventually be put or the reason why the price of maple went up or down on a given day, to efficiently choose whether to work cutting maple or oak. Indeed, there exists a theorem due to Leonid Hurwicz that prices in a competitive market are the *minimal* amount of information necessary to generate efficient trade; if information is decentralized, there is no way for a central planner to coordinate production efficiently using less information than that contained in a vector of prices for various goods.[2] In a sense, the fact that individuals acting independently in a market generate social efficiency is the fundamental theoretic argument for capitalism. An oft-quoted description of the market economy from the early 20th century, indeed, is that

> The normal economic system works itself. For its current operation it is under no central control, it needs no central survey. Over the whole range of human activity and human need, supply is adjusted to demand, and production to consumption, by a process that is automatic, elastic and responsive.[3]

Why, then, are firms strange? The polymath Herbert Simon[4] puts it nicely:

> A mythical visitor from Mars, not having been apprised of the centrality of markets and contracts, might find the new institutional economics rather astonishing. Suppose that it (the visitor I'll avoid the question of its sex) approaches the Earth from space, equipped with a telescope that reveals social structures. The firms reveal themselves, say, as solid green areas with faint interior contours marking out divisions and departments. Market transactions show as red lines connecting firms, forming a network in the spaces between them. Within firms (and perhaps even between them) the approaching visitor also sees pale blue lines, the lines of authority connecting bosses with various levels of workers. As our visitor looked more carefully at the scene beneath, it might see one of the green masses divide, as a firm divested itself of one of its divisions. Or it might see one green object gobble up another. At this distance, the departing golden parachutes would probably not be visible.

[2] See the Hurwicz chapter in the book Studies in Resource Allocation Processes (1977). Hurwicz won a Nobel for this and similar work.

[3] Sir Arthur Salter quoted in Coase, Economica (1937).

[4] Simon, "Organizations and Markets", Journal of Economic Perspectives (1991); Simon is another Nobel laureate.

> No matter whether our visitor approached the United States or the Soviet Union, urban China or the European Community, the greater part of the space below it would be within the green areas, for almost all of the inhabitants would be employees, hence inside the firm boundaries. Organizations would be the dominant feature of the landscape. A message sent back home, describing the scene, would speak of "large green areas interconnected by red lines." It would not likely speak of "a network of red lines connecting green spots."

So here is the puzzle. We quite literally organize the economic basis of our society around market transactions, largely due to a theoretical belief that decentralized markets are more efficient than Soviet-style command and control. And yet, a great deal of the economic activity in a market economy happens within firms, with decisions of whether to use input A or B, or whether a given employee should work on project C or D today, being determined by fiat by some manager, exactly as in a communist economy! D. H. Robertson as far back as 1928 referred to the economy as being made up of firms who are "islands of conscious power in oceans of unconsciousness like lumps of butter coagulating in buttermilk."[5] The existence of firms, these conscious islands in the market sea, is a strange state of affairs, and one that was questioned in the early 1930s by a young student at the London School of Economics, Ronald Coase. If we are to understand when firms should merge and when they should split, when antitrust authorities should interfere in firm relationships and when they should leave firms alone, when firms should outsource and when they should vertically integrate, then we surely need to understand why firms exist in the first place and what size they ought optimally be.

There are many such theories, broadly called *the theory of the firm*. In these notes, we will discuss five: the transaction cost theory, the extension of transaction costs due to Williamson, the property rights (or "residual control rights") theory, the resource-based theory of the firm, and the knowledge-based view which refines the resource-based view. We will briefly mention a handful of other theories – the agency theory, the economies of scale theory, and the argument against incomplete contracts. Which of these theories is "correct"? That question is ill-posed. Each theory can be useful in some circumstances for understanding firm behavior, and when analyzing the size of firms, it can be often be useful to draw on more than one perspective.

The Transaction Costs Approach

The transaction costs theory, due to Coase, is straightforward to state: firms perform actions within the firm because it can be costly to perform those

[5] From the 1928 book "Control of Industry" cited in Coase (1937) and very often thereafter.

actions in the marketplace.[6] Writing contracts with suppliers, suing them when things go wrong, searching for employees, and many other tasks are not free. When it is too expensive to perform these activities over and over, we will instead form a firm, a more stable set of relations between the means of production and output, where decision by fiat replaces the costs of transacting in a market.

Why, then, does the firm not continue to grow and grow until the economy consists of only one firm? First, there are legal restrictions like antitrust law. But more generally, as firms grow, the costs of organizing the firm itself - the costs of bureaucracy - increase as well. For Coase, one ought integrate activities into the firm as long as the marginal transaction cost of organizing and performing those activities outside the firm is higher than the marginal bureaucratic cost of bringing them inside the firm.

The idea of transaction costs is in some sense enticing. Why, for instance, do we see so many more conglomerates in developing countries compared to developed countries? Because horizontal and vertical integration is particularly important when transaction costs like enforcing contracts are very high, as might be the case in a country with a weaker legal system.

But transaction cost approaches suffer two major drawbacks. The first is that it is not at all clear what a "transaction cost" and a "bureaucratic cost" are, and why in particular they are of differential importance within the firm than in interfirm relationships, a question we will return to briefly when we discuss the property rights theory. The second problem is one that was actually introduced by Coase himself in a 1960 paper.

The famous Coase Theorem says the following. *If* there are no transaction costs, then it does not matter who owns what or what externalities exist: we should expect efficiency. Coase introduced this idea with his famous cattle-versus-crops example. A farmer wishes to grow crops, and a rancher wishes his cattle to roam where the crops grow. Should the rancher be liable for damage to the crops, or ought we to restrain the farmer from building a fence where the cattle wish to roam?

Coase points out that in some sense both parties are causally responsible for the externality. There is some socially efficient amount of cattle grazing and crop planting, and if a bargain can be reached costlessly - if there are no transaction costs - then there is some set of side payments where the rancher and the farmer are both better off than having the crops eaten or the cattle fenced. Further, this bargain is theoretically identical whether you give grazing

[6]The theory of the firm is from Coase, Economica (1937), whereas the broader idea of the Coase Theorem to be introduced shortly comes from Coase, Journal of Law and Economics (1960). Coase continued writing past his 100th birthday; a short summary of his ideas written by the author of the present notes following Coase's death can be found at http://www.voxeu.org/article/economic-ideas-ronald-coase.

rights to the cattle and force the farmer to pay for the right to fence and grow crops, or whether you give farming rights and force the rancher to pay for the right to roam his cattle.

This basic principle applies widely in law, where Coase had his largest impact. He cites a case where confectioner machines shake a doctor's office, making it impossible for the doctor to perform certain examinations. The court restricts the ability of the confectioner to use the machine. But Coase points out that if the value of the machine to the confectioner exceeds the harm of shaking to the doctor, then there is scope for a mutually beneficial side payment whereby the machine is used (at some level) and one or the other is compensated.

This is a very powerful idea. What it says is that we cannot ascribe the existence of firms solely to misaligned incentives between a producer and its suppliers, or between an owner and her employees. If it were possible to freely bargain, these parties would all prefer to take whatever action maximizes their joint payoff, then split the proceeds in some mutually agreeable way. This tells us that firms exist not to align incentives among different parties, but merely to remove transaction costs in bargaining and hence permit mutually beneficial bargains to appear among parties in an economic transaction.

The Williamsonian Extension of Transaction Costs

Oliver Williamson, another Nobel laureate, has done more than any thinker to try to operationalize Coase's transaction costs idea. In his extension, for firms to appear, there must be *appropriable quasi-rents* that appear as a result of the relationship-specific joint actions of two parties. Quasi-rents means that the joint actions generate more surplus than their joint cost, hence there is something to bargain over. Appropriable means that the surplus can be captured by one or both parties. When two firms generate AQRs from their joint relationship-specific actions, we say that a *fundamental transformation* has occurred. There is no fundamental transformation, and hence no AQRs, in perfect competition since if one supplier does not cooperate, you just buy from next best supplier at precisely the same price, and profit for every participant in these transactions is driven by competition down to zero. That is, in perfect competition world, there is no reason to care about specific relationships.

AQRs become important to the existence of firms when, for whatever reason, firms cannot contract for every eventuality (so called *unprogrammed adaptation*). The reason contracts may be incomplete, not specifying exactly what will happen to the joint investment in every state of the world, is unclear, but a common explanation is simply bounded rationality: no one has enough foresight to anticipate literally every potential state of the world that might matter for how two parties prefer to carry on their relationship. When unprogrammed adaptation occurs, there will be debate about what to do, and therefore some haggling. If a firm is integrated, when unprogrammed adap-

tation occurs, the CEO simply specifies what is to be done, and no haggling costs are incurred.

In Williamson's formulation, then, transaction costs have a very specific interpretation. In relationships where relationship-specific investments generate AQRs, and where contracts are incomplete, integration solves the problem of haggling over what to do following an unprogrammed adaptation. Therefore, integration will be common when bureaucracy costs are low, or when interactions are frequent and the value of specific investments is high, or when transactions are complex hence unprogrammed adaptations are common.

An interesting implication is that the only activities which are done inside the firm are the ones where there is frequent adjustment to unforeseen contingencies, hence it should not be surprising that activities within the firm appear highly regimented and bureaucratic compared to free-flowing, easy-going transactions in the marketplace. The *whole reason* certain activities are done inside the firm is *because* they necessitate a lot of management oversight following unforeseen contingencies; moving these activities from within the firm to outside the firm will simply result in frequently haggling, inefficient bargaining and contract rewrites with suppliers or other outside agents.

Haggling costs and bureaucratic costs can be made even more specific. First, it has been noted that unforeseen contingencies when firms are integrated do not simply result in a CEO deciding what to do by fiat. Rather, division managers and other interested parties will try to influence the decision of the CEO in accordance with their individual contracts, wasting resources on lobbying instead of producing something of value. High *influence costs* mitigate the benefits of integration, and may provide a check on the existence of very large firms. Second, an important type of haggling cost is *hold-up*. If I make a relationship-specific investment, and have not contracted on how I will be compensated by you tomorrow for making our relationship more valuable, then what stops you from simply thanking me for my kind investment, profiting from that investment, and leaving me with nothing? For instance,

> Suppose that an electricity generator has strong cost-based incentives to locate near a coal mine. Building a new generator involves a sunk investment. However, once this investment is sunk, the generator firm will find itself in a bilateral monopoly situation vis-à-vis the coal mine. The electricity generator can sign a contract with the coal mine before investing in a generator. However, after the generator is built, it becomes a sunk cost. The coal mine will have an incentive to seek some reinterpretation or renegotiation of the contract that would allow it to receive a higher price for coal. If the electricity generator anticipates this "hold-up" situation, it may simply decide not to make the investment.[7]

[7]Example due to Aghion and Holden, Journal of Economic Perspectives (2011)

The Property Rights Theory

The fundamental problem with transaction costs theories is figuring out why exactly haggling or bureaucracy costs are different inside the firm than in interfirm relationships. Firms, after all, are not dictatorships.

> It is common to see the firm characterized by the power to settle issues by fiat, by authority, or by disciplinary action superior to that available in the conventional market. This is delusion. The firm does not own all its inputs. It has no power of fiat, no authority, no disciplinary action any different in the slightest degree from ordinary market contracting between any two people. I can "punish" you only by withholding future business or by seeking redress in the courts for any failure to honor our exchange agreement. This is exactly all that any employer can do. He can fire or sue, just as I can fire my grocer by stopping purchases from him or sue him for delivering faulty products.[8]

Whether my supplier doesn't do what I want them to do, or my employee doesn't do what I want her to do, there is no difference in what action I can take as an owner: I can fire the supplier or the employee, or I can keep them on. The big difference between a supplier and an employee is that when the supplier is fired, they get to keep their machines and other assets, but when the employee is fired, my firm gets to keep the machines. This is the fundamental idea of the property rights, or *residual control rights*, theory of Grossman, Hart and Moore.[9]

In the property rights theory, the fundamental question is not "Should a firm make or buy?", but rather, "What is a firm?" For Grossman and Hart, the fundamental feature of a firm is that, following some unexpected event, whatever "unprogrammed adaptation" happens will have as its starting point that the firm owns certain assets and can use them as it wishes. The Coase Theorem seems to suggest that it doesn't matter who owns an asset, but this will not hold under two conditions. First, there must be incomplete contracts, or states of the world where we cannot simply contract about will be done with the assets we are using in our relationship. Second, there must be investments we would make which change the value of these assets, hence situations where the party making the investment in an asset they don't control is worried about hold-up in these unexpected states of the world. A supplier will consider locating its factory next to the car manufacturer it supplies only if it is

[8] Alchian and Demsetz, American Economic Review (1972), an incredibly influential and well-written paper about what firms are or could be.

[9] Grossman and Hart, Journal of Political Economy (1986) and Hart and Moore, Journal of Political Economy (1990) are the seminal results. The Grossman is Sandy Grossman. When someone asks, "if these economic theorists are so smart, why aren't they rich?", you can point them to Sandy Grossman, who left academia to run QFS Asset Management, a hedge fund that has earned him on the order of a billion dollars.

sure that the manufacturer will actually buy the input once this new factory has been built. A car manufacturer will only build an assembly line specific to that supplier's part if it sure that the supplier will be able to provide the needed inputs tomorrow. Whether either or both of these parties will invest in the factory or the supply chain depends on who gets to decide whether to use the input, and how much each firm will earn, when something unexpected like the 2008 recession happens. This is easiest to see in an example.[10]

Let a supplier S produce a widget which is used in the production of a sprocket by producer X. The widget costs \$16 to produce unless the supplier invests \$5 in a cost-reducing technology, after which the widget costs \$10 to produce. Customers value the sprocket at \$32 unless the producer X invests \$5 in quality improvements, in which case consumers value the sprocket at \$40. By assumption, only S can make the cost-reducing investment and only X can make the value-enhancing investment. The most profit that can be made jointly is \$40-10-5-5=\$20, which involves both investments being made. If complete contracts are possible, they will simply agree to contract on both making the investment, or on sharing costs, and an efficient relationship is possible even if the firms do not integrate.

What if it is not possible, for whatever reason, to write a contract which specifies a cost sharing rule or a required investment? In that case, after the investments are made, S and X will bargain over what price X will pay for the widget. One assumption might be that whatever surplus is generated from a transaction, the bargain gives half the surplus to the supplier S and half to the producer X.[11] If X invests, she raises the consumer value, and hence the surplus, by \$8 (from \$32 to \$40). Since under the bargaining assumption, she only earns half of that, \$4, in additional profit at a cost to herself of \$5, she will not make the quality enhancing investment. Likewise, if S invests by himself, he increases joint profits \$6 by reducing costs from \$16 to \$10, at a cost to himself of \$5, but his own profits only increase by \$3 under the bargaining assumption, and hence he will also not invest. If firms are not integrated, neither is willing to make a productive investment.

What happens under different forms of integration, which under the property rights theory means different ownership of means of production following uncontracted events? If supplier S owns the machines that can produce the widget and sprocket (with X remaining unique in that only it can make the quality-improving investment) then S finds it worthwhile to make the cost-reducing investment, improving his profits by \$6. However, S can still not compel X to make the value-enhancing investment, nor guarantee that X won't be "held up" if it makes such an investment, so total profit generated is 0 for X and \$32-10-5=\$17 for S.

[10]The example is drawn nearly exactly from Aghion and Holden, Journal of Economic Perspectives (2001).

[11]In this example, that would be the so-called "Nash bargain", but the exact bargaining rule does not matter.

What if the firms instead backward-integrate, with producer X owning the machines that make the widget and sprocket, but S remaining the only one capable of making the cost-reducing investment? Analogous to the previous paragraph, now X will find it worthwhile to make the value-improving investment, but S will not make the cost-reducing investment. Profits will be 0 for S and \$40-16-5=\$19 for X. Note that total profits are higher when X instead of S has the residual control rights.

This simple example gives the two major insights of the property rights theory. First, if contracts are incomplete, it matters who owns assets that are useful in production. Second, assets should be owned by the person whose investments in those assets is most sensitive to the ownership structure. Independent decisionmakers, when deciding to make investments, only take into account the profits *they themselves* earn from that investment, which can either efficiently be described in a contract via the Coase Theorem, or which simply affects the terms of a future bargaining agreement if contracts are incomplete. In the example, when S owns the widget producing machine and X owns the sprocket producing machine, and investments cannot be contracted, neither firm is willing to make valued investments. But it is the value-enhancing investment by X which increases the value of the final product by \$8 at a cost of \$5 that is really important to incentivize, compared to the cost-reducing investment by S which decreases the cost of the final product by \$6 at a cost of \$5. Hence optimal firm structure involves X having residual control over the means of production.

The problem of residual control rights comes up frequently in firm and supplier or firm and worker interaction. As an example, long distance truckers face two important problems in their contracts. First, if the truck is not owned by the driver, then the driver may not adequately care for the truck in terms of maintenance or safe driving; just consider how hard the average taxi driver works the brakes and gearbox! Second, if the truck *is* owned by the driver, the trucking company may not care enough about efficiently utilizing that truck by, for instance, ensuring that the truck always has a full return load after a long drive. A pair of economists showed that when monitoring computers became feasible to install in trucks, allowing trucking companies to carefully watch how safe their drivers operated, trucking companies became more likely to own the previously driver-owned trucks.[12] Why? The most important, hard-to-contract-upon investment became ensuring full truck utilization, because the problem of ensuring safe driving could now be included in a contract. Hence, in line with the property rights theory, trucks became more efficiently owned by the trucking company.

[12]Baker and Hubbard, Quarterly Journal of Economics (2004).

The Resource Based View

"Firm resources include all assets, capabilities, organizational processes, firm attributes, information, knowledge, etc., controlled by a firm that enable the firm to conceive of an implement strategies that improve its efficiency."[13] Imagine that firms - a collection of people and assets, but also of intangibles like a company culture and a set of brands - possesses certain resources from among the above set. If these resources differ between different firms, and if they are "immobile", in the sense that one cannot simply purchase those resources, then firms may exist as a bundle of unique, advantage-granting resources. In transaction cost and residual control rights theory, the important idea is that unforeseen contingencies introduce inefficiencies which cannot be handled by the Coase Theorem. There is no role in those theories for "corporate culture" or other firm resources which exist only when the firm is in a specific form.

It goes without saying that for firm resources to be a reason why firms exist, the resources in question must be rare and not easily substituted for, and must also be valuable. For instance, "good management" is surely a valuable resource, but if I can just work with suppliers who are also properly managed, then the fact that my firm is managed well does not tell me anything about what activities I should pursue and what I shouldn't. The deeper question is why certain advantages are not imitable. Three common arguments are that a resource is history dependent, that the reason the resource is useful is casually ambiguous, and that the resource operates in a socially complex way.

History dependence means that the specific time and place a firm originates or grows offers conditions permitting certain organizational features, and once that time has passed, it is impossible to recreate those features in other firms. For instance, in some industries network effects matter. Once Facebook has built up a large social network, it is very difficult for an alternative firm to enter successfully even if it has a slightly better technology. No diamond extraction firm will ever have the concessions DeBeers possesses as a result of the colonial history under which it was founded. An engineering firm founded right after large layoffs at Research in Motion will be able to hire better Ontario engineers for a given wage than a firm founded five years earlier. Once the resource is acquired at some unique and specific point in time, it should be exploited fully.

Casual ambiguity means that, even though if a rival firm is very productive, other firms may not totally understand how. Why was Bell Labs able to produce so many fundamental scientific breakthroughs? Was it "good culture"? Was it specific managers? Was it blind luck? How would one go about replicating what they are doing? Causal ambiguity in particular requires that even *successful* firms are unable to specify exactly why they are so successful - if they could do so, a rival could simply hire away some manager that knows the

[13] Barney, Journal of Management (1991) summarizes well the resource-based view, a research program beginning largely with Penrose's 1959 book "The Theory of the Growth of the Firm".

secret sauce.[14] If causal ambiguity explains why certain firms "just are" good at performing some set of tasks efficiently, then even the successful firm needs to take care when expanding production or integrating into related activities. Without knowing why the firm is successful, it is very difficult to know whether they will continue to be successful as they grow or differentiate.

Social complexity is related to causal ambiguity. Even in some cases where we know how a given resource affects firm productivity, the exact circumstances under which that resource is beneficial may be tough to specify. For instance, we may know that information technology is profitably used by the average firm in a given industry, but the exact benefit of IT may depend on aspects of firm organization or corporate culture or communication chains which are difficult to understand or state. When that is the case, we may say the firm has a resource advantage in "using IT", and the firm may want to integrate with suppliers where IT is important.

With any resource argument, great care needs to be taken. In particular, if a firm is good at doing something, or possesses some beneficial "resource" or culture, this is *only* relevant to the optimal size and structure of the firm when that resource cannot be imitated, when the resource is valuable, and when we understand the particular reason imitability is hard. It is the height of hubris to argue that one's firm ought do more simply because it is "better" than other firms. Beyond the potential for hubris to outweigh analysis, resource-based theories of the firm tend to be stated verbally, and tend to be underexplored formally, hence the precise relation of unique, non-imitable resources to the size of the firm is not as well understood as other theories.

The Knowledge Based View

How do firms develop nonreplicable resources? Some are real - such as DeBeers links to government in diamond-producing nations. Others, however, are dynamically acquired as a firm faces and solves problems over time.[15] It is not the case that individuals alone know how to solve past and future problems, but rather that routines, practices, norms, and hierarchies develop over time which in a sense allow *the firm* the learn. The knowledge based view of the firm explicitly built off of the resource based view to argue that firms have a special ability to learn, combine, and organize new knowledge into difficult to replicate resources, making them a valuable mode of organizing above and beyond pure market forces.

Knowledge in the firm is composed of know-how and information. Know-how

[14] The work of Chad Syverson discusses many empirical examples of productivity differences across firms where the cause of the difference is difficult to explain. Matouschek and Callender's working paper "Managing on Rugged Landscapes" provides a formal model.

[15] Kogut and Zander, Organization Science (1992) provides the foundation of the knowledge based view.

is "the accumulated practical skill or expertise that allows one to do something smoothly and efficiently".[16] Information, on the other hand, can be easily be communicated between individuals once rules for deciphering the information are known. Firms are useful in that they provide a forum for individuals to develop and recombine these two types of knowledge into useful and difficult-to-replicate, on-the-job routines. Since these socially and informationally complex routines are not entirely owned by any one individual, they are able to be maintained, altered, and transmitted over time under the aegis of the firm. This process of maintaining and updating routines serves to organize the firm's production process, allows the firm to innovate, and provides natural advantages over purely market-based organizing.

At a micro level, this social process of individuals recombining and storing valuable know-how and information in routines helps firms address different problems they face.[17] As in other theories of the firm, firms should outsource simple problems to the cheapest vendor in the market ("Should I use cedar or pine for this table?"). For problems of moderate difficulty, a straightforward directional search led by a traditional firm hierarchy quickly drives the firm to a satisfactory solution while curtailing unnecessary exploration. For the most complex search processes, an exploratory heuristic search conducted by group consensus tends to produce more innovative and useful solutions by fully engaging the firm's exploratory capabilities. That is, firms exist to solve the kind of difficult problems which frequently arrive in novel, slightly different variations.

This view suggests that the semi-permanence of the firm, in contrast to market transactions, permit a unique set of routines to be built up. If future problems a firm faces are efficiently diagnosed and solved with similar routines to those which were successful in the past, then those routines become a strategic advantage for the firm. Since the development of knowledge is dynamic, firms want to ensure that outsourcing decisions today do not limit the firm's ability to solve unforeseen problems tomorrow.

That is, the knowledge based view essentially says that "the ability to solve a particular class of problem" is the most important resource a firm cannot buy in the market. You can buy products, or physical inputs, efficiently in the market. You can sometimes buy information or ideas.[18] You may even be able to buy complex sets of transactions when you purchase something at the end of a supply chain. But it is not clear how good markets are at selling the organizational routines which have proved useful to your firm in your industry

[16]This definition comes from Eric Von Hippel's book The Sources of Innovation, which explores how companies source innovative ideas from the market, customers, and partners. See also Joel Mokyr's "The Gifts of Athena" on prescriptive versus propositional knowledge in economic history.

[17]Nickerson and Zenger, Organization Science (2004) provides insight into this "problem solving perspective".

[18]See Kenneth Arrow, "Economic Welfare and the Allocation of Resources for Invention" (1962) for some limitations on this idea, though!

with your specific problems. Hence, "firms learn".

Empirically, firms' knowledge and routines appear to affect their boundary choices. Integrated semiconductor firms tend to perform better in product categories that require complex knowledge to solve problems, while non-integrated firms tend to perform better when producing products that require workers to solve straightforward problems. Similar arguments have been made among pharmaceutical research scientists, who tend to produce more significant patents when they are part of a large, diversified firm than a smaller, more specialized firm.[19]

While these results present a positive view of firm-level knowledge, knowledge diversity is often expensive. The market has much deeper and broader expertise than the collective employees of any one firm. Firms do not know precisely what problems they will face in the future, hence do not know what type of problem-solving expertise to prioritize. In many fast-moving industries, firms that tend to spurn external knowledge can quickly lose their edge and lock themselves into low performing path dependent outcomes.[20] Worried about having their knowledge based strategic advantage stolen, firms sometimes avoid useful partnerships that can help fill knowledge gaps.[21]

The knowledge based view, like the resource based view, has an important insight, but one that is often misunderstood. Just as the only resources which matter for the size of the firm are those which are nonreplicable, immobile, and known to be useful to the particular markets a firm will move into, the only way knowledge matters for the size of the firm is when it satisfies those same features. That "routinized problem solving in a particular domain" can be hard to replicate or buy should be clear. That firms are going to face the same types of problems in the future, or have the ability to know when their particular routines are better suited for solving future problems than the collective knowledge of the market, is much less obvious.

Alternative Arguments

There have been many other arguments for why firms exist and why they are the size they are. These brief notes are not sufficient to cover every explanation, but briefly:

[19] See Macher (Management Sciences, 2006) on semiconductors, and Henderson and Cockburn (PNAS 1996) on pharma.

[20] Rosenkopf and Nerkar (Variations in Organization Science 1999).

[21] Oxley and Wada, Management Science (2009).

Increasing Returns to Scale

In many industries, there are fixed costs like R D or the construction of a factory. This large scale production requires assets more expensive than those which can be bought by any single person, and requires the coordination of large numbers of workers. Hence the optimal size of a firm needs to be bigger than one person. This was an argument for firms in the early part of the 20th century, but it is now generally seen as wrongheaded. First, the size of firms and the size of the minimum efficient scale of a factory are often completely different: firms own many factories in some cases. Second, economies of scale do not in any way imply that every piece of that scale economy need be owned and controlled centrally. Consider Uber: the profitability of the company certainly comes from the scale of the network, yet the cars driven are not owned by Uber, nor is the schedule of the drivers controlled by Uber.

The Agency View

The "principal-agent problem" is that employees or suppliers need to be monitored because they possess private information about their effort or their skill. It is uncontroversial that large parts of how firms are organized once they exist involves solutions to agency problems, such as whether to pay bonuses versus fixed salaries, how to use hierarchies of management to control and transmit information, or whether to rely on sole suppliers or whether to rotate to the one offering the best deal at a given time. Some of these contracts are not even formally written down, but rather are so-called relational contracts, where workers or divisions cooperate today because they expect to be paid or rewarded in the future.

The agency theory of the firm views a firm as a nexus of contracts among the firm and its members.[22] These long-term contracts, whether legal or relational, help solve agency problems which are difficult to solve in one-time or short-term relationships. The major problem with the agency theory is that it is not totally clear why one cannot use legal or relational contracts outside the firm as well as inside, or why these contracts inside the firm are in some sense more efficient.

The most interesting such theories, due to Bengt Holmstrom and Paul Milgrom, show that under a problem called multitasking, it can be difficult to provide incentives when people have more than one goal which trades off with another goal.[23] In particular, suppliers care about their performance today and

[22] Jensen and Meckling, Journal of Financial Economics (1976) is the most famous statement of the agency theory, and Milgrom and Roberts, Canadian Journal of Economics (1988) provides an accurate description of exactly when agency problems or risk-sharing can get past the Hurwicz criterion discussed in the first section.

[23] Holmstrom and Milgrom, Journal of Law, Economics and Organization (1991) and Holmstrom and Milgrom, American Economic Review (1994).

how they are paid, as well as about the value of their assets which they will not want to run down. Holmstrom and Milgrom have shown that if supplier effort is not observable or contractible, there sometimes do not exist contracts under multiple goals such as these which generate efficiency. When incentive problems of this type are particularly severe, a firm must integrate so that the ownership of assets and their long-term value is no longer a concern of workers, and hence workers can be more efficiently incentivized. It will be noted that property rights theories are in some ways a more refined version of a general agency theory, in that they specify specific circumstances where asset ownership distorts not only investment or effort.[24]

The Maskin and Tirole Critique

Recall that many of the theories of the firm we have discussed rely on the existence of incomplete contracts. If all events in the future can be contracted upon, the Coase theorem applies, and hence the transaction cost, Williamsonian and property rights theories can no longer explain why firms do things internally instead of just contracting everything out.

In a seminal paper, Eric Maskin and Jean Tirole (both of whom are now Nobel winners) extended a theory called subgame implementation to show that even if firms in a relationship may be unaware of some future contingencies, they can write a contract that *as if* they were aware.[25] Showing precisely how to write this contract is beyond the level of these notes, but the essential point is that instead of writing a contract that states "If X happens, we will change what we do in way Y", parties can write a contract that says "If something we don't expect happens, and the maximal value of the relationship in the future changes to Z, we will pay each other such that you get some portion of Z and I get the rest."

The only reason that it matters that parties in a relationship do not know about some future state is that they will still care about the payoff they get in that state. But this means that a complete contract can, instead of telling each of us what to do when it rains, when it snows, and when it is sunny, just tell us that whatever the weather is tomorrow, when working together earns us thirty dollars jointly, then I get ten and you get twenty. The difficulty is that the future profitability of a relationship is often not publicly verifiable, hence how can a contract of this type be legally enforced? This is the clever trick of Maskin and Tirole: it is possible to write a legally verifiable contract where each party finds it in their interest to truthfully reveal their private

[24] Gibbons' paper "Four formal(izable) theories of the firm" in the Journal of Economic Behavior and Organization (2005) provides a nice formal reckoning of agency theories, Williamsonian transaction costs, and the property rights theory. It is somewhat technical.

[25] Maskin and Tirole, Review of Economic Studies (1999) gives the theory, drawing on Moore and Repullo (1988) in Econometrica. The latter paper in particular is technically challenging.

knowledge of the value of the relationship. The trick turns out to be rather convoluted, and it is still up for debate how realistic Maskin-Tirole contracts are in actually-existing contracts. Nonetheless, any time a defence is made of firms or their size on the basis of incomplete contracting, it now must be explained what keeps parties from contracting on the split of future profits instead of on unforeseen future contingencies.

Legal Theories

Of less interest to economists, though surely important, is that firms are *legal entities*. A contract between two firms, and a contract between a firm and its employees, or a contract between two individuals, is *not* treated symmetrically in the courts. Indeed, there are very specific laws (generically referred to as "forbearance") about how different parts of a single firm or organization can sue each other for failing to uphold some agreement. While it is beyond doubt that some aspects of firm organization have to do with differential tax treatment of suppliers versus integrated divisions, or of the legal possibilities of contracts using prices for a downstream buyer versus internal (or "transfer") prices to a downstream division of the same firm, these issues are well beyond the purview of notes of this type. We should nonetheless see that firms - organizations in a market economy that exist with some permanence and which perform some market activities within the firm's boundaries - exist across many different types of legal systems, and hence it can fairly be said that the *fundamental* reasons for why firms exist and why they are not of infinite size depend on more than simply idiosyncratic laws.