1. Theory of the Firm

1.1 Nature of the Firm

What is a Firm? How does it relate to Market and Hierarchies, Incomplete Contracts, Property Rights?

The theory of the firm is a crucial topic in economics as well as an integral part of industrial organisation and law. The study of the firm is preliminary and complementary to the study of markets. It coincides with monopoly. We enquire into its nature and objectives.

We look at different definitions of a firm, focusing on what determines the size of the firm.

Next we will take firms as well defined entities and investigate how firms behave. We will discuss the profit-maximisation hypothesis to check whether it is actually the main objective of the firm and if not whether it can still be considered a useful guide for predicting strategic interactions. We will examine in detail the various mechanisms that put constraints on managerial discretion.

From the next topic on, to simplify matters, we will then treat firm as single decision makers that maximise profit. Problems of managerial control by shareholders or capital markets will be ignored. A fair treatment of agency problems would require another course and here we will just mentioned these issues as far as our topic is concerned.

1.2 Objectives of the Firm and Managerial Incentives

Here we will examine issues as: Separation of Ownership and Control, Agency Problems, Constraints on Managerial Discretion.

Let us start with a few simple question: What is a Firm?
Which activities are carried out within the firm and which are carried out across markets?
What determines the boundaries of the firm?

Spectrum of possibilities for carrying out activities:

\[
\text{Spot market} \quad \leftrightarrow \quad \text{Contracts} \quad \rightarrow \quad \text{Firm}
\]

For instance consider the electricity supply industry.

Before being privatised, the generation and the transmission of bulk electricity were carried out by one public corporation. The privatisation of this utility has brought a split up between generation and transmission. This shows as we can create a market in alternative with a firm to solve economic efficiency issues.

In dealing with such questions we are mainly concerned with efficiency reasons for integration or disintegration. Efficiency motives are associated with the cost minimising organisation of economic activity. We will consider three views of the firm as a cost-minimising devise.

1) Technological view
Firm = synergy between different units to exploit economies of scale and scope.
This leads to the discussion of the various form of cost and demand complementarities and the U form of internal organisation.

2) Contractual view
firm = long run arrangements of its units. This brings us to the theory of Williamson, which warns us from the danger of hold up or opportunism (the confiscation of the gains associated with one
party's investment by the other party). A long term contract must ex post guarantee the parties a fair return in order to ex ante encourage specific investment.

3) Incomplete-contracting view
These two views have little to do with the legal definition of a firm. The third view emphasises that firms and contracts are rather different models of governance.
Firm = way of specifying what is to be done in the event of contingencies not foreseen in a contract. It starts from the idea that contracts are necessarily incomplete because some contingencies are unforeseeable or there are too many of them to specify in advance.

Let us now examine in detail these alternative views.

1) Technological view
Firm = sinergy between different units to exploit economies of scale and scope. Very roughly, economies of scale exist when the production cost of a single product decreases with the number of units produced; economies of scope are cost saving externalities between product lines (the production of a given good reduces the production cost of another good).
One of the main determinants of the size of a firm is the extent to which it can exploit economies of scale and scope, the idea being that increases in production levels allow the use of more efficient techniques.
Economies of scale related to a single product are called product specific economies. Concerning the volume of a single product a plant with a higher number of machines can sustain a flow of output proportionally more than one with a small number (i.e. the random breakdown of a machine has less impact on output, as the flow of production can be reallocated to other machines). Similarly a firm serving several markets faces less uncertainty than a firm serving these market independently and can save on peak load investments. The very same reasoning applies to multiproduct firm sharing the same production techniques.

The gathering of activities may be also related not to production in a narrow sense but to all other functions such as Finance, Marketing etc. Such a gathering avoids duplication of fixed costs or at least decrease costs on the average.

Let see the formalisation of return to scale for a single product firm.
Let $C(q)$ denote a firm total cost of producing $q$ units of output, where $q = \sum q_i$. Subadditivity means that it costs less to produce the various output together than to produce them separately.

$\Sigma C(q_i) > C(\Sigma q_i)$

(An industry is said to be a natural monopoly if the cost function is subadditive over the relevant range of output. A natural monopoly arises when the regulator prefers production by a single firm. The regulator who knows the cost function perfectly has no incentive to have several firms produce the output when the aggregate output can be produced more cheaply by a single firm. Under incomplete information about the cost function there is a trade off between returns to scale and the extraction of the relevant information through competition).

Decreasing average costs imply subadditivity.

$C(q_j)/q_i > C(q)/q \quad \Sigma C(q_i) > \Sigma q_i C(q)/q = C(q)$

The reverse is not true, in fact in Fig.2 we have subadditivity even after MES, till the AC of additional quantity is not equal between the incumbent and a new firm.
For a multiproduct firm the same definition of subadditivity applies if we consider \( q \) as the production vector \( q = (q_1, q_2, \ldots, q_m) \).
From the last formulation we can also define economies of scale: if we let \( q_1 \) and \( q_2 \) be the quantities of two different goods.

\[
C(q_1, 0) + C(0, q_2) > C(q_1, q_2)
\]

where \( C(q_1, 0) \) and \( C(0, q_2) \) are called stand-alone costs.

A railroad company producing both passenger travel and freight transport is technologically more efficient than two companies specialised in one of the two productions.

In sum the technological view aims at defining the size of a firm. The limit to the size stems from the fact that AC rise at high output (as in fig. 2).

We are now able to assess this view as a theory of the firm.

First, notice how returns to scale have their limits: machines or functional divisions can be pulled together only if they are not employed to their capacities. Savings in peak load capacity become smaller and smaller as the size of the firm grows. Moreover there are many factors such as human capital or managerial talent that cannot be duplicated as the firm expands.

But why can't one firm set up independent divisions to produce the high output in order to avoid diseconomies of scale (for instance if producing twice the output were to cost more than producing the output separately a firm could set up two independent divisions. Its production technology would exhibit no diseconomies of scale)? Why there should be limits to firm size stemming from diseconomies of scale?

On the other side, applying the same reasoning why should economies of scale be exploited within the firm and not through contracting between separated divisions?

Why there cannot be something in the middle between the spot market and the firm? Electric utilities do engage in arrangements to pool electric power.

These are two reasons why this view does not constitute a self-contained theory of firm size.

2) Contractual view

firm = long run arrangements of its units (set of contracts).
We already mentioned some reasons why some units might want to merge or coordinate their activities through a "static" contract. We now study the problems associated with long run relationships.

Amongst the hazards associated to long run relationships are the needs to invest \( \text{ex ante} \) in idiosyncratic or specific investment with the prospects of future trading rather than current trading. A supplier must design equipment the characteristic of which are specific to a buyer's order. Site specificity is associated with the gain in trading with a nearby supplier or buyer, specific investment in human capital involve the learning of processes and team work. It is important that the gains from trade be divided properly in order to induce the efficient amount of specific investment \( \text{ex ante} \).

The crucial aspect of specific investment is that even though the supplier and the buyer may select each other \( \text{ex ante} \) in a pool of competitive suppliers and buyers they end up in a bilateral monopoly \( \text{ex post} \) in that they have an incentive to trade between themselves rather than with outside parties. Under bilateral monopoly each party wants to appropriate the common surplus (the gains from trading) thus jeopardising the efficient realisation of trade \( \text{ex post} \).

Hence there are two problems (inefficiency) associated with repeated long run spot market transactions:

- \( \text{ex post problem: volume of trade} \)
- \( \text{ex ante investment} \)

Suppose there are 2 periods \( t=1,2 \)

A supplier and a buyer may or may not sign a contract in period 1. To focus on the \( \text{ex post} \) issue we ignore the first period investment. The volume of trade is either 0 or 1 (1 unit of the good). The value of the good to the buyer is \( v \), its production cost to the supplier is \( c \). The gain from trade is \( v-c \), divided between the buyer \( v-p \) and the supplier \( p-c \), where \( p \) is the trading price. Suppose at time 1 no contract is signed: some bargaining occurs in period 2 to determine whether to trade and at what price. If \( v \) and \( c \) are common knowledge we expect the efficient amount of trade to occur. If \( v \) is greater than \( c \) any trading at a price \( p \in (c, v) \) that leaves a surplus to both is preferred by both parties to no trade at all (in which case the surplus is equal to zero for both).

If \( v \) is less than \( c \) one of them makes a negative surplus hence is better off refusing to trade. More generally under perfect information bargaining is efficient (version of Coase theorem).

Under asymmetric information a wrong allocation of the bargaining power can lead to inefficiency. Suppose \( c \) is common knowledge, but \( v \) is known only to the buyer. The supplier's belief is represented by a uniform distribution \( F(v) \) on \([0, 1]\).

Suppose the supplier has all bargaining power, so that he can make a take it or leave it offer to the buyer: the buyer accepts only and only if \( v \) is greater or equal to \( p \). The probability of trade is \( 1-F(p) = 1-p \).

The supplier expected profit is

\[
\Pi = (p-c) \cdot [1-F(p)] = (p-c) \cdot (1-p)
\]

FOC \( [1-F(p)] - (p-c)f(p) = (1-p)- (p-c) \)

Increasing the price gives rise to a trade off between an increase profit with probability \([1-F(p)]\) and a loss of trade which give rise to a net loss of \( (p-c)f(p) \).

Familiar monopoly pricing formula for the demand curve \( q=D(p)=1-F(p)=1-p \)

\[
p^* = (1+c)/2
\]
The volume of trade is suboptimal, since \( p^* \) is greater than \( c \).
Hence ex post bargaining does not lead to the efficient volume of trade.
A solution might involve giving the informed party the right to decide the price. \( p = c \) that leaves the
seller indifferent between accepting or refusing to trade.
It is also superior to set in advance a fixed price contract determining in advance the amount of
trade and the price (this works also in the case of bilateral asymmetric information).
Let us now focus on the ex ante problem.
Suppose the production cost is a function of the supplier's investment: i.e. \( c(I) \) with \( c' < 0 \) (investment
lowers cost).
Let the price be determined by ex post bargaining \( p = \frac{[c(I)+v]}{2} \) i.e. according to the Nash
bargaining solution, so that \( v-p = p-c(I) \).
The supplier's profit is

\[
\Pi = p-c(I) - I = \frac{v}{2} - \frac{c(I)}{2} - I
\]

\[
\text{FOC} \quad c'(I) = -2 \quad \text{Privately optimal investment}
\]

The socially optimal investment

\[
\max v-c(I) - I
\]

\[
\text{FOC} \quad c'(I) = -1
\]

Hence bargaining yields underinvestment in specific assets.
The problem is that the party investing does not capture all the cost saving generated by his
investment. The other party can use the threat of not trading to appropriate some of his savings.
As the threat of termination (no trade) is often the vehicle that allows parties to appropriate a share
of common gains from trade in bargaining it is desirable to contract ex ante (to trade at a given
price no matter what) to impose penalties for breach. As in the previous problem the party investing
should be given the authority over the price.
Bilateral investment or asymmetric information about \( v \) calls for a different solution: specific
investment, valuation and cost are commonly observable but not verifiable. Auditing prevents
inefficiencies associated with incomplete information. Integrated firms might be more easily to
auditing, though integration per se is unlikely to change information.
A limitation of long run is their lack of flexibility: the loss of advantageous outside opportunities. There is also a risk of collusion.

In sum long term contracts are likely to be more efficient than short term contract (relying on the spot market) where is feasible and not too costly; that is, with a high degree of specific investment and in the absence of outside opportunities.

Joskow has analysed contracts between coal mines and electric utilities in the USA. In the East there are a large number of suppliers and buyers transportation facilities are competitive. The hazards of expropriating specific investment are small and actually spot market transactions were very relevant. In the West there is less competition between mines and also for transportation; as a consequence contracts are of much longer duration.

3) Incomplete-contracting view
Firm = way of specifying what is to be done in the event of contingencies not foreseen in a contract. It starts from the idea that contracts are necessarily incomplete because of the presence of transaction costs. Transaction costs might occur at the contracting date (i.e. some contingencies are unforeseeable or there are too many of them to specify in advance) or later (i.e. costs of monitoring the contract or legal costs of enforcing contracts).

Intermediate form of contracting (between no contract i.e. relying on ex-post bargaining and a complete contract) can save on transaction costs relative to complete contracts avoiding the perverse effects of unconstrained bargaining.

A third party might be employed in the settling of disputes. This solution is known as arbitration. The arbitrator's ex post trading and transfer decision must yield ex post the efficient volume of trade and encourage ex ante the right amount of specific investments. An arbitrator must be able to learn and understand the situation at relatively low cost and must be independent. The first condition may restrict the use of external arbitrators. The second requires the internal arbitrator to be trusted for settling disputes fairly.

The optimal arrangement is the one that best protects the specific investments. In the absence of complete contracts ownership is a second best solution. The power to fill unspecified contingency (authority) may be given to one of the concerned parties. Authority puts the party that has it in a better bargaining position. Grossman and Hart trace authority to the ownership of some physical assets (though other authors Kreps extend it to intangible assets). Integration is defined as the allocation of the residual rights of control to one of the parties. Vertical integration, like long term contracts, is more likely the more specific the investments are. This may make it difficult to distinguish between the two solutions empirically. The main conclusion of the analysis is that integration is more likely when transaction costs are high.

Empirically it has been found that relationships between firms tend to be more informal than was predicted by the theory. Efficiency is then sustained by the firm's reputation. A firms that cheat at some date runs the risks of losing future profitable deals with its partner. Reputation allows a firm to save on the cost of writing the contract or of distributing the authority. Informality exposes the firm to the threat of opportunism.

Internal organisation
Historically, the technological view dominated in the early 20s which witness the emergence of large multi-functional firms organised in an U form. The U form can be seen as an attempt to exploit potential economies of scale.

U form
The technologically rational organisation is the unitary form with its specialization by function (manufacturing, sales, finance and engineering). The collapse of this type of organisation is mainly due to the loss of control by the top management. There are basically two methods that can be employed to control the functional divisions. The first method is to supervise directly the divisions and can be employed only if the firm is small. Another method involves the use of incentive schemes based on the output (i.e. rewarding each functional division for good performance). Each division's performance may be hard to measure. Technological non-separability and the concomitant problem of identifying individual performance create a problem of moral hazard in teams. The difficulty of measuring individual performance is that it requires careful supervision and a good understanding of each functional division by top managers. The latter become easily overloaded as the firm grows. And indeed the U-form collapsed with the expansion of firms. It was replaced by the M-form (multidivisional form) which resembles a collection of scaled down U-form structures.

In the M-form divisions are organised so that their performance can be reliably measured. To achieve this aim the divisions (A, B and C) are defined by product, brand, or geographic lines are fairly authonomous. The role of the top management is then reduced to advise, auditing and allocating resources between the competing divisions. Within a division, by contrast the supervisory mode is more prevalent and allows some assessment of the relative contributions of functional subdivisions.

**M form**

The switch from the U-form to M-form was partly triggered by a changing environment. A more recent example of this phenomenon is the matrix organisation which try to promote horizontal communication and decision-making. The need for joint decision making was due to the gradual shortening of the life cycle of products (in an industry where products become obsolete within a year firms must be particularly quick in finding market segment). Some discontent has been
recorded concerning the high number of authority conflicts between the joint decision-makers. We need more formal agency-related work modelling these organisational forms.

1.2 Objectives of the Firm and Managerial Incentives
Separation of Ownership and Control, Agency Problems, Constraints on Managerial Discretion.

The shareholders of a firm are the claimants for its revenue net of costs. If they were able to run the firm they would minimise costs and maximise profit. The emergence of different objectives (i.e. non profit maximisation) is mainly associated with the separation of ownership and control. According to the principal agent theory the concern for size and growth (rather than profit) is due to conflict of interest between managers and shareholders rather than their different preferences for such objectives.

The basic moral hazard problem derive from a trade off between insurance and incentive. We summarise the problem as the division of a random size pie (profit) between two parties: a risk neutral shareholder and a risk averse manager. One of the crucial assumption is that the principal is risk neutral, whereas the agent is risk averse. The principal is the claimant of the profit and must design a wage contract \( w(\Pi) \) for the agent, who runs the firm.

Let first specify their expected utility:

- Shareholder expected utility: \( E[\Pi - w(\Pi)] \)
- Manager expected utility: \( E[u(w(\Pi))] \)

If incentive issues are left aside and we consider only the problem of insurance the optimal solution implies that the risk averse party should get full insurance, i.e. should be given a constant income over all states of nature.

Suppose now the manager takes some unobservable action (in general will exert some effort) to influence the size of the profit and this action is costly for him. The shareholder observes only the realisation of profit and not the level of effort. If the manager (risk averse party) is given a fixed income that does not depend on this realisation he has no incentive to exert effort. So full insurance conflicts with incentives. The trade off between insurance and incentives leaves the parties with suboptimal insurance and profit.

This trade off would not arise if both parties were risk neutral so that the manager's need not to be insured. The shareholder can ensure that the manager takes the optimal action by selling the firm to him, making him the residual claimant. Because the manager's expected income is the expected size of the profit he has all the incentives to maximise it. The manager bears all the risk but it doesn't matter because he is risk neutral.
Finding the optimal incentive scheme when the agent is risk averse is a complex task.

Suppose the profit can take two values $\Pi_L$ and $\Pi_H$.
A manager can choose between two level of effort: high (work) and low (shirk). The utility function of the manager when he works is $u(w-\phi)$ and $u(w)$ when he shirks. Assume that working outside the manager can get a salary equal to $w_0$ giving him a reservation utility $u(w_0)$. If the manager works the profit is $\Pi_L$ with probability $\alpha$ and $\Pi_H$ with probability $1-\alpha$. If the manager shirks the profit is $\Pi_L$ with probability $\beta$ and $\Pi_H$ with probability $1-\beta$, with $\alpha > \beta$.

If effort is not observed a high effort cannot be induced by a constant wage structure. The shareholder must reward the manager when the profit is high.

The wage structure must satisfy the incentive compatibility constraint:

$$\alpha u(w_L-\phi) + (1-\alpha) u(w_H-\phi) \geq \beta u(w_L) + (1-\beta) u(w_H)$$

where $w_L$ and $w_H$ are the wage paid when the realised profit is $\Pi_L$ and $\Pi_H$.

$$\alpha u(w_L-\phi) + (1-\alpha) u(w_H-\phi) \geq u(w_0) \quad \text{Partecipation constraint}$$

The expected utility of the shareholder is maximised subject to the two constraints above.

$$\text{Max } \alpha (\Pi_L - w_L) + (1-\alpha) u(\Pi_H - w_H)$$

General results are exceptional. Even for the simplest specification of the utility function of the manager quasi separable in the effort and the wage one can prove only that the wage function must be increasing at some level of the profit and has slope less than unity at some level of the profit. The optimal wage structure is a profit sharing scheme a compromise between a fixed wage which yield optimal insurance and residual claimancy which yields optimal incentives.

In practice managerial compensation is contingent on the value of the firm as well as on its current profits. Managers are rewarded on the basis of the stock value rather than the profit of the firm. Stock options can be considered as an incentive for a manager to care about the firm's future profits as well as its current profits.

**Limits to managerial discretion**

**Yardstick competition**

We have seen how a firm's low profit may be due to factors other than to managerial slack (factors such as decrease in the demand increase in costs). Such effects can be detected to some extent by comparing the agent's performance with that of other agents placed in similar conditions. Assume a shareholder has two managers in charge of two divisions serving two distinct geographic markets whose demand are perfectly correlated.

In this setting the shareholder can use contracts based on the relative performance of the managers, in the sense that each manager's wage depends on the other manager's performance as well as on his own.

If both managers reach the same level of profit they both receive the same wage (the full information wage) otherwise the high profit manager gets this wage and the other gets punished. Under perfect correlation of shocks at the equilibrium both managers are expected to work. The intuition goes as follows: if a manager is expected to work and yield the high profit the other
manager automatically reveals that he did not work by obtaining the low profit. He cannot attribute his poor performance to adverse circumstances and he is heavily fined. The idea of yardstick competition carries over to environment with imperfect correlation. However it relies on the correlation of the agents' technologies. Comparison can be made with competing firm's manager.

**Takeover threats**

Another limit to managerial discretion is coming from the stock market. Failure to maximise profit leads to an undervaluation of the firm (i.e. to a decrease of the stock value from $V^*$ to $V$) and may induce outside entrepreneurs to buy the firm replace its management restoring the value $V^*$. In theory the threat of takeovers may impose discipline on managers. Is this a credible threat? The argument according to which managers are punished is not really convincing because of limited liability in reality far from being punished managers usually receive golden parachutes after a raid. More credible is the threat of being removed in losing all the rents the enjoyed within the firm. In particular the loss of prestige and reputation, also the loss of slackness (on job leisure due to asymmetric information).

However some limits to the takeover threat derives from the fact that costly information must be collected about the firm's inefficiencies. Outsiders must judge that they can derive substantial profits from the takeover enough to recover the takeover costs.

The shareholder has an incentive to accept as far as $p > V$. The raider gain $V^* - p$ per share. If there are $n$ shares the raid is profitable if and only if $n \cdot (V^* - p) > c$. To avert the risk of takeover the existing management must maintain at least a value $V'$ such that $n \cdot (V^* - V') < c = \text{cost of a bid}$. That is $V' > V^* - c/n$

Then any bid that would be acceptable to the shareholders ($p > V'$) would not be profitable for the raider. If $c$ is relatively small then takeover would act as a very powerful force for profit maximisation.

This optimistic view has been challenged by Grossman and Hart. Grossman and Hart pointed out also a potential free rider problem: in the event of a takeover a shareholder may not want to tender his shares because if he keeps them he can enjoy the increase in the stock price ($V^* - p$) freeriding on the back of the raider. For the shareholder rejecting the bid is a dominant strategy. The underlying assumption is that individual shareholders are atomistic (shareholding is so dispersed that an individual decision on whether to accept or reject the bid does not affect the outcome).

On the other hand the raider can only make a profit if the tender price of the shares is less than the post raid price ($V^* > p + c$).