

Chapter 1

Introduction

1.1 What is Inframarginal Economics?

Business decisions can be categorized into two classes: marginal decisions of resource allocation and inframarginal decisions of economic organization. Marginal decisions concern with the extent to which resources are allocated to a pre-determined set of activities. Inframarginal decisions concern with what activities to engage in (or whether or not to engage in an activity). To illustrate, before a student enrolls in a university, she needs to choose her major (field of study). If she chose economics as her major, then typically she would take microeconomics, macroeconomics classes and not chemistry or physics classes. Her choice of major and the associated choices of courses to take are inframarginal decisions since they involve deciding *what* activities to engage in (which is a series of yes-no decisions). Once she has chosen her major and classes, she then decide how to allocate her time to the chosen courses. These allocation decisions are marginal decisions since they involve deciding the *amount* of resources devoted to each activity given the activities that have been chosen.

In the context of social division of labor, inframarginal decisions are perhaps more important than marginal decisions. If each individual chose to be self-sufficient (which is an inframarginal decision involving saying yes to production of all essential consumption goods and saying no to all trade activities), there would be no trade connection between individuals and no social division of labor. If each individual chose to specialize in producing a single good and to buy all other goods that he consumed

(which is an inframarginal decision), then the network of division of labor would be very large. There are many intermediate network sizes of division of labor between the two extremes. As an individual becomes more specialized, he must have more trade connections with other specialists to obtain goods that he needs but does not produce. Thus an individual's specialization decision determines his trade connections with others, and all individual's specialization decisions jointly determine the network size and pattern of social division of labor. For this reason we refer to individual's specialization decisions as inframarginal network decisions.

Inframarginal analysis concerns with optimal inframarginal network decisions and the outcome of these decisions. The optimization of inframarginal network decisions involves both total cost benefit analysis across different network patterns of specialization and trade connections as well as marginal analysis of resource allocation for a given network pattern.

In mathematical terms, inframarginal analysis includes all non-classical mathematical programming (linear and non-linear programming, mixed integer programming, dynamic programming, and control theory) that allows corner solutions. Inframarginal analysis was developed by mathematicians in the 1950s and has been applied to economics by Koopman, Arrow and other economists in the 1950s and 1960s. Coase (1946, p. 173) noted "a consumer does not only have to decide whether to consume additional units of a product; he has also to decide whether it is worth his while to consume the product at all rather than spend his money in some other direction". He applies this inframarginal analysis to criticize marginal cost pricing rule and Pigou's (1940) marginal analysis of externality. Buchanan and Stubblebine (1962) coined term inframarginal analysis. Koopman (1957) and Arrow, Enthoven, Hurwica, and Uzawa (1958) are among those economists who initiated formal inframarginal analysis in economics. Inframarginal economics is a combination of inframarginal analysis and a Smithian framework of consumer-producers.¹ It focuses on individuals' inframarginal decisions. The outcome of inframarginal networking

¹ Yang (2001) called the framework "New Classical Framework".

decisions forms the “topological properties of an organism”, which can be represented by a graph consisting of vertices (or nodes, points) and edges (or lines, curves). All information about marginal decisions of resources allocation form non-topological properties of an organism and can be represented by weights attached to edges and vertices of the graph. A weighted directed graph (or digraph) can describe topological as well as non-topological properties of an economic organism. Marginal analysis focuses solely on non-topological properties of economic organisms, while inframarginal analysis focuses on topological properties of economic organisms, taking into account non-topological properties as well.

1.2 Economic Analysis of Division of Labor – A Literature Review

The classical insight of the economies of the division of labor can be traced back to the writings of Ancient Greeks. Ancient Greek philosophers Democritus (460BC-370BC), Xenophon (431BC-354BC) and Plato (427BC-347BC) were the earliest classical thinkers who noted the pattern of division of labor and specialization. For example, Xenophon (1886, pp. 244-5) observed that large cities tend to have high level of division of labor, which leads to a greater diversity of occupations and more production of goods. Plato, in his famous work “The Republic”, considered the welfare implications and the connection between the division of labor, the market, and money (380BC, pp.102-6). Similar observations were documented sporadically by scholars and philosophers from countries, ranging from Ancient Chinese, Islam scholars to Anglo-European, and these studies existed long before the publication of Adam Smith’s celebrated work “The Wealth of Nation”, which was widely considered as the forerunner in the economic analysis of division of labor.² About a century before Smith’s writings, British Philosopher William Petty observed how specialization in clothmaking,

² Sun (2005) presented a comprehensive and systematic examination on the emergence of the doctrine of division of labor, based on hundreds of studies, written by ancient Greeks, ancient Chinese, medieval Islam scholars, Medieval Latin scholasticists and Anglo-European.

watch manufacturing and shipping improves productivity. Petty (1683, pp. 471-2) observed that cities could promote the division of labor by reducing transaction costs. This coincides with Josiah Tucker (1755, 1774), who examined the productivity implications of the division of labor, and the positive relations between variety of goods, degree of production roundaboutness, and the level of division of labor. Following Petty, numerous works, notably Denis Diderot (1713), Henry Martyn (1701, who's authorship was uncovered by MacLeod, 1983), Henry Maxwell (1721, p. 33) and Josiah Tucker (1755, 1774) recognized three advantages of division of labor: First, it improves the skill (or human capital, in modern terms) of individual workers. Second, it saves the time and effort spent on switching from one task to another. Third, it facilitates the invention of machinery. They have also recognized the role of the market and population size in stimulating specialization. Turgot (1751, pp. 242-3) observed the association between the development of division of labor and the increased in living standard for even the humblest member of society and the increased in inequality of income distribution. Turgot (1766, pp. 44-6, 64, 70) further noted the association between the division of labor and the introduction of money, the extension of commerce, and the accumulation of capital.

The most influential classical work on the division of labor is Adam Smith's "The Wealth of Nation" first published in 1776. Smith (1776) systematically investigated the implications of division of labor for wealth creation and prosperity of nations. One of Smith's well-known insights was that the division of labor is limited by the extent of the market and the extent of the market is affected by transportation efficiency (1776, chapter 3 of book I). Smith proposed that a theory of capital in which capital is a vehicle for increasing division of labor in roundabout productive activities (p. 371). He also observed that economies of specialization and division of labor may exist even if all individuals are ex ante identical and that the differences in productivities between various specialists are consequences rather than causes of the division of labor (p. 28).

Smith contended that the main reason for the productivity difference between the industrial sector and the agricultural sector was that the former gained more from specialization (relative to the seasonal

adjustment costs associated with specialization) than the later. A decline in income share of the agricultural sector occurs not because of a change in tastes, in income, or in exogenous technical conditions, but because the agricultural sector has a higher coordination cost of the division of labor compared to the benefits derived from the division of labor. The productivity of the agricultural sector can be improved only by importing an increasingly larger number of industrial goods whose production takes advantage of a high level of division of labor where coordination costs are more likely outweighed by the economies of division of labor.

Alfred Marshall recognized the importance of the classical insights on the division of labor in his influential book “Principles of Economics” (1890, chapters 8-12 of book IV). While he received enormous success and popularity in his mathematical framework of resource allocation (marginal analysis), which later became mainstream neoclassical microeconomics, he failed to comprehend Smith’s classical insights consistently in his framework because his partial equilibrium analysis is incapable of formalizing general equilibrium flavor of network effects of division of labor. As Buchanan (1994, p. 6) observes, “with one part of his mind always in classical teachings, Marshall recognized that this genuinely marvelous neoclassical construction requires that the Smithean proposition on labor specialization be abandoned”. Unfortunately, Marshall’s shortfall had not inspired neoclassical mainstream economists to explore and include classical economic thinking on the division of labor, partly due to the overwhelming influence of Marshall’s neoclassical economic theory with a focus on the marginal and partial equilibrium analysis of the problem of resource allocation. The modern literature on the implications of division of labor and the associated problem of economic organization is limited. Following the success of Samuelson (1948) as the prototype for principle textbooks in economics, there has not been a place for problems of specialization and division of labor in textbooks. Most principles textbooks pay only brief symbolic respect to classical economic thinking on these issues, with little discussion of classical insights on the division of labor. However, as many would have agreed with Houthakker’s (1956, p. 182) view that “there is hardly any part of economics that would not be advanced by a further analysis of specialization”, some distinguished economists have expressed concern

over the glaring omission of formal theory on the division of labor in mainstream economics. As observed by Stigler (1976, pp. 1029-1210),

The last of Smith's regrettable failures is one for which he is overwhelmingly famous – the division of labor. How can it be that the famous opening chapters of his book, and the pin factory he gave immortality, can be considered a failure? Are they not cited as often as any passages in all economics? Indeed, over the generations they are. The failure is different: almost no one used or now uses theory of division of labor, for the excellent reason that there is scarcely such a theory. ... there is no standard, operable theory to describe what Smith argued to be the mainspring of economic progress. Smith gave the division of labor an immensely convincing presentation – it seems to me as persuasive a case for the power of specialization today as it appeared to Smith. Yet there is no evidence, so far as I know, of any serious advance in theory of the subject since his time, and specialization is not an integral part of the modern theory of production, which may well be an explanation for the fact that the modern theory of economies of scale is little more than a set of alternative possibilities.

Since Marshall (1890), there have been some notable contributions to the understanding of the driving forces and implications of the division of labor, and to the development of techniques necessary for studying the division of labor in an internally consistent mathematical framework. First and most important of all, is Allyn Young (1928). Young argued that demand and supply are two sides of the division of labor. He revived Smith's (1776) famous theorem that division of labor is limited by the extent of the market and asserted that “[n]ot only the division of labor depends upon the extent of the market, but the extent of the market also depends upon the division of labor” (p. 539). The extent of the market is determined not only by population size, but also by purchasing power, which is determined by productivity, which is in turn dependent on the extent of division of labor. This ingenious insight implies a circular causation between the division of labor and the extent of the market, which is a common feature of network effects. This is similar to the circular causation where the value of a telephone set to a user is dependent on the number of telephone users within the network, and the number of users is dependent on the value of telephone set to each user. In this sense, the

Smith-Young theorem is about the network effect of division of labor. It implies that individuals' decisions to choose their levels of specialization are determined by the benefits of division of labor, which are dependent on the number of participants in the network of division of labor (the extent of the market). Meanwhile, the number of participants in the network is determined by all individuals' specialization decisions. The circular causation is also a typical feature in general equilibrium models. In a conventional general equilibrium model, the optimum quantities demanded and supplied are dependent on market prices while market prices are determined by all individuals' decisions of quantities of demand and supply. Hence, the Smith-Young theorem captures some important linkages between the division of labor, network effects and general equilibrium.³

Young (1928) used three concepts to describe the economies of specialization, which differs from the economies of scale foreshadowed in Marshall's neoclassical framework. The first is individual's level of specialization. An individual's level of specialization increases as he/she narrows down his/her scope of activities. The second is the length of a roundabout production chain, or the roundaboutness of production. The third is the number of intermediate goods in each link of the production chain. He argued that increasing returns with which he was concerning are not caused by the scale of a firm or an industry. It was caused by specialization and division of labor. As he claimed, "[t]he mechanism of increasing returns is not to be discerned adequately by observing the effects of variations in the size of an individual firm or of a particular industry, for the progressive division of labor and specialization of industries is an essential part of the process by which increasing returns are realized. What is required is that industrial operations be seen as an interrelated whole". Young's contribution foreshowed two promising lines of research on division of labor. The first is to formalize the concept

³ Young's (1928) criticism of the notion of static general equilibrium does not mean that he rejected the notion of general equilibrium, but rather it means we need a notion of dynamic equilibrium to describe spontaneous evolution in division of labor. He called the dynamic general equilibrium "moving equilibrium". Indeed, comparative statics of general equilibrium may, to a certain extent, capture the essence of his moving equilibrium.

of division of labor which includes individuals' specialization, production roundaboutness, and variety of occupations and goods. The second is to model how the size and features of the market network are determined by impersonal networking and specialization decisions in a decentralized system.

Early models of economies of specialization since Smith and Young began with Houthakker (1956), which studied the trade-off between economies of specialization and transaction costs. Houthakker noted that if transaction costs are high, the economies of specialization will be outweighed by transaction costs associated with specialization, and the equilibrium level of division of labor and aggregate productivity will be low. As the trading efficiency is improved, the equilibrium level of division of labor and productivity will increase, and the extent of the market and demand and supply in the marketplace will also increase. He argued that economies of specialization is different from the economies of scale of a firm because of time saving between task-switching and internal coordination when several activities are combined into one. This gives rise to greater scope for workers to accumulate experience and as a result, the total output of combined specialized activities must necessarily be greater than the combined output of several activities when they are not specialized. Houthakker recognized that the formal analysis of the division of labor "involves the use of methods that are rather unlike those by which the classical questions of economics are discussed. These classical questions are treated with the aid of traditional calculus methods, but the latter are not suited to deal with indivisibility. It is in fact from indivisibility that the division of labor takes its start, and the basic indivisibility is that of the individual." (p. 182) Falling short of developing a mathematical model, Houthakker used a graph (reproduced below) to illustrate the economies of specialization assuming production technologies with fixed learning costs.

Suppose there are two ex ante identical individuals. The production functions for the two goods and the endowment constraint are:

$$x_1 = \max\{l_1 - a, 0\}, \quad (1.1)$$

$$x_2 = \max\{l_2 - a, 0\}, \quad (1.2)$$

the simple two-good-two-person case, the aggregate transformation curve for the division of labor, where at least one individual produces only one good, is MCAKBJL.

It is obvious that the aggregate transformation curve for the division of labor is higher than the aggregate transformation curve for autarky, even if the two persons are ex ante identical or even if exogenous comparative advantage is absent. This is because each individual's total learning cost is $2a$ if she produces both goods and her total learning cost is reduced to a if she produces only one good. That is, her time for production increases from $1 - 2a$ to $1 - a$ as she reduces the number of goods produced from two to one. Hence, the total learning cost for the economy with two individuals is $4a$ in autarky, $3a$ for partial division of labor (segment CA or BJ), and $2a$ for complete division of labor (point K) in an economy with two individuals. The economies of division of labor are represented by the difference between the transformation curve for the division of labor, MCABJL (which is also the PPF), and the transformation curve for autarky, DI. As we indicate previously, the economies of division of labor are network effects since a person's decision determines not only her productivity, but also the extent of the market for others' produce, thereby setting up a constraint on others' decisions in choosing their levels of specialization, which affect their productivity. As transaction efficiency is improved, the equilibrium aggregate production schedule jumps from line DI to point K, generating positive network effects on aggregate productivity. Also, there is an ex post difference in productivity between ex ante identical individuals if they choose different levels of specialization.

The economies of division of labor are, of course, generated by *endogenous comparative advantage*. When ex ante identical individuals choose different levels of specialization in an activity, a specialist endogenously acquires a higher productivity than a novice. Consider the two-person-two-goods example in Figure 1.1, where the two individuals choose complete division of labor. If person 1 specializes in producing good 1 (accordingly choosing l_1), her labor productivity in that good is $\max\{(l_1 - a)/l_1, 0\} = 1 - a$. For person 2, specializing in good 2 (and thus choosing $l_1 = 0$), labor productivity in good 1 is $\max\{-\infty, 0\} = 0$. The ex

post difference in productivity arises due to endogenous comparative advantage when individuals specialize.

Like Houthakker, Stigler (1951) investigated the economies of specialization by using a graph to show that a firm's productivity increases as it narrows down its range of production activities. Stigler demonstrated that a firm's cost function will endogenously and discontinuously change as the firm makes different decisions about its level of specialization. Stigler's work followed Marshall's approach of separating the analysis of demand from the analysis of decision making regarding the level of specialization, but it emphasized increasing returns to specialization rather than Marshall's concept of external scale of economies.

Among others, Sherwin Rosen and Gary Becker were two pioneers in the research of modeling individuals' decisions to determine their levels of specialization, which is an important subject on Young's research agenda. Rosen (1978) extended the Ricardo model to include many goods and many individuals. He used a managerial decision model and applied linear programming to examine individuals' specialization decisions which involve corner solutions. He showed that economies of division of labor are endogenously determined individuals' decisions on their levels and patterns of specialization. As individuals choose different levels and patterns of specialization, resource allocation may jump from one transformation curve to another, generating changes in productivity (see Figure 1 in Rosen, 1978, p. 236). More interactions among individuals imply a larger scope for productivity improvement. Rosen called this social complementarity the "superadditivity".

Superadditivity is different from economies of scale or technical complementarity which relates to economies of scope. Economies of scale and economies of scope relate to pure technical relations between outputs and inputs and they are independent of the degree of interpersonal and social interdependence. Superadditivity is closely dependent on the interdependence between individuals which is a result of individuals' specialization decisions. Since the aggregation of individuals' decisions can generate many possible patterns of division of labor and each pattern is associated with a certain size of network of exchanges, Rosen's superadditivity is a concept that captures the essence

of the Smith-Young conjecture about the interdependence between the extent of the market and the division of labor.

Becker (1981) developed a model to endogenize individuals' decisions on specialization within a family. The solution of this model involved inframarginal analysis of many corner and interior solutions. The positive interaction between labor and human capital allocated to produce a certain good leads to complete specialization for all members of the family except one when an integer condition for the numbers of different specialists is not satisfied. Becker's model endogenized the pattern of specialization within a family and comparative advantages as a result of specialization (endogenous comparative advantage). It also formalized the classical idea that an advantage of specialization is the avoidance of duplicated learning and training costs.

Following Becker's (1981) idea of modeling learning costs, Rosen (1983) developed a model to explain individuals' specialization pattern. In his model, there is a learning cost associated with activity i ($i = 1, 2$). An individual chooses his time (t) allocation to maximize the net benefit specified as

$$V = w_1 k_1 t + w_2 k_2 (1 - t) - C(k_1, k_2),$$

where k_i is learning and training level in activity i , C is the total learning and training cost, $1 - t$ is the amount of time allocated to activity 2, and w_i is a given benefit coefficient for activity i . Since V is linear in t , the optimum value of t may involve corner solution. Rosen used marginal analysis to solve for the two corner solutions which involve specialization in different activities, and the interior solution. He then compared the total cost-benefit across all possible solutions. The result is that non-specialization is optimal if and only if the economies of technical complementarity between two learning activities outweigh the economies of specialization generated by a higher utilization level of a particular learning and training investment. His result again showed that social complementarity is different from technical complementarity. If $\partial^2 C / \partial k_1 \partial k_2 = 0$, which means no technical complementarity exists, two individuals can still take advantage of social complementarity by specializing in different activities.

Another important work on individuals' level of specialization is Becker and Murphy (1992). Becker and Murphy investigated the relation between the division of labor, coordination costs and human capital. Their model is a decision model where the optimum level of division of labor is determined by the condition that the benefit and cost of the division of labor are equalized at the margin. In their model, if coordination costs increase more rapidly than economies of division of labor as the number of different professions increases, the optimum level of division of labor will evolve as the human capital parameter and/or the coordination cost coefficient change.

The Becker-Murphy model shows that the efficient level of division of labor is determined not only by the population size which is usually considered as the extent of the market, but also by the efficient balance between economies of division of labor and coordination transaction costs. This suggests that the concept of extent of the market needs to be more accurately defined. There are three aspects of the extent of the market: population size; the number of goods; and the number of traded goods relative to the number of all goods. Generally a large population size means more scope for the division of labor and a larger market. A greater number of goods can also mean a larger market – if each individual buys all goods that he consumes, then the number of consumption goods determines his trade volume which in turn affects the extent of the market. Many new trade and growth models (see, for example, Dixit-Stiglitz (1977), and Ethier (1982), Grossman and Helpman (1989)) have modelled the extent of the market by endogenizing the number of goods. However, these models do not endogenize the third aspect of the extent of the market which is determined by individuals' levels of specialization. If individuals' levels of specialization are not endogenized, then population size becomes the ultimate driving force of the number of goods, division of labor, and productivity. This result is inconsistent with evidence of a negative correlation between the population size and productivity in some less developed economies. The Becker-Murphy model shows that even if the population size is fixed, a trade-off between economies of specialization and coordination costs can endogenize the level of division of labor – the

driving force behind the division of labor can be falling coordination costs.

Tamura (1992) develops a dynamic version of the Becker-Murphy model. Economies of specialization due to a higher utilization rate of training and learning investment are specified to endogenize the interval of activities of each individual specialist. He specifies a CES production function and the trade-off between current and future consumption to tell an endogenous growth story. However, the novelty of endogenization of specialization is lost in his dynamic macroeconomic model, due to aggregation. This model cannot predict evolution of individuals' specialization and of division of labor in society. Coordination cost is specified as an aggregate function of the population size. No trade-off between economies of specialization and transaction (or coordination) costs exists. Each specialist's interval of activities is directly determined by the population size. This is a setback from the B-M model which can explain individuals' levels of specialization by the trade-off between economies of specialization and coordination costs even if the population size is fixed.

1.3 Inframarginal Economics vs. Neoclassical Economics

As discussed in the previous section, the focus of the classical mainstream of economics was on the implications of specialization and division of labor for the wealth of the nations. According to classical economic thinking, the most important decisions of individuals involve their choice of occupation and level of specialization. The aggregate outcome of these decisions determines the community's level of division of labor (or, in modern language, the network size of division of labor). Demand and supply are two sides of division of labor. The most important function of the invisible hand is to coordinate individuals' decisions in choosing levels and patterns of specialization to enable utilization of the positive network effects of division of labor.

Since Leon Walras (1874), Carl Menger (1871), and Marshall (1890), the focus of economic analysis has been shifted from the function of the

price system in coordinating specialization and division of labor to the function of the price system in allocating resources.

Marshall's neoclassical framework is characterized by the dichotomy between pure consumers and firms, the replacement of the concept of economies of specialization with the concept of economies of scale, and the marginal analysis of demand and supply.⁴ Marshall's neoclassical framework cannot be used to analyze individuals' decisions in choosing levels and patterns of specialization, so that structure of division between pure consumers and firms is exogenously given. His marginal analysis of demand and supply, which would become the basis of neoclassical microeconomics in the mainstream, has nothing to do with individuals' decisions in choosing their levels of specialization.

In order to get more intuition about the distinction between Marshall's neoclassical framework and the framework used in the new literature of endogenous specialization, let us look at Figure 1.2. Figure 1.2(a) illustrates Marshall's neoclassical framework. The two circles with numbers represent two consumers who do not make production decisions. The two circles with x and y represent two firms producing respectively goods x and y . The solid lines represent flows of goods sold by firms and the broken lines denote flows of labor and other factors from consumers to firms.

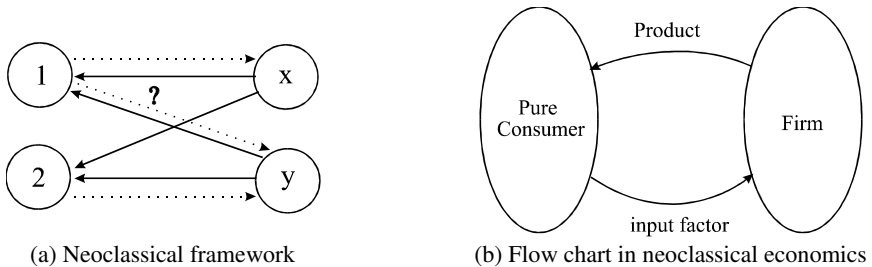


Figure 1.2: Neoclassical Analytical Framework.

⁴ The modern Arrow-Debreu model of general equilibrium, which features the first two properties of Marshall's framework, has generalized and consolidated that framework. Arrow and Debreu use the concept of non-convex production set to generalize the concept of economies of scale.

In the *neoclassical framework* all goods consumed by consumers are bought from the firms, consumers cannot choose their levels of self-sufficiency. Therefore, the existence of the market and the institution of the firm is exogenously given. Productivity of the firm relates only to its operation scale and has nothing to do with the levels of specialization of workers and managers within the firm. Suppose for simplicity that each consumer sells one unit of labor, which is the sole production factor, to the firms. There are at least two patterns of specialization. In one pattern, consumer 1 sells her all labor to the firm producing x and consumer 2 sells her all labor to the firm producing y . Hence, the scale of labor employed by each firm is of one unit and each individual is completely specialized. In another pattern of specialization, each consumer sells 0.5 unit of her labor to each of the two firms, so that individuals are not specialized, but the size of labor employed by each firm is still one unit. According to the neoclassical notion of economies of scale, productivity in the pattern of complete specialization and in the pattern with no specialization is the same. But according to the classical notion of economies of specialization, the pattern with specialization should generate a higher productivity.

In a neoclassical model, with or without economies of scale, the two patterns of specialization are associated with the same general equilibrium. In other words, individuals' patterns and levels of specialization are not well defined and make no difference. According to classical economic thinking on the other hand, the productivity implications of the level of division of labor of society which is determined by individuals' levels of specialization are the focus of the economic analysis. The pattern of division of labor is related to the number of transactions, which is referred to mathematically as one of topological properties of the graph of an organism. Thus we may say that the focus of classical economics is on the implications of the topological properties of economic organisms. Quantities of goods consumed and produced can be denoted by the thickness of lines representing flows of goods and factors. Mathematically this characteristic is referred to as a non-topological property of the graph of an organism. Thus, because of its preoccupation with relative prices and quantities, we may say that neoclassical economics focuses on the implications of the non-

topological properties of economic organisms. Since topological properties are not the focus of neoclassical economics, the flow chart in Figure 1.1(b), which ignores these, becomes a common representation of the neoclassical economy in standard textbooks.

Now consider Figure 1.3. In panel (a) a circle with A denotes a consumer-producer self-providing all goods that she need. In other panels a circle with number i ($=1, 2, 3, 4$) denotes an individual selling good i , while an arrowed line with number i denotes a flow of good i . This figure illustrates the features of the *Smithian framework* in the modern literature of endogenous specialization. In each panel there is an economy with four consumer-producers, each of whom consumes four goods and can choose to produce either one, two, three, or four goods. In panel (a), each consumer-producer is self-sufficient in all goods; thus there is no market, and the economy is divided into four separate components. Clearly the degree of commercialization (or marketization) and the degree of integration in this economy is low. Accordingly, the degree of concentration of production is low (there are four producers of each good), as is each person's level of specialization. If it is true that specialization can speed up the process of learning by doing, or can reduce total fixed learning costs by avoiding duplicated learning and training and by increasing the utilization rate of those costs in the society, then it follows that each person's productivity is low in the autarchic structure of panel (a). But on the other hand this structure does not have transaction costs. Since each person has the same configuration of production, the degree of diversification in autarky is also low.

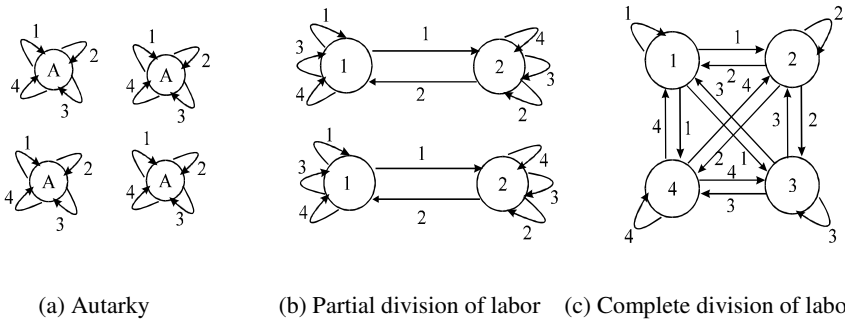


Figure 1.3: Smithian Framework.

In the organizational structure depicted in panel (b), the number of goods produced by each person has been reduced from four to three, so that relative to panel (a) each individual's level of specialization and productivity has increased. Markets for two goods have emerged from the division of labor, and the number of transactions and related transaction costs have increased. The economy now comprises two, rather than four, separate local business communities, so that the degree of market integration has increased. The number of producers of the (now traded) good 1 or 2 has decreased from four to two, so that the degree of production concentration is higher than in autarky. Since two different professions have emerged from the division of labor, the degree of diversity of the economic structure has increased as has each individual's level of specialization. The degree of interpersonal dependence, the degree of interaction between individuals, the degree of each person's trade dependence, the degree of commercialization for society, and the number of markets have all increased compared with autarky.

In panel (c), each individual is completely specialized in producing one good. Accordingly, the degree of market integration, the number of markets, the diversity of the economic structure, the number of transactions, the concentration of production, the degree of commercialization, productivity, and transaction costs are all higher than in panel (b).

We will show later, because of the trade off between the positive network effects of division of labor and transaction costs, as a transaction cost coefficient for a unit of goods traded falls in a static model, transaction costs caused by division of labor will be more likely to be outweighed by positive network effects of division of labor. Hence, the economy will evolve from autarky structure in (a) to that in (b), followed by that in (c). In a dynamic equilibrium model such evolution of division of labor may occur spontaneously in the absence of exogenous improvements in trading efficiency. The above concurrent phenomena will take place as different aspects of the evolution of division of labor. In this process, the emergence of the market is endogenously determined by individuals' self-interested decisions in choosing their levels and

patterns of specialization and by the interactions of the self-interested decisions in the market place. Demand and supply are two aspects of the division of labor. The main features that distinguish the Smithian framework from the neoclassical one can be summarized as follows.

Firstly, in the Smithian framework, there is no *ex ante* dichotomy between pure consumers and firms. Individuals make *inframarginal* decisions about their level and pattern of specialization and their optimal decisions are corner solutions. In the neoclassical framework, consumers and firms are separated. Consumers choose their consumption patterns and their allocation of endowments; their optimum choices are usually interior solutions; corner solutions are exceptional. It might be argued that the consumer-producer framework makes the algebra much more complicated than that in the neoclassical framework, and that the framework appears to be suited for an economy characterized by prior industrial revolution home production. However, as shown in the graphs, the framework of consumer-producers is essential for defining equilibrium labor allocation of each person. Moreover, if labor trade and a roundabout production chain are introduced into Smithian models, a very complicated social production involving firms may occur and home production does not take place in equilibrium as division of labor evolves to a very high level.

Secondly, in the Smithian framework, transaction cost (transaction efficiency) has important implications for the network properties of the equilibrium structure. As transaction efficiency improves, the equilibrium network size of division of labor is enlarged, aggregate productivity and social welfare increase, and total transaction costs increase.

Thirdly, in the Smithian framework, productivity is determined by individuals' specialization decisions and is described using the concept of economies of specialization and network effect of the division of labor. In the neoclassical framework, productivity positively relates to scale of a firm. As shown by Liu and Yang (2000), the level of division of labor is positively related to the average size of firms if division of labor evolves within each firm and negatively related to the latter if division of labor evolves among increasingly more specialized firms.

Fourthly, in the Smithian framework, the institution of the firm is not given; it emerges from the division of labor only if individuals have decided to use labor market to coordinate the division of labor. Production functions are specified for each consumer-producer for all possible production activities. A firm cannot simply pool all employees' labor into its own production function; rather the production function of a firm is a combination of production functions of all employees and the employer. This implies that aggregate production set may be non-convex even if all individuals' production sets are convex. As a result, general increasing returns and network effects of division of labor are compatible with a competitive market.

In inframarginal economics, there are two kinds of decisions that each individual has to make. The first involves the choice of one's occupation and the level of specialization in that chosen occupation, which relates to how many (professional and self-sufficient) activities one engages in, or to choose whether one does it or not. This kind of decision is inherently associated with corner solutions (cases in which zero values attach to some activity levels). Marginal analysis for interior solutions does not work for this kind of decision making. The second type of decision is to choose how much one allocates her limited resources to all activities that one has decided to undertake. This kind of decision features marginal analysis of a corner or interior solution. Inframarginal analysis is a combination of these two types of decision making.

1.4 Inframarginal Analysis of Division of Labor

The basic approach of inframarginal analysis of division of labor started with the assumption of *ex ante* identical individual consumer-producers who have the same utility function and production functions for each goods. The production functions exhibit economies of specialization – as an individual devotes a larger share of his labor to producing a good increases (i.e., as his level of specialization increases), his productivity in that good increases. Since each individual has a fixed labor endowment, the economies of specialization are individual-specific and increasing returns are localized, and consequently simply pooling labor without an

increase in individuals' levels of specialization cannot increase their productivities.

An individual's decision is presented by a non-linear programming problem – each individual is assumed to maximize his utility by choosing his quantities of goods produced, traded, and consumed, and level and pattern of specialization, subject to the production functions, endowment constraint, and the budget constraint. An individual's specialization decision can be thought as choosing a profile of zero and positive values of the decision variables associated with each good: whether or not to self-provide, to sell or to buy. Consequently, there are three decision variables for each good: quantity self-provided, quantity sold, and quantity purchased. For a model with three goods, there are $2^{3 \times 3} = 512$ possible profiles of zero and positive values of decision variables.

The large number of combinations of choices makes the individual's decision problem difficult to solve especially when there are more than three goods. To solve this problem, Kuhn-Tucker theorem is used to systematically rule out a large number of profiles. For example, when there are economies of specialization and transaction costs, a consumer-producer never simultaneously sells and buys the same good, never simultaneously buys and produces the same good, and sells at most one good. In the model with three goods, the candidates of possible profiles are reduced from 512 to 10. This result was originally obtained by Yang (1988) and Wen (1998) generalized it for a general specification of quasi-concave utility function and separable production functions with economies of specialization, and non-increasing transaction cost coefficient functions. Wen's generalized result is referred to as the *theorem of optimum configuration*. This theorem is extremely powerful as it significantly narrows the number of candidates for the optimal decision. Later, Yao (2002) showed that the theorem may not hold in the case with linear production functions with fixed learning costs. He improved the theorem and proved that although selling two goods or more is possible, the optimal outcome for an individual can be achieved by selling at most one good.

We refer to the combination of zero and non-zero variables that is compatible with theorem of optimum configuration as a *configuration*;

and an aggregation of configurations that is compatible with the market clearing conditions for traded goods as a *structure*.

General equilibrium is defined as a state of the economy that satisfies the following conditions: (1) each individual maximizes her utility with respect to configurations and quantities of each good produced, traded, and consumed for a given set of relative prices of traded goods and numbers of individuals selling different goods; (2) The set of relative prices of traded goods and numbers of individuals selling different goods clear the markets for traded goods and equalize utility for all individuals.

A two step approach is employed to solve for the general equilibrium. First, for each structure, “corner equilibrium” prices and utilities are calculated using the market clearing conditions and utility equalization conditions. Secondly, the corner equilibrium in which individuals derive the highest utility is selected as the general equilibrium.⁵ For each configuration the individual’s indirect utility can be calculated as a function of the given prices, and the individual chooses the configuration that gives him the highest utility. Clearly which configuration pattern brings about the highest real income depends on the relative prices. The price set can therefore be partitioned into several sub-sets in each of which a particular configuration (i.e., specialization patterns) brings about the highest real income. This together with the market clearing condition (and utility equalization if applicable) then allows the identification of subsets of parameters in which a particular structure of specialization occurs in equilibrium.

There are two types of comparative statics of the general equilibrium. The first type involves shifts between general equilibrium structures as parameters in the model reach certain critical values. With the general equilibrium structural shift, demand and supply functions, and indirect

⁵ However, there are limitations with the two step approach, mainly due to its reliance on the assumption of *ex ante* identical individuals, and the possibility of missing some structures. To overcome these limitations, Sun (2003) proposed a unified approach to identify the general equilibrium structure. This approach can identify a complete set of equilibrium structures and the associated set of parameter subspaces in one step and in most cases, it significantly simplifies the algebraic computation.

utility function will (often discontinuously) shift.⁶ The second type of comparative statics of the general equilibrium involve continuous changes in equilibrium relative prices, quantities and other endogenous variables in response to continuous changes of the parameters within each parameter subsets that define equilibrium structures. The second type of comparative statics reveal resource allocation implications for a given network pattern of division of labor, and is the same as those in the neoclassical framework. The first type of comparative statics based on inframarginal analysis focus on implications of the network pattern of division of labor. They can be used to explain changes in the patterns of social division of labor. It implies that the efficient pattern and level of division of labor is determined by the efficient trade-off between transaction cost and the positive network effect of the division of labor. If transaction efficiency is low, the positive network effect of the market is outweighed by transaction costs, so autarky or low levels of division of labor occurs in equilibrium. If transaction efficiency is improved, the efficient and equilibrium level of division of labor and related efficient size of market network will increase.

1.5 Structure of this Book

In this book, we provide systematic and comprehensive materials for applying inframarginal analysis to study of a wide range of economic phenomena. A large amount of materials is drawn from the emerging literature of inframarginal economics, pioneered by Yang (1988) and Yang and Ng (1993). Based on a systematic and integrated framework, we resurrect the spirit of classical economic thinking on network effects of division of labor and general equilibrium mechanisms that simultaneously determine interdependent benefits of specialization and number of participants in the network of division of labor (extent of the market) in a modern body of inframarginal economics. Through revision and generalization of the core of mainstream economics, this framework

⁶ The discontinuous jump of the supply function is consistent with Stigler's (1951) conjecture that a change in the level of division of labor will lead to discontinuous change of the cost function.

is able to encompass, within the one overarching framework, many areas of the discipline that have been customarily treated as separate branches. These include, for example: microeconomics, macroeconomics, development economics, international economics, urban economics, growth theory, industrial organization, applications of game theory in economics, economics of property rights, economics of transaction costs, economics of institutions and contract, economics of organization, economics of states, managerial economics, theory of hierarchy, new theory of the firm, theory of money, theory of insurance, theory of network and reliability, and so on.

This book is divided into eight parts. Part I develops simple new Smithian models with only two final goods to illustrate the basic concepts of inframarginal economics and to introduce the techniques required to manage the inframarginal analysis of Smithian models. Chapter 4 introduces the theoretical foundation of inframarginal economics. A general Smithian model with no explicit specification of functional forms is used to investigate the existence of general equilibrium network of division of labor and the first welfare theorem. The notion of topology and weighted digraph of economic organisms are used to define general equilibrium network of division of labor and pattern of resource allocation. It is shown that the equilibrium organism which comprises the equilibrium network of division of labor and the equilibrium resource allocation are Pareto optimal. Hence, the most important function of “the invisible hand” is to coordinate individuals’ decisions in choosing their levels and patterns of specialization and to utilize network effects of division of labor.

Part II, Institution of the Firm and Inframarginal Analysis of Endogenous Transaction Costs, introduces intermediate goods into the analysis to explain why and how the institution of the firm emerges from the evolution of division of labor. Various game models and models of moral hazard are then used to explore the effects of endogenous transaction costs on the equilibrium network size of division of labor, and on the equilibrium structure of residual rights and ownership. Particular attention is paid to the connection between strategic interactions and network effects of division of labor. Also the implications of sequential equilibrium models for investigating

economies of information asymmetry and endogenous transaction costs are explored. Chapter 8 covers principal-agent models and Grossman-Hart-Moore models of two-sided moral hazard and optimum ownership structure. The Smithian framework is used to explore the implications of endogenous transaction costs caused by moral hazard for simultaneous endogenization of specialization, principal-agent relationship, and structure of residual rights.

Part III, *Inframarginal Analysis of Trade and Globalization*, applies the new framework to develop endogenous trade theory and the notion of endogenous comparative advantage. The results provide interesting support for Smith's view that differences in productivity between different specialists are the consequence rather than the cause of the division of labor. Then conventional exogenous comparative advantages in productivity and endowment are introduced to investigate the implications of the coexistence of endogenous and exogenous comparative advantage for economic analysis. Also, implications of *inframarginal comparative statics of general equilibrium* for discontinuous jumps of trade pattern between corner solutions are explored in connection to neoclassical trade models (the Ricardo model and the Heckscher-Ohlin model). Chapter 10 shows that since each country is both a consumer and a producer in the trade models, *inframarginal analysis* becomes essential and results based on marginal analysis may be misleading.

Part IV, *Urbanization, Population, Coordination Reliability, and the Trade off Between Working and Leisure*, explores the implications of the network effect of division of labor for emergence of cities and for the equilibrium differential of land prices between urban and rural areas. We show how the trade off between working and leisure in the Smithian framework can explain concurrent increases in both leisure and working time for the market as a consequence of the evolution of division of labor that reduces the working time for self-sufficient production. The professional infrastructure sector and primary resources are then introduced into the Smithian framework in order to endogenize transaction efficiency, and to explore the effects of population size and shortage of primary resources on the evolution of division of labor. Chapter 14 examines the economics of property rights and the theory of

insurance, in the light of transaction costs caused by coordination failure risk of a large network of division of labor. A new theory of endogenous externality is developed as part of inframarginal economics of property rights and transaction costs.

Part V, Hierarchical Structure of Division of Labor, develops the analysis to endogenize simultaneously the number of links, and the level of division of labor in the roundabout production chain. Also, the number of layers of the hierarchical structure of transactions and cities based on a high level of division of labor is endogenized and the intrinsic relationship between the hierarchical structure of organization, industrialization, and the evolution of division of labor is investigated.

Part VI, Inframarginal Analysis of Economic Development and Growth, uses various dynamic Smithian models to explain many development phenomena as different aspects of the evolution of division of labor. In particular, a sequential equilibrium model is used to explore the implications of spontaneous social experiments, through the price system, with various structures of the network of division of labor. This model features bounded rationality, adaptive behavior, uncertainty of the direction of the evolution of economic organisms, and spontaneous evolution of the information acquired by society through the social experiments. Chapter 18 discusses Smithian endogenous growth models based on spontaneous evolution of division of labor and the associated empirical observations.

Part VII, Macroeconomic Phenomena and Endogenous Size of Network of Division of Labor, develops a Smithian model explaining the emergence of money from the evolution of division of labor. We also develop a Smithian model of capital, which formalizes classical ideas about the intimate relationship between capital and the evolution of division of labor in roundabout production. A Smithian model of endogenous business cycles and unemployment is then used to explore the implications of efficient long-run business cycles and long-run regularly cyclical unemployment for long-run endogenous growth. Also, the implications of all kinds of Smithian models for explaining macroeconomic phenomena, such as an endogenous and efficient risk for mass unemployment caused by the trade off between the positive network effects of division of labor on aggregate productivity and

coordination reliability of a larger network of division of labor, are explored. Chapter 24 discusses two Smithian models that examine, respectively, the role of government infrastructure expenditure and financing policies of public spending on the evolution of division of labor.

Part VIII, Political Economics, the Economics of the State, New Economy, and Endogenous Network Size of Division of Labor, covers two Smithian models that look at, respectively, the emergence of the state as the protector of property rights and the effect of monopolistic state, on the level of division of labor and aggregate productivity. Chapter 26 introduces two Smithian models that capture the widespread phenomena of franchising networks and product bundling. The Smithian model of product bundling is particularly relevant to many e-business where many bundled goods are provided free of charge.

Further Reading

Inframarginal economics and inframarginal analysis: Smith (1776), Marshall (1890, chapters 8-12 of Book IV), Young (1928), Stigler (1951, 1976), Houthakker (1956), Rosen (1977, 1983), Becker (1982), Becker and Murphy (1992), Tamura (1992), Borland and Yang (1992), Yang and Y-K. Ng (1993, Ch. 0, 1), Yang (1994, 1996), Ben-Ner (1995), Yang and S. Ng (1998), Cheng and Yang (2004) and references therein, Sun (2005); *Neoclassical microeconomics*: Marshall (1890), Samuelson (1955), Henderson and Quandt (1980), Kreps (1990), Debreu (1991), Stiglitz (1993), Varian (1993), Mas-Colell, Whinston, and Green (1995).

Questions

1. Classical economists never used the concept of economies of scale. They tended rather to think in terms of the benefits of division of labor, and of the transportation and seasonal adjustment costs caused by division of labor. In the last quarter of the 19th century, economists started to use the concept of economies of scale to describe benefit of division of labor on the basis of the neoclassical dichotomy between pure

consumers and firms. Marshall was acutely aware of the inadequacy of the concept of economies of scale for describing phenomenon of economies of division of labor. He suggested the notion of external economies of scale to differentiate the economies of division of labor from the economies of scale of a firm. He also used this notion to try to salvage the classical concept of the invisible hand. However, Allyn Young's student, Frank Knight (1925), pointed out that the notion of external economies of scale involves a logical inconsistency, since economies of scale that are external to all firms is an empty box. Young (1928) indicated that the notion of external as well as internal economies of scale is simply a misrepresentation of the economies of division of labor. He argued (1928, p. 531) that "the view of the nature of the processes of industrial progress which is implied in the distinction between internal and external economies is necessarily a partial view. Certain aspects of those processes are illuminated, while, for that very reason, certain other aspects, important in relation to other problems, are obscured." According to Young, the notion of economies of scale misses qualitative aspect of economies of division of labor, though it may capture the quantitative aspect. Young evidently had in mind what today we call the network effect. He spelt this out as follows (1928, p. 539). "The mechanism of increasing returns is not to be discerned adequately by observing the effects of variations in the size of an individual firm or of a particular industry, for the progressive division of labor and specialization of industries is an essential part of the process by which increasing returns are realized. What is required is that industrial operations be seen as an interrelated whole." Suppose there are three ex ante identical individuals who prefer diverse consumption and specialized production of each of three goods x , y , and z . If an individual chooses to specialize completely in producing x , then she will demand y and z . If she chooses to self-provide x and y and not to produce z , then she has no demand for x and y from the market and will demand z . But if two individuals choose self-sufficiency of all goods, then the other cannot choose specialization. This implies that each person's decision as to her level of specialization not only determines her productivity, but also determines the extent of the market for the produce of others, thereby imposing a constraint on their decisions on their levels of

specialization and productivity. The Young theorem (1928, p. 534, p. 539) explores a typical feature of network effects of the division of labor and related market. The theorem states that “the securing of increasing returns depends on the progressive division of labor. ... not only the division of labor depends upon the extent of the market, but the extent of the market also depends upon the division of labor. ... demand and supply are two sides of the division of labor.” Young also believed that Marshall’s marginal analysis of demand and supply cannot be used to explain network size of division of labor. Discuss why were not Young’s insights into the limitations of the concept of economies of scale formalized and taught in the mainstream textbooks until recently?