ECONOMICS 2535 LECTURE NOTES

Advanced Topics in International Trade: Firms and International Trade

Pol Antràs
Harvard University Department of Economics

Spring 2004

Contents

Preface

These lecture notes review some of the material that I cover in the advanced graduate course in the International Trade that I teach at Harvard University. The course focuses on a firm-level approach to international trade and on selected topics in trade policy. I am teaching this class for the first time this Spring, so the notes are likely to contain several typos and mistakes. Comments, suggestions, and corrections would be most welcome.

Pol Antràs
Department of Economics
Harvard University
January 2004

Chapter 1

Introduction and Basic Facts

- In **Neoclassical Trade Theory**, firms are treated as a black box. The supply side of the economy is characterized by a set of production functions according to which the factors of production (capital, labor) are transformed into consumption goods.

- Moreover, for the most part, the literature assumes constant returns to scale, under which the size of the firm is indeterminate (the general equilibrium only pins down the size of the sector or industry to which the firm belongs).

- **New Trade Theory** introduced increasing returns and imperfect competition in international trade. This resolved the indeterminacy of the size of the firm. As an example, take a Helpman-Krugman type of model with product differentiation. The unique producer of a particular variety ω faces demand given by:

  $$q(\omega) = Ap(\omega)^{-\varepsilon}, \quad \varepsilon > 1$$

  and hence sets $q(\omega)$ to maximize:

  $$\pi(\omega) = A^{1/\varepsilon} q(\omega)^{(\varepsilon-1)/\varepsilon} - \frac{q(\omega)}{\varphi} - f,$$

  where $1/\varphi$ is the marginal cost of production and f is a fixed cost. Profits are strictly concave in $q(\omega)$, so there is a well-determined profit-maximizing level of output $q^*(\omega)$.

- Still, as discussed below, New Trade Theory cannot account for important facts in the data.

A. Firms and the Decision to Export

- The Helpman-Krugman models feature complete specialization: each industry variety is produced by a single firm in just one country, which **exports its output everywhere else in the world**. Adding transport costs could potentially invalidate this, but *not* if transport costs are of the iceberg type. In that case, we still get a similar result (the elasticity of demand remains unaffected). The transport cost inflates the marginal cost and reduces profits on foreign sales:

$$\pi_j(\omega) = A_j^{1/\varepsilon} q_{jj}(\omega)^{(\varepsilon-1)/\varepsilon} + \sum_{k\neq j} A_k^{1/\varepsilon} q_{jk}(\omega)^{(\varepsilon-1)/\varepsilon} - \frac{1}{\varphi}\left(q_{jj}(\omega) + \tau \sum_{k\neq j} q_{jk}(\omega)\right) - f,$$

but one can show that the optimal $q_{jk}^*(\omega)$ satisfies

$$A_k^{1/\varepsilon} q_{jk}^*(\omega)^{(\varepsilon-1)/\varepsilon} - \frac{\tau}{\varphi} q_{jk}^*(\omega) = \frac{1}{\varepsilon} A_k \left(\frac{(\varepsilon-1)\varphi}{\varepsilon\tau}\right)^{\varepsilon-1} > 0 \text{ for all } k \neq j.$$

Hence, even in the presence of transport costs, a firm continues to export everywhere else in the world. As we will see, two features of this example are crucial: (i) that transport costs affect only the marginal cost, and (ii) that foreign competition does not affect the markup the firm can charge over marginal cost.

- In reality, not all domestic producers export to foreign markets. And, more importantly, the literature has "uncovered stylized facts about the behavior and relative performance of exporting firms and plants which hold consistently across a number of countries" (Bernard et al., 2003, BEJK hereafter).

  - Exporters are in the minority. In 1992, only 21% of U.S. plants reported exporting anything.
  - Exporters sell most of their output domestically: around 2/3 of exporters sell less than 10% of their output abroad.

  - Exporters are bigger than non-exporters: they ship on average 5.6 times more than nonexporters (4.8 times more domestically).
  - Plants are also heterogeneous in measured productivity; Figures 2A and 2B in BEJK.
  - Exporters' productivity distribution is a shift to the right of the nonexporter's distribution. Exporters have, on average, a 33% advantage in labor productivity relative to nonexporters.
  - This suggests that the most productive firms self-select into export markets, but it could also reflect learning by exporting (Clerides et al., 1998)

- Furthermore, micro-level studies have also found evidence of substantial reallocation effects within an industry following trade liberalization episodes.

  - Exposure to trade forces the least productive firms to exit or shut-down (Bernard and Jensen, 1999; Aw, Chung and Roberts, 2000; Clerides et al., 1998).
  - Trade liberalization leads to market share reallocations towards more productive firms, thereby increasing aggregate productivity (Pavcnik, 2002, Bernard, Jensen and Schott 2003).

- These studies suggest that successful theoretical frameworks for studying firms and the decision to export should include two features:

  1. Within sectoral heterogeneity in size and productivity.
  2. A feature that leads *only* the most productive firms to engage in foreign trade:
     - This could be a sunk cost of exporting as documented by Roberts and Tybout (1997) and Bernard and Jensen (2004), and formalized by Melitz (2003);
     - Or a limitation on product differentiation (i.e., a fixed measure of goods) that leads to worldwide (price) competition in the production of a particular good, which in turn gives rise to variable markups (BEJK).

- We will study each of these two approaches and revisit the empirical evidence in light of the theories.

B. Firms and the Decision to Invest Abroad

- Another important fact that traditional trade theory neglects is that firms have (at least) two modes of servicing a foreign market. The first mode is the exporting option, which was discussed above. An alternative mode, however, is to set up multiple production plants to service the different foreign markets (i.e. engage in foreign direct investment, FDI hereafter). This trade-off was first formalized by Markusen (1984).

- Multinational firms may also arise when, in the presence of factor price differences across countries, a producer may find it optimal to fragment the production process and undertake different parts of the production process in different countries. This "vertical" approach to the multinational firm was first developed by Helpman (1984).

- Why should we care about multinational firms? Because they play a key role in the global economy:

  - One-third of the volume of world trade is **intrafirm trade**. In 1994, 42.7 percent of the total volume of U.S. imports of goods took place within the boundaries of multinational firms, with the share being 36.3 percent for U.S. exports of goods (Zeile, 1997).
  - About another third of the volume of world trade is accounted for by transactions in which multinational firms are in one of the two sides of the exchange.
  - Still, is this large? Rugman (1988) estimates that the largest 500 multinational firms account for around one-fifth of world GDP.

- Furthermore, some **stylized facts** about multinational firms and FDI (Markusen 1995, 2003) provide foundations for theorizing:

  I. Macro Facts

1. FDI has grown rapidly throughout the world, especially in late 1980s and late 1990s.
2. The bulk of FDI flows between developed countries. In 2000, developed countries were the source of 91 percent of FDI flows and also the recipient of 79 percent (UNCTAD, 2001). Furthermore, 80 percent of the inflows into developing countries went exclusively to Hong Kong, China and Korea.
3. Two-way FDI flows are common between pairs of developed countries.
4. There exists little evidence that FDI is positively related to differences in capital endowments across countries; see, however, Yeaple (2003).
5. Political risk and instability deter inward FDI.

II. Firm and Industry Characteristics

1. The relative importance of multinational firms varies by industry. The significance is higher in sectors that:
    - have high levels of R&D expenditures over sales
    - employ large number of nonproduction workers
    - produce new and/or complex goods
    - have high levels of product differentiation and advertising
    - feature high productivity dispersion (Helpman, Melitz and Yeaple, 2003)
2. At the firm level, multinationality is:
    - negatively associated with plant-level scale economies
    - positively associated with size, up to a threshold size level
    - positively associated with trade barriers.

- We will study different theoretical approaches to explaining these facts. I will refer to these as *technological theories of the multinational firm.*

- Of particular interest will be the contribution by Helpman, Melitz and Yeaple (2003), which combines insights from this branch of the literature together with insights from the literature on within sectoral heterogeneity and the exporting decision discussed above.

- We will also briefly discuss another branch of the literature that has focused on studying the *effects* of FDI.

C. Firm Boundaries: Trade and Organizational Form

- In recent years, we have witnessed a spectacular increase in the way firms organize production on a global scale. Feenstra (1998), citing Tempest (1996), describes Mattel's global sourcing strategies in the manufacturing of its star product, the Barbie doll:

> The raw materials for the doll (plastic and hair) are obtained from Taiwan and Japan. Assembly used to be done in those countries, as well as the Philippines, but it has now migrated to lower-cost locations in Indonesia, Malaysia, and China. The molds themselves come from the United States, as do additional paints used in decorating the dolls. Other than labor, China supplies only the cotton cloth used for dresses. Of the $2 export value for the dolls when they leave Hong Kong for the United States, about 35 cents covers Chinese labor, 65 cents covers the cost of materials, and the remainder covers transportation and overheads, including profits earned in Hong Kong. (Feenstra, 1998, p. 35-36).

- A variety of terms have been used to refer to this phenomenon: the "slicing of the value chain", "international outsourcing", "fragmentation of the production process", "vertical specialization", "global production sharing", and many more.

- One-third of world trade is intrafirm trade, but notice that **multinational firms choose not to internalize an equally sizeable volume of their transactions**. In developing their global sourcing strategies, firms not only decide on where to *locate* the different stages of the value chain, but also on the extent of *control* they want to exert over these processes.

- The internalization issue is nothing more than the classical **"make-or-buy"** decision in industrial organization. Firms may decide to keep the production of intermediate inputs within firm boundaries, thus engaging in FDI when the integrated supplier is in a foreign country, or they may choose to contract with arm's length suppliers for the procurement of these components. An example of the former is *Intel Corporation*, which assembles most of its microchips in wholly-owned subsidiaries in China, Costa Rica, Malaysia, and Philippines. Conversely, *Nike* subcontracts most of the manufacturing of its products to independent producers in Thailand, Indonesia, Cambodia, and Vietnam, while keeping within firm boundaries the design and marketing stages of production.

- The decision to internalize an international transaction also seems to be systematically related to certain industry and country characteristics. For instance, Antràs (2003*a*) reports that the share of intrafirm imports in total U.S. imports is larger in R&D and capital intensive sectors. In a cross-section of exporting countries, this share is also significantly larger in imports from capital-abundant countries.

- Antràs (2003*b*) also reviews some evidence from firm-level studies that suggests that the choice between intrafirm and market transactions is significantly affected by both the degree of standardization of the good being produced abroad and also by the domestic firm's resources devoted to product development.

- The previously discussed approaches to the multinational firm share a common failure to properly model the crucial issue of internalization. These models can explain why a domestic firm might have an incentive to undertake part of its production process abroad, but they fail to explain why this foreign production will occur within firm boundaries (i.e., within multinationals), rather than through arm's length subcontracting or licensing. In the same way that a theory of the firm based purely on *technological* considerations does not constitute a satisfactory theory of the firm (cf., Tirole, 1988, Hart, 1995), a theory of the multinational firm based solely on economies of scale and transport costs cannot be satisfactory either.

- In this section we will instead discuss purely organizational or contractual theories of the multinational firm. We will also review the theories of the firm that serve as basis for these new approaches to the multinational firm.

- Of particular interest will be the contribution by Antràs and Helpman (2003), which combines insights from this branch of the literature together with insights from the literature on intraindustry heterogeneity.

Part I

Firms and the Decision to Export

Chapter 2

Intraindustry Heterogeneity with Fixed Costs of Exporting: Melitz (2003)

- As argued in the Introduction, the available empirical studies suggest that successful theoretical frameworks for studying firms and the decision to export should incorporate intraindustry heterogeneity in size and productivity. This chapter and the next present two recent theoretical frameworks that elegantly introduce such heterogeneity in otherwise standard models of international trade.

- I follow Melitz in discussing first the closed economy model and then moving on to the open economy model.

The Closed Economy Model

- On the **demand** side, there is a representative consumer with preferences:

$$U = \left[\int_{\omega \in \Omega} q(\omega)^{\rho} \, d\omega \right]^{1/\rho}, \quad 0 < \rho < 1, \tag{2.1}$$

where Ω denotes the measure of available products and $\sigma = 1/(1-\rho) > 1$ is the constant elasticity of substitution. We will focus on stationary equilibria, so we drop time subscripts. Consumers maximize (2.1) subject to the budget constraint

$$\int\limits_{\omega\in\Omega} p(\omega)\, q(\omega)\, d\omega = R.$$

It is well-known (prove it yourselves!) that this leads to the following demand function for a particular variety ω:

$$q(\omega) = \frac{R}{P}\left(\frac{p(\omega)}{P}\right)^{-\sigma}, \tag{2.2}$$

where

$$P = \left[\int\limits_{\omega\in\Omega} p(\omega)^{1-\sigma}\, d\omega\right]^{1/(1-\sigma)}.$$

Because consumers value variety, they are willing to consume positive (although lower) amounts of even relatively expensive varieties.

- The **supply** side is characterized by monopolistic competition. Each variety is produced by a single firm (so we hereafter index varieties by φ) and there is free entry into the industry. Firms produce varieties under a technology that features a constant marginal cost and a fixed overhead cost in terms of the unique composite factor of production (labor), which we take as numeraire. The fixed cost is assumed identical across firms and we denote by f. So far the set up is identical to Krugman (1980). Here are the distinguishing features:

1. The **marginal cost is assumed to vary across firms** and is denoted by $1/\varphi$, i.e.

$$TC(\varphi) = f + \frac{q(\varphi)}{\varphi} \tag{2.3}$$

Firms with higher φ are therefore more productive, in the sense that they need to hire fewer workers to attain a given amount of output.[1] Higher productivity firms also charge lower prices, produce more output, and obtain both higher revenues $r(\varphi)$ and higher profits $\pi(\varphi)$. To see this, notice that with CES preferences, the

[1] A higher φ can also be interpreted as higher quality varieties (see Melitz, 2003).

profit-maximizing price is a constant mark-up over marginal cost:

$$p(\varphi) = \frac{1}{\rho\varphi}, \tag{2.4}$$

which by way of (2.2) implies:

$$\begin{aligned} q(\varphi) &= RP^{\sigma-1}(\rho\varphi)^{\sigma} \\ r(\varphi) &= p(\varphi)\,q(\varphi) = R(P\rho\varphi)^{\sigma-1} \qquad (2.5) \\ \pi(\varphi) &= \frac{1}{\sigma}r(\varphi) - f, \qquad (2.6) \end{aligned}$$

where remember that R and P are common across firms.

2. The other additional assumption is that prior to entry, **firms face uncertainty as to how productive they will turn out to be**. In particular, to start producing a particular variety a firm needs to bear a fixed cost consisting of f_e units of labor. Upon paying this sunk cost, the firm draws its productivity level φ from a known distribution with pdf $g(\varphi)$ and associated cdf $G(\varphi)$. After observing this productivity level, the producer decides whether to exit the market immediately or start producing according to the technology in (2.3). In the latter case, in every period, the firm faces a probability δ of exogenous exit, which is common across firms.

- Let us next turn to the **equilibrium** of the closed economy. Consider first **firm behavior**. Since we focus on steady state equilibria, a firm with productivity φ earns profits $\pi(\varphi)$ in each period, until it is hit by a shock, at which point it is forced to exit. Hence, a firm that is contemplating starting production expects a (probability) discounted value of profits of

$$v(\varphi) = \max\left\{0, \sum_{t=s}^{\infty}(1-\delta)^{t-s}\pi(\varphi)\right\} = \max\left\{0, \frac{1}{\delta}\pi(\varphi)\right\}, \tag{2.7}$$

where we impose that if a firm anticipates stationary negative operating profits, it will choose to exit the market upon observing φ. It is clear from (2.6) and (2.7) that there is a unique threshold productivity φ^* such that $v(\varphi) > 0$ if and

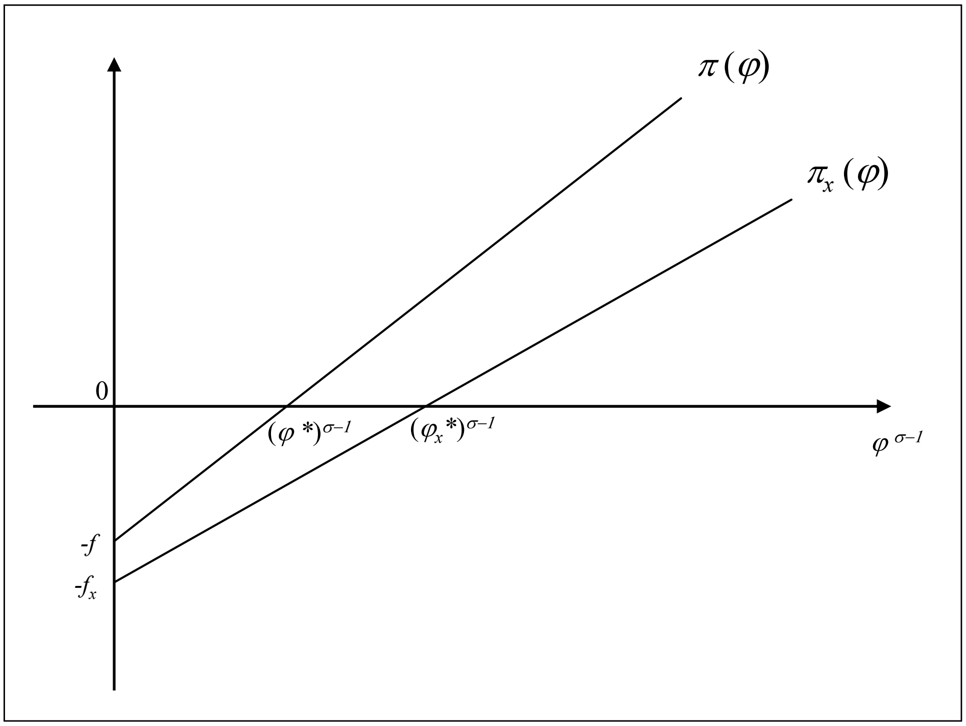


Figure 2.1: Firm Behavior

only if $\varphi > \varphi^*$. This implies that a firm will remain in the market and produce if and only if it is sufficiently productive. Following Helpman, Melitz and Yeaple (2003) and Antràs and Helpman (2003), Figure 2.1 illustrates the equilibrium. Notice that profits are proportional to $\varphi^{\sigma-1}$, and that $\pi(0) = -f$.

- Consider next the **industry equilibrium**, where we solve for the endogenously determined measure M of firms (and varieties), as well as for the distribution of (active firms') productivities in the economy $\mu(\varphi)$. We follow Melitz (2003) in expressing all the equilibrium conditions in terms of the cutoff φ^* and then obtaining the remaining variables of interest from it. For that purpose, it is useful to start by defining the weighted average productivity measure,

$$\widetilde{\varphi} = \left[\int_0^\infty \varphi^{\sigma-1}\mu(\varphi)\, d\varphi\right]^{1/(\sigma-1)},$$

which, as we will see, completely summarizes the relevant information in the distribution of probabilities. Notice that the conditional distribution $\mu(\varphi)$ equals:

$$\mu(\varphi) = \begin{cases} \frac{g(\varphi)}{1-G(\varphi^*)} & \text{if } \varphi \geq \varphi^* \\ 0 & \text{otherwise} \end{cases},$$

from which

$$\widetilde{\varphi}(\varphi^*) = \left[\frac{1}{1-G(\varphi^*)}\int_{\varphi^*}^{\infty} \varphi^{\sigma-1} g(\varphi)\, d\varphi\right]^{1/(\sigma-1)}, \tag{2.8}$$

and hence $\widetilde{\varphi}$ is uniquely pinned down by φ^* and the exogenous (unconditional) distributions $g(\varphi)$ and $G(\varphi)$.

Next, we can define average profits $\overline{\pi} = \pi(\widetilde{\varphi})$ as

$$\overline{\pi} = \frac{r(\widetilde{\varphi})}{\sigma} - f = \left(\frac{\widetilde{\varphi}(\varphi^*)}{\varphi^*}\right)^{\sigma-1} \frac{r(\varphi^*)}{\sigma} - f = f\left(\left(\frac{\widetilde{\varphi}(\varphi^*)}{\varphi^*}\right)^{\sigma-1} - 1\right), \qquad \text{(ZCP)} \tag{2.9}$$

where we have used (2.5), (2.6) and $\pi(\varphi^*) = 0$.[2]

Finally, free entry ensures that, in the industry equilibrium, the *expected* discounted value of profits for a potential entrant equal the fixed cost of entry, or[3]

$$\int_0^{\infty} v(\varphi)\, g(\varphi)\, d\varphi = f_e \Leftrightarrow \overline{\pi} = \frac{\delta f_e}{1-G(\varphi^*)}. \qquad \text{(FE)} \tag{2.10}$$

Notice that (2.9) and (2.10) form a system of two equations in two unknowns $\overline{\pi}$ and φ^*. Because $G'(\varphi^*) > 0$, it is clear that along the FE schedule $\overline{\pi}$ is an increasing function of φ^* and satisfies $\overline{\pi}(0) = \delta f_e$ and $\lim_{\varphi^* \to \infty} \overline{\pi}(\varphi^*) = \infty$. Intuitively, for a given expected value of entry f_e, the probability of success should be decreasing in the average profit level $\overline{\pi}$. Hence, an increase in $\overline{\pi}$ should be matched by an increase in φ^*. On the other hand, Melitz (2003) shows that the FE curve is cut by the ZCP curve only once from above, thus ensuring the existence and uniqueness of the equilibrium. Furthermore, under common distributions, the ZCP schedule is downward sloping in the space $(\overline{\pi}, \varphi^*)$ (see Figure

[2] Notice that we refer to these as average profits because $\overline{\pi} = \int_0^{\infty} \pi(\varphi)\, \mu(\varphi)\, d\varphi$ (go ahead and prove it!).

[3] Notice that $\int_0^{\infty} v(\varphi)\, g(\varphi)\, d\varphi = \int_{\varphi^*}^{\infty} \frac{1}{\delta}\pi(\varphi)\, g(\varphi)\, d\varphi = [1-G(\varphi^*)]\frac{1}{\delta}\int_{\varphi^*}^{\infty} \pi(\varphi)\, \mu(\varphi)\, d\varphi = \frac{1}{\delta}[1-G(\varphi^*)]\,\overline{\pi}$.

2.2 below). Intuitively, an increase in φ^* will increase the average productivity of the surviving firms

$$\widetilde{\varphi}(\varphi^*)' = \frac{g(\varphi^*)\,\widetilde{\varphi}^{2-\sigma}}{(\sigma-1)(1-G(\varphi^*))^2}\int_{\varphi^*}^{\infty}\left(\varphi^{\sigma-1}-(\varphi^*)^{\sigma-1}\right)g(\varphi)\,d\varphi > 0.$$

Because profits tend to increase with a firm's productivity, an increase in φ^* will have a direct positive effect on profits $\overline{\pi}$. But because firm profits are decreasing in the productivity of rivals, there is also an additional effect that goes in the opposite direction. If the distribution $G(\varphi)$ has a fat enough right tail, the latter effect will dominate and the ZCP will be downward sloping. An interesting case is that of a Pareto distribution, i.e., $G(\varphi) = 1-\left(\frac{b}{\varphi}\right)^k$, which yields

$$\begin{aligned}\widetilde{\varphi}(\varphi^*) &= \left[\frac{1}{\left(\frac{b}{\varphi^*}\right)^k}\int_{\varphi^*}^{\infty}\varphi^{\sigma-1}kb\left(\frac{b}{\varphi}\right)^{k-1}d\varphi\right]^{1/(\sigma-1)} = \\ &= \left[k(\varphi^*)^k\int_{\varphi^*}^{\infty}\varphi^{\sigma-k}d\varphi\right]^{1/(\sigma-1)} = \left(\frac{k(\varphi^*)^{\sigma-1}}{\sigma-k+1}\right)^{1/(\sigma-1)},\end{aligned}$$

and the ZCP schedule is flat.

- Once we have the equilibrium values of φ^* and $\overline{\pi}$, we can easily solve for the equilibrium number of firms. Notice that the identical price index in (2.2) becomes simply:

$$P^{1-\sigma} = \int\limits_{\omega\in\Omega} p(\omega)^{1-\sigma}\,d\omega = \int\limits_0^{\infty}(\rho\varphi)^{\sigma-1}M\mu(\varphi)\,d\varphi = M(\rho\widetilde{\varphi})^{\sigma-1},$$

and hence,

$$\overline{\pi} = \frac{1}{\sigma}\frac{R}{M} - f.$$

Finally notice that the equality of income and expenditure ($R = L$) implies that:[4]

$$M = \frac{L}{\sigma(\overline{\pi}+f)}, \tag{2.11}$$

[4] In particular, $R = \Pi + L_p = M\overline{\pi} + L_p = \frac{\delta M}{1-G(\varphi^*)}f_e + L_p = M_e f_e + L_p = L_e + L_p = L.$

which completes the characterization of the stationary equilibrium of the closed economy.

- Notice the following **features** of the equilibrium:
  - $\widetilde{\varphi}, \varphi^{*}$, $\overline{\pi}$ and $\mu(\varphi)$ are independent of L, while M is proportional to country size.
  - Welfare is given by

$$U=\left(\int\limits_{\omega\in\Omega} q(\omega)^{\rho}\,d\omega\right)^{1/\rho}=\left(\int\limits_{\omega\in\Omega}\left(\frac{L}{M(\rho\widetilde{\varphi})^{\sigma-1}}(\rho\varphi)^{\sigma}\right)^{\rho}\mu(\varphi)\,d\varphi\right)^{1/\rho}=LM^{1/(\sigma-1)}\rho\widetilde{\varphi}$$

  - Notice that the aggregate outcome predicted by the model is identical to that generated by a Krugman (1980) model with homogenous firms with productivity $\widetilde{\varphi}$. This shows how nicely the model aggregates the sectoral heterogeneity.

The Open Economy Model

- With this machinery at hand, we can now move to the **open economy** version of the model and analyze the exporting decision as well as the reallocation effects generated by trade. If trade opening is just an increase in the relevant size of the economy, then we know that all firms will export and also, from the equilibrium above, that trade will have no impact on average productivity (see, however, footnote 16 as well as Melitz and Ottaviano, 2003, for the importance of CES preferences for these results). Melitz (2003) thus introduces trade frictions. These are of two types:

  1. A standard per-unit iceberg costs, so that τ units need to be shipped for 1 unit to make it to any foreign country;
  2. An initial fixed cost of f_{ex} units of labor to start exporting, *which is incurred once the firm has learned φ.*

It is also assumed that the domestic economy can trade with $n \geq 1$ other countries and that all countries are of equal size, which implies that factor price equalization will hold and the wage will equal 1 everywhere.

- Let us consider the implications of this extended set-up for **firm behavior**. It is well-known (remember Chapter 1!) that the iceberg transport cost does not affect the elasticity of demand faced by each producer. It follows that firms will again charge a constant markup over marginal cost, but notice that the latter will be higher for exports. Notice that, as in the closed economy, revenues from domestic sales are:

$$r_d(\varphi) = R(P\rho\varphi)^{\sigma-1}$$

where as revenues from foreign sales in country k are:

$$r_x(\varphi) = \tau^{1-\sigma} R_k (P_k \rho\varphi)^{\sigma-1}.$$

As we will see later, the assumption of factor price equalization will imply that $RP^{\sigma-1} = R_k P_k^{\sigma-1}$ for all k, so following Melitz we can express firm revenues by export status as

$$r(\varphi) = \begin{cases} r_d(\varphi) & \text{if the firm does not export} \\ (1 + n\tau^{1-\sigma})\, r_d(\varphi) & \text{if the firm exports to all countries.} \end{cases}$$

As before, profits from domestic sales are simply

$$\pi_d(\varphi) = \frac{r_d(\varphi)}{\sigma} - f, \tag{2.12}$$

while profits from exporting to a particular country are given by

$$\pi_x(\varphi) = \frac{r_x(\varphi)}{\sigma} - f_x = \frac{\tau^{1-\sigma} r_d(\varphi)}{\sigma} - f_x, \tag{2.13}$$

where f_x is amortized per-period portion of the initial fixed cost (i.e., δf_{ex}).[5] Notice that eq. (2.13) is independent of the importing country k, and hence a

[5] Remember that we focus on stationary equilibrium and that the sunk cost of exporting is incurred after φ has been revealed. Hence, the firm will either not export or export in every period.

firm does not export at all or it exports to all countries. Per period profits are therefore $\pi(\varphi) = \pi_d(\varphi) + \max\{0, n\pi_x(\varphi)\}$ while the present discounted value of profits is given by again by (2.7), i.e., $v(\varphi) = \max\{0, \pi(\varphi)/\delta\}$. This now defines two thresholds:

$$\varphi^* = \inf\{\varphi : v(\varphi) > 0\}$$

and

$$\varphi_x^* = \inf\{\varphi : \varphi \geq \varphi^* \text{ and } \pi_x(\varphi) > 0\}.$$

Importantly, because $RP^{\sigma-1}$ is identical in all country, φ^* will also be identical everywhere. Notice that firms with $\varphi \geq \varphi^*$ will remain in the market after learning their productivity, while those with $\varphi \geq \varphi_x^*$ will not only produce domestically, but also export. So long as $\varphi_x^* > \varphi^*$ the model is able to replicate the micro-level findings that the more productive firms within an industry self-select into the export market. This will hold true whenever $\tau^{\sigma-1} f_x > f$, a case illustrated in Figure 2.1.

- **Important:** It is clear that in the model, a higher φ is associated with a higher productivity level. But is it also associated with a higher *measured* productivity level? In Chapter 1, we saw that the evidence indicates that exporters feature a higher value added per worker. One is tempted to identify this with the firm's mark-up, which in Melitz's (2003) model is independent of φ. His model would then not be able to account for heterogeneity in measured productivity. But, in fact, taking account of the fixed costs, one can easily show that:

$$\frac{r_d(\varphi) + nr_x(\varphi)}{q_d(\varphi)/\varphi + n\tau q_x(\varphi)/\varphi + f + f_x} > \frac{r_d(\varphi)}{q_d(\varphi)/\varphi + f} \text{ if and only if } \tau^{\sigma-1} f_x > f,$$

  and hence the model is consistent with the evidence that uses the available measures of productivity. Notice that fixed costs are crucial for this. An alternative route explored by Bernard et al. (2003) is to dispense with fixed costs but introduce a theory that generates variable markups. We will study this alternative approach in Chapter 3.

- We next solve for the industry equilibrium to prepare the ground for the study of how the model can account for the type of trade-induced reallocations stressed by the empirical literature. We are again going to follow Melitz's approach of expressing all the relevant equilibrium conditions in terms of the cut-off φ^*. For that purpose, notice first that from (2.5), (2.13) and the definition of these thresholds,

$$0 = \frac{\tau^{1-\sigma} r_d\left(\varphi_x^*\right)}{\sigma} - f_x = \frac{\tau^{1-\sigma}\left(\varphi_x^*\right)^{\sigma-1}}{\sigma}\frac{r\left(\varphi^*\right)}{\left(\varphi^*\right)^{\sigma-1}} - f_x = \frac{\tau^{1-\sigma}\left(\varphi_x^*\right)^{\sigma-1}}{\left(\varphi^*\right)^{\sigma-1}} f - f_x$$

or

$$\varphi_x^* = \varphi^* \tau \left(\frac{f_x}{f}\right)^{1/(\sigma-1)} .$$

The equilibrium distribution of productivity levels for incumbent firms $\mu\left(\varphi\right)$ is again given by $\mu\left(\varphi\right) = g\left(\varphi\right) / \left[1 - G\left(\varphi^*\right)\right]$ for $\varphi \geq \varphi^*$, while the probability that a surviving firm exports is given by $p_x = \left[1 - G\left(\varphi_x^*\right)\right] / \left[1 - G\left(\varphi^*\right)\right]$. Next, we can define $\widetilde{\varphi}\left(\varphi^*\right)$ and $\widetilde{\varphi}_x\left(\varphi_x^*\right)$ as in (2.8), and again using the same type of weighted average to define

$$\widetilde{\varphi}_t = \left\{\frac{1}{M_t}\left[M\widetilde{\varphi}^{\sigma-1} + nM_x\tau^{1-\sigma}\widetilde{\varphi}_x^{\sigma-1}\right]\right\}^{1/(\sigma-1)} \tag{2.14}$$

where M is the measure of domestic producers, nM_x is the measure of foreign firms that sell in the domestic country, and $M_t = M + nM_x$. $\widetilde{\varphi}_t$ is the average productivity of *all* firms competing in a country. As was the case in the closed economy, the aggregates R and P can be expressed in terms of $\widetilde{\varphi}_t$. Notice the importance of the symmetry assumption, which will ensure that the cutoff φ^*, as well as M and M_x, are identical for all countries, which in turn implies that $\widetilde{\varphi}_t$ is also identical across countries.

Next, we can define average expected profits as[6]

$$\begin{aligned}\overline{\pi} &= \pi_d\left(\widetilde{\varphi}\right)+p_x n\pi_x\left(\widetilde{\varphi}_x\right)= \\ &= f\left(\left(\frac{\widetilde{\varphi}\left(\varphi^*\right)}{\varphi^*}\right)^{\sigma-1}-1\right)+p_x n f_x\left(\left(\frac{\widetilde{\varphi}_x\left(\varphi^*\right)}{\varphi_x^*\left(\varphi^*\right)}\right)^{\sigma-1}-1\right), \quad (\text{ZCP}_t)(2.15)\end{aligned}$$

which is the open-economy analog to (2.9).

Finally, the free entry condition requires the *expected* operating profits for a potential entrant to equal the sunk entry cost

$$\int_0^\infty v\left(\varphi\right)g\left(\varphi\right)d\varphi=f_e\Leftrightarrow\overline{\pi}=\frac{\delta f_e}{1-G\left(\varphi^*\right)}, \qquad (\text{FE}_t) \qquad (2.16)$$

and hence this relationship remains unaltered in the open economy. We again have a system of two equations in two unknowns $\overline{\pi}$ and φ^*, which we plot in Figure 2.2.

To solve for the equilibrium number of firms M, M_x and M_t notice that the M domestic producers together collect a revenue equal to R, while their average revenue is given by

$$\overline{r}=\int_0^\infty r\left(\varphi\right)\mu\left(\varphi\right)d\varphi=\sigma\left(\overline{\pi}+f+p_x n f_x\right)$$

[6] To see this note:

$$\begin{aligned}\overline{\pi} &= \int_0^\infty \pi\left(\varphi\right)\mu\left(\varphi\right)d\varphi= \\ &= \int_0^\infty\left(R\left(P\rho\varphi\right)^{\sigma-1}-f\right)\mu\left(\varphi\right)d\varphi+n\left(\int_{\varphi_x^*}^\infty R\tau^{1-\sigma}\left(P\rho\varphi\right)^{\sigma-1}-f_x\right)\mu\left(\varphi\right)d\varphi= \\ &= R\left(P\rho\right)^{\sigma-1}\widetilde{\varphi}^{\sigma-1}-f+nR\tau^{1-\sigma}\left(P\rho\right)^{\sigma-1}\int_{\varphi_x^*}^\infty\varphi^{\sigma-1}\mu\left(\varphi\right)d\varphi-nf_x\int_{\varphi_x^*}^\infty\mu\left(\varphi\right)d\varphi= \\ &= \pi_d\left(\widetilde{\varphi}\right)+nR\tau^{1-\sigma}\left(P\rho\right)^{\sigma-1}\int_{\varphi_x^*}^\infty\varphi^{\sigma-1}\frac{g\left(\varphi\right)}{1-G\left(\varphi^*\right)}d\varphi-nf_x\int_{\varphi_x^*}^\infty\frac{g\left(\varphi\right)}{1-G\left(\varphi^*\right)}d\varphi= \\ &= \pi_d\left(\widetilde{\varphi}\right)+nR\tau^{1-\sigma}\left(P\rho\right)^{\sigma-1}\frac{1-G\left(\varphi_x^*\right)}{1-G\left(\varphi^*\right)}\widetilde{\varphi}_x^{\sigma-1}-\frac{1-G\left(\varphi_x^*\right)}{1-G\left(\varphi^*\right)}nf_x \\ &= \pi_d\left(\widetilde{\varphi}\right)+p_x n\pi_x\left(\widetilde{\varphi}_x\right)\end{aligned}$$

and thus, imposing the equality of income and spending, we get

$$M = \frac{R}{\overline{r}} = \frac{L}{\sigma\left(\overline{\pi} + f + p_x n f_x\right)} \tag{2.17}$$

and $M_t = (1 + np_x) M$.[7] This completes the characterization of the stationary equilibrium of the open economy.

The Impact of Trade

- Let's follow Melitz and analyze the impact of trade by comparing the stationary equilibria of the closed and open economy. Let φ_a^* and $\widetilde{\varphi}_a$ denote the cut-off and average productivities under autarky (as computed in the closed-economy model). From simple inspection of (2.9) and (2.15), it follows that the ZCP schedule in the open economy is an upward shift of the ZCP schedule under autarky. It thus follows that $\varphi^* > \varphi_a^*$ and $\widetilde{\varphi} > \widetilde{\varphi}_a$, as illustrated in Figure 2.2. Firms with productivity between φ_a^* and φ^* are not able to earn positive operating profits under trade. Consistently with the findings in the empirical literature, exposure to trade thus forces the least productive firms to exit or shut-down (see Chapter 1 and Chapter 4 later on).

- It is important to understand the intuition for this result. Remember that the elasticity of demand is unaffected by trade opening, so the fall in profit for domestic producers is not explained by a fall in mark-ups driven by increased foreign competition.[8] The *actual* channel operates through the domestic factor market. In particular, trade translates into increased profitable opportunities for the rel-

[7] Notice also that the ideal price index can also be computed as follows:

$$\begin{aligned} P^{1-\sigma} &= \int\limits_{\omega\in\Omega} p(\omega)^{1-\sigma} d\omega = \int\limits_{0}^{\infty} (\rho\varphi)^{\sigma-1} M\mu(\varphi)\, d\varphi + \int_{\varphi_x^*}^{\infty} \tau^{1-\sigma} (\rho\varphi)^{\sigma-1} nM_x\mu(\varphi)\, d\varphi = \\ &= M(\rho\widetilde{\varphi})^{\sigma-1} + nM_x\tau^{1-\sigma}(\rho\widetilde{\varphi}_x)^{\sigma-1} = M_t(\rho\widetilde{\varphi}_t)^{\sigma-1}. \end{aligned}$$

[8] Departing from the CES preferences assumption, Melitz and Ottaviano (2003) develop a model in which trade liberalization leads to an increase in the toughness of competition and an associated fall in markups. We will not go into the details of this paper, but this should be required reading for anyone interested in this area.

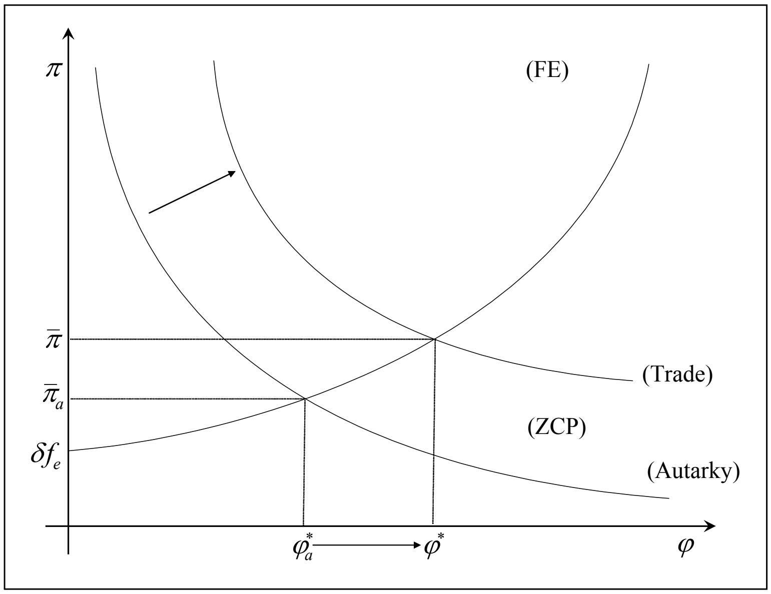


Figure 2.2: The Impact of Trade on the Industry Equilibrium

atively productive firms that can afford the fixed exporting cost. This translates into more entry, thereby increasing labor demand and (given the fixed supply of labor) leading to a rise in the real wage (w/P). This, in turn, brings down the profit level of the least productive firms to a level that forces them to exit.

- Notice also that $\overline{\pi} > \overline{\pi}_a$, which from (2.11) and (2.17), implies that $M < M_a$ and the number of domestic producers will fall. However, as long as τ is not too high, $M_t = (1 + np_x) M > M_a$. And even when this does not hold, Melitz shows that welfare unambiguously goes up (due to aggregate productivity gain, see p. A-3).

- Finally, we are interested in showing that the model can replicate the type of market shares reallocations found in the data. In particular, we want to show that:

$$r_d(\varphi) < r_a(\varphi) < r_d(\varphi) + nr_x(\varphi) \text{ for all } \varphi \geq \varphi^*$$

so that exposure to trade reallocates market shares (which are this figures divided by R) from purely domestic producers to exporters. Notice that since $\varphi^* > \varphi_a^*$,

$$r_d(\varphi) = \sigma f \left(\frac{\varphi}{\varphi^*}\right)^{\sigma-1} < \sigma f \left(\frac{\varphi}{\varphi_a^*}\right)^{\sigma-1} = r_a(\varphi),$$

which proves the first inequality. The second inequality is more cumbersome to establish as its proof requires an analysis of the elasticity of the equilibrium φ^* with respect to τ (see Appendix E in the paper for details).

- Melitz also describes the effects on firm profits and shows that the most efficient firms are those that stand to gain the most from exposure to trade, while a range of exporters see their profits squeeze in spite of the increased market share (exporting brings positive profits but not large enough to compensate for the loss in profits from domestic sales *and* the fixed cost of exporting).

- In the last section, Melitz demonstrates that similar reallocation effects arise in response to smooth variations in τ and n. This is important because there is some tension between the previous comparison of steady states equilibria (which captures long-run consequences of trade) and the type of short-run adjustments unveiled by the empirical literature. The added appeal of these smooth comparative statics comes at the cost of substantially more cumbersome algebra.

Chapter 3

Intraindustry Heterogeneity and Bertrand Competition: Bernard, Eaton, Jensen, and Kortum (2003)

- Bernard et al. (hereafter, BEJK) develop an alternative model of firm heterogeneity along the lines of the probabilistic model of comparative advantage of Eaton and Kortum (*Econometrica*, 2002). Because the firm-level facts that motivate the paper have been discussed in Chapter 1, I focus here on their theoretical framework and simulation results.

Set-up

- On the **demand side**, preferences are symmetric across goods, CES and identical in all N countries, but unlike in Melitz (2003), the measure of goods is fixed at one. Following their notation, expenditure on good j in country n is given by

$$X_n(j) = x_n \left(\frac{P_n(j)}{p_n} \right)^{1-\sigma},$$

where $P_n(j)$ is the price of good j in country n, x_n is total expenditure in n and p_n is the ideal price index

$$p_n = \left[\int_0^1 P_n(j)^{1-\sigma} \, dj \right]^{1/(1-\sigma)}.$$

Notice that unlike Melitz (2003), lower case letters denote aggregates, while upper case letters refer to good-specific variables.

- On the **supply side**, each country has multiple potential producers of good j with varying levels of technical efficiency. As in Melitz (2003), productivity heterogeneity is driven by differences in the marginal cost of production. The kth most efficient producer of good j in country i needs to hire $1/Z_{ki}(j)$ units of the unique composite factor of production (e.g., workers) to produce one unit of the good. There are no fixed costs of production so the technology features constant returns to scale.

  All goods are tradable but $d_{ni} \geq 1$ units of the good need to be shipped from country i for 1 unit to make to country n. It is assumed that $d_{ni} = 1$ and that $d_{ni} \leq d_{nk} d_{ki}$. Notice that the first (second) letter of subscripts denotes the destination (origin) country.

  The composite input is perfectly mobile within countries, but not between them. The cost of such input will therefore generally vary across countries (remember that in Melitz's model countries were identical), and will be denoted by w_i.

  From the previous assumptions, it follows that the kth most efficient producer of good j can deliver the good in country i at unit cost:

$$C_{kni}(j) = \frac{w_i}{Z_{ki}(j)} d_{ni} \tag{3.1}$$

- It is assumed that potential sellers in country n compete à la Bertrand. As with perfect competition, the most efficient (lowest-price) firm captures the market and becomes the only seller in n, but with productivity heterogeneity, the price it can charge will generally be above marginal cost. In particular, this optimal price is

$$P_n(j) = \min\left\{C_{2n}(j), \overline{m} C_{1n}(j)\right\},$$

where

$$C_{1n}(j) = \min_i \{C_{1ni}(j)\} = C_{1ni^*}(j) \leq \min\left\{C_{2ni^*}(j), \min_{i \neq i^*} \{C_{1ni}(j)\}\right\} = C_{2n}(j) \tag{3.2}$$

and

$$\overline{m} = \begin{cases} \sigma/(\sigma-1) & \text{if } \sigma > 1 \\ \infty & \text{otherwise} \end{cases}.$$

Notice that $P_n(j) = C_{2n}(j)$ is more likely the higher the ratio $C_{2n}(j)/C_{1n}(j)$ and the lower the elasticity of substitution σ. If this is the case, the markup will be a function of $C_{2n}(j)/C_{1n}(j)$.

From equations (3.1) and (3.2), it is clear that $C_{2n}(j)/C_{1n}(j)$ will depend on the ratio of these two producers Z's and, when $i \neq i^*$ in eq. (3.2), it will also depend on relative input costs and transport costs.

- Following Eaton and Kortum (2002), BEJK next adopt a probabilistic representation of the relevant efficiency parameters $Z_{1i}(j)$ and $Z_{2i}(j)$. This is a very useful trick because it allows to derive implications for trade flows in terms of the small number of parameters that characterize the underlying probability distribution from which the efficiency parameters are drawn. Notice that is similar in spirit to Melitz's approach, but the BEJK approach is more general in certain aspects (for instance, countries are asymmetric and so are the probability distributions from which the Z's are drawn).

  On the other hand, the BEJK approach is a bit less general in that they focus on a particularly convenient probability distribution. In particular, they assume that for a particular country i, the joint distribution of $Z_{1i}(j)$ and $Z_{2i}(j)$ takes the generalized Fréchet form

  $$F_i(z_1, z_2) = \Pr[Z_{1i} < z_1, Z_{2i} < z_2] = \left[1 + T_i\left(z_2^{-\theta} - z_1^{-\theta}\right)\right] e^{-T_i z_2^{-\theta}}. \tag{3.3}$$

  Notice that:

  - Countries with higher values of T_i will tend to draw higher average values of Z_{1i} and Z_{2i}. T_i is therefore a measure of *absolute advantage*.

  - $\theta > 1$ measures the amount of variability in the distribution. A lower θ implies more variability and thus will strengthen the potential gains from *comparative advantage*.
  - The distribution $F_i(z_1, z_2)$ is independent of j and of $i' \neq i$. This means that unlike in the classical Ricardian world, here countries are not inherently better at producing particular types of goods. Actual comparative advantage is here stochastic.

The Beauty of the Fréchet Distribution

- We will see below that the model is able to account for several stylized facts emphasized in the literature on exporting and productivity, which we mentioned in Chapter 1 and will review at greater length in Chapter 4. Although for some of the results below the Fréchet assumption is actually unnecessary, some derivations below do require that we first spend some time discussing a few properties of the equilibrium distributions of costs and markups across countries.

- Let us compute first the implied joint distribution of the lowest cost C_{1n} and second-lowest cost C_{2n} of supplying some good to country n. It is worth following the proof step by step (the details are not in the paper, but can be found in their Mathematical Appendix, which I closely follow). First, for a destination country n *and* origin country i, notice that for $c_2 \geq c_1$,

$$\begin{aligned} G^c_{ni}(c_1, c_2) &= \Pr\left[C_{1ni} \geq c_1, C_{2ni} \geq c_2\right] = \Pr\left[Z_{1i} \leq \frac{w_i d_{ni}}{c_1}, Z_{2i} \leq \frac{w_i d_{ni}}{c_2}\right] = \\ &= F_i(\frac{w_i d_{ni}}{c_1}, \frac{w_i d_{ni}}{c_2}) = \left[1 + T_i\left(\left(\frac{w_i d_{ni}}{c_2}\right)^{-\theta} - \left(\frac{w_i d_{ni}}{c_1}\right)^{-\theta}\right)\right] e^{-T_i\left(\frac{w_i d_{ni}}{c_2}\right)^{-\theta}} = \\ &= \left[1 + T_i\left(w_i d_{ni}\right)^{-\theta}\left(c_2^\theta - c_1^\theta\right)\right] e^{-T_i(w_i d_{ni})^{-\theta} c_2^\theta} \end{aligned} \tag{3.4}$$

Next, from the definitions in (3.1), as well as from (3.3) and (3.4),

$$\begin{aligned}
G_n^c(c_1,c_2) &= \Pr\left[C_{1n}\geq c_1, C_{2n}\geq c_2\right] = \\
&= \underbrace{\prod_{i=1}^{N} G_{ni}^c\left(c_2,c_2\right)}_{\substack{\text{Prob all countries have}\\ C_{1ni}\geq c_2\ \&\ C_{2ni}\geq c_2}} + \sum_{i=1}^{N} \underbrace{\left[G_{ni}^c\left(c_1,c_2\right) - G_{ni}^c\left(c_2,c_2\right)\right]}_{\text{Extra Prob } C_{1n}\geq c_1 \text{ (remember } c_2>c_1)} \times \underbrace{\prod_{k\neq i}^{N} G_{nk}^c\left(c_2,c_2\right)}_{\substack{\text{Prob } k\neq i \text{ still have}\\ C_{1nk}\geq c_2\ \&\ C_{2nk}\geq c_2}} = \\
&= \prod_{i=1}^{N} e^{-T_i(w_i d_{ni})^{-\theta} c_2^{\theta}} + \sum_{i=1}^{N}\left[T_i\left(w_i d_{ni}\right)^{-\theta}\left(c_2^{\theta}-c_1^{\theta}\right) e^{-T_i(w_i d_{ni})^{-\theta} c_2^{\theta}}\right] \prod_{k\neq i}^{N} e^{-T_k(w_k d_{nk})^{-\theta} c_2^{\theta}} = \\
&= e^{-\Phi_n c_2^{\theta}} + e^{-\Phi_n c_2^{\theta}}\left(c_2^{\theta}-c_1^{\theta}\right)\Phi_n
\end{aligned}$$

where,

$$\Phi_n = \sum_{i=1}^{N} T_i\left(w_i d_{ni}\right)^{-\theta}.$$

Finally,

$$\begin{aligned}
G_n(c_1,c_2) &= \Pr\left[C_{1n}\leq c_1, C_{2n}\leq c_2\right] = \\
&= 1 - G_n^c\left(0,c_2\right) - G_n^c\left(c_1,c_1\right) + G_n^c\left(c_1,c_2\right) = \\
&= 1 - e^{-\Phi_n c_1^{\theta}} - \Phi_n c_1^{\theta} e^{-\Phi_n c_2^{\theta}}. \qquad (3.5)
\end{aligned}$$

The simple form of (3.5) illustrates the usefulness of the Fréchet distribution in characterizing the joint distribution of extreme values, such as C_{1n} and C_{2n} (the same of course is true in a univariate set up, such as in Eaton and Kortum, 2002). Furthermore, as in Melitz (2003), the choice of functional forms permits an elegant aggregation of the inherent heterogeneity in these models. In particular, Φ_n captures all the relevant information on the efficiency distributions, input costs, and trade costs around the world, and the joint distribution of c_1 and c_2 is independent of the *actual* sources of supply to country n.

- In the derivations below, we will also make use of the following results, which are relatively easy to derive using the joint cdf in eq. (3.5) (see their Mathematical Appendix for details).

- The marginals distribution of the lowest cost and second lowest suppliers to country n are given by:

$$G_{1n}\left(c_1\right) = \lim_{c_2 \to \infty} G_n(c_1, c_2) = 1 - e^{-\Phi_n c_1^\theta},$$

and

$$G_{2n}\left(c_2\right) = \lim_{c_1 \to c_2} G_n(c_1, c_2) = 1 - \left(1 + \Phi_n c_2^\theta\right) e^{-\Phi_n c_2^\theta}.$$

- The markup of the unique seller in country n is the realization of a random variable M_n drawn from a Pareto distribution truncated at $\overline{m}$. To see this define $M_n' = C_{2n}/C_{1n}$ so that $M_n = \min\left\{M_n', \overline{m}\right\}$. Notice first that:

$$\begin{aligned} \Pr\left[M_n' \leq m' | C_{2n} = c_2\right] &= \Pr\left[c_2/m' \leq C_{1n} < c_2 | C_{2n} = c_2\right] = \\ &= \frac{\int_{c_2/m'}^{c_2} \left(\frac{dG_n(c_1,c_2)}{dc_1 dc_2}\right) dc_1}{dG_{2n}\left(c_1\right)/dc_2} = \\ &= 1 - m'^{-\theta}, \end{aligned}$$

which is Pareto and independent of c_2. It follows that the unconditional distribution $H_n\left(m\right) = \Pr\left[M_n \leq m\right]$ is also Pareto but truncated at $\overline{m}$. Notice that the distribution of the markup is again independent of the *actual* supply source.

Implications for Productivity, Exporting and Size

- We are now ready to study how the model is able to account for some of the stylized facts on productivity and exporting.

1. **Productivity:** Because technology is CRTS and there are no fixed costs of production, differences in measured productivity across firms (differences in value added per worker) reflect only differences in their markups (remember the importance of fixed costs in Melitz's CES setup). In particular, notice value added

per worker of a firm selling in country n from country i^* is given by:

$$\frac{P_n(j)\,Q_n(j)}{Q_n(j)\,C_{1n}(j)/w_{i^*}} = M_n(j)\,w_{i^*}$$

It easy to see that the model implies that, on average, plants that are more efficient charge a **higher markup**. In particular, conditional on a productivity level z_1, the distribution of the markup is:[1]

$$H_n(m|z_1) = \Pr\left[M_n \leq m | Z_{1n} = z_1\right] = \begin{cases} 1 - e^{-\Phi_n w_n z_1^{-\theta}\left(m^{\theta}-1\right)} & 1 \leq m < \overline{m} \\ 1 & m \geq \overline{m} \end{cases}$$

Hence, actual productivity Z_{1n} and measured productivity M_n are in line. Intuitively, the more efficient the lowest-cost firm, the more likely it is that the difference between C_{1n} and C_{2n} is relatively large, and hence the more likely it is that the charged markup is relatively high.

2. **Exporting:** Consider the lowest-cost producer of good j in country i. To sell domestically, its $Z_{1i}(j)$ need only satisfy:

$$Z_{1i}(j) \geq \varphi^* = \max_{k \neq i}\left\{Z_{1k}(j)\,\frac{w_i}{w_k d_{ik}}\right\}.$$

On the other hand, to sell in some other market n requires

$$Z_{1i}(j) \geq \varphi^*_x = \max_{k \neq i}\left\{Z_{1k}(j)\,\frac{w_i d_{ni}}{w_k d_{nk}}\right\}.$$

It follows from the *triangle inequality* that $\varphi^*_x > \varphi^*$. In words, **only a fraction** of those that sell at home **will also export abroad**. This is simply explained by the fact that, because of transports costs, exporting anywhere imposes a higher efficiency hurdle than selling only at home. Notice that **this is independent of the Fréchet assumption**. The crucial features of the model for this result

[1] This follows from,

$$\Pr\left[M'_n \leq m' | C_{1n} = C_1\right] = \frac{\int_{c_1}^{m'c_1}\left(\frac{dG_n(c_1,c_2)}{dc_1 dc_2}\right)dc_2}{dG_{1n}(c_1)/dc_1} = 1 - e^{-\Phi_n c_1^{\theta}\left(m'^{\theta}-1\right)},$$

and $C_{1n} = \frac{w_n}{Z_{1n}(j)}$ (the efficiency $Z_{1n}(j)$ is inclusive of the transport cost).

are that (i) it is too costly for firms to differentiate their products and that (ii) potential producers of a good compete in prices, thus leading to sales by only the lowest-cost deliverer in a given country.

Furthermore, the assumption that the Fréchet distribution is independent of j, also implies that **exporting firms** in a given country will, on average, appear to be **more productive** than firms that sell only domestically.

3. **Size:** The distribution of the second-lowest cost, which if $M_n < \overline{m}$ determines the price, conditional on the lowest cost is given by:

$$\Pr\left[C_{2n} \leq c_2 | C_{1n} = c_1\right] = \frac{\int_{c_1}^{c_2} \left(\frac{dG_n(c_1,c_2)}{dc_1 dc_2}\right) dc_2}{dG_{1n}\left(c_1\right)/dc_1} = \frac{dG_n(c_1,c_2)/dc_1}{dG_{1n}\left(c_1\right)/dc_1} = 1 - e^{-\Phi_n\left(c_2^\theta - c_1^\theta\right)},$$

and is therefore stochastically increasing in c_1 – remember that $F(x)$ first order stochastically dominates $G\left(x\right)$ if and only $F(x) < G(x)$ for all x. This means that the more efficient is the least-cost deliverer to country n, the higher will on average be the productivity of the second least-cost firm. This, in turn, implies that because, on average, firms that export from a given country i are more productive than firms in country i that do not export, exporters will on average charge lower prices to domestic buyers than nonexporting firms. For a higher-than-one elasticity of substitution, $\sigma > 1$, it follows that **exporters will on average be bigger** (have higher domestic sales) than non-exporters.[2]

One might find confusing that, on the one hand, exporters tend to charge relatively high markups, whereas, on the other hand, they tend to charge relatively low prices. Notice, however, that this just reflects that the lower is C_{1n} (i.e., the more efficient is the least-cost producer), the lower is C_{2n} also (and hence the lower is the price), but the higher is C_{2n}/C_{1n} (and thus the higher is the markup).

Quantification and Counterfactuals

- As we have just seen, the model is able to qualitatively replicate some of the findings in the empirical literature on productivity and exporting. BEJK next

[2] This discussion has presumed $M_n < \overline{m}$, but it is straightforward to show that the same is true when $M_n = \overline{m}$.

explore whether the model also does a good **quantitative** job. For this purpose, the authors first show that simulating the model only requires data on bilateral trade shares π_{ni} between any two countries, total consumption (or absorption) x_n for each country n, as well as values for the parameters σ and θ.[3]

Defining the transformations

$$\begin{aligned} U_{1i}(j) &= T_i Z_{1i}(j)^{-\theta} \\ U_{2i}(j) &= T_i Z_{2i}(j)^{-\theta}, \end{aligned}$$

one can show that $U_{1i}(j)$ and $U_{2i}(j)$ are random variables drawn from a parameter-free distribution characterized by:

$$\begin{aligned} \Pr\left[U_{1i}(j) \leq u_1\right] &= 1 - e^{-u_1} \\ \Pr\left[U_{2i}(j) \leq u_2 | U_{1i}(j) = u_1\right] &= 1 - e^{-u_2+u_1}. \end{aligned} \tag{3.6}$$

The authors then implement the following algorithm (see paper for more details):

1. They draw $U_{1i}(j)$ and $U_{2i}(j)$ from (3.6) for 47 countries and 1,000,000 goods j.

2. For each destination country n and good j, they identify the source country i^* from:
$$i^* = \arg\min_i \left\{ \frac{U_{1i}(j)}{\pi_{ni}} \right\},$$
where π_{ni} is the ratio of i's exports to n divided by n's total absorption.

3. Letting $i^* = USA$ and identifying a good j with a plant, this delivers a simulated sample of active U.S. active plants and their **export status**.

4. For each plant j, information on $U_{1i}(j)$ and $U_{2i}(j)$ (as well as π_{ni} for all i) is sufficient to compute the markup charged in country n, from which **sales**, **exports**, and **total production costs** can easily be computed (see eq. 16 and 17 in their paper).

[3] In particular, data on the other parameters of the model w_i, T_i and d_{ni} are not needed.

5. Finally, assuming that the composite factor of production is a Cobb-Douglas aggregator of wages and intermediate inputs (which are themselves an aggregate of all the j's), permits computation of **employment** and **value added per worker**.

- The procedure is repeated for different values of σ and θ searching for the values that deliver the same productivity advantage of exporters (remember from Chapter 1 that their value added per worker is 33% higher) and the same size advantage (on average, exporters ship domestically 4.8 times more than non-exporters). This yields $\sigma = 3.79$ and $\theta = 3.60$.

  With these parameter values, it is assessed how well the model fits the other facts regarding exporting and productivity. The model does a pretty good job and in particular the simulations match the skewness of the distribution of export intensity, with most exporters selling only a small fraction of their output abroad. Nevertheless, the model tends to overpredict the fraction of firm that export and underpredict the variability in productivity and in size.

- The final section of the paper performs a couple of counterfactual experiments: (i) a 5% worldwide decline in trade costs, and (ii) a 10% increase in U.S. wages (U.S. dollar real appreciation). The results of (i) indicate that the model is also able to replicate the positive aggregate productivity effects of trade liberalization documented by Pavcnik (2002) and others. In particular, as in Melitz (2003), the model features reallocation effects by which the fall in trade costs leads to exit of relatively unproductive domestic producers and expansion of relatively productive exporters.[4] Unlike in Melitz (2003), however, the model also generates substantial productivity gains within surviving firms driven by the decline in the price of intermediate inputs.

Limitations and Extensions

- An implication of the framework is that the number of country i's exporters

[4] Remember, however that the mechanism is different. Here, exit is a direct consequence of intensified price competition from foreign suppliers.

to country n should vary proportionately with the market share of country i in country n's imports. Eaton, Kortum and Kramarz (2003) show that this feature is not borne by data on French firms. In particular, they find that for a given level of French market share in country n, the number of exporters is significantly higher, the higher country n's market share. To account for this interesting finding, they modify the BEJK set-up by including fixed costs of exporting as well as Cournot competition.

Chapter 4

Firms and the Decision to Export: Empirics

- In this chapter, we will first briefly discuss a few recent empirical studies on the link between exporting and plant-level performance. We will then study in more depth a particularly interesting recent paper by Pavcnik (2002) on the reallocation effects of trade liberalization. In the next chapter, we will cover additional empirical papers on the relevance of sunk costs of exporting.

4.1 Exporting and Plant-Level Performance

- Several studies have documented the **superior performance** characteristics of exporting plants and firms relative to non-exporters. As argued in Chapter 1, BEJK report that, on average, U.S. exporting plants sell 4.8 times more than non-exporting U.S. plants domestically, and have, on average, a 33% advantage in labor productivity relative to non-exporting plants. Similar results are reported in Bernard and Jensen (1999), where it is also shown that U.S. exporters tend to employ more workers, pay higher wages, operate at a higher capital-labor ratio and record higher TFP levels.

- Other studies have shown that similar patterns emerge in other countries. Bernard and Wagner (2001) show that, in a sample of German plants, exporters are significantly bigger and have higher labor productivity than non-exporters in the same region (Lower Saxony). Similarly, Aw, Chung and Roberts (2000) compute significantly higher multifactor productivity levels for Taiwanese and Korean plants that export than for plants that do not export.

These findings raise at least three issues:

1. First, the majority of the studies fail to measure appropriately plant-level productivity. Certainly, labor productivity is an informative measure, but differences in labor productivity could simply reflect differences in capital intensity between exporting and non-exporting firms, which is precisely what a Hecksher-Ohlin model would predict if the U.S. or Germany are capital-abundant countries. TFP measures do not suffer from this problem but simple "Solow-residual type" computations still yield biased estimates of plant-specific productivity, as they fail to account for (potentially) important simultaneity and selection biases. We will elaborate on this when we discuss the paper by Pavcnik (2002) below.

2. Second, even when productivity is appropriately measured, exporters could be more productive because of other plant-specific characteristics that make them both more productive and more likely to export. Clerides, Lach and Tybout (1998) acknowledge this problem and provide evidence that, even when productivity levels are purged of industry-wide time effects and observable plant-specific characteristics, exporting plants in Colombia, Mexico and Morocco, tend to have lower *residual* average costs and higher *residual* labor productivity than non-exporting plants.

3. Still, the correlations between productivity and exporting do not necessarily reflect a causal link *from* productivity *to* exporting. Indeed, an older literature reviewed in Clerides et al. (1998) stressed the potential productivity enhancing effects of exporting. In the presence of learning by exporting, exporting firms might be more productive *because* they export. Clerides et al. (1998), Bernard and Jensen (1999), and Aw et al. (2000) have proposed different methodologies for studying more systematically the causal link between exporting and productivity. Interestingly, all these studies find substantial support for the self-selection mechanism formalized by Melitz (2003) and BEJK (2003), and find little evidence

of the existence of significant learning-by-exporting. In particular, although the authors use different methodologies to disentangle the direction of causation, their tests basically reveal that past performance levels significantly impact current export market participation, whereas past export market participation has no significant effect on current measures of productivity.

To be fair, Aw et al. (2000) find that in certain Taiwanese industries, past export status has a significant positive effect on current productivity levels. Clerides et al. (1998) also find a significant effect of past export status in some cases, but of the wrong sign! Finally, Bernard and Jensen (1999) find that other measures of plant perfomance (e.g., survival rates) respond positively to past export market participation.

4.2 Evidence on Reallocation Effects: Pavcnik (2002)

- We consider next Pavcnik's (2002) careful analysis of the effects of trade liberalization in Chile on plant-level and industry-level productivity. Pavcnik's contribution consists of three parts:

  1. The construction of measures of plant-level productivity using the methodology developed by Ericson and Pakes (1995) and Olley and Pakes (1996). This technique is a close cousin of the Solow-residual type of computations used by other authors, but uses regression analysis to structurally correct for potential biases caused by the simultaneity of input choice and by the non-random nature of entry and exit. As Pavcnik shows, these biases turn out to be important in her sample.

  2. With these measures at hand, Pavcnik next attempts to identify the effect of trade liberalization on productivity using both time-series *and* cross-sectoral variation in the extent to which different sectors were affected by the dismantling of trade barriers. This allows to separate the effects of trade liberalization on plant productivity from the effects of other policies. Her findings suggest significant productivity improvements related to liberalized

trade.

3. Finally, Pavcnik aggregates productivity levels across plants in a given industry and finds that the reallocation of market shares from less to more efficient plants accounts for a significant fraction of the industry-wide productivity increases.

Let us study each of these parts in more detail.

The Measurement of Productivity

- Pavcnik considers the following model of firm behavior. Profits of a plant i in industry j at time t are given by

$$\Pi_{ijt} = f\left(k_{ijt}, \omega_{ijt}\right)$$

where k_{ijt} is its capital stock and ω_{ijt} is a plant-specific productivity level that is known by the plant but not by the econometrician. The plant solves the problem

$$V_t\left(\omega_t, k_t\right) = \max\left\{L_t, \sup \Pi_t\left(\omega_t, k_t\right) - c\left(i_t\right) + dE\left[V_{t+1}\left(\omega_{t+1}, k_{t+1}\right)|\Omega_{it}\right]\right\},$$

where L_t is the liquidation value of the plant, $c\left(i_t\right)$ is the cost of investment, d is the discount factor, and Ω_{it} reflects the information available to the plant at time t. The capital stock evolves according

$$k_{t+1} = \left(1-\delta\right)k_t + i_t.$$

Assuming that ω_{ijt} evolves according to a first-order Markov process, Ericson and Pakes (1995) show that the solution of this problem takes the form of:

a. a threshold exit rule, by which a plant will decide to exit if

$$\omega_t < \underline{\omega}_t\left(k_t\right), \tag{4.1}$$

where notice the dependence of the threshold productivity on the plant's capital stock; and

b. an investment rule of the form

$$i_t = i_t(\omega_t, k_t). \tag{4.2}$$

- Consider the implications of this set up for the measurement of productivity. Let i's plant technology at time t be given by

$$y_t = \beta_0 + \beta x_t + \beta_k k_t + e_t, \tag{4.3}$$

where x_t is a vector of variable intermediate inputs,

$$e_t = \omega_t + \mu_t,$$

and μ_t has zero mean. The standard method for computing ω_{it} is to run eq. (4.3) under OLS and then compute

$$\widehat{\omega}_t = y_t - \widehat{\beta}_0 - \widehat{\beta} x_t - \widehat{\beta}_k k_t.$$

The problem with this approach is that the model developed above points out that the estimate of $\widehat{\beta}_k$ is likely to be biased because the investment rule depends on productivity (simultaneity bias) and because the exit rule depends on the stock of capital (selection bias). Pavcnik corrects for these biases in the following way:

1. To deal with the simultaneity bias, she starts by inverting (4.2), plugging it back in (4.3) to obtain

$$y_t = \beta x_t + \lambda_t(k_t, i_t) + \mu_t, \tag{4.4}$$

and approximating $\lambda_t(\cdot)$ with a polynomial series expansion in capital and investment. This yields consistent estimates of the coefficient on variable inputs $\widehat{\beta}$.

Notice that in order to invert (4.2) it is necessary for i_t to be positive and strictly monotonic in ω_t. In her sample, it turns out that many plants report zero investment. In order to check that this does not significantly biases her results, she reports empirical estimates obtained when using only observations with non-zero investment and compares them to the benchmark estimates. The estimates indeed seem to be similar in both cases.

2. Next, in order to identify the coefficient $\widehat{\beta}_k$ she notices that

$$\begin{aligned} y_{t+1} - \beta x_{t+1} &= \beta_0 + \beta_k k_{t+1} + \omega_{t+1} + \mu_{t+1} = \\ &= \beta_0 + \beta_k k_{t+1} + E\left[\omega_{t+1} | \omega_t, k_{t+1}\right] + \xi_{t+1} + \mu_{t+1} = \\ &= \beta_k k_{t+1} + g\left(\omega_t\right) + \xi_{t+1} + \mu_{t+1} = \\ &= \beta_k k_{t+1} + g\left(\lambda_t\left(k_t, i_t\right) - \beta_k k_t\right) + \xi_{t+1} + \mu_{t+1}, \end{aligned} \tag{4.5}$$

where it has been used that ω_t follows a first-order Markov process. Because ξ_{t+1} is the unanticipated part of ω_t, this methodology yields consistent estimates of β_k.

3. Still this methodology needs to be modified to correct for the selection bias. Notice that conditional on a plant staying in the market, the expectation of future productivity is in fact

$$E\left[\omega_{t+1} | \omega_t, k_{t+1}, \omega_{t+1} > \underline{\omega}_{t+1}\left(k_{t+1}\right)\right] = \Phi\left(\omega_t, \underline{\omega}_{t+1}\right) - \beta_0.$$

Because $\underline{\omega}_{t+1}$ is likely to increase in k_{t+1}, estimates of β_k from (4.5) are likely to be downward biased. To deal with this, Pavcnik first estimates the probability of staying in the market

$$P_t = \Pr\left[\omega_{t+1} > \underline{\omega}_{t+1}\left(k_{t+1}\right)\right] = p_t\left(\underline{\omega}_{t+1}\left(k_{t+1}\right), \omega_t\right) = p_t\left(\underline{\omega}_{t+1}\left(k_t, i_t\right), \omega_t\right) = p_t\left(k_t, i_t\right),$$

where the capital accumulation equation and (4.2) have been used. In particular, she runs a Probit regression where the function $p_t\left(k_t, i_t\right)$ is again approximated with a polynomial series expansion in capital and investment. With this probability at hand, she then runs

$$y_{t+1} - \beta x_{t+1} = \beta_k k_{t+1} + \Phi\left(\lambda_t\left(k_t, i_t\right) - \beta_k k_t, P_t\right) + \xi_{t+1} + \mu_{t+1}, \tag{4.6}$$

where again $\Phi(\cdot)$ is approximated with a polynomial series.

- Pavcnik estimates equation (4.6) at an industry disaggregation between the two and three digit ISIC levels. Her results suggest that correcting for these biases is indeed important as her semiparametric estimates of β_k are between 40% to over 300% higher than those obtained under simple OLS.

Trade Liberalization and Plant-Level Productivity

- Pavcnik next attempts to identify the effect of trade liberalization on plant-level productivity by estimating the following regression:

$$Prod_{it} = \alpha_0 + \alpha_1 (Time) + \alpha_2 (Trade) + \alpha_3 (Trade * Time) + \alpha_4 Z_{it} + \upsilon_{it},$$

  where $Prod_{it}$ is the plant-level productivity measure computed from the estimated input coefficients in (4.4) and (4.6); $Time$ is a vector of time effects (1979 is excluded); $Trade$ is a vector of dummies indicating the trade orientation of a plant, i.e., whether it belongs to an export-oriented or import-competing sectors (non-traded goods sector are the excluded category); $Trade * Time$ is a vector of interactions; and Z_{it} is a set of plant-specific controls, that includes a dummy indicating whether a plant ceased to produce in that particular year.

- The idea is that firms in the tradable sector are more likely to experience improvements in productivity as a result of trade liberalization. This is clear in the case of firms in import-competing industries, which we would expect to "trim their fat" once they cease to be shielded from foreign competition. For the case of firms in the export-oriented sector, one might expect a differential impact of trade liberalization relative to the non-tradable sector, but the sign of the difference is far from clear.

- Although her sample starts in 1979, a year already within the time of the trade liberalization period, Pavcnik appeals to lags in the response of firms to justify the expectation of a positive sign in the sign of the coefficient α_3.[1] Notice that

[1] Notice that it is also necessary to appeal to lags in order to avoid complications in the inter-

α_1 ideally purges out the effects of other policies or shocks that might have affected all plants, *independently* of the industry in which they are classified. Similarly, α_2 leaves out permanent differences in productivity across different types of industries.

- Her results suggest that the interaction $ImportCompeting * Time$ is indeed positive and significant, and furthermore the coefficients seem to increase through time, indicating divergent productivity trends between firms in the import-competing sector and firms in the nontradable sector. In particular, her estimates suggests that "the productivity gains for plants in the import-competing sector attributable to liberalized trade range from 3% to 10.4 %". On the other hand, she finds much weaker evidence of a positive coefficient on the interaction $ExportOriented * Time$, as one might have expected.

  She also finds that the coefficient on exit is significant and negative, and its magnitude implies that exiting plants are on average 8.1% less productive than surviving plants.

The Significance of Reallocation Effects

- The third main contribution of Pavcnik's paper consists in documenting the significance of reallocation effects in accounting for growth in productivity in Chilean plants following trade liberalization. In particular, with the measure of plant-level productivity constructed using the Ericson and Pakes (1995) methodology, she computes a weighted aggregate productivity measure in a particular industry, which can be decomposed into an unweighted aggregate measure and a covariance term:

$$W_t = \sum_i s_{it} Prod_{it} = \overline{Prod}_t + \sum_i \left(s_{it} - \overline{s}_t\right)\left(Prod_{it} - \overline{Prod}_t\right).$$

  The covariance term will be positive whenever firms with above average productivity gain market share. To the extent that reallocation effects are important,

pretation of the interaction terms for the years 1983 and 1984, in which tariffs in fact increased substantially.

we would expect the second term to play an important role in explaining the growth in aggregate productivity. This is indeed confirmed by the data. For instance, for overall manufacturing, the covariance term accounts for almost 2/3 of total productivity growth between 1979 and 1986. In particular, aggregate productivity grew 19.3%, of which 12.7 percentage points are explain by reshuffling of resources from less productive to more productive firms (see Table 3). Similar results are found for most two-digit ISIC industries, as well as for the different types of industries classified by their trade orientation. Interestingly, these raw measures also present a picture similar to the regression results discussed above. While in import-competing sectors aggregate productivity grew by 31.9% between 1979 and 1986 (with 21.3%, that is around 2/3, being explained by the covariance term), in the export-oriented and nontradable sectors productivity grew by 25.4% and 6.2% respectively.

- Recent work by Bernard, Jensen and Schott (2003) also attempts to find evidence for reallocation effects using data from U.S. plants. Although they choose to proxy productivity by simple labor productivity, with no correction for simultaneity or selection biases, their approach has the virtue of exploiting cross-industry variation in the extent to which trade barriers have fallen through time. Interestingly, they find evidence that productivity growth has been faster in industries with falling trade costs. They also report other cross-sectional correlations between falling trade barriers and certain firm level facts that are consistent with the predictions of the Melitz and BEJK models.

Chapter 5

The Relevance of Sunk Costs

- In Melitz's (2003) model, the assumption of the existence of sunk costs of exporting is crucial to predict the self-selection of the most productive firms into export markets, as well as the reallocation effects within an industry following trade liberalization. Intuitively, after trade opening, only the most productive firms will obtain revenues large enough to be able to amortize the sunk cost of exporting. Furthermore, the increased profit opportunities generated by exporting lead to increased entry and labor demand, which push up the real wage and force the least productive firms to shut down.

  Melitz's approach is very rich in several dimensions, but leaves aside certain features that make sunk costs of exporting relevant. In particular, the fact that (after the initial draw) the firm's productivity remains constant through time implies that operating profits are also constant through time and therefore a firm does not export or exports in every period.

- In this section, we will focus on situations in which an exporter's operating profits are stochastic and will show that this generates additional interesting predictions. In particular, we will sketch how the coupling of sunk costs and uncertainty gives rise to **hysteresis** in export markets.

A Simple Model of Hysteresis: Dixit (1989*a*)

- Consider the following illustrative example from Dixit (1989*a*), which I adapt to the case of an exporter using Melitz's (2003) notation. Imagine that, after paying a sunk cost f_{ex}, a firm can obtain an exogenous flow π_x of profits per unit of time by selling their product in foreign markets. Assume that, provided that the firm exports in every period, this investment is required just once. Assume, however, that if the firm ceases to export in a particular year, an additional sunk cost will be required in order to resume exporting. Let δ be the rate of interest.

- Suppose firm that the exporter has yet not incurred the sunk cost and believes that π_x will persist unchanged forever. As in Melitz (2003), the firm will then make the investment if $\pi_x > \delta f_{ex} = f_x$. Conversely, if the firm has *already* incurred the fixed cost, notice that the firm will cease exporting only if $\pi_x < 0$.[1] Hence, there is a range of profit levels for which a potential entrant decides **not to enter** the export market, while an actual exporter decides **not to cease** exporting. This is known as the range or band of inaction.

- To see how this band of inaction generates hysteresis we can think about the following simple dynamics. Initially, exporting profits are $f_x > \pi_x > 0$, so the firm does not export. At some point in time, the firm changes its expectations and anticipates $\pi_x > f_x$ ad infinitum. This leads the firm to enter the exporting market. Then at some later point in time, it reverts back to the expectation of the initial constant profit flow $f_x > \pi_x > 0$. Notice, however, that this profit flow will now not induce this firm to exit the exporting market.

- If you think about profits as being a monotonic function of productivity, this example illustrates the fact that, *conditional on productivity*, the exporting decision at any point in time is likely to be a function of **past exporting status**. Notice also, that this is the case in the presence of sunk costs, i.e., only when $f_x > 0$.

- One could think that unless f_x is unreasonable high, the band of inaction will be relatively small and hysteresis might not be expected to be empirically significant. But in fact, the band of inaction can be much larger than the previous example

[1] Notice that in Melitz (2003) this never occurs because he focuses on stationary equilibria in which $\pi_x > f_x > 0$.

suggests. Notice, that in the previous example, the exporter's expectation were irrational: how can you expect constant profits in the presence of frequent shocks!

- To illustrate how the proper modelling of expectations can amplify the band of inaction, consider the following example. Suppose that the initial profit flow is $\pi_x = f_x$ and that, from then on, at each point in time it will take equal steps up or down with equal probabilities. Assume that the firm forms his expectations on the evolution of π_x rationally. In such case, if the firm invests immediately and continues active forever its expected present value of profits net of sunk costs is zero. It is easy to see, however, that by waiting one period the firm can do better in expected terms. In particular, if profits go up in this next period, the firm can start exporting and expect a positive expected present value of profits net of sunk costs. What about if profits go down? Notice that in such case the firm will decide not to invest, so its expected *payoff* will remain at zero. Weighting each of these possible outcomes by 1/2, delivers a strictly positive expected payoff from not exporting in the initial period.

- Notice, however, that for $\pi_x > f_x$, the expected payoff from exporting also becomes positive, and for some $\overline{\pi_x} > f_x$, it will necessarily be the case that the firm will export provided that $\pi_x > \overline{\pi_x}$. An analogous argument can be used to argue that a firm that has incurred the sunk cost will only cease exporting if $\pi_x < \underline{\pi_x} < 0$. With rational expectations, the band of inaction is therefore given by $\overline{\pi_x} - \underline{\pi_x} > f_x$.

- The Dixit (1989*a*,*b*) papers formalize this logic assuming that π_x evolves exogenously over time as a Brownian motion with drift. He also present simulations that indicate that hysteresis can be significant even when the sunk costs f_x are small.

Empirical Evidence: Roberts and Tybout (1997)

- As pointed out before, if some of the costs involved in exporting are sunk in nature, we should expect the decision to export to be a function of prior exporting status even after controlling for a whole set of firm or plant-specific characteristics.

The contribution of Roberts and Tybout (1997) consists precisely in formally testing this hypothesis with data on a large sample of manufacturing plants in Colombia in the period 1981-1989.

- Roberts and Tybout (1997) develop a very simple partial equilibrium framework in which the export-market participation of a given plant i at time t is modelled using the dynamic discrete-choice equation (see the paper for details):

$$Y_{it} = \begin{cases} 1 & \text{if } \mu_t + \boldsymbol{\beta}\mathbf{Z}_{it} + \gamma^0 Y_{i,t-1} + \sum_{j=2}^{J} \gamma^j \widetilde{Y}_{i,t-j} + \varepsilon_{it} \geq 0 \\ 0 & \text{otherwise.} \end{cases}, \tag{5.1}$$

where

$$\widetilde{Y}_{i,t-j} = Y_{i,t-j} \prod_{k=1}^{j-1} (1 - Y_{i,t-k})$$

and:

  - $Y_{it} = 1$ only if plant i exports in period t;
  - μ_t is a time effect meant to capture variations in export profitability that are common across all plants (e.g., exchange rate movements, trade policy).
  - $\mathbf{Z}_{it}$ is a vector of plant-specific variables that includes industry dummies, capital stock and age, but not a direct measure of productivity.
  - $\widetilde{Y}_{i,t-j}$ takes a value of one if the plant was last on the export market j years earlier and 0 otherwise.

- Robert and Tybout run equation (5.1) and then test for the joint significance of γ^0 and the γ^j's, while dealing skillfully with a whole set of econometric issues (see the paper for details). Their results suggest that exporting history indeed matters (the Wald statistic is well above its critical value). In particular, a plant that exported in the previous year is up to 60 percentage points more likely to export in the current year than a comparable plant that has never exported. Their estimates also suggest that the effects of prior exporting quickly vanish through time, and a plant that has been out of the exporting market for two or more years is not significantly more likely to export in the current year. Interestingly, they also find that the probability of exporting is increasing in the age of the plant and also in its size, as proxied by the plant's capital stock (notice that this is consistent with the self-selection mechanism in Melitz, 2003).

- In recent work, Bernard and Jensen (2004) have used a similar methodology to study the relevance of sunk costs for U.S. manufacturing plants. They posit a linear probability model analogous to the specification in (5.1) and they also find that past export status is a significant predictor of current export status. In particular, in their sample, exporting today raises the probability of exporting tomorrow by as much as 66%. Contrary to Roberts and Tybout (1997), their vector of plant-specific controls also includes a direct measure of productivity which they construct using the Olley-Pakes methodology (as in Pavcnik, 2002). This productivity measure turns out to have a very significant positive effect on the probability of exporting. Bernard and Jensen (2004) also run their linear probability model with plant fixed effects and find that this reduces the effect of prior exporting history, which however remains highly significant (with fixed effects, exporting today increases the probability of exporting tomorrow by 39%).

- These papers provide evidence of the statistical significance of sunk costs. But how large are these costs in U.S. dollars? Answering this question requires laying out a structural model from which these values can be backed out. Recently, Das, Tybout and Roberts (2001) have estimated that, in their sample of Colombian plants, the expected sunk costs from breaking into exporting markets may well exceed \$ 1 million.

Part II

Firms and the Decision to Invest Abroad

Chapter 6

Horizontal FDI: Brainard (1997)

- In the models we have discussed so far, a firm was allowed to serve the foreign market only through exports. In addition, it was assumed that the whole production process was undertaken in the domestic economy, so that all trade was in final goods.

- The evidence suggests instead that firms frequently choose to service foreign markets through local production by a subsidiary, thus becoming multinational firms. The literature refers to this arrangement as **horizontal** foreign direct investment (FDI hereafter). Furthermore, multinational corporations account for a very significant fraction of world trade flows, with trade in intermediate inputs between divisions of the same firm constituting an important portion of these flows (c.f., Hanson, Mataloni and Slaughter, 2001). People refer to this phenomenon as **vertical** FDI.

- Why do some firms choose to engage in foreign direct investment (FDI) to service a foreign market rather than focus on simple exporting? Why do firms sometimes choose to break up the production process across borders rather than keeping all stages in the home country and simply exporting the final good? The next three Chapters will discuss possible answers to the first question.

- Let us start with a simple theoretical model along the lines of Brainard (1997), which I extend and modify to facilitate a comparison with the work of Helpman, Melitz and Yeaple (2003) discussed below.

Set-up

- The world consists of two countries, H and F, that use labor to produce goods in $M+1$ sectors. One sector produces a homogenous good z, which we take as the numeraire, while the remaining M sectors produce a continuum of differentiated products.

- On the **demand** side, each country is inhabited by a representative consumer with identical preferences:

$$U = \left(1 - \sum_{m=1}^{M} \beta_m\right) \log z + \sum_{m=1}^{M} \frac{\beta_m}{\alpha_m} \log \left(\int_{v \in V_m} x_m (v)^{\alpha_m} \, dv \right), \quad 0 < \alpha_m < 1, \tag{6.1}$$

  where $x_m(v)$ is consumption of variety v in sector m, V_m denotes the measure of available products in that sector, and $\varepsilon_m = 1/(1-\alpha_m)$ is the elasticity of substitution across varieties. Notice that (6.1) implies that consumers spend a fraction β_m of their income on sector m's varieties, and the remaining fraction $1-\sum_m \beta_m$ on the homogenous z. Because preferences feature a unit elasticity of substitution across varieties in different sectors, we can focus on sector-by-sector analysis and safely drop the subscript m. As we have seen in previous papers, maximizing (6.1) subject to the budget constraint $z + \sum_m \int_{v \in V_m} p_m(v) x_m(v) \, dv \leq E^i$ yields the following demand for each variety in a given sector

$$x(v) = \frac{\beta E^i}{\int_{v \in V} p(v)^{1-\varepsilon} \, dv} p(v)^{-\varepsilon} = A^i p(v)^{-\varepsilon}, \quad i = H, F.$$

- In this section, we will focus on the case in which both countries are endowed with L units of labor. Furthermore, both countries have access to an identical constant-returns-to-scale technology for producing good z. By choice of units, producing one unit of good z requires exactly one worker. Assume that $\sum_m \beta_m$ is small enough so as to ensure that good z is produced in every country. As a result, the wage rate is equal to one in every country.

- On the **supply** side, the invention (or process of differentiation) of a particular variety requires a fixed cost of f_E units of labor. f_E will be a measure of firm-level economies of scale and is unaffected by the number of plants producing the good (see Markusen, 1984). These are also plant-level economies of scale. In particular, setting up a production plant entails a fixed cost of f_D units of labor, regardless of where this plant is set up. Later on, we will generalize this to the case in which it might be more costly to set up a plant in a foreign country. Furthermore, there is also a marginal cost of production equal to 1 for every plant in every sector and every country.

- Goods that are exported are subjected to iceberg costs by which τ units of the good need to be shipped for each unit actually delivered.

- The differentiated-good sectors are characterized by monopolistic competition. Each variety is produced by a single firm and there is free entry into the industry, so that profits net of all fixed costs are driven down to zero.

Exports vs. FDI

- Given this setup, a firm producing a particular variety will service its domestic market from a domestic plant. On the other hand, this faces a choice between servicing the foreign market through exports or rather by setting up a plant in that other country. The first option entails higher transport costs, while the second option saves on plant-level fixed costs.

- Notice that because the marginal cost of production (excluding the transport cost) is identical in all these arrangements, the optimal price will be $p = 1/\alpha$ for goods sold in the country where they are produced and $p = \tau/\alpha$ otherwise.

- Defining
$$B^i = (1-\alpha)\,\alpha^{\varepsilon-1}A^i$$
an exporter in country i will obtain profits equal to
$$\pi_X^i = B^i + \tau^{1-\varepsilon}B^j - f_E - f_D, \tag{6.2}$$

while a firm from i that engages in horizontal FDI in j will instead obtain

$$\pi_I^i = B^i + B^j - f_E - 2f_D. \tag{6.3}$$

- Let us consider different possible types of equilibria and analyze under which parameter configuration they apply:

1. **Equilibrium with Pervasive Exporting in Each Country**

   In this case no firm engages in FDI, and equation (6.2) yields

   $$\pi_X^H = B^H + \tau^{1-\varepsilon} B^F - f_E - f_D$$

   and

   $$\pi_X^F = B^F + \tau^{1-\varepsilon} B^H - f_E - f_D.$$

   Imposing free entry yields the equilibrium demand levels

   $$B^H = B^F = B_X = \frac{f_E + f_D}{1 + \tau^{1-\varepsilon}}, \tag{6.4}$$

   from which the measure of firms can be derived using $B^i = (1-\alpha)\beta L / \left(n^i + \tau^{1-\varepsilon} n^j\right)$:

   $$n^H = n^F = \frac{(1-\alpha)\beta L}{f_E + f_D} > 0.$$

   Remember from Chapter 1, that without fixed costs of exporting, n^H and n^F will also be the number of exporters in the corresponding country.

   In order to ensure that this is indeed an equilibrium, we need to check that no firm has an incentive to deviate from this equilibrium and set up a plant in the foreign country. Given the continuum assumption, this deviatior would have no impact on the demand level and hence, using (6.3) and (6.4), her profits would be given by

   $$\pi_I^i = 2\left(\frac{f_E + f_D}{1 + \tau^{1-\varepsilon}}\right) - f_E - 2f_D$$

which is negative if and only

$$\frac{f_D}{f_E + f_D} > \frac{1 - \tau^{1-\varepsilon}}{1 + \tau^{1-\varepsilon}}, \tag{6.5}$$

which is more likely to hold the higher are plant-specific economies of scale relative to firm-specific economies of scale, f_D/f_E, and the lower are transport costs τ.

2. **Equilibrium with Pervasive FDI in Each Country**

In this case no firm exports from the home country and equation (6.3) yields

$$\pi_I^H = B^H + B^F - f_E - 2f_D$$

and

$$\pi_I^F = B^F + B^H - f_E - 2f_D.$$

Free entry then imposes

$$B^H + B^F = f_E + 2f_D,$$

while, from $B^i = (1-\alpha)\,\alpha^{\varepsilon-1}A^i$, it follows that $B^i = (1-\alpha)\,\beta L/\left(n^H + n^F\right)$ and hence,

$$B^H = B^F = B_I = \frac{f_E + 2f_D}{2}.$$

The number of firms in each country is given by

$$n^H = n^F = \frac{(1-\alpha)\,\beta L}{f_E + 2f_D} > 0.$$

For this to be an equilibrium, we again need to ensure that no firm in any country will find it profitable to switch to exporting. In either case, this would yield profits equal to

$$\pi_X^i = \left(1 + \tau^{1-\varepsilon}\right)\left(\frac{f_E + 2f_D}{2}\right) - f_E - f_D,$$

which are negative if and only

$$\frac{f_D}{f_E + f_D} < \frac{1 - \tau^{1-\varepsilon}}{1 + \tau^{1-\varepsilon}},$$

which is the converse of (6.5) and is therefore more likely to hold the lower f_D/f_E

and the higher trade costs.

3. **Mixed Equilibria in Each Country**

   Consider next the case in which some firms export and some firms engage in FDI in each country. For all firms to break even, as dictated by free entry, it would need to be the case that:

$$\begin{aligned}
\pi_X^H &= B^H + \tau^{1-\varepsilon} B^F - f_E - f_D = 0 \\
\pi_X^F &= B^F + \tau^{1-\varepsilon} B^H - f_E - f_D = 0 \\
\pi_I^H &= B^H + B^F - f_E - 2f_D = 0 \\
\pi_I^F &= B^F + B^H - f_E - 2f_D = 0
\end{aligned}$$

   which requires

$$B^H = B^F = \frac{f_E + f_D}{1 + \tau^{1-\varepsilon}} = \frac{f_E + 2f_D}{2},$$

   which can only hold in a knife-edge case, i.e. when (6.5) holds with equality. Using $B^i = (1-\alpha)\,\alpha^{\varepsilon-1} A^i$ and imposing labor market clearing, it is also possible to show that in this equilibrium, the number of exporters and multinational firms needs to be identical in both countries. But the actual number of firms of each type remains indeterminate.

- **Other Equilibria?**

  One may wonder whether additional equilibria involving pervasive exporting in one country and pervasive FDI in the other are possible. In particular, suppose now that all firms in country H export and all firms in country F undertake FDI. Then the relevant profit functions are

$$\pi_X^H = B^H + \tau^{1-\varepsilon} B^F - f_E - f_D$$

  and

$$\pi_I^F = B^H + B^F - f_E - 2f_D.$$

Free entry now implies

$$\begin{aligned} B^H &= f_E + f_D - \frac{\tau^{1-\varepsilon} f_D}{1-\tau^{1-\varepsilon}} \\ B^F &= \frac{f_D}{1-\tau^{1-\varepsilon}} \end{aligned}$$

This in turn implies the following conditions determining the measure of firms in each country:

$$\begin{aligned} B^H &= \frac{(1-\alpha)\beta L}{n^H + n^F} = f_E + f_D - \frac{\tau^{1-\varepsilon} f_D}{1-\tau^{1-\varepsilon}} \\ B^F &= \frac{(1-\alpha)\beta L}{\tau^{1-\varepsilon} n^H + n^F} = \frac{f_D}{1-\tau^{1-\varepsilon}} \end{aligned}$$

which yield

$$\begin{aligned} n^H &= \left(\frac{f_D\left(1+\tau^{1-\varepsilon}\right) - (f_D+f_E)\left(1-\tau^{1-\varepsilon}\right)}{(f_D+f_E)\left(1-\tau^{1-\varepsilon}\right) - f_D\tau^{1-\varepsilon}} \right) \frac{(1-\alpha)\beta L}{f_D} \\ n^F &= \left(\frac{(f_D+f_E)\left(1-\tau^{1-\varepsilon}\right) - 2f_D\tau^{1-\varepsilon}}{(f_D+f_E)\left(1-\tau^{1-\varepsilon}\right) - f_D\tau^{1-\varepsilon}} \right) \frac{(1-\alpha)\beta L}{f_D}. \end{aligned}$$

That the denominator of this expressions is positive is implied by $B^H > 0$. Hence, for $n^H > 0$ it need be the case that

$$\frac{f_D}{f_E + f_D} > \frac{1-\tau^{1-\varepsilon}}{1+\tau^{1-\varepsilon}}.$$

On the other hand, in order to rule out a deviation by a firm in F wishing to switch to exporting it need be the case that

$$\begin{aligned} \pi_X^F &= B^F + \tau^{1-\varepsilon} B^H - f_E - f_D = \\ &= \frac{f_D}{1-\tau^{1-\varepsilon}} + \tau^{1-\varepsilon}\left(f_E + f_D - \frac{\tau^{1-\varepsilon} f_D}{1-\tau^{1-\varepsilon}} \right) - f_E - f_D = \\ &= \left(1+\tau^{1-\varepsilon}\right) f_D - (f_E + f_D)\left(1-\tau^{1-\varepsilon}\right) < 0 \end{aligned}$$

or simply

$$\frac{f_D}{f_E + f_D} < \frac{1-\tau^{1-\varepsilon}}{1+\tau^{1-\varepsilon}},$$

which is inconsistent with $n^H > 0$. Hence, this can't be an equilibrium. The case with firms engaging in FDI in country H and exporting in country F is entirely symmetric.

Empirical Implementation

- The model above has the stark prediction that in a given sector, the share of exports over foreign affiliate sales will be:

$$\frac{X}{S+X}=\begin{cases} 0 & \text{if } \frac{f_D}{f_E+f_D}<\frac{1-\tau^{1-\varepsilon}}{1+\tau^{1-\varepsilon}} \\ [0,1] & \text{if } \frac{f_D}{f_E+f_D}=\frac{1-\tau^{1-\varepsilon}}{1+\tau^{1-\varepsilon}} \\ 1 & \text{if } \frac{f_D}{f_E+f_D}>\frac{1-\tau^{1-\varepsilon}}{1+\tau^{1-\varepsilon}} \end{cases}.$$

- To smooth out this prediction, Brainard (1997) focuses on the range of parameters for which the equilibrium is mixed and discusses how the share of firms that choose to export varies with certain parameters of the model. Her discussion here is actually quite puzzling. On the one hand, in a mixed equilibrium the share of firms that export is actually indeterminate (see above). On the other hand, given that the parameter space for which this equilibria exists is of measure zero, her comparative statics are not meaningful.

- An alternative way to smooth out the prediction is to simply assume that the statistician disaggregates the industry data under a criterion different from the one dictated by the model (see the NBER working paper version of Antràs, 2003*a*, for details). It is straightforward to show that, in in the recorded data, the fraction of firms that export in a given industry will smoothly increase in f_D/f_E and smoothly fall in τ. The latter result will also apply to the ratio of exports to foreign affiliate sales, because the ratio of export revenues to FDI sales is simply given by:

$$\frac{f_E+f_D-B_X}{f_E+2f_D-B_I}=\frac{2\tau^{1-\varepsilon}\left(f_E+f_D\right)}{\left(1+\tau^{1-\varepsilon}\right)\left(f_E+2f_D\right)}$$

which is decreasing decreasing in τ. Notice, however, that this ratio is also decreasing in f_D/f_E, which complicates the particular comparative static with respect to f_D/f_E. Intuitively, industries with a higher ratio f_D/f_E will have more exporters, but foreign affiliates will be relatively larger.

- One last alternative would be to appeal to firm-level heterogeneity within an industry (where here the theorist and the statistician agree on what constitutes an industry). In particular, firms could have idiosyncratic *preferences* over exporting and FDI. This would generate a non-degenerate distribution of exporters and FDI within an industry but the average number of exporters would tend to increase in f_D/f_E and decrease in τ. Furthermore, so long as all firms within an industry share the same parameters f_D/f_E and τ, the relative size (sale revenues) of exporters and foreign affiliates could well be independent of these parameters. This similar to the approach followed by Helpman, Melitz and Yeaple (2003), which will be discussed in Chapter 8.

Empirical Results

- Leaving these caveats aside, Brainard (1997) starts by proposing an econometric model in which the share of total U.S. sales (exports + affiliate sales) in industry m and country i that is accounted for by exports is regressed on: (i) industry measures of plant-specific and firm-specific economies of scale; (ii) industry and country specific measures of trade costs (tariffs and freight costs); and (iii) a set of controls related to the importing country, such as GDP per capita or corporate tax rates. Her data is from 1989 and she can exploit both the cross-industry as well as cross-country variation.

- Her results lend support to the proximity-concentration hypothesis. She first presents OLS, country random effects, industry random effects, and generalized Tobit estimates, the latter to address the large number of zero affiliate sales in her sample. The results are quite robust to the specification and indicate that:

  - The coefficient on both tariffs and freight costs appear to be negative and significant;
  - Her measure of plant-level economies of scale (number of nonproduction employees in the median U.S. *plant* ranked by value added in each industry)

turns out with a positive and significant coefficient, while the converse os true for her measure of *firm*-level economies of scale (number of nonproduction workers in the average U.S.-based firm in each industry).

  - The coefficient on the difference in GDP per capita between the foreign country and the U.S. is positive and significant. This suggest that differences in factor proportions tend to foster exporting more than FDI.

- She also runs the specification with country and industry fixed effects, thus dropping all variables but the measures of transport costs. The results become weaker, but the estimates are still for the most part of the right sign.

- Next, Brainard repeats the exercise with the share of U.S. imports over U.S. imports plus sales of U.S. affiliates of foreign firms. Her results are again broadly consistent with the theory, although the transport cost measures lose some of their significance (especially tariffs, but notice this measure now varies only across industries).

- Finally, Brainard runs regressions of exports and affiliates sales in levels, although the results cannot easily be interpreted in light of the theoretical model developed above, in which the effect of f_D/f_E and τ applied only to the ratio of exports to total sales.

Chapter 7

Exports vs. FDI with Asymmetric Countries: Markusen and Venables (2000)

- Brainard's (1997) set up elegantly illustrates the trade-off between proximity and concentration and provides a theoretical foundation for the negative impact of trade costs and the positive impact of plant-level economies of scale on the ratio of export sales to foreign affiliate sales. Nevertheless, the simplified framework in Chapter 6 cannot account for Brainard's (1997) finding that this ratio is also increasing in factor endowment differences.

- Markusen and Venables (2000) generalize the previous set up to account for this important fact. In particular, they develop a $2 \times 2 \times 2$ (two-factor, two-sector, two-country) Helpman-Krugman model of international trade with monopolistic competition and product differentiation, which they extend to include positive transport costs and endogenous multinational firms.[1]

Set-up

- Markusen and Venables (2002) consider a setup similar to that in the simple

[1] They also build on Markusen and Venables (1998), who consider an oligopolistic set-up with homogeneous goods.

model above, but with the following modifications (I choose to stick to the notation above, rather than to the one in their paper):

1. There are only two sectors, i.e., $M = 1$. Preferences are assumed to be homothetic.
2. Production of both the homogeneous good z and the varieties in the differentiated sector X require capital and labor, with production of z being relatively labor intensive.
3. Country $i = H, F$ is endowed with L^i units of labor and K^i units of capital. Let L^W and K^W denote the world supplies of each of these factors.
4. An exporter incurs all its fixed costs in its home country, and this fixed costs have the same factor intensity as variable costs. These are denoted by $b(w^i, r^i) f$, where $b(w^i, r^i)$ is the marginal cost of production (hence, f is equal to $f_E + f_D$ in the model above).
5. A multinational firm incurs a portion of the fixed costs in the home country and another portion in the foreign country (notice that because of FPE this was immaterial in the previous model). These fixed costs are denoted by $\varphi\left(b\left(w^H, r^H\right), b\left(w^F, r^F\right)\right) g$.

- The analysis is somewhat more cumbersome than in the symmetric model in Chapter 6, so in the following discussion I make the following additional assumptions:
  - Preferences are Cobb-Douglas and given by

$$U = (1-\beta)\log z + \beta \log \left(\int\limits_{v \in V_m} x_m(v)^{\alpha_m} dv \right), \quad 0 < \alpha_m < 1. \qquad (7.1)$$

  - Production of one unit of z requires only one unit of labor, while production of differentiated varieties requires only capital and $b\left(w^H, r^H\right) = r^H$.
  - The fixed costs for a MNE are $\varphi\left(r^H, r^F\right) = a\left(r^H + r^F\right)/2 + (1-a)\left(r^H r^F\right)^{1/2}$. Notice that this specification treats the two countries symmetrically, thus

implying that the "home" and "host" countries are both indeterminate and irrelevant. We will thus denote by n_I the measure of multinational plants.

Firm Behavior

- Imagine the case in which market i, H or F, is supplied by n_X^i domestic producers who also export to country j, n_X^j foreign producers and n_I plants belonging to multinational firms. After straightforward manipulations, profits for an exporter from country i can be expressed as:

$$\pi_X^i = \left(r^i\right)^{1-\varepsilon} B^i + \left(\tau r^i\right)^{1-\varepsilon} B^j - r^i f,$$

where

$$B^i = (1-\alpha)\frac{\beta E^i}{\left(r^i\right)^{1-\varepsilon} n_X^i + \left(\tau r^j\right)^{1-\varepsilon} n_X^j + \left(r^i\right)^{1-\varepsilon} n_I}$$

while those of a multinational firms are

$$\pi_I = \left(r^i\right)^{1-\varepsilon} B^i + \left(r^j\right)^{1-\varepsilon} B^j - \varphi\left(r^i, r^j\right) g.$$

If both exporters from each country are active, this implies

$$\begin{aligned} \left(r^H\right)^{1-\varepsilon} B^H + \left(\tau r^H\right)^{1-\varepsilon} B^F &= r^H f \\ \left(r^F\right)^{1-\varepsilon} B^F + \left(\tau r^F\right)^{1-\varepsilon} B^H &= r^F f, \end{aligned}$$

from which B^H and B^F need to satisfy

$$B^H = \frac{\left(\left(r^H\right)^{\varepsilon} - \tau^{1-\varepsilon}\left(r^F\right)^{\varepsilon}\right) f}{1-\tau^{2(1-\varepsilon)}} \tag{7.2}$$

$$B^F = \frac{\left(\left(r^F\right)^{\varepsilon} - \tau^{1-\varepsilon}\left(r^H\right)^{\varepsilon}\right) f}{1-\tau^{2(1-\varepsilon)}}. \tag{7.3}$$

On the other hand, for MNE to be active in equilibrium, it need to be the case that:

$$\left(r^H\right)^{1-\varepsilon} B^H + \left(r^F\right)^{1-\varepsilon} B^F \geq \varphi\left(r^H, r^F\right) g.$$

Plugging the values of B^H, B^F and $\varphi\left(r^H, r^F\right)$ yields

$$\frac{r^H\left(1-\tau^{1-\varepsilon}\left(\frac{r^F}{r^H}\right)^{\varepsilon}\right)+r^F\left(1-\tau^{1-\varepsilon}\left(\frac{r^H}{r^F}\right)^{\varepsilon}\right)}{\left(1-\tau^{2(1-\varepsilon)}\right)\left(a\left(r^H+r^F\right)/2+(1-a)\left(r^H r^F\right)^{1/2}\right)} \geq \frac{g}{f}$$

or

$$\Phi\left(\tau, r^H/r^F\right)=\frac{\frac{r^H}{r^F}\left(1-\tau^{1-\varepsilon}\left(\frac{r^H}{r^F}\right)^{-\varepsilon}\right)+\left(1-\tau^{1-\varepsilon}\left(\frac{r^H}{r^F}\right)^{\varepsilon}\right)}{\left(1-\tau^{2(1-\varepsilon)}\right)\left(a\left(\frac{r^H}{r^F}+1\right)/2+(1-a)\left(\frac{r^H}{r^F}\right)^{1/2}\right)} \geq \frac{g}{f}. \qquad (7.4)$$

Proposition 1 *Multinational firms are more likely to emerge in equilibrium when:*

(i) firm-level economies of scale are high (f is high);

(ii) plant-level economies of scale are low (g is low);

(iii) transport costs are high (τ is high);

(iv) factor price differences are small (r^F/r^H is close to one).

Proof. Points (i) and (ii) are straightforward. For (iii), it suffices to show that $\partial\Phi\left(\tau, r^F/r^H\right)/\partial\tau > 0$, while for (iv) we want to show that $\Phi\left(\tau, r^H/r^F\right)$ attains a maximum at $r^H/r^F = 1$. Letting, $\rho = \tau^{1-\varepsilon}$ and $x = r^H/r^F$, notice that:

$$\Phi(\rho, x)=\frac{x\left(1-\rho(x)^{-\varepsilon}\right)+\left(1-\rho(x)^{\varepsilon}\right)}{\left(1-\rho^2\right)\left(a(x+1)/2+(1-a)(x)^{1/2}\right)}.$$

Simple differentiation yields

$$\frac{\partial\Phi(\rho, x)}{\partial\rho}=-\frac{\left(2\rho\left(1+x^{2\varepsilon-1}-x^{\varepsilon-1}(1+x)\right)+(1-\rho)^2\left(1+x^{2\varepsilon-1}\right)\right)}{\left(1-\rho^2\right)^2 x^{\varepsilon-1}\left(a(x+1)/2+(1-a)(x)^{1/2}\right)}<0,$$

where the inequality follows from $1+x^{2\varepsilon-1}-x^{\varepsilon-1}(1+x)>0$ for all $x>0$ and $\varepsilon>1$.[2] It thus follows that $\partial\Phi\left(\tau, r^F/r^H\right)/\partial\tau > 0$. Finally, for point (iv) it suffices to show that $\Phi(\rho, x)$ is strictly concave in x and that $\partial\Phi(\rho, x)/\partial x = 0$ if $x = 1$. The latter is straightforward to check; the former is left as an exercise. ■

[2] To see this, notice that $1+x^{2\varepsilon-1}-x^{\varepsilon-1}(1+x)$ increases in ε for $\varepsilon \geq 1$, and takes a value of 0 at $\varepsilon = 1$.

- The intuition for (i), (ii) and (iii) is as in Brainard (1997). In particular, (iii) follows from the fact that for exporting to be profitable when transport costs are high, then the demand levels B^H and B^F must be large (notice that B^H and B^F are increasing in τ in (7.2) and (7.2)). But this high demand levels make it more likely for multinationals to also be profitable.

- As for (iv), the crux is that whereas exporting firms only use factors (capital here) from one country, multinational firms use factors from both countries. It thus follows that the higher are factor price differences, the higher is the cost-disadvantage faced by multinational firms with respect to exporters from the low factor-price country.

General Equilibrium

- Markusen and Venables (2000) next solve for the general equilibrium in which income equals spending and factor markets clear. The main purpose is to link the likelihood of observing multinational firms to relative factor endowment differences. First, however, they derive some interesting results even in the case without multinational firms.

- Notice that if we let $g \to \infty$ and $\tau \to 1$, the model converges to a standard Helpman-Krugman model of international trade with extreme factor intensity (cf, Ventura, 1997). This implies that for any endowment vector $\left(K^H, L^H, K^F, L^F\right)$ such that $K^H + K^F = K^W$ and $L^H + L^F = L^W$, international trade in goods will necessarily bring about factor price equalization (FPE).

- Markusen and Venables (2000) show that with transport costs ($\tau > 1$) the FPE set becomes one-dimensional. To see this in our simplified version of their model, notice that FPE will attain whenever, for a given endowment vector, both the factor market *and* the goods market clear with the same factor prices. Imposing free entry, one can show that total revenue per firm is given by $r^i f / (1 - \alpha)$, and that the factor market conditions are given by:

$$\begin{aligned} r^i K^i &= n_X^i r^i f / (1-\alpha) \\ w^i L^i &= z^i \end{aligned}$$

for $i = H, F$. On the other hand, Markusen and Venables, derive the following goods-market clearing conditions:

$$\begin{aligned} n_X^H r^H f / (1-\alpha) &= \frac{\beta \left(w^H L^H + r^H K^H \right) - \beta \tau^{1-\varepsilon} \left(w^F L^F + r^F K^F \right)}{1 - \tau^{1-\varepsilon}} \\ n_X^F r^F f / (1-\alpha) &= \frac{\beta \left(w^F L^F + r^F K^F \right) - \beta \tau^{1-\varepsilon} \left(w^H L^H + r^H K^H \right)}{1 - \tau^{1-\varepsilon}} \end{aligned}$$

Factor price equalization then requires:

$$\left(1 - \tau^{1-\varepsilon} - \beta \left(1 + \tau^{1-\varepsilon} \right) \right) r K^H = \beta \left(1 + \tau^{1-\varepsilon} \right) w L^H - \tau^{1-\varepsilon} \beta \left(w L^W + r K^W \right). \tag{7.5}$$

In the endowment space, the FPE set is thus a (one-dimensional) straight line with a slope:

$$\frac{dL^H}{dK^H} = \frac{1 - \tau^{1-\varepsilon} - \beta \left(1 + \tau^{1-\varepsilon} \right)}{\beta \left(1 + \tau^{1-\varepsilon} \right)} \frac{r}{w}.$$

Remember that in moving around the endowment box we are holding L^W and K^W fixed!

- It can be shown that the FPE line goes through the midpoint of the endowment box. To see this, substitute $K^H = K^W/2$ and $L^H = L^W/2$ and notice that in the FPE equilibrium, $wL^W = (1-\beta) \left(wL^W + rK^W \right)$ and $rK^W = \beta \left(wL^W + rK^W \right)$, $wL^W / rK^W = (1-\beta) / \beta$.

- This indicates that factor prices differences will tend to be small whenever **both** relative factor **and absolute factor endowment differences** are small.

- Combining this result with the discussion on firm behavior above, Markusen and Venables solve for the general equilibrium with multinational firms and find that the emergence of multinationals is more likely, the smaller are relative factor and absolute factor endowment differences.

- Furthermore, Markusen and Venables partition the endowment space into different regions according to which types of firms emerge in equilibrium.[3]

- For instance, their analysis indicates that if firm-level economies of scale are high, plant-level economies of scale are low, or transport costs are high, there will be a two-dimensional region in the endowment space in which only multinational firms will operate. And furthermore, this region is centered around the midpoint of the endowment space.

- To see this, notice from equation (7.4) that if factor prices differences are negligible, multinationals will emerge in equilibrium whenever

$$\Phi\left(\tau,1\right)=\frac{2}{1+\tau^{1-\varepsilon}}\geq\frac{g}{f}.$$

  If the inequality is strict, horizontal FDI strictly dominates exporting, and an equilibrium with only exporting does not exist.

  We next consider whether an equilibrium with only multinationals exists around the midpoint of the endowment space. Notice that free entry implies that multinationals break even thus implying:

$$\pi_I=\left(r^H\right)^{1-\varepsilon}B^H+\left(r^F\right)^{1-\varepsilon}B^F-\varphi\left(r^H,r^F\right)g=0.$$

  With full symmetry, we get

$$B^H=B^F=\frac{\varphi\left(r,r\right)g}{2r^{1-\varepsilon}}.$$

  Finally, we need to ensure that no firm finds exporting profitable:

$$\pi_X^i=\left(r^i\right)^{1-\varepsilon}B^i+\left(\tau r^i\right)^{1-\varepsilon}B^j<r^if,$$

[3] This is left as an exercise, but presumably, with our extreme factor intensity assumption, the conditions determining the size of these regions should be rather straightforward to derive.

or

$$\frac{2}{1+\tau^{1-\varepsilon}} > \frac{g}{f},$$

which is the same condition that rules out a pure exporting equilibrium. Markusen and Venables show that for small deviations from the midpoint of the endowment space, the equilibrium is still one with only MNEs. Intuitively, even when $r^H \neq r^F$, the deviation payoffs vary smoothly with r^H/r^F, and hence, if $2/\left(1+\tau^{1-\varepsilon}\right) > g/f$, exporting will not be profitable in some neighborhood of this midpoint.

- Markusen and Venables also show that there are two-dimensional regions in which the equilibrium is mixed, with both MNEs and exporting firms being active. This contrasts with Brainard's model with symmetric countries, in which mixed equilibria occurred only in knife-edge cases.

- Interestingly, Markusen and Venables also study the implications of horizontal FDI for trade flows. Contrary to the predictions of the Helpman-Krugman model of international trade, and consistently with the Heckscher-Ohlin model, the volume of trade may well be minimized when the two countries are identical. The intuition is that, in spite of the presence of economies of scale and product differentiation, in that region of the endowment space only multinational firms are active.

Chapter 8

Exports vs. FDI with Heterogenous Firms: Helpman, Melitz and Yeaple (2003)

- In the previous two chapters, we have developed models of the decision of exporting vs. FDI in which firms within an industry were treated as entirely symmetric.
- Helpman, Melitz and Yeaple (2003) incorporate intraindustry heterogeneity of the Melitz (2003) type in an otherwise standard proximity-concentration model of horizontal FDI. Their theoretical model delivers testable implications that go beyond the predictions from horizontal FDI models with homogeneous goods.
- For instance, the model predicts that, in a cross-section of industries, the ratio of exports to FDI sales should be higher in industries with higher productivity dispersion. Helpman et al. (2003) present regressions analogous to those in Brainard (1997), which they extend to include a measure of productivity dispersion. The econometric results provide strong evidence in support of the model.

Set-up

- Consider the following variant of the model presented in Chapter 6.
- The world consists of N countries that use labor to produce goods in $M+1$ sectors. One sector produces a homogenous good z, which we take as the numeraire, while

the remaining M sectors produce a continuum of differentiated products.

- On the **demand** side, each country is inhabited by a representative consumer with identical preferences:

$$U = \left(1 - \sum_{m=1}^{M} \beta_m\right) \log z + \sum_{m=1}^{M} \frac{\beta_m}{\alpha_m} \log \left(\int_{v \in V_m} x_m(v)^{\alpha_m} dv \right), \quad 0 < \alpha_m < 1, \tag{8.1}$$

  where $x_m(v)$ is consumption of variety v in sector m, V_m denotes the measure of available products in that sector, and $\varepsilon_m = 1/(1-\alpha_m)$ is the elasticity of substitution across varieties. Because preferences feature a unit elasticity of substitution across varieties in different sectors, we can focus on sector-by-sector analysis and safely drop the subscript m. As in Chapter 6, maximizing (8.1) subject to the budget constraint yields the following demand for each variety in a given sector

$$x(v) = \frac{\beta E^i}{\int_{v \in V} p(v)^{1-\varepsilon} dv} p(v)^{-\varepsilon} = A^i p(v)^{-\varepsilon}, \quad i = H, F. \tag{8.2}$$

- Country i is endowed with L^i units of labor. Unlike Brainard (1997) or Melitz (2003), we will not impose full symmetry, but we shall see that the equilibrium described below will require that cross-country differences in L^i be sufficiently small.

- On the **supply** side, both countries have access to an identical constant-returns-to-scale technology for producing good z. By choice of units, producing one unit of good z requires exactly one worker. For now, assume that $\sum_m \beta_m$ is small enough so as to ensure that good z is produced in every country, and the wage rate is equal to one in every country (an extension with cross-country factor price differences is discussed below).

- The differentiated-good sectors are characterized by monopolistic competition. Each variety is produced by a single firm and there is free entry into the industry.

- Firms produce varieties under a technology that features:

    1. A fixed cost of entry of f_E units of labor.
    2. A fixed overhead costs of f_D units of labor if the firm produces a positive amount.
    3. A fixed cost of exporting of f_X units of labor per foreign market where the firm exports.
    4. A fixed cost of FDI of f_I units of labor per foreign market served through a plant in that foreign market.
    5. A marginal cost that varies across firms and is denoted by a. As in Melitz (2003), it is assumed that firms face ex-ante uncertainty on their productivity, and that the actual marginal cost a is drawn, upon paying the fixed cost of entry, from a distribution $G(a)$.

- After observing this productivity level, the producer decides whether to exit the market immediately or incur f_D (and perhaps f_X and f_I) and start producing.

- Notice that while fixed costs are common across firms, there is intraindustry heterogeneity in productivity originated from differences in the marginal cost of production.

- The difference $f_I - f_X$ indexes plant-level economies of scale, i.e., the extra fixed cost associated with opening a plant in a foreign market, rather than simply exporting part of the output produced in the domestic economy.

- Goods that are exported are subjected to iceberg costs by which $\tau^{ij} > 1$ units of the good need to be shipped from country i for each unit actually delivered in country j.

- For reasons that will become clear below, it is assumed that:

$$f_I > \left(\tau^{ij}\right)^{\varepsilon-1} f_X > f_D$$

(Hint: the second inequality should look familiar from Chapter 2).

Firm Behavior

- Remember that with CES preferences, firms set a price that is a constant markup over marginal cost. In particular, a firm with productivity a will set a price equal to a/α for domestically produced goods – both by domestic producers and foreign affiliates) – and a price of $\tau^{ij}a/\alpha$ for exports from country i to country j.

- As in Chapter 6, define
$$B^i = (1-\alpha)\,\alpha^{\varepsilon-1}A^i. \tag{8.3}$$

- Combining (8.2), (8.3), and plugging the optimal prices, we can express operating profits from serving the domestic market as
$$\pi_D^i = a^{1-\varepsilon}B^i - f_D.$$
Similarly, the additional operating profits from exporting to country j are
$$\pi_X^{ij} = \left(\tau^{ij}a\right)^{1-\varepsilon}B^j - f_X,$$
whereas the additional profits from servicing country j through FDI are
$$\pi_I^{ij} = a^{1-\varepsilon}B^j - f_I,$$
and are thus independent of i.

- These profit levels are depicted in Figure 8.1 for the case in which $B^i = B^j$ so that the profit lines π_D^i and π_I^i are parallel. In such case, $\pi_D^i > \pi_I^i$ for all a, because $f_I > f_D$. It is also clear that so long as $\tau^{ij} > 1$, the profit lines π_D^i and π_I^i is steeper than π_X^i.

- It is also apparent from the figure that if $\left(\tau^{ij}\right)^{\varepsilon-1} f_X > f_D$, the line π_X^i will cross the horizontal axis to the right of the point at which π_D^i crosses that axis. Furthermore, if $f_I > \left(\tau^{ij}\right)^{\varepsilon-1} f_X$, then π_I^i intersects π_X^i to the right of the point at which π_X^i crosses the horizontal axis.

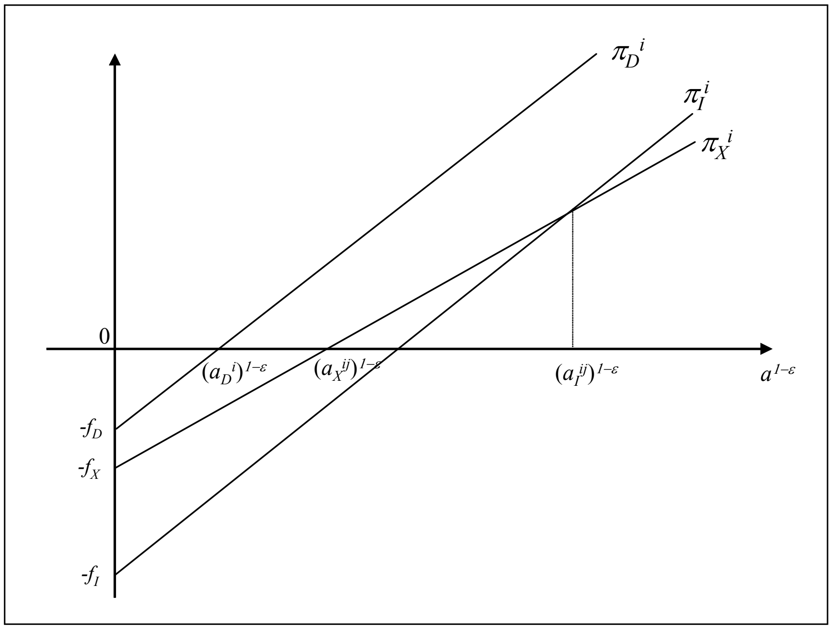

π_D^i
π_I^i
π_X^i
0
$(a_D^i)^{1-\varepsilon}$
$(a_X^{ij})^{1-\varepsilon}$
$(a_I^{ij})^{1-\varepsilon}$
$a^{1-\varepsilon}$
$-f_D$
$-f_X$
$-f_I$

Figure 8.1: Firm Behavior

- It thus follows that as long as $f_I > (\tau^{ij})^{\varepsilon-1} f_X > f_D$, the following sorting emerges:
  - the least productive firms, $a > a_D^i$, exit upon observing their productivity;
  - firms with productivity $a \in \left(a_X^{ij}, a_D^i\right)$ stay in the market but sell only domestically;
  - firms with productivity $a \in \left(a_I, a_X^{ij}\right)$ not only sell domestically but also export to country j;
  - the most productive firms, $a < a_I^{ij}$, not only sell domestically but also services country j through FDI.
- The thresholds are determined by

$$\begin{aligned}
\left(a_D^i\right)^{1-\varepsilon} B^i &= f_D && \text{for all } i \\
\left(\tau^{ij} a_X^{ij}\right)^{1-\varepsilon} B^j &= f_X && \text{for all } j \neq i \\
\left(1 - \left(\tau^{ij}\right)^{1-\varepsilon}\right) \left(a_I^{ij}\right)^{1-\varepsilon} B^j &= f_I - f_X && \text{for all } j \neq i.
\end{aligned}$$

Industry Equilibrium

- In the industry equilibrium, free entry into sector m ensures that the expected operating profits for a potential entrant,

$$\int_0^{a_D^i} \left(a^{1-\varepsilon}B^i - f_D\right) dG\left(a\right) + \sum_{j\neq i} \int_{a_I^{ij}}^{a_X^{ij}} \left[\left(\tau^{ij}a\right)^{1-\varepsilon} B^j - f_X\right] dG\left(a\right) +$$

$$+\sum_{j\neq i} \int_0^{a_I^{ij}} \left[a^{1-\varepsilon}B^j - f_I\right] dG\left(a\right),$$

  equal the fixed cost of entry f_E.

- Defining

$$V\left(a\right) = \int_0^a y^{1-\varepsilon} dG\left(y\right),$$

  we can express this condition as

$$V\left(a_D^i\right)B^i + \sum_{j\neq i}\left[1-\left(\tau^{ij}\right)^{1-\varepsilon}\right]V\left(a_I^{ij}\right)B^j + \sum_{j\neq i}\left(\tau^{ij}\right)^{1-\varepsilon}V\left(a_X^{ij}\right)B^j$$

$$-\left[G\left(a_D^i\right)f_D + \sum_{j\neq i}G\left(a_I^{ij}\right)\left(f_I - f_X\right) + \sum_{j\neq i}G\left(a_X^{ij}\right)f_X\right] = f_E \quad \text{for all } i.$$

General Equilibrium

- In the working paper version of their paper, Helpman et al. (2003) also solve for the general equilibrium of the model in a particular case in which, for a given sector,

  1. the fixed cost coefficients are identical in all countries;
  2. the distribution function $G\left(a\right)$ is identical in all countries;
  3. transport costs are identical for every pair of countries, that is, $\tau^{ij} = \tau$ for all $i \neq j$;
  4. the endowment of labor is not *too* different across countries.

- Under these assumptions, there will be a positive measure of entrants in all countries and the demand level B^i will be equalized across countries, that is, $B^i =$

B for all i. As a result of this, together with assumptions 1-3, the equilibrium cutoffs will be independent of i and j.

- In particular, the equilibrium B, a_D, a_X, and a_I are the solution to the following system

$$\begin{aligned} (a_D)^{1-\varepsilon} B &= f_D && (8.4) \\ (\tau a_X)^{1-\varepsilon} B &= f_X && (8.5) \\ \left(1-\tau^{1-\varepsilon}\right)(a_I)^{1-\varepsilon} B &= f_I - f_X && (8.6) \end{aligned}$$

$$\begin{aligned} V(a_D) B + (N-1)\left[\left(1-\tau^{1-\varepsilon}\right) V(a_I) B^j + \tau^{1-\varepsilon} V(a_X) B\right] - G(a_D) f_D - \\ - (N-1)\left[G(a_I)(f_I - f_X) + G(a_X) f_X\right] = f_E. \qquad (8.7) \end{aligned}$$

- Finally, with the value of B, we can then use (8.2) and (8.3) to obtain the number of entrants in each country. Interestingly, Helpman et al. (2003) illustrate how the model gives rise to home-market effects, by which larger countries (i) attract a disproportionately larger measure of entrants and sellers, and (ii) are disproportionately served by domestically-owned firms.

- For our purposes, it suffices to point out that from (8.5) and (8.6),

$$\frac{a_X}{a_I} = \left(\frac{f_I - f_X}{f_X} \frac{1}{\tau^{\varepsilon-1}-1}\right)^{1/(\varepsilon-1)} \qquad (8.8)$$

Exports vs. FDI

- Under the symmetry assumptions above, in a given industry m, the ratio of exports from country i to country j relative to i's FDI sales in j are given by

$$\frac{s_X^{ij}}{s_I^{ij}} = \frac{\int_{a_I}^{a_X} (\tau a)^{1-\varepsilon} B}{\int_0^{a_I} a^{1-\varepsilon} B} = \tau^{1-\varepsilon}\left[\frac{V(a_X)}{V(a_I)} - 1\right]. \qquad (8.9)$$

- Remember that a_X and a_I are determined by the system (8.4)-(8.7) and thus will be a function of the different fixed costs, of transport costs, and of the parameters of the function $V(a)$. Hence, s_X^{ij}/s_I^{ij} will also depend on these parameters.

- Helpman et al. (2003) discuss comparative statics that hold regardless of a particular choice of a functional form for $G(a)$. In particular, they show that s_X^{ij}/s_I^{ij} is increasing in f_I and decreasing in f_X and τ.

- For expositional purposes and in order to explore the effects of productivity dispersion on the ratio s_X^{ij}/s_I^{ij}, consider the case in which $\theta = 1/a$ is characterized by a Pareto distribution with shape parameter k, i.e.,

$$F(\theta) = 1 - \left(\frac{b}{\theta}\right)^k, \quad \text{for } \theta \geq b > 0,$$

where it is assumed that $k > \varepsilon + 1$. It is straigthforward to show that the higher is k the higher is the variance of both productivity and sale revenues.

- Remember the following statistical result. If $\theta \sim F(\theta)$ and $a = h(\theta)$ with $h'(\theta) < 0$ in the relevant range, then $a \sim G(a) = 1 - F(h^{-1}(\theta))$. In our particular case, $h(\theta) = 1/\theta$, and thus $a \sim G(a) = 1 - F(1/a) = (ba)^k$ for $a \leq b$. Notice, in turn, that

$$V(a) = \int_0^a y^{1-\varepsilon} dG(y) = ca^{k-(\varepsilon-1)},$$

where c is some constant.

- Plugging back in (8.9),

$$\begin{aligned} \frac{s_X^{ij}}{s_I^{ij}} &= \tau^{1-\varepsilon}\left[\left(\frac{a_X}{a_I}\right)^{k-(\varepsilon-1)} - 1\right] = \\ &= \tau^{1-\varepsilon}\left[\left(\frac{f_I - f_X}{f_X}\frac{1}{\tau^{\varepsilon-1}-1}\right)^{\frac{k-(\varepsilon-1)}{(\varepsilon-1)}} - 1\right]. \end{aligned}$$

- It is then straightforward to see that the ratio of exports to FDI sales is:
  - decreasing in transport costs;
  - increasing in plant-level economies of scale $f_I - f_X$;

  - decreasing in productivity dispersion, as parametrized by k.

- The first two points are a restatement of the proximity-concentration hypothesis, but notice that the results now follow from aggregating across producers *within* the same industry that choose different modes for servicing the foreign market.

Empirical Implementation

- Helpman et al. (2003) impose a lot of symmetry in the model to be able to close the general equilibrium of the model. In order, to derive testable implications for the ratio of exports to FDI sales, we need not impose so much structure.

- Helpman et al. allow for cross-sectoral differences in plant-levels economies of scale, transport costs, productivity dispersion and the elasticity of demand. And also for cross-country differences in transport costs, fixed costs of exporting, factor prices (wage rates) and again transport costs.

- In particular, denoting by w^j the wage rate in country j and defining $f_{hP} \equiv f^i_{hI} - f^i_X$, one can use generalizations of (8.5) to (8.6) to derive the following expression for the ratio exports to FDI sales in sector h for $i = U$ (or U.S.):

$$\frac{s_X^{Uj}}{s_I^{Uj}} = \left(\frac{w^U}{w^j}\tau_h^{Uj}\right)^{1-\varepsilon_h} \left\{ \left[\frac{f_{hP}}{f_X^j} \frac{1}{\left(\frac{w^U}{w^j}\tau_h^{Uj}\right)^{\varepsilon_h - 1} - 1} \right]^{\frac{k_h^U - (\varepsilon_h - 1)}{\varepsilon_h - 1}} - 1 \right\}, \tag{8.10}$$

which delivers the same comparative statics with respect to transport costs, plant-economies of scale and productivity dispersion, provided that

  - $w^U \tau_h^{Uj}/w^j < \left(f^j_{hI}/f^j_X\right)^{1/(\varepsilon_h - 1)}$, which ensures that some U.S. firms export to country j.
  - $w^U \tau_h^{Uj}/w^j > 1$, which ensures that some firms locate in country j.
  - $w^j \tau_h^{jU}/w^U > 1$, which ensures that some firms locate in country the U.S.

- Helpman et al. (2003) run a log-linearized version of eq. (8.10):

$$\log\left(s_X^{Uj}/s_I^{Uj}\right)=\alpha+\beta_\tau\log\left(\tau_h^{Uj}\right)+\beta_P\log\left(f_{hP}\right)+\beta_k\left(k_h^U-\left(\varepsilon_h-1\right)\right)+\beta_Z Z_h+\mu_j+v_{ij},$$

where the country fixed effects are used to control for f_X, ω, and other unobservable country variables, and Z_h is a vector of industry controls. In light of the model, we expect $\beta_\tau<0$, $\beta_P>0$, and $\beta_k<0$.

Data and Empirical Results

- The left-hand-side variable is constructed using FDI sales data from the BEA for 1994, as well as U.S. export data from Feenstra (1997). Measures of tariffs and freight costs, which proxy for τ_h^{Uj} are also obtained from the Feenstra dataset. As in Brainard (1997), plant-economies of scale are proxied by the number of nonproduction workers.[1]

- Helpman et al.'s (2003) construction of the proxy for productivity dispersion is a contribution in its own right. They note that, in the model, if productivity is drawn from a Pareto distribution with shape parameter k, then the implied distribution of firm sales will also be Pareto with shape parameter $k-(\varepsilon-1)$, which incidentally is consistent with evidence.

- Making use of the properties of the Pareto distribution, the parameter $k-(\varepsilon-1)$ can then be recovered from a regression of the logarithm of an individual firm's rank within the distribution on the logarithm of firm's size. Notice that this does not permit disentangling the separate effects of productivity dispersion k from the elasticity of demand ε, but this is not a problem because the theory predicts that the ratio of exports to FDI is affected precisely by the object that is being estimated.

- This measure is computed for both U.S. and European firms using data from the U.S. Census of Manufactures and the Amadeus database. Helpman et al. also

[1] In particular, by the average number of nonproduction workers in a particular industry (remember that Brainard used instead the number of nonproduction employees in the median U.S. plant ranked by value added in each industry).

experiment with the alternative of simply proxying k with the standard deviation of the logarithm of firm sales, computed from the same databases.

- Finally, in order to control for omitted industry characteristics, Helpman et al. include measures of capital intensity and R&D intensity.

- Their results strongly support the predicitions of the model. In particular, they find that $\beta_\tau < 0$, $\beta_P > 0$, and $\beta_k < 0$, and in most specifications the coefficients are significantly different from zero. Only when industry random effects are included does the coefficient on tariffs become insignificantly different from zero.

- The results are robust to the use of several measures of productivity dispersion and to instrumenting the U.S. productivity dispersion measure with the European one (make sure you convince yourself why this is an appealing identification strategy).

- Finally, Helpman et al. (2003) find that capital intensity has a negative and significant effect on the ratio of exports to FDI sales, while the ratio is essentially unaffected by R&D intensity.

Chapter 9

Vertical FDI: Theory and Evidence

- In recent years, we have witnessed a spectacular increase in the way firms organize production on a global scale. Feenstra (1998), citing Tempest (1996), describes Mattel's global sourcing strategies in the manufacturing of its star product, the Barbie doll:

  > The raw materials for the doll (plastic and hair) are obtained from Taiwan and Japan. Assembly used to be done in those countries, as well as the Philippines, but it has now migrated to lower-cost locations in Indonesia, Malaysia, and China. The molds themselves come from the United States, as do additional paints used in decorating the dolls. Other than labor, China supplies only the cotton cloth used for dresses. Of the $2 export value for the dolls when they leave Hong Kong for the United States, about 35 cents covers Chinese labor, 65 cents covers the cost of materials, and the remainder covers transportation and overheads, including profits earned in Hong Kong. (Feenstra, 1998, p. 35-36).

- A variety of terms have been used to refer to this phenomenon: the "slicing of the value chain", "international outsourcing", "fragmentation of the production process", "vertical specialization", "global production sharing", and many more.

- As argued in Chapter 6, multinational corporations account for a very significant fraction of world trade flows, with trade in intermediate inputs between divisions of the same firm constituting an important portion of these flows. The literature refer to this phenomenon as **vertical** FDI. In this chapter, we will first discuss the seminal paper by Helpman (1984), which was the first to formalize the rational for this type of FDI. His theory associates multinational corporations with the ability of firms to exploit cross-country differences in factor prices by shifting activities to the cheapest locations. We will then review some evidence that supports the empirical relevance of this type of FDI.

Theory: Helpman (1984)

- Consider the following version of Helpman's (1984) model, as laid out in Chapter 12 in Helpman and Krugman (1985).

Set-up

- The world consists of two countries (Home and Foreign) that use two factors of production (capital and labor) to produce goods in two sectors. One sector produces a homogenous good Y, which we take as the numeraire, while in the other sector a continuum of firms produce differentiated products.

- On the **demand** side, each country is inhabited by a representative consumer with identical homothethic preferences $u(Y, U_x)$, where Y is consumption of the homogeneous good and U_x is the subutility level attained in the consumption of differentiated products. U_x could be a Dixit-Stiglitz CES function, as in the previous chapters, but nothing below depends on this assumption. Denote by α_Y the share of expenditure that consumers allocate to sector Y.[1]

- Home is endowed with K units of capital and L units of labor. Foreign's endowments are K^* and L^*. Let $\overline{K} = K + K^*$ and $\overline{L} = L + L^*$. Factor of production are internationally immobile. Goods can be costlessly traded.

- On the **supply** side, both countries have access to an identical constant-returns-to-scale technology for producing good Y. Let the unit cost function for this

[1] Remember that, unless $u(Y, U_x)$ features a unit elasticity of substitution between its arguments, α_Y will depend on the price and measure of varieties in the differentiated-goods sector.

good be:

$$c_Y\left(w_L, w_K\right) = 1, \tag{9.1}$$

where the equality follows from our choice of numeraire. Assume that a producer in good Y needs to employ all inputs in the same location.

- The differentiated-good sectors are characterized by monopolistic competition. Each variety is produced by a single firm and there is free entry into the industry.

- Firms produce varieties under a technology that combines labor, capital, and headquarter services H (e.g., management, distribution, product-specific R&D). The technology is as follows:

  1. H is a firm-specific, differentiated product that is itself produced with capital and labor according to the total cost function $C^H\left(w_L, w_K, h\right)$, where h is the number of units of H produced and $C^H\left(\cdot\right)$ is associated with a nondecreasing returns to scale production function. A particular h is valuable to only one firm, but within the firm, it can serve many plants, **regardless of where the plants are located**.

  2. Let $C^P\left(w_L, w_K, h, x\right)$ be the costs required to produce x units of a particular variety in a particular plant, once the costs for making h firm-specific have been incurred. $C^P\left(\cdot\right)$ is associated with an increasing returns to scale production function in which h is essential. For example, as in previous chapters, we could let

     $$C^P\left(w_L, w_K, h, x\right) = f\left(w_L, w_K\right) + g\left(w_L, w_K, h, x\right),$$

     where $f\left(\cdot\right)$ is a fixed cost and $g\left(\cdot\right)$ is linear homogenous in $\left(h, x\right)$.

     Because trade is costless and there are increasing returns associated with $C^P\left(\cdot\right)$, all firms will be single-plant firms.

  3. We can thus express the firm's single-plant cost function as:

     $$C\left(w_L, w_K, x\right) = \min_h \left\{C^P\left(w_L, w_K, h, x\right) + C^H\left(w_L, w_K, h\right)\right\}. \tag{9.2}$$

Firm Behavior and Industry Equilibrium

- When maximizing profits, firms in the differentiated sector choose the level of output at which the marginal cost of production equals the marginal revenue, which is different from the price, because of the downward sloping demand curve associated with product differentiation. Letting $R(p,\overline{n}) = p/MR(p,\overline{n})$, this condition is
$$C_x(w_L,w_K,x) = \frac{R(p,\overline{n})}{p},$$
where $\overline{n}$ is the total measure of available varieties.

- On the other hand, free entry implies that firms will break even and, in the industry equilibrium, all firms charge a price equal to the average cost of production:
$$p = \frac{C(w_L,w_K,x)}{x} \equiv c(w_L,w_K,x). \tag{9.3}$$

- Using this expression, condition $R(p,\overline{n}) = p/MR(p,\overline{n})$ can be written as
$$R(p,\overline{n}) = \frac{c(w_L,w_K,x)}{C_x(w_L,w_K,x)} \equiv \theta(w_L,w_K,x). \tag{9.4}$$

General Equilibrium in an Integrated Economy

- Consider the general equilibrium of an integrated economy in which the endowments of capital and labor are freely mobile so that factor price equalization holds.

- Because of all firms have access to the same technology, the equilibrium is symmetric with all firms setting up a single plant and choosing the same values for h, x and p. For given factor endowments, these three variables, together with factor prices (w_L,w_K), the total number of producers $\overline{n}$, and total output $\overline{Y}$ in the homogenous good sector, are pinned down by equations (9.1)-(9.4), as well as the factor market clearing conditions
$$\begin{aligned} a_{LY}(w_L,w_K)\overline{Y} + a_{LX}(w_L,w_K,x)\overline{n}x &= \overline{L} \qquad (9.5)\\ a_{KY}(w_L,w_K)\overline{Y} + a_{KX}(w_L,w_K,x)\overline{n}x &= \overline{K} \qquad (9.6)\end{aligned}$$

and the goods market clearing condition

$$\alpha_Y\left(p, \overline{n}\right) = \frac{\overline{Y}}{\overline{Y} + p\overline{n}x}. \tag{9.7}$$

- Remember that by Shepard's lemma, the cost-minimizing input level of factor i per unit of good j is given by $a_{ij}\left(\cdot\right) = \partial c_j\left(\cdot\right)/\partial w_i$.

- Defining

  $$a_{iH}\left(w_L, w_K, h\right) = \frac{\partial C^H\left(w_L, w_K, h\right)/\partial w_i}{h}$$

  and

  $$a_{iP}\left(w_L, w_K, h, x\right) = \frac{\partial C^P\left(w_L, w_K, h, x\right)/\partial w_i}{x},$$

  we can write

  $$a_{iX}\left(w_L, w_K, x\right) = a_{iP}\left(w_L, w_K, h, x\right) + a_{iH}\left(w_L, w_K, h\right)\frac{h}{x}.$$

- In what follows, we will assume that, for all w_L, w_K, h, and x, the following inequalities hold

  $$\frac{a_{KH}\left(\cdot\right)}{a_{LH}\left(\cdot\right)} > \frac{a_{KP}\left(\cdot\right)}{a_{LP}\left(\cdot\right)} > \frac{a_{KY}\left(\cdot\right)}{a_{LY}\left(\cdot\right)}.$$

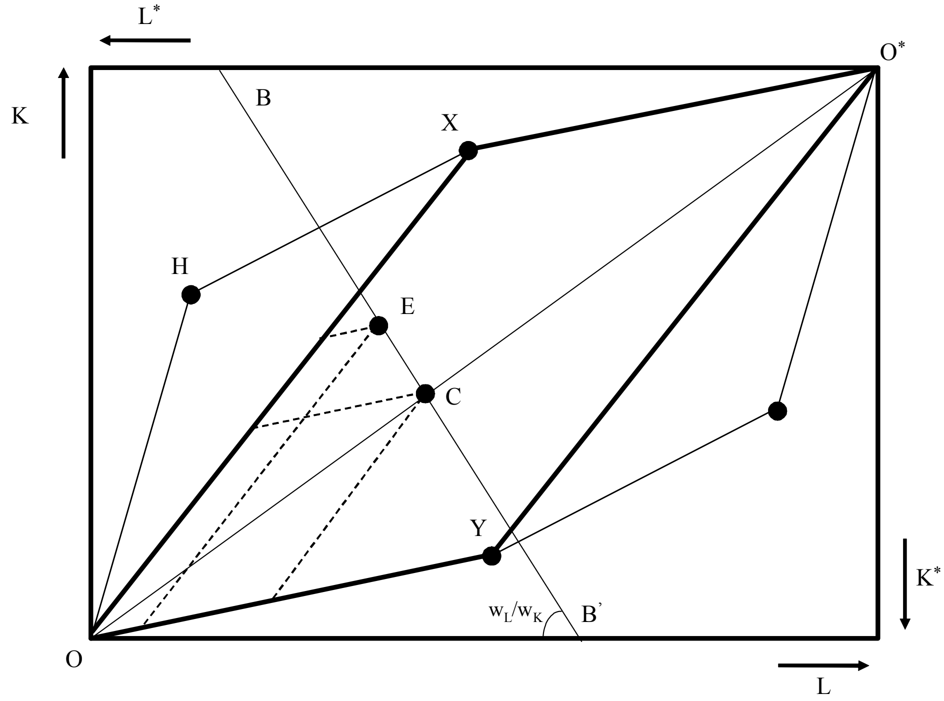
  In words, headquarter services is the most capital-intensive production process, whereas production of the homogeneous good is the most labor intensive. Furthermore, there are no factor intensity reversals.

- Under this assumption, Figure 9.1 depicts the general equilibrium of the integrated economy. The vectors OX and OY represent the employment of factors in the differentiated and homogenous good sectors. Furthermore, within sector X, the vector OH represents factor employment in the production of headquarter services, while the vector HX corresponds to factor employment in plant production.

Pattern of Production

- Now imagine that the endowments of the integrated economy are divided between Home and Foreign. Assume that factors of production are immobile across

Figure 9.1: Pattern of Production of the Integrated Economy



countries. The model is essentially a $2 \times 2 \times 2$ Helpman-Krugman model of international trade. But there is one *twist*: the production of differentiated varieties has two stages, and these stages need not be undertaken in the same country. In particular, headquarter services are perfectly mobile and can be applied to plant production in foreign countries.

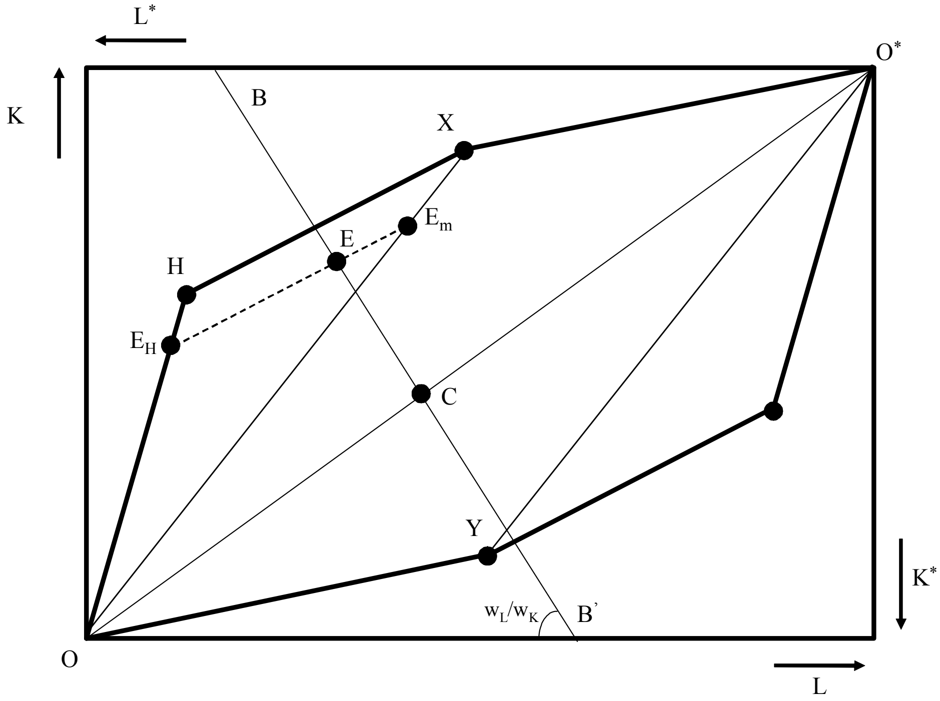
- Without loss of generality, we can focus on endowment points for which $K/L > K^*/L^*$, i.e., for which Home is relatively capital abundant.

- Consider first endowment points in the set OXO^*, such as point E in Figure 9.1. From the analysis in Helpman and Krugman (1985), it is well known that, even if we impose that h and x need to be produced in the same country, the equilibrium will still feature factor price equalization (FPE hereafter). In this equilibrium firms based in one country have no incentive to open subsidiaries (production plants) in the other country. Furthermore, we can rule out the existence of multinational firms in the set OXO^* by adopting the following criterion:

Criterion 2 *For endowment points in the set OXO^*, unless there are factor price differences across countries, headquarter services and plant production will be undertaken in the same country.*

- Notice that this is equivalent to assuming that there is some positive (managerial) cost γ to fragmenting the production process. In the analysis we then focus on the limiting case $\gamma \rightarrow 0$.
- So far, we have shown that for endowment points in the set OXO^*, there exists an equilibrium without multinational firms. In fact, it is possible to show that this is the *unique* equilibrium. In particular, one can show that even if firms chose to fragment the production process and, for instance, Home specialized in the production of headquarter services and good Y, we would still attain FPE, thereby reaching a contradiction in light of Criterion 1.
- Intuitively, for FPE to hold in our two-factor model, we only need both countries to employ a positive amount of factors in at least two common production processes. Because, complete specialization is inconsistent with endowment points in the set OXO^*, FPE will hold and no multinational firms will emerge.
- Remember that along the line BB' in Figure 9.1, the relative size (in terms of income) of the two countries is held constant. It is then apparent from Figure 9.1 that, for given relative size, multinational firms will not emerge unless relative factor endowment differences between countries are large enough.
- Consider next endowment points in the set OHX, such as point E in Figure 9.2. From the discussion above, if h and x are produced in the same country, then FPE will fail to hold. In particular, given $K/L > K^*/L^*$, the wage-rental ratio at Home will be higher than in Foreign.
- This implies that if factor price differences and/or factor intensity differences are high enough, firms will find it optimal to fragment the production process and

Figure 9.2: Pattern of Production in the set OHX



produce headquarter services at Home and undertake plant production in Foreign. Naturally, this fragmentation will tend to push up labor demand in Foreign and reduce at Home, thus creating an additional force towards convergence in factor prices.

- Helpman (1984) shows that for endowment points in the set OHX this forces will bring about FPE. The intuition is analogous to that used to argue for the uniqueness of equilibrium in the set OXO^*. For endowment points in OHX too, both countries will employ a positive amount of factors in at least two common production processes and this is sufficient for FPE to hold.

- Now, according to Criterion 1, the emergence of multinational firms in the set OHX is inconsistent with FPE. The difference here is that without multinational firms, FPE will not attain. Hence, we consider here a different criterion to determine the equilibrium in the set OHX:

Criterion 3 *For endowment points in the set OHX, we consider equilibria with the smallest number of multinational corporations.*

- Again, it is useful to appeal to positive but negligible managerial costs γ in order to justify this criterion.

- Figure 9.2 depicts such an equilibrium. The vectors OE_H and E_HE correspond to factor employment at Home in the production of headquarter services and in plant production, respectively. Home does not produce homogenous goods. On the other hand, Foreign produces the whole world demand for good Y (vector O^*X), and also employs a positive amount of factors in the production of both h and x.

- Notice that the vector EE_m measure factor usage in the production of x corresponding to firms for which headquarter services are produced at Home, whereas x is produced in Foreign. The length $\|EE_m\|$ is thus a measure of the extent of multinationality in the model.

- It is clear from Figure 9.2 that for a constant relative size of the two countries **the measure of multinational firms in the model is increasing in relative factor endowment differences**.

The Volume of Trade

- Under the present assumptions, the model does not feature intrafirm trade in physical goods. Notice, however, that in the model there are invisible exports of headquarter services from the parent to its subsidiaries. Assuming that these services are valued at average cost, Helpman (1984) derives some interesting results that complement and qualify the predictions of the benchmark Helpman-Krugman for the volume of international trade and its components.

1. The larger the role of multinational firms in the world economy, the weaker the effects of relative country size dispersion on the volume of trade.

- Remember that in the benchmark Helpman-Krugman model, for given relative factor endowments, the volume of trade is maximized when countries are of equal size. In the model with multinational firms, the intrafirm component of trade is larger the larger is Home relative to Foreign, and this tends to weaken the above prediction.

2. For a given relative size of countries, the share of intrafirm trade in the total volume of trade is increasing in relative factor endowment differences.

    - This follows directly from the discussion above of the length of the vector $\|EE_m\|$.

3. The larger the role of multinational firms in the world economy, the weaker the effects of relative factor endowments on the share of intraindustry trade in the total volume of trade.

    - In the benchmark Helpman-Krugman model, this share monotonically falls with relative factor endowment differences. As described by Helpman (1984), the emergence of multinational firms may alter the pattern of trade in the differentiated good sector. In particular, if Home is sufficiently capital abundant, Foreign may become a net exporter in sector X. This in turn, gives rise to an area of endowment space in which the share of intraindustry trade is increasing in relative factor endowment differences.

Evidence: Yeaple (2003) and others

- The Helpman (1984) model of vertical FDI predicts that the size of multinational activity should be increasing in relative factor endowment differences.

- The results in Brainard (1997) suggest otherwise. As discussed in Chapter 6, her regression results indicate that the ratio of exports to outward FDI sales is increasing in relative factor endowments differences (see Table 1) as measured by differences in income per capita between the US and the recipient/host country (variable PWGDP). Furthermore, she also finds that the *level* of outward FDI sales is decreasing in relative factor endowment differences (Table 7, column I).

- Some people have interpreted this evidence as suggesting that the bulk of FDI is of the horizontal type. In a recent paper, Carr, Markusen and Maskus (2001) find similar results. In particular, the authors claim that bilateral affiliate sales between the U.S. and 36 other countries for the period 1986-1994 appear to be better explained by horizontal FDI measures (transport costs, plant-level economies of scale...) than by vertical FDI measures (relative factor endowment differences).

- In recent years, some researchers have challenged this wisdom. Let us briefly review their contributions.

Yeaple (2003)

- Yeaple (2003) starts by noting that most of the evidence against the vertical nature FDI comes from econometric studies that use data aggregated across industries to the country level. The Helpman (1984) model does not predict that that FDI will be increasing in relative factor endowment differences. It predicts that in industries that are intensive in a particular factor, FDI flows should be flowing to countries that are abundant in that particular factor.

- Focusing on the particular case of skilled labor, the model predicts that, in industries with high skilled-labor intensities, U.S. MNEs should favor skilled-labor-abundant countries over skilled-labor-scarce countries, but that, in industries with low skilled-labor intensities, U.S. MNEs should favor skilled-scarce-abundant countries over skilled-labor-abundant countries.

- In econometric terms, in order to test for the empirical relevance of vertical FDI, it is *not* sufficient to run regressions of the type:

$$FDI_{ij} = \beta_1 T_{ij} + \beta_2 ScaleEco_i + \beta_3 MKTSIZE_j + \beta_4 RelFactEnd_j + \beta_5 FactIntens_i + \varepsilon_{ij}$$

(where i indexes industries, and j, countries) and test for the sign and significance of β_4. The model should instead be specified as

$$\begin{aligned} FDI_{ij} &= \beta_1 T_{ij} + \beta_2 ScaleEco_i + \beta_3 MKTSIZE_j + \beta_4 RelFactEnd_j + \\ &\quad + \beta_5 FactIntens_i + \beta_6 RelFactEnd_j * FactIntens_i + \varepsilon_{ij}, \end{aligned} \tag{9.8}$$

The relevant coefficient for assessing the vertical motive of FDI is the β_6 and the predicted sign is positive.

- Notice that if β_4 is sufficiently negative, a feature that is *not* inconsistent with vertical FDI, then regressions that omit the interaction term will tend to find a negative effect of relative factor endowment differences on FDI flows.

- Yeaple (2003) runs regressions of the type in equation (9.8), which exploit both cross-industry as well as cross-country variation in the importance of FDI. He experiments with different measures of the extent of FDI, including Brainard's (1997) and Carr et al.'s (2001). His results are strongly supportive of the vertical dimension of FDI. For instance, in Table 3, he reports that the ratio of exports to FDI sales is decreasing in the interaction of a measure of Human capital abundance and a measure of skilled-labor intensity. Similarly, the interaction term has a positive effect on the level of FDI sales (both local sales as well as export sales).

- Leaving aside the regression results, Figure 1 in his paper is self-explanatory. It shows how in skill-labor-scarce host countries, FDI flows are concentrated in low-skill-intensive industries, whereas in skill-labor-abundant host countries, FDI flows are concentrated in high-skill-intensive industries.

Others

- At least two other studies have shed some light on the empirical relevance of vertical FDI.

- On the one hand, Hanson, Mataloni and Slaughter (2001) use detailed information on the foreign operations of U.S.-based multinational firms to analyze directly the form that FDI flows take. One of their main findings is that a large and increasing fraction of FDI flows is related to exports of intermediate inputs to foreign affiliates for *further processing*. For instance, imported inputs for further processing account for over 30 percent of affiliate sales for affiliates in Canada and Mexico. This finding is a clear indication of the empirical relevance of vertical FDI.

- On the other hand, Antràs (2003) has recently argued that even if Helpman's (1984) model might be the right model for understanding the recent trend of increasing fragmentation of the production process, it is far less clear that the model will deliver the right predictions for FDI flows and the intrafirm component of trade. In particular, Antràs (2003) argues that a substantial part of the recent fragmentation of the production process has occured at arm's length and will not be recorded in FDI or intrafirm trade data. Furthermore, he develops a theoretical model in which a proper modelling of the internalization decision leads him to substantially different predictions for how the share of intrafirm trade should correlate with relative factor endowment differences. We will study this paper in more detail in Chapter 12.

Firms and FDI: Colophon

- Before we conclude this set of Chapters on firms and FDI, it is useful to briefly mention other branches of the literature that we will not have time to discuss because of time constraints.

- In the last few chapters, we have studied horizontal and vertical models separately. A very recent literature has started to study some interesting complementarities between these two forms of FDI. This literature is motivated by the fact that in the data we see that firms use both types of strategies simultaneously, thus engaging in what people refer to as "complex integration strategies". Contributions to this literature include Yeaple (2003, JIE) and Grossman, Helpman, and Szeidl (2003), as well as Ekholm, Forslid, and Markusen's (2003) work on export-platform FDI.

- Another interesting branch of the literature that we will not cover tries to go beyond a study of the determinants of FDI, and instead focuses on the *effects of FDI*. This literature is mostly empirical and has tried to identify the effect of FDI flows on the productivity of firms in the countries that are recipients of FDI. Aitken and Harrison (1999) searched for these effects in a sample Venezuelan firms and found an almost negligible effect of FDI. On the other hand, Haskel, Pereira and Slaughter (2002) studied the effects of FDI on a sample of U.K. manufacturing firms and found substantial evidence of positive FDI spillovers, although the size of these effects was not too large.

Part III

Intermission: The Boundaries of The Firm

Chapter 10

The Theory of The Firm: Transaction-Cost Approaches

- In the previous chapters we have studied several theories of the multinational firm. In those models, the emergence of multinational firms was determined by some combination of location advantages related to the host country (distance as captured by transport costs, factor prices, factor endowments) and by some technological factors (firm vs. plant-economies of scale, transport costs) that favored or hindered a fragmentation of the production process.

- These theories enhance our understanding of trade and FDI flows, but they share a common failure to properly model the crucial issue of internalization. These models can explain why a domestic firm might have an incentive to undertake part of its production process abroad, but they fail to explain why this foreign production will occur within firm boundaries, rather than through arm's length subcontracting or licensing. And there is some evidence that suggests that the growth of foreign outsourcing by U.S. firms might have outpaced the growth of their foreign intra-firm sourcing (cf, Antràs and Helpman, 2004).

- To address issues that arise from the choice of outsourcing versus integration and home versus foreign production, we need to develop theoretical frameworks in which companies make endogenous organizational choices. As a necessary first step, in this chapter we will review some of the main theories of the firm.

The Technological Approach: Some Caveats

- In the Neoclassical theory of the firm (see for instance, Mas-Colell et al., 1995, Chapter 5), the size of the firm is determined by firms' cost-minimization. The problem can be thought of as consisting of two stages.

- In the first stage, firms minimize total costs subject to output reaching a particular amount. In particular, letting x denote a vector of inputs with associated price vector w, and letting $y = f(x)$ denote output, the program is

  $$\begin{aligned} \min \quad & w \cdot x \\ s.t. \quad & f(x) > y. \end{aligned}$$

  This gives rise to a total cost function $C(y)$ with associated marginal cost $C'(y)$.

- In the second stage, the level of output is chosen to maximize profits, $py - C(y)$, which under perfect competition, gives rise to the well-known condition $p = C'(y^*)$, which pins down optimal firm size. If marginal costs are strictly increasing, $p > C'(0)$ and $p < C'(y)$ for high enough y, then an optimal firm size exists and is unique. The associated profits are $[p - C(y^*)/y^*]\, y^*$, thus implying that in a long-run equilibrium with zero profits $p = C'(y^*) = C(y^*)/y^*$, and y^* also minimizes average costs.

- Hart (1995) identifies three caveats with this technological view of the firm:

  1. It ignores incentive problems inside the firm by treating the firm as a perfectly efficient black box. For instance, the profit-maximizer perfectly controls the level of inputs x.
  2. The theory has nothing to say about the internal organization of firms: their hierarchical structure, the extent of authority and delegation...
  3. The theory does *not* pin down firm boundaries. It is better thought of as a theory of plant size than as a theory of firm size.

- Elaborating on point 3, in general, the source of the diseconomies of scale within the firm is unclear. Why are marginal costs increasing? If this is explained by limited span of control by managers, why not set up a second plant and hire a second manager? Why is it assumed that this second plant/manager is outside the firm? As pointed by Coase (1937), neoclassical theory is perfectly consistent with the existence of just one big firm carrying all production in the world (p. 394).

- Incidentally, notice that in the monopolistically competitive setups we discussed earlier in the course, a similar problem arose. Although marginal costs were constant, marginal revenue was decreasing in output (because of downward sloping demand curves). This led to a strictly concave maximization problem and, consequently, to a unique profit-maximizing level of output for each variety. But why should a firm be producing only one variety? Those models are perfectly consistent with the existence of just one firm in each country, so long as each plant producing a different variety does not internalize the effect of their own price on the demand for the other plants' varieties (Coase, 1937, makes a similar point on page 402).

- In the same way that a theory of the firm based purely on *technological* considerations does not constitute a satisfactory theory of the firm, the theories of the multinational firm described in Chapters 6 through 9 cannot be satisfactory either.

The Transaction-Cost Approach: Coase and Williamson

Coase (1937)

- Coase's starting point is that there are are substantial transaction costs associated with running the economic system. Importantly, the size of these transaction costs may vary in market transactions and in intrafirm ones.

- In his view, firms emerge when certain transactions can be undertaken with less transaction costs inside the firm than through the market mechanism: "The main reason why it is profitable to establish a firm would seem to be that there is a cost of using the price mechanism" (p. 390).

- Coase mentions the following as transaction-cost disadvantages of the price mechanism:
  - costs of discovering what the relevant prices are;
  - costs of negotiating and concluding a separate contract for each exchange transaction;
  - costs of specifying all possible contingencies in a long-term contract;
  - taxes on market transactions.
- Coase is a bit more vague in proposing factors that limit the size of the firm. He suggests that increases in firm size will tend to lead to:
  - decreasing returns to the entrepreneur function;
  - increasing cost to allocating factors of production to their best use;
  - increasing supply price of some factors.
- The relative size of these transaction costs will ultimately determine the size of the firm.

Williamson (1985)

- Coase's view of the firm did not instantly become part of mainstream economics. It was criticized for its vagueness and was dubbed tautological. Between 1940 and 1970, the literature focused instead on exploring technological theories of the firm.
- Williamson brought transaction-cost considerations back into the spotlight by making this approach much more operational. Williamson contributed particularly to our understanding of the source of transaction costs associated with using the price system.

- His theory is based on three concepts: (1) bounded rationality, (2) opportunism and (3) asset specificity:

1. Following Herbert Simon, Williamson assumes that economic actors are "*intendedly* rational, but only *limitedly* so". The assumption of **bounded rationality** provides a foundation for the incompleteness of contracts.

   - In particular, in a complex and unpredictable world, boundedly rational agents will be unable to plan ahead for all the contingencies that may arise.
   - Furthermore, even when contingencies are foreseen, it may be hard for contracting parties to negotiate about these plans because of limited capability of describing these possible states.
   - Finally, even when parties can plan and negotiate these contingencies, it may be hard for a third party to verify them and enforce the contract.

   As a result, ex-ante contracts will tend to be incomplete and will tend to be renewed or renegotiated as the future unfolds.

2. By **opportunism**, Williamson means that economic actors are "self-interest seeking with guile" (p. 47). The fact that agents are opportunistic is a necessary condition for the incompleteness of contracts to lead to inefficiencies. If agents could credibly pledge at the outset to execute the contract efficiently, then although the contract would have gaps, renegotiation would always occur in a joint profit maximizing manner.

3. Finally, Williamson points out that certain assets or investments are **relationship-specific**, in the sense that the value of these assets or investments is higher inside a particular relationship than outside of it. This is important because it implies that, at the renegotation stage, parties cannot costlessly switch to alternative trading partners and are partially locked in a bilateral relationship. This is what Williamson calls the "*fundamental transformation*" from an ex-ante competitive situation to one of bilateral monopoly.

- As in Coase, the firm will replace the price system when transaction costs are minimized by transacting inside the firm. Williamson posits that this is more likely to occur (1) the larger the specificity of the assets involved in the transaction, (2) the larger the uncertainty surrounding the transaction, and (3) the more frequent the transactions between the parties.

- Williamson is quite clear in his description of the transaction costs associated with market transactions between two non-integrated firms. His description of intrafirm transactions is somewhat vaguer. What limits the size of the firm?

- Williamson seems to suggest that lock-in effects are less important in intrafirm transactions. He instead appeals to incentive and bureaucratic costs to limit the size of the firm. As long as these "governance costs" are unrelated to specificity, his claim that market transactions dominate integration at low levels of asset specificity is a valid one.

- To fix ideas and to illustrate the effect of relationship-specificity in the choice between intrafirm and market transactions, consider the following model.

A Simple Transaction-Cost Model

- Consider a situation in which the manager of a firm F has a access to a technology for converting a specialized intermediate input into a final good. If the specialized input is of high quality, final-good production generates sales revenues equal to $R(x)$, where x refers to the amount of high quality intermediate input used in production. If the input is of low quality, sale revenues are zero.

- Assume $R'(x) > 0$, $R''(x) < 0$, $\lim_{x\to 0} R'(x) = +\infty$, and $\lim_{x\to\infty} R'(x) = 0$.

- The manager F has two options for obtaining intermediate inputs. It can either manufacture them herself at a marginal cost of $\lambda > 1$ or obtain them from an independent supplier.

- Assuming no frictions inside the firm, the problem of an **integrated structure** is straightforward to solve. In particular, x^V units of high-quality intermediate input will be produced, where x^V is implicitly defined by

$$R'\left(x^V\right) = \lambda.$$

Naturally, the larger λ, i.e., the larger governance costs, the lower x^V. The net profit for the final good producer is

$$\Pi^V = R\left(x^V\right) - \lambda x^V,$$

and by the envelope theorem, $d\Pi\left(\lambda\right)/d\lambda = -x^V < 0$. Hence, net profit are decreasing in λ.

- An **independent supplier** manager S has access to a technology for producing a specialized, high-quality intermediate inputs at a marginal cost of 1. It can also produce low-quality intermediate inputs at a negligible cost.

- The intermediate input is specialized in the sense that the independent supplier tailors it specifically to the final-good producer. In particular, if the contractual relationship between the two parties broke down, the supplier would have access to a technology for converting that input into a final good herself, but in that case sale revenues would be $(1-s)\,R\left(x\right) < R\left(x\right)$. The higher is s, the higher the degree of specifity in the model.

- The setting is one of incomplete contracts. The managers F and S are boundedly rational so they are unable to write an ex-ante enforceable contract specifying the purchase of a specialized intermediate input of a particular quality for a certain price. In addition, the parties cannot sign contracts contingent on the volume of sales revenues obtained when the final good is sold.

- The source of the contract incompleteness could be related to boundedly rational managers failing to write the ex-ante contract in a way that would allow a third party to distinguish between a high-quality and a low-quality intermediate input. The parties could indeed sign an ex-ante contract, but a *self-interested* input supplier would have every incentive to produce a low-quality input at the negligible cost, still cash the price specified in the contract, and face no risk of being penalized by a third party.

- The last assumption is that the initial contract includes an upfront fee for participation in the relationship that has to be paid by S. The purpose of the fee is to secure the participation of S in the relationship at minimum cost to F. When the supply of managers S is infinitely elastic, S's profits from the relationship net of the participation fee are equal in equilibrium to its ex-ante outside option, which we set to zero without loss of generality. This implies that the net profit to F equals joint surplus and the choice of ownership structure is ex-ante efficient.

- The lack of an enforceable ex-ante contract creates a classical **hold-up problem**. The price of the intermediate input will only be determined ex-post, that is, after uncertainty has been resolved and both parties perfectly observe the quality of the input. At this point, the final-good producer manager realizes that the investment incurred by the supplier has a relatively lower value outside the relationship and will thus try to lower the purchase price as much as possible. Foreseeing this, the supplier will have lower incentives to ex-ante invest in x, which will tend to reduce joint surplus.

- To see this formally, assume that in the ex-post bargaining symmetric Nash Bargaining leaves each party with its outside option plus an equal share of the ex-post gains from trade. Because at this point the ex-ante investment as well as the quality of the input are observable to both parties, costless bargaining will yield an ex-post efficient outcome.

- For simplicity, assume that the outside option for the final-good producer is zero (this assumption can easily be relaxed). Then if π_i denotes the Nash bargaining payoff of agent i, the final-good producer manager will obtain,

$$\pi_F = \frac{1}{2}\left(R\left(x\right) - \left(1-s\right)R\left(x\right)\right) = \frac{s}{2}R\left(x\right)$$

On the other hand, the supplying firm manager obtains:

$$\pi_S = \left(1 - \frac{s}{2}\right) R\left(x\right).$$

- Before the bargaining, the manager will set x to maximize $\pi_S - x$, where we use the fact that the marginal cost is equal to 1. This produces

$$\left(1 - \frac{s}{2}\right) R'\left(x^O\right) = 1.$$

  Notice that x^O is decreasing in s, i.e, in the level of specificity, and that $R'\left(x^O\right) > 1$.

- Ex-ante, the upfront fee ensure that the final-good producer obtains all of the surplus and thus.

$$\Pi^O = R\left(x^O\right) - x^O.$$

  Notice that

$$d\Pi^O/ds = \left[R'\left(x^O\right) - 1\right] \frac{dx^O}{ds} < 0,$$

  and hence, net profits for the final good producer are decreasing in s.

- The final-good producer will thus choose intrafirm sourcing versus market procurement whenever $\Pi^V > \Pi^O$. From the results above it is clear that $\Pi^V - \Pi^O$ is decreasing in bureaucratic costs λ and increasing in the degree specificity s.

- Furthermore, if s goes to zero and $\lambda > 1$, it is easy to see that $\Pi^V < \Pi^O$, and market transactions are the preferred mode of organization for transactions with little asset specificity.

- Conversely, if $\lambda \to 1$ and $s \in (0, 1)$, then $\Pi^V > \Pi^O$ and integration is chosen for low incentive and bureaucratic costs of running an integrated structure.

An Example

- Consider the following example, to which we will return in future chapters. Imagine that demand for the good is given by $y = Ap^{-1/(1-\alpha)}$ – looks familiar? – and provided that x is of high quality, $y = x$. Then sale revenues py are given by $R(x) = A^{1-\alpha}x^{\alpha}$. You should convince yourself that this formulation is consistent with the assumptions we made about $R(x)$ above.

- Using the expressions above, we find that under vertical integration we now have

$$x^V = A\left(\frac{\alpha}{\lambda}\right)^{1/(1-\alpha)}$$

and

$$\Pi^V = (1-\alpha)\,A\left(\frac{\alpha}{\lambda}\right)^{\alpha/(1-\alpha)}.$$

- On the other hand, if the input is purchased from an independent supplier, we find

$$x^O = A\left(\left(1-\frac{s}{2}\right)\alpha\right)^{1/(1-\alpha)}$$

and

$$\Pi^O = A\left(1-\left(1-\frac{s}{2}\right)\alpha\right)\left(\left(1-\frac{s}{2}\right)\alpha\right)^{\alpha/(1-\alpha)}.$$

- It is straightforward to see Π^V is decreasing in λ and that Π^O is decreasing in s – remember that $(1-x)\,x^{\alpha/(1-\alpha)}$ is increasing in x for $x < \alpha$. Furthermore, when $\lambda \to 1$, $\Pi^V > \Pi^O$, while when $s \to 0$, $\Pi^V < \Pi^O$.

Chapter 11

The Theory of The Firm: The Property-Rights Approach

- As we discussed in Chapter 10, Coase (1937) emphasized the existence of substantial transaction costs associated with market transactions. Building on his work, Williamson (1985) provided a theory of the boundaries of the firm in which the endogenous benefits of integration are explained appealing to bounded rationality, opportunism and asset specificity.

- The seminal paper by Grossman and Hart (1986) provides the first unified theoretical framework that features both endogenous benefits **and endogenous costs** of integration. Remember that the costs of integration were exogenous in the simple model above.

- The Grossman-Hart approach starts by arguing that it is not satisfactory to assume that the contractual frictions that plague the relationship between two nonintegrated firms disappear when these firms integrate. After all, inside firms too agents are boundedly rational and opportunistic, and it is not likely that integration will change asset specificity. What defines then the boundaries of the firm?

- Grossman and Hart suggest that **ownership is a source of power** when contracts are incomplete. What does this mean?

  - First notice that when a particular firm decides to integrate another firm say a supplier, it is acquiring the suppliers' assets. These consist of physical and other nonhuman assets (machines, buildings, inventories, patents, copyrights......). Absent slavery, the human capital of workers in the supplying firm belong to them both before and after the acquisition.
  - Remember that when contracts are incomplete, the parties will often encounter contingencies that were not foreseen in the initial contract. In those situations, who decides on the usage of the physical assets? According to the property-rights approach, the owner of the asset has these residual rights of control.
  - These residual rights of control are important because they are likely to affect how the surplus is divided ex-post. In particular, in the presence of unforseen contingencies, an opportunistic asset owner will tend to decide on the use of the asset that maximizes his payoff in his ex-post bargaining with the supplier. This is the sense in which ownership is a source of power.

- Grossman and Hart then show that in the presence of relationship-specific investments, these considerations lead to a theory of the boundaries of the firm in which both the benefits and the costs of integration are endogenous.
  - As we saw in the simple model above, in the presence of incomplete contracts, parties have reduced incentives to undertake relationship-specific investments.
  - Furthermore, the incentives to invest for a particular party are increasing in the share of the surplus that accrues to that party. In the model above, the fraction of the surplus obtained by the supplier was decreasing in s, and so was x^O.
  - In a set up in which both the integrating and integrated parties undertake relationship-specific investments, the **benefit of integration** is that increases the incentives of the integrating firm to make investments that are partially specific to the integrated firm.

  - On the other hand, the cost of integration is that it reduces the the incentives of the integrated firm to make investments that are partially specific to the integrating firm.

- Let us look at a formal version of their model, as described in Chapter 2 of Hart (1995).

The Formal Model

Basic Set-up

- There are two managers M1 and M2 who operate two assets a1 and a2. M2 uses a2 to produce a single unit of input (a widget) which he supplies to M1. M1 uses a1 and this widget to produce a final good. Our interest is in solving for the optimal ownership structure, that is, for the optimal allocation of assets to managers.

- We are going to focus on the following three cases:

  - *Non-integration:* M1 owns a1 and M2 owns a2.
  - *Backward integration:* M1 owns a1 and a2.
  - *Forward integration:* M2 owns a1 and a2.

- The relationship between M1 and M2 lasts for two periods. At date 0, M1 and M2 make relationship-specific investments and, at date 1, M2 supplies the widget to M1. We can think of these investments as making the assets a1 and a2 more productive *within* that specific relationship.

- The parties have symmetric information throughout and there is no uncertainty about costs and benefits. Nevertheless, at date 0 there is uncertainty about the specific type of asset that M1 will require. This uncertainty is resolved at date 1, but remember that at this point the investments have already been made.

- This uncertainty implies that any ex-ante contract is infeasible, because the agents are unable to describe this widget in a contract, and thus no third party could verify that the contract has indeed been honored.

- The parties thus bargain at date 1 over the terms of trade. Because M1 and M2 have symmetric information this ex-post bargaining will deliver an efficient outcome. As in Chapter 10, let us assume that in the bargaining each party obtains an equal share of the ex-post gains from trade.

- As in Chapter 10, it is also assumed that the parties are not cash constrained, which ensures that the ownership structure chosen at date 0 maximizes joint surplus.

Investments and Payoffs

- Denote by i M1's relationship-specific investment at date 0. This investment affects the payoff of M1 both when a trade with M2 occurs and when it does not. But the effect is different in both cases. If trade occurs, M1 is left with an ex-post payoff of

  $$R(i) - p$$

  where p is the price paid for the widget. If instead trade does not occur, M1 can still buy a non-specific widget from the spot market at price $\overline{p}$, and obtain a

  $$r(i; A) - \overline{p},$$

  where A refers to the set of assets available to M1 at date 1, and hence $A \subset \{\{a1, a2\}, \{a1\}, \varnothing\}$.

- Similarly, let e denote M2's relationship-specific investment at date 1. Assume that e affects the unit cost for producing the widget at date 1: the higher is e, the lower $C(e)$ is. If trade occurs, M2 thus obtains a payoff equal to

  $$p - C(e).$$

  If trade does not occur, M2 will still be able to sell the widget in the spot market, but the unit costs will change, as the widget will need to be made less specific. Again, the set of assets available to M2 in case of a failure to trade are likely to affect this ex-post unit cost. We thus express this ex-post payoff to M2 as

$$\overline{p} - c(e; B),$$

where B refers to the set of assets available to M2 at date 1.

- We assume that there are ex-post gains from trade, i.e., the total surplus if trade occurs is higher than if trade does not occur:

$$\begin{gathered} R(i) - C(e) > r(i; A) - c(e; B) \text{ for all } i, e \\ \text{and all } A, B, \text{ such that} \\ A \cap B = \varnothing \text{ and } A \cup B = \{a1, a2\}. \end{gathered}$$

This reflects that the investments i and e are relationship specific.

- Furthermore it is assumed that the marginal returns to the investments i and e are weakly increasing in the amount of assets available to the corresponding party, and that these *marginal* products are strictly higher if trade occurs:

$$R'(i) > r'(i; a1, a2) \geq r'(i; a1) \geq r'(i; \varnothing) \text{ for all } i \tag{11.1}$$

and

$$|C'(e)| > |c'(e; a1, a2)| \geq |c'(e; a2)| \geq |c'(e; \varnothing)| \text{ for all } e. \tag{11.2}$$

- Assume also that $R' > 0$, $R'' < 0$, $C' < 0$, $C'' > 0$, $r' \geq 0$, $r'' \leq 0$, $c' \leq 0$, $c'' \geq 0$.

- R, r, C, c, i, e are observable to all parties, but not verifiable by a third party, and thus are non-contractibles.

First-Best Choice of Investments

- Suppose that M1 and M2 were able to sign an ex-ante enforceable contract. Then the contract would stipulate the level of investments i and e that maximizes:

$$R(i) - C(e) - i - e.$$

This yields:

$$\begin{aligned} R\left(i^{*}\right) &= 1 \\ \left|C'\left(e^{*}\right)\right| &= 1. \end{aligned}$$

Second-Best Choice of Investments

- When no contract is signed at date 0, the parties bargain over the terms of trade at date 1. For a given distribution of assets A and B, M1 anticipates obtaining

$$\pi_1 = r\left(i;A\right) - \overline{p} + \frac{1}{2}\left[R\left(i\right) - C\left(e\right) - r\left(i;A\right) + c\left(e;B\right)\right] \tag{11.3}$$

  in the symmetric Nash bargaining. On the other hand, M2 anticipates obtaining[1]

$$\pi_2 = \overline{p} - c\left(e;B\right) + \frac{1}{2}\left[R\left(i\right) - C\left(e\right) - r\left(i;A\right) + c\left(e;B\right)\right]. \tag{11.4}$$

- At date 0, M1 and M2 set i and e to maximize (11.3) and (11.3), respectively, net of investment costs. This yields:

$$\frac{1}{2}R'\left(i\right) + \frac{1}{2}r'\left(i;A\right) = 1$$

  and

$$\frac{1}{2}\left|C'\left(e\right)\right| + \frac{1}{2}\left|c'\left(e;B\right)\right| = 1.$$

- Using (11.1) and (11.2), $R'' < 0$ and $C'' > 0$, it is clear from these first order conditions that the second best investments satisfy $i < i^{*}$ and $e < e^{*}$.

- Intuitively, with incomplete contracts, the threat of contractual breach coupled with the specificity of assets imply that parties only capture a fraction of the marginal return to their investments in the ex-post bargaining.

- Furthermore, using (11.1) and (11.2) one easily show that:

$$i^{*} > i_B > i_N > i_F \tag{11.5}$$

[1] Notice that the implied price of the widget is $p = \overline{p} + \frac{1}{2}\left[R\left(i\right) - C\left(e\right) - r\left(i;A\right) + c\left(e;B\right)\right]$.

and

$$e^* > e_F > e_N > e_B, \tag{11.6}$$

where a subscript B denotes backward integration, a subscript N denotes no integration, and a subscript F denotes forward integration.

- Inequalities (11.1) and (11.2) illustrate how the model features endogenous benefits and costs of integration. For instance, let us analyze a shift from nonintegration to backward integration. This raises the bargaining power of M1, thus increasing M1's incentives to make relationship specific investments. As a result, i moves closer to its first-best level. On the other hand, integration reduces the share of the surplus that M2 obtains ex-post, and e *moves further away* from its first-best level. In sum, although integration reduces the hold-up problem faced by M1, it increases M2's hold-up.

Choice of Ownership Structure

- As argued above, it is assumed that the ownership structure chosen at date 0 maximizes joint surplus.

  - This can be justified, as in Chapter 10, through an ex-ante lump-sum transfer paid by M2 to M1 to participate in the relationship. Because at date 0 no relationship-specific investment has been made, it is possible to specify that M1 faces an ex-ante perfectly elastic supply of M2 agents (remember Williamson's *fundamental transformation*!). In such case, if the outside option of M2 agents is normalized to 0, M1 is able to appropriate all the surplus ex-ante and, hence, self-interested payoff maximization ensures that the ownership structure also maximizes joint surplus.

- Hence the optimal ownership structure $k \in \{N, B, F\}$ is the solution to:

$$\max_k R(i_k) - C(e_k) - i_k - e_k.$$

Analysis of the Optimal Ownership Structure

- In order to derive predictions for which factors will affect the optimal ownership structure, Grossman and Hart's approach is to define certain concepts and see how they affect the integration decision:

  Definition 4 *M1's investment is* ***inelastic*** *in the range* $1/2 \leq \rho \leq 1$ *if the solution to* $\max_i \rho R(i) - i$ *is independent of* ρ*. Similarly, M2's investment is* ***inelastic*** *in the range* $1/2 \leq \sigma \leq 1$ *if the solution to* $\min_e \sigma C(e) + e$ *is independent of* σ *in this range.*

  This is equivalent to assuming that the choices of i and e are independent of firm boundaries (see eq. 11.3 and 11.4). It is thus not surprising that:

  Proposition 5 *If M2's (M1's) investment is inelastic, then backward (forward) integration is optimal.*

  Definition 6 *M1's will be said to become* ***relatively unproductive*** *if* $R(i)$ *is replaced by* $\theta R(i) + (1-\theta) i$ *and* $r(i; A)$ *is replaced by* $\theta r(i; A) + (1-\theta) i$ *for all* A*, where* $\theta > 0$ *is small. Similarly, M2's will be said to become* ***relatively unproductive*** *if* $C(e)$ *is replaced by* $\theta C(e) - (1-\theta) e$ *and* $c(i; B)$ *is replaced by* $\theta c(i; B) - (1-\theta) e$ *for all* B*.*

  This implies that M1's (respectively, M2's) net social return becomes $\theta(R(i) - i)$ (respectively, $\theta(C(e) + e)$), and is thus lower the lower θ is. As an agent's net social return decreases, the ownership structure will focus on ensuring that the other agent (the relatively productive one) has the right incentives to invest. Formally,

  Proposition 7 *If M2's becomes relatively unproductive and* $r'(i; a1, a2) > r'(i; a1)$ *then backward integration is optimal. If M1's becomes relatively unproductive and* $|c'(e; a1, a2)| > |c'(e; a2)|$ *then forward integration is optimal.*

Definition 8 *Assets a1 and a2 are* ***independent*** *if* $r'(i; a1, a2) = r'(i; a1)$ *and* $c'(e; a1, a2) = c'(e; a2)$*.*

This means that access to a2 (a1) does not strengthen M1's (M2's) ex-post outside option. Nonsurprisingly, this implies that:

Proposition 9 *If assets a1 and a2 are independent then nonintegration is optimal.*

Definition 10 *Assets a1 and a2 are* ***strictly complementary*** *if* $r'(i; a1) = r'(i; \varnothing)$ *and* $c'(e; a2) = c'(e; \varnothing)$*.*

This implies that access to a1 (a2) does not strengthen M1's (M2's) ex-post outside option unless M1 (M2) also has access to a2 (a1). It this follows that:

Proposition 11 *If assets a1 and a2 are strictly complementary then some form of integration is optimal.*

Definition 12 *M1's human capital is* ***essential*** *if* $c'(e; a1, a2) = c'(e; \varnothing)$*. M2's human capital is essential if* $r'(e; a1, a2) = r'(e; \varnothing)$*.*

This means that ownership of assets has no effect on ex-post outside options. Hence,

Proposition 13 *If M1's (M2's) human capital is essential, then backward (forward) integration is optimal. If both M1's and M2's human capital are essential then all ownership structures result in the same joint surplus.*

Discussion

- All the results above seem quite intuitive. Furthermore, Grossman and Hart (1986) and Hart (1995) discuss how these predictions seem to be in line with real-life phenomena (see, especially, Hart, 1995, pp. 49-55).

- The Grossman-Hart approach has been criticized for focusing exclusively on the incentives of top executives to make relationship-specific investments. Hart and Moore (1990) develop a property-rights theory of the boundaries of the firm in which ownership of nonhuman assets affects the incentives of workers.

  - Their approach is based on the idea that the difference between integration and subcontracting is that in the former case the integrating party can selectively fire the workers of the supplying firm, whereas under subcontracting it can only "fire" the entire firm (i.e., terminate the relationship).
  - As Hart and Moore show, this has implications for how the surplus is divided between managers and workers, and thus, in an incomplete-contracting setting, firm boundaries have an effect on the incentives of workers to undertake relationship-specific investments.
  - In particular, they develop a multi-agent bargaining model (using tools from cooperative game theory) and derive a series of interesting results. For instance, they show that ownership of assets should reside in the hands of those with important human capital that is complementary to these assets.

- The Grossman-Hart-Moore approach to the theory of the firm has had a huge impact in the profession. It has generated a lot of research both applying similar concepts to other fields (e.g., the incomplete-contracting approach of Aghion and Bolton, 1992, in corporate finance) and also studying the robsutness of some of the predictions of the model (e.g., De Meza and Lockwood, 1998, Rajan and Zingales, 1998).

- A related, much more theoretical, literature has formally studied the foundations of incomplete contracts:

  - Among other people, Maskin and Tirole have argued that in the property-rights approach there is a tension between the fact that certain objects are assumed to be non-verifiable to outside parties but observable to both parties in the transaction. This raises the issue of why the inside parties are not able to truthfully reveal this information to outside parties.

- This idea has been formalized by Maskin and Tirole (1999), who borrow tools from the mechanism design literature to illustrate this point.
- Segal (1999) and Hart and Moore (1999) have in turn replied to this criticism. In particular, Hart and Moore (1999) show how Maskin and Tirole's (1999) critique relies heavily on the assumption that the parties can commit not to renegotiate an ex-ante contract.

A Variant of the Grossman-Hart-Moore Model (Antràs, 2003)

- Recently, Antràs (2003) has developed a variant of the Grossman-Hart-Moore model that delivers a result analogous to Proposition 2. Because we will cover the paper later, it might be useful to briefly discuss how a simplified version of his set-up relates to the one discussed above.

- There are two agents, H and M. H controls the provision of a relationship-specific input h, whereas S controls the provision of another relationship-specific input m. When combined, these inputs produce a final good y according to the technology:
$$y = \left(\frac{h}{\eta}\right)^{\eta} \left(\frac{m}{1-\eta}\right)^{1-\eta}.$$

- Demand for the final good is
$$y = Ap^{-1/(1-\alpha)},$$
and, hence, sale revenues are given by:
$$R(h,m) = A^{1-\alpha} y^{\alpha} = A^{1-\alpha} \left(\frac{h}{\eta}\right)^{\alpha\eta} \left(\frac{m}{1-\eta}\right)^{\alpha(1-\eta)}.$$

- Intermediate inputs are produce with a composite factor of production (which we take as the numeraire) with an input-output coefficient of one.

- The relationship between H and M lasts for two periods. At date 0, H and M produce the relationship-specific inputs h and m and at date 1, they bargain over the division of the surplus (e.g., a price paid by H for the use of m).

- Ex-ante enforceable contracts are not enforceable, so the surplus is divided after the costs of production in h and m have been incurred. For simplicity, assume that these costs are fully specific to the relationship, in the sense that their outside value is zero.

- We will focus on the choice between non-integration and backward integration. H is essential for the final good to generate positive sale revenues, so forward integration is a weakly dominated choice (alternatively, we could appeal to financial constraints to rule out this organizational mode endogenously).

- The only difference between integration and nonintegration is that only in the former case can H selectively fire M in case trade fails to occur.
  - To see how this can have an impact on the outside options in the ex-post bargaining, notice that under integration the production facility where M works and where m is sitting are owned by H. Hence, if M is refusing to trade, H has the option of firing M and seizing the amount of m that has already been produced. Assume that this comes at the cost of a loss of a fraction δ of final-good production.
  - On the other hand, under nonintegration, if trade fails to occur, H is left with nothing.

- For simplicity, let the outside option of M be zero regardless of ownership structure. This is equivalent to assuming that (1) m is fully tailored to H and is useless to anybody else, and (2) M does not have a technology for converting m into y.

- Assuming symmetric Nash bargaining, one can show that the optimal ownership structure solves

$$\begin{aligned}
\max_{k\in\{V,O\}} \quad & \Pi_k = R\left(h_k, m_k\right) - h_k - m_k \\
s.t. \quad & h_k = \arg\max_h \beta_k R\left(h, m_k\right) - h \\
& m_k = \arg\max_m \left(1-\beta_k\right) R\left(h_k, m\right) - m
\end{aligned}$$

where $\beta_V = \frac{(1+\delta^\alpha)}{2} > \frac{1}{2} = \beta_O$.

- One can show that this leads to:

$$\Phi(\eta) = \frac{\Pi_V}{\Pi_O} = \left(\frac{1 - \alpha \left(\frac{\eta(1+\delta^\alpha)}{2} + (1-\eta) \frac{(1-\delta^\alpha)}{2} \right)}{1 - \frac{\alpha}{2}} \right) \left((1+\delta^\alpha)^\eta (1-\delta^\alpha)^{1-\eta} \right)^{\alpha/(1-\alpha)}.$$

- It is possible to show that $\Phi(0) < 1$, $\Theta'(\eta) > 0$ and $\Theta(1) > 1$. Hence, there is a threshold $\widehat{\eta}$ under which nonintegration is chosen and over which integration is chosen.

- As in Grossman and Hart, it is optimal to assign residuals rights of control to the party undertaking a relatively more important, productive investment in the relationship. But the "importance" of production is here directly linked to the output elasticity of that agent's investment.

Chapter 12

The Theory of The Firm: Alternative Approaches

- The purpose of this chapter is to provide a brief introduction to three alternative approaches to the theory of the firm. The first attempts to study the boundaries of the firm in a theoretical framework in which workers' incentives play a central role. The second and third focus more on the internal organization of firms, emphasizing the importance of the allocation of authority and of the assignment of personnel to hierarchial positions.

- It is important to note that Chapters 10 through 12 do not attempt to provide a comprehensive discussion of *all* the available theories of the firm. Rather, I have selected only those theories of the firm that have already been applied to the study of the international organization of production.

12.1 The Firm as an Incentive System: Holmstrom and Milgrom (1994)

- Holmstrom and Milgrom (1994) – start by emphasizing that most approaches to the theory of the firm tend to be *unidimensional.*

  - As we saw in Chapter 11, Grossman and Hart (1986) focus on **ownership of assets** as a source of power when contracts are incomplete.

- Alchian and Demsetz (1972) and Holmstrom (1982) stress instead issues related to **monitoring and worker compensation** as determining the boundaries of the firm.
- Coase (1937) and Simon (1951) emphasize instead the **discretion** that the employer has to direct his employee's activities.

- Holmstrom and Milgrom do not deny that each of these views captures important factors determining the make-or-buy decision. In their view, asset ownership, contingent rewards, and job restrictions, all have an influence on workers' incentives and, in particular, they affect how workers divide their attention or effort among different tasks.

- But they make the following interesting observation:

> "Why does inside procurement tend to involve production by a worker who is supervised by the firm *and* uses the firm's tools *and* is paid a fixed wage? Why does outside procurement tend to involve purchases from a worker who chooses his or her own methods *and* hours and owns the tools used *and* is paid only for quantities supplied?" (p. 972)

- This suggests that the optimal organization of production tries to keep the various incentives to the worker *in balance.* Weak incentives for maintaining asset values should go with weak incentives to exert effort in narrowly measured performance and with weak incentives (or rather *no* incentives) related to certain job restrictions.

- Why should the different incentives to the worker be kept in balance? In Holmstrom and Milgrom's theory, this feature comes from the assumption that workers view the different tasks they perform as **substitutes**. Increasing the incentive (or reward) for just one task will thus tend to cause the worker to devote too much attention or effort to that particular task, while neglecting other tasks in his job. Balancing incentives constitutes a simple way to ensure that workers exert similar effort on all tasks.

- In other words, due to task substitutability, in the incentive problem that delivers the optimal organizational structure, the levels of incentives provided for the different tasks of a worker tend to be **complementary**.

- Building on previous work of theirs on the optimality of linear contracts (Holmstrom and Milgrom, 1987), they develop a theoretical framework in which the optimal incentive problem is solved in terms of a set of exogenous parameters that tend to favor internal or external procurement. Using statistical concepts that generalize the concept of covariance, they show how exogenous changes in these parameters can plausibly create comovements in the incentive instruments of the sort identified in the quote above.

- I next develop a simplified variant of the Holmstrom-Milgrom set up, in order to illustrate how workers' task substitutability leads to the levels of incentives being complementary in the incentive problem.

A Simple Model

General Set-up

- Assume a situation in which a worker (or Agent) allocates effort among two activities: t_1 and t_2. The worker's utility function is given by

$$U_A = -\exp\left(-rA\right),$$

  where $r > 0$ is the worker's coefficient of absolute risk aversion and A denotes his income. The employer (or Principal) is risk neutral.

- Effort is not directly observable but it can be monitored indirectly via a performance measure:

$$X\left(t_1, t_2\right) = F\left(t_1, t_2\right) + \varepsilon_X,$$

  where $\varepsilon_X \sim N\left(0, \sigma_X^2\right)$.

- In a remarkable paper, Holmstrom and Milgrom (1987) develop a dynamic principal-agent model in which, under certain stationarity assumptions, the optimal incentive contract coincides with that of a reduced-form static model. Furthermore, the optimal incentive scheme takes the linear form:

$$s(X) = \alpha X + \beta,$$

where α can be interpreted as a commission rate, while β is salary.

- In addition, there is an asset associated with a transferrable return $Y(t_1, t_2)$. This return can be allocated between the employer and the worker. We denote ownership structure by λ, and so in general

  $$Y(t_1, t_2) = Y_A(t_1, t_2; \lambda) + Y_P(t_1, t_2; \lambda),$$

  where $Y_A(t_1, t_2; \lambda)$ can be interpreted (following Grossman and Hart, 1986) as the share of returns that the agent is able to appropriate *ex-post* under ownership structure λ.

- The asset returns are random and

  $$Y(t_1, t_2) = G(t_1, t_2) + \varepsilon_Y,$$

  where $\varepsilon_Y \sim N(0, \sigma_Y^2)$, and we assume that ε_X and ε_Y are jointly normally distributed.

- Exerting effort is costly to the worker. Let $C(t_1, t_2)$ denote this cost and assume that $C_1(t_1, t_2) > 0$, $C_2(t_1, t_2) > 0$, $C_{11}(t_1, t_2) > 0$, $C_{22}(t_1, t_2) > 0$, and (crucially) $C_{12}(t_1, t_2) > 0$.

- The worker has an outside opportunity that delivers an income stream of $\overline{w}$ with certainty.

A Particular Case

- To illustrate the intuition behind the results, it is useful to follow Holmstrom and Milgrom and concentrate on the case in which $F(t_1, t_2)$ is independent of

t_2, while $Y_A(t_1, t_2; \lambda)$ is independent of t_1. In particular, we are going to impose $F(t_1, t_2) = t_1$, $G(t_1, t_2) = t_2$ and $Y_A(t_1, t_2; \lambda) = \lambda t_2$.

- Let us also assume that ε_X and ε_Y are independent.

Optimal organizational design

- An *organizational design* therefore consists of a choice of a commission rate α, a salary β, and an ownership structure λ.

- Given our assumptions, the worker's expected utility has a certainty equivalent equal to
$$ACE = \alpha t_1 + \lambda t_2 + \beta - C(t_1, t_2) - \frac{r}{2}\left(\alpha^2 \sigma_X^2 + \lambda^2 \sigma_Y^2\right).$$

- The optimal organizational design then solves the problem:

$$\begin{aligned} \max_{\alpha,\beta,\lambda} \quad & (1-\alpha) t_1 + (1-\lambda) t_2 - \beta \\ s.t. \quad & (t_1, t_2) = \arg\max_{t_1', t_2'} \left\{\alpha t_1' + \lambda t_2' + \beta - C(t_1', t_2')\right\} \\ & ACE \geq \overline{w} \end{aligned}$$

- The employer has every incentive to make the incentive rationality constraint hold with equality, i.e., $ACE = \overline{w}$, which plugging in the objective function delivers an equivalent formulation of the problem:

$$\begin{aligned} \max_{\alpha,\lambda} \quad & t_1 + t_2 - C(t_1, t_2) - \frac{r}{2}\left(\alpha^2 \sigma_X^2 + \lambda^2 \sigma_Y^2\right) \\ s.t. \quad & (t_1, t_2) = \arg\max_{t_1', t_2'} \left\{\alpha t_1' + \lambda t_2' - C(t_1', t_2')\right\}, \end{aligned}$$

with β being set such that $ACE(\alpha, \lambda) = \overline{w}$.

- The incentive compatibility constraint imposes the following constraints on the levels of incentives:

$$\alpha = C_1(t_1, t_2) \tag{12.1}$$
$$\lambda = C_2(t_1, t_2) \tag{12.2}$$

- We can express the solution of (12.1) and (12.2) as $t_1(\alpha, \lambda)$ and $t_2(\alpha, \lambda)$. Defining, $\Delta = C_{11}C_{22} - (C_{12})^2$, total differentiation of (12.1) and (12.2) delivers:

$$\begin{aligned} \frac{dt_1}{d\alpha} &= \frac{C_{22}}{\Delta} > 0; \frac{dt_1}{d\lambda} = \frac{-C_{12}}{\Delta} < 0 \\ \frac{dt_2}{d\lambda} &= \frac{C_{11}}{\Delta} > 0 \ ; \frac{dt_2}{d\alpha} = \frac{-C_{12}}{\Delta} < 0. \end{aligned}$$

- These inequalities reflect the substitutability of tasks. Increasing the incentive for one task (say task 1) increases the worker's effort on that task, but reduces the worker's effort on the other task (say task 2). It should be clear that substitutability is driven by the assumption $C_{12} > 0$.

- The optimal incentives α and λ thus maximizes

$$TCE = t_1(\alpha, \lambda) + t_2(\alpha, \lambda) - C(t_1(\alpha, \lambda), t_2(\alpha, \lambda)) - \frac{r}{2}\left(\alpha^2\sigma_X^2 + \lambda^2\sigma_Y^2\right).$$

Proposition 14 *(Holmstrom and Milgrom) If* $\frac{d^2t_1}{d\alpha d\lambda} \geq 0$ *and* $\frac{d^2t_2}{d\alpha d\lambda} \geq 0$*, the function* TCE *is supermodular on the domain where* $\sigma_X^2 \geq 0$*,* $\sigma_Y^2 \geq 0$*,* $\alpha \leq 1$ *and* $\lambda \leq 1$*.*

- This implies that the two kinds of incentives are *complementary* in the relevant range. In other words, a change in an exogenous parameter (say a fall in σ_X^2) that tends to increase the optimal incentive level for a particular task, simultaneously lowers the opportunity cost of raising the incentives for the other task.

- This simple model is thus able to predict comovements in α and λ of the type described in the quote before, by which outside procurement tends to be characterized by both relatively high commission rates (high α) and by worker's ownership of assets (high λ), whereas inside procurement is done by workers who earn a fixed rate (low α) and use the firm's tools (low λ).

- The model can easily be extended to include additional tasks and instruments for affecting incentives, such as job restrictions (see the paper for details).

12.2 Formal and Real Authority: Aghion and Tirole (1997)

- Remember that in the Grossman-Hart-Moore theory of the firm only the owner of a particular asset has control rights over such asset. In other words, the owner of an asset has *formal authority* over decisions concerning the use of the asset.

- Aghion and Tirole's (1997) starting observation is that in practice ownership of assets does not necessarily confer *real authority* or effective control over decisions. To illustrate this point they develop a theoretical framework that stresses the role of informational asymmetries between a Principal (or manager) and an Agent (or worker).

- In their model, a separation between formal authority and real authority emerges when the Agent is much better informed than the Principal about the best way to use a particular asset.

- Interestingly, information acquisition is endogenous in the model. This allows an analysis of how the allocation of formal authority affects the incentives of parties to acquire information, and thus endogenously determines real authority within organizations.

Set-up

- Consider a hierarchy composed of a principal (P herafter) and an agent (A hereafter), who is hired to collect information and potentially implement a project.

- There are $n \geq 3$ possible projects to choose from. With each project $k \in \{1, ..., n\}$ is associated a verifiable monetary benefit B_k for P and a private benefit b_k for A. If no project is implemented, P and A obtain B_0 and b_0, respectively.

- For each party, at least one project delivers a very large negative payoff, so that uninformed parties (in a sense to be discussed below) will have an incentive to recommend inaction or rubber-stamp decisions from informed parties, rather than pick projects at random.

- Each agent has a preferred project. P's preferred project leaves P with a benefit of B and A with a payoff of βb, where $\beta \in (0,1]$. On the other hand, A's preferred project yields αB to P and b to A, where again $\alpha \in (0,1]$. The parameters α and β can be interpreted as *congruence* parameters.

- P is risk neutral and has utility $B_k - w$ if project k is chosen. For simplicity, it is assumed that A is infinitely averse to income risk and earns a fixed wage equal to his reservation wage, which is normalized to zero. Hence, A's payoff consists only of the private benefit b_k.

- Information acquisition works as follows. At private cost $g_A(e)$, A learns the payoffs of all projects with probability e and remains completely uninformed with probability $1-e$. Similarly, at private cost $g_P(E)$, P becomes perfectly informed with probability E and learns nothing with probability $1-E$. For simplicity, it is assumed that P and A acquire information simultaneously.

- Assume that $g_A(\cdot)$ and $g_P(\cdot)$ are increasing, strictly convex and satisfy $g_i(0) = 0$, $g_i'(0) = 0$, and $g_i'(1) = \infty$, $i = A, P$.

- We distinguish between two organizational forms:

  - In the P-Organization (or *integration*), P has formal authority in the sense that it can always overrule A, and will of course do so when P is informed. In such case, P has both formal *and* real authority. Conversely, an uninformed P will rubber-stamp a suggestion from A because $\alpha > 0$ (remember that A will only make a suggestion if he is informed!).

  - In the A-Organization (or *delegation*), A has formal authority in the sense that P cannot overrule A.

- The setting is one of incomplete contracts. The initial contract only specifies an allocation of formal authority.

Payoffs under the P-Organization and the A-Organization

- Consider first the P-Organization. With probability E, P will be informed and will thus pick the project that yields him B. With probability $(1-E)\,e$, P will be uninformed but A will be informed and will thus suggest his preferred project, which P will rubber-stamp because $\alpha B > 0$. In sum, P's expected payoff is:

$$u_P = EB + (1-E)\,e\alpha B - g_P(E).$$

  Similarly, A's expected (private) payoff is

$$u_A = E\beta b + (1-E)\,eb - g_A(e).$$

  Notice that although A is infinitely averse to income risk, he is risk neutral in terms of non-monetary payoffs.

- In a similar way, it is straightforward to compute the payoffs under the A-Organization or delegation:

$$\begin{aligned} u_P^d &= e\alpha B + (1-e)\,EB - g_P(E) \\ u_A^d &= eb + (1-e)\,E\beta b - g_A(e). \end{aligned}$$

Information Acquisition under the P-Organization and the A-Organization

- In the case of P-Organization, e_P and E_P are determined by the intersection of the reaction curves

$$\begin{aligned} (1-\alpha e)\,B &= g_P'(E) \\ (1-E)\,b &= g_A'(e). \end{aligned}$$

  Notice from the first equation that E is higher (P supervises more) the higher is B, the lower is α and the lower e. On the other hand, from the second equation, e is decreasing in E: A shows more initiative the lower P's interference.

- On the other hand, under the A-Organization, the analogous reaction functions are

$$
\begin{aligned}
(1-e)\,B &= g_P'(E) \\
(1-\beta E)\,b &= g_A'(e)\,,
\end{aligned}
$$

which deliver e_A and E_A.

- It is straightforward to check that $E_P > E_A$ and that $e_P < e_A$. In words, delegation increases the initiative of the agent, because he holds formal authority and the principal cannot overrule him. This is the benefit of delegation. On the other hand, the cost of delegation is that the principal loses control, in the sense that if both agents are informed his preferred project is not picked. This is costly because anticipating this the principal invests less in information acquisition.

- The optimal organization of the firm (the allocation of formal authority) is thus the result of a **trade-off between loss of control and initiative**.

- The different parameters of the model affect the choice between the P-Organization and the A-Organization in non-surprising ways. Higher B and higher β lead to more integration, whereas higher b and higher α favor delegation.

- Aghion and Tirole (1997) then go on to study extensions of the model that shed some light on the internal organization of the firm. For instance, in a multi-agent extension of the model, it is shown that the principal has an incentive to run the firm in a situation of overload (where the marginal profit of an extra employee is negative) so as to credibly commit to reward initiative. You are most encouraged to look at these extensions in detail.

12.3 Authority and Hierarchies: Rosen (1982)

- The theories of the firm we have discussed so far stress the role of incentives and downplay the role of technological factors. The rationale for this focus on incentives was laid out in Chapter 10. In particular, the neoclassical, technological theory of the firm is better thought of as a theory of plant size than as a theory of firm size.

- Still, in the last few years there have been interesting developments in the technological view of the firm. Lucas (1978) and Rosen (1982) develop simple theories that explictly model diminishing returns to the entrepreneurial function and deliver endogenous limits to the span of control. On top of this, Rosen (1982) looks at the internal organization of firms and, in particular, provides an endogenous theory of hierarchies.

- Recent contributions to this technological view of the firm include the work of Garicano (2000) and Garicano and Rossi-Hansberg (2003).

- Let us briefly describe Rosen's (1982) contribution.

Set-up

- Consider a multilevel or hierarchical firm, where output of levels below the top is an intermediate input that is improved by the activities of workers in the next highest level. The ouput of the highest level is sold in the open market.

- Workers are divided into these ranks or hierarchies. Let R_j index rank, where j is one plus the number of ranks below workers in rank R_j. Hence, R_1 corresponds to production workers, R_2 corresponds to heads of two-layer firms or to second-line subordinates of larger organizations, etc.

- Some more notation:

  - Let q_i denote the skill or productivity of worker i in rank R_1. This productivity is allowed to vary across workers.

  - Let r denote the skill of a second-line manager, which can also differ across managers in R_2.

  - Let t_i be the time r allocates to monitoring q_i.

- The product attributable to r controlling q_i is given by

  $$x_i = g(r) f(rt_i, q_i),$$

  where $f(\cdot)$ is a standard constant-returns-to-scale neoclassical function and $g'(r) \geq 0$.

- A particular second-line manager can have various production workers under his supervision. Aggregating across these workers, a particular manager in R_2 produces

  $$X = g(r) \sum_i f(rt_i, q_i), \tag{12.3}$$

  where both i and t_i are endogenous.

- Notice that through the function $g(r)$, the skill of r increases the marginal product of all workers below him, irrespective of how large i is. This introduces a scale economy or increasing-returns element in the model. On the other hand, the fact that r needs to spend time monitoring each of the workers i – this is the first argument of $f(\cdot)$ – is the force that will ultimately limit the scope of control and the size of the firm.

- Consider next a manager in R_3 with talent s. Let y_j, the output generated by this manager when it manages the output X_j of a second-line manager r_j, be given by

  $$y_j = G(s) F(sv_j, X_j),$$

  where v_j is the time allocated to monitoring or supervising r_j. Total output of the R_3 firm is thus

  $$Y = G(s) \sum_j F(sv_j, X_j),$$

  where again both j and v_j are endogenous.

- Production at higher levels is defined analogously.

- To complete the model we need to specify factor supplies. Each person is endowed with a vector of *latent* skills $(q, r, s, ...)$. By assumption, each person ends up being assigned to a unique rank, and hence only one of these skills will be used in equilibrium.

- Latent skills are distributed in the population according a nonhomogenous one-factor structure:

$$\begin{aligned} q &= a_q + b_q \xi \\ r &= a_r + b_r \xi \\ s &= a_s + b_s \xi \\ &\ldots \end{aligned}$$

  where a_i and b_i are positive constants, common across people, and ξ is general ability and its distribution in the population is given by the cdf $m(\xi)$.

- This completes the description of the model. We can succintly state the problem to be solved:

  Problem 15 *Find an assignment from the distribution of latent talents $m(\xi)$ to ranks and firms that maximizes the total output of all persons to be assigned.*

- Associated with this solution will be some prices that sustain and decentralize such the equilibrium assignment.

Two-level firms

- Consider first a particular firm with only two ranks, where a person with talent r controls n workers. Given n, r choses a vector of t_i's, $i = 1, ..., n$, to maximize (12.3) subject to a time constraint

$$\sum_{i=1}^{n} t_i = T.$$

  This produces the first-order condition

$$rg(r) f_1(rt_i, q_i) = \lambda,$$

where λ is the Lagrange multiplier on the time constraint. From the properties of $f(\cdot)$ – namely, $f_{11} < 0$ and $f_{12} > 0$ – it follows that more time is spent monitoring more able workers.

- Furthermore, with constant returns to scale, this first order condition implies that the ratio t_i/q_i is constant across workers of different skill. Denoting this ratio by k and plugging in the time constraint, we obtain

$$\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} kq_i = k\sum_{i=1}^{n} q_i \equiv kQ = T.$$

  and thus $t_i/q_i = T/Q$.

- Letting $\theta(rt_i/q_i) \equiv f(rt_i, q_i)/q_i$ (remember constant-returns-to scale!) we can write aggregate production of r as

$$X = g(r)\sum_{i=1}^{n} q_i\theta(rt_i/q_i) = g(r)\,Q\theta(rT/Q),$$

  and thus X depends only on the total amount of "effective" labor (as measured by Q) available to r, and not on the particular number of workers n or how the skills are distribution across these workers.

- This in turn implies that q_i and q_j are perfect substitutes, so a competitive production labor market implies a single price for Q. Let this price be w. Then r will choose Q to maximize profits and will obtain:

$$\pi_r(r) = \max_Q \{pg(r)\,Q\theta(rT/Q) - wQ\}, \tag{12.4}$$

  where p is the market price for output.

- Because $Q\theta(rT/Q)$ features decreasing returns to Q, the program in (12.4) delivers a unique solution. In particular, letting $T = 1$, the maximum of (12.4) is implictly defined by:

$$g(r)\left[\theta(r/Q) - (r/Q)\,\theta'(r/Q)\right] = w/p. \tag{12.5}$$

- Define $\beta \equiv Q/r$ as the span of control of a second-level manager with skill r, i.e. the total amount of effective labor he controls per unit of skill. Notice that equation (12.5) pins down β as a function of w/p, r, and the properties of $g(\cdot)$ and $\theta(\cdot)$.

- Letting ϵ be the elasticity of $g(r)$, σ be the elasticity of substitution between rt and q in $f(\cdot)$, and κ the ratio of wQ to total sales of the firm, differentiation of (12.5) delivers
$$\frac{d\ln\beta}{d\ln r} = \frac{\epsilon\sigma}{1-\kappa} \geq 0,$$
$$\frac{d\ln Q}{d\ln r} = 1 + \frac{\epsilon\sigma}{1-\kappa} \geq 1$$
and
$$\frac{d\ln X}{d\ln r} = 1 + \epsilon + \frac{\epsilon\sigma}{1-\kappa} \geq 1.$$

- Notice that the span of control is nondecreasing in talent, and nonsurprisingly it increases with ϵ. Furthermore, labor hired and output increase more than proportionately with r. In other words, larger firms have more talented people at the top, and size differences are increasingly larger than the inherent differences in the quality of their managers.

- Furthermore, using the envelope theorem on (12.4), we find
$$\frac{d\ln\pi}{d\ln r} = 1 + \frac{\epsilon}{1-\kappa}.$$
This implies that the reward to talent is convex in talent and, hence, the distribution of income is more skewed to the right than the distribution of talent.

- Although r is not observable in the data, output is, and it can be shown that
$$\frac{d\ln\pi}{d\ln r} = \frac{(1-\kappa)(1+\epsilon)+\kappa\epsilon}{(1-\kappa)(1+\epsilon)+\kappa\epsilon\sigma},$$
which to match the available empirical estimates – 0.3 – requires $\sigma >> 1$.

Market equilibrium and assignements with two-level firms

- Notice that under constant returns to scale, we need only worry about assigning workers to ranks, because within R_1 the allocation of different workers to different second-level managers is both indeterminate and irrelevant.

- Hence, the problem can be treated as one of occupational choice. A person with latent skills $q = a_q + b_q\xi$ and $r = a_r + b_r\xi$ realizes that it can earn a wage of $w\left(a_q + b_q\xi\right)$ if he becomes a production worker in R_1. Alternatively, he can become a manager an earn $\pi\left(a_r + b_r\xi\right)$.

- From the linearity of the wage and the convexity of the reward to managers $\pi\left(\cdot\right)$, it follows that in equilibrium only higher-skilled people (those with ξ higher than some threshold ξ^*) will become managers, while lower skilled people will become production workers (see Figure 2 in Rosen's paper). Furthermore, the overall earnings distribution must be more skewed to the right than the underlying distribution of talent ξ.

- Rosen finally shows how to pin down the threshold ξ^*. Solving for the ξ^* that maximizes the total value of output, he ends up with a condition that can be expressed as
$$\pi^* = X^* - wQ^* = wq^*,$$
which implies the absence of rent at the margin.

Multilevel firms

- Rosen shows that in the constant returns to scale case, hierarchical structures with more than two layers can be studied in an analogous manner.

- Higher-ranked managers now face a trade off between selling their output in the open market or transfering it internally to a manager with higher rank. Again, this has the flavor of an occupational choice.

- Rosen shows that, under plausible assumptions, a similar rank-ability sorting emerges, with people with higher talent being assigned to higher levels. Furthermore, the convexity of the reward to talent also increases with rank, and thus the earnings distribution is more skewed to the right the larger the number of levels in firms.

- The decreasing returns implicit in (12.4) apply to all ranks, and this is what makes the model deliver a nondegenerate distribution of firm sizes – i.e., there is not just one firm and there are firms with different size, depending on the distribution $m(\xi)$.

Part IV

Trade and Organizational Form

Chapter 13

Early Transaction-Cost Approaches

- As argued in chapter 10, traditional theories of FDI enhance our understanding of trade and FDI flows, but they share a common failure to properly model the crucial issue of internalization. In recent years, the literature has acknowledged this caveat and has brought in tools from the theory of the firm to study the boundaries of multinational firms.

- In the next two chapters we will discuss the recent work of McLaren (2000), Grossman and Helpman (2002, 2004), Antràs (2003a,b), and Antràs and Helpman (2004). In this chapter we briefly discuss two important predecessors to this literature, which can both be viewed as applications of the transaction-cost approach of Coase and Williamson to the study of internalization.

13.1 Ethier (1986)

- Ethier (1986) is the seminal paper in the study of the internalization decision by multinational firms. In Ethier's view, the main difference between transacting within the boundaries of multinational firms and transacting at arm's length is that in the latter case certain types of (complex) contracts are infeasible, thus leading to inefficiencies in market transactions whenever attaining efficiency would require the use of these infeasible contracts. This clearly has a Williamsonian flavor.

- The model, however, does not consider explicitly the costs of integration (e.g., governance costs). Instead, it is posited that multinationals (internalization) will emerge only when they strictly dominate market transactions, which occurs only in a subset of the endowment space (this is not too different from Helpman's 1984 criteria discussed in Chapter 9).

- Interestingly, Ethier finds that modelling the internalization decision has important consequences for the link between vertical FDI and relative factor endowment differences. Unlike Helpman (1984), the model predicts a predominance of FDI when relative factor endowment differences are small. The intuition is that, in his model, the first best requires the use of the type of complex contracts that are infeasible in market transactions, only when relative factor prices differences between countries are small. When these differences are large, simpler contracts are sufficient to attain efficiency and thus internalization is unnecessary.

- Let us discuss a partial equilibrium version of his model. We will then sketch the general equilibrium.

A Simplified Version of the Model

- Consider a world with two countries, Home and Foreign, and (for now) a single good.

- The production process for this good consists of three stages: research, upstream production and downstream (or final-good) production. It is convenient to discuss them in reverse order.

  1. Downstream production of one unit of final good requires, in fixed proportions, q units of labor and one unit of a specialized intermediate input.
  2. Upstream, the specialized intermediate input can be produced at a choice of quality levels indexed by Q, $0 \leq Q \leq Q_1$. Upstream production uses only labor and the variable cost of production of one unit of an input of quality Q is equal to aQw.

3. Research determines the value of the parameter a, which can take one either of two values, $a_H > a_L$. If R workers are employed in research, a will take a value of a_L with probability $p(R)$, where naturally $p'(R) > 0$ and $p''(R) < 0$. It is assumed that labor must be committed to research and to downstream production before the uncertainty about the value of a is resolved.

- On the demand side it is assumed that world demand for the good is equal to one if the price of the good is equal (or lower) than its quality Q, and zero otherwise. Furthermore, it is assumed that Home consumes a fraction μ of the good and Foreign consumes the remaining fraction $1-\mu$.

- It is assumed that downstream production is nontradable. It thus follows that downstream production will employ $q\mu$ workers at Home and $q(1-\mu)$ in Foreign. On the other hand, the specialized intermediate input is freely tradable and upstream production and research can be concentrated in one country and be used for downstream production in another country. It thus follows that these stages of production will be located in the country with the lower (efficiency-adjusted) wage.

- Consider the case in which the wage in Home is lower, $w < w^*$ (the other case is symmetric).

Integration

- Following Ethier, it is useful to consider first the case in which all stages of production are integrated within a single multinational firm. If this firm is risk neutral, then we can express its expected profits as:

$$p(R)\,Q_L(1-a_Lw) + (1-p(R))\,Q_H(1-a_Hw) - wR - q(\mu w + (1-\mu)w^*), \tag{13.1}$$

  where Q_L (Q_H) is the upstream choice of quality when $a = a_L$ $(a = a_H)$.

- The firm choose R, Q_L and Q_H to maximize (13.1). The choices crucially depend on how large w is:

1. If $w > 1/a_L > 1/a_H$, then it is clear that $Q_L = Q_H = 0$, and consequently, the firm will also set $R = 0$. Intuitively, if the Home wage is too large, the firm will lose money regardless of the particular realization of a and thus will optimally choose to not produce.

2. If $w < 1/a_H < 1/a_L$, $Q_L = Q_H = Q_1$ and R is implicitly defined by $p'(R) = 1/Q_1 (a_H - a_L)$. Here again the optimal quality is independent of the realization of a because profits in both cases are increasing in quality.

3. If $w \in (1/a_H, 1/a_L)$, $Q_L = Q_1 > 0 = Q_H$, while R is implicitly defined by $p'(R) = w/Q_1 (1 - wa_L)$. Hence, the firm will wish to produce with positive (and highest) quality only if $a = a_L$. Notice also that, by the convexity of $p(\cdot)$, the optimal ex-ante level of research is decreasing in the wage w. Intuitively, conversely to case 2 above, the marginal benefit of research is now not proportional to w.

Non-Integration

- Now consider the case in which research and upstream production are still undertaken by the same firm, say the headquarters, while downstream production is controlled by an independent firm. We will label this the non-integration case, and notice that in such case the equilibrium will feature no multinational firms, since research and upstream production are always located in the same country.

- Under this arrangement the headquarters still control the choice of R, Q_L and Q_H, but sale revenues are collected by two independent final good producers. Depending on the realization of a, the foreign downstream firm will obtain a profit of

$$P_L^* = Q_L - qw^*$$

or

$$P_H^* = Q_H - qw^*,$$

per unit of final good. The expressions for the Home downstream firm are identical with w replacing w^*.

- Assuming that the ex-ante supply of downstream firms is perfectly elastic with an outside option of zero, the headquarters in the Home country will be able to extract all surplus from the downstream firms *through a quality-contingent contract*. For instance, the headquarters will ask a price equal to P_L^* to the foreign downstream firm for the use of an input of quality of Q_L, thus leaving the foreign firm with a net profit of zero. Overall, the headquarters will obtain

  $$p(R)(\mu P_L + (1-\mu) P_L^* - a_L Q_L w) + (1-p(R))(\mu P_H + (1-\mu) P_H^* - a_H Q_H w) - wR$$

  which simplifies to equation (13.1).[1]

- Hence, in the presence of state-dependent contracts, the outcome under non-integration will be identical to that under integration. Intuitively, the contract effectively makes the headquarters the residual claimant.

- Ethier (1986) argues, however, that in reality quality-contingent contracts are likely to be too complex to be feasible. The implicit assumption is that, as in Williamson (1985), specifying a quality-contingent contract is too complicated a task for boundedly rational agents. Of course, this raises the issue of why such contracts are feasible within firm boundaries (cf., Grossman and Hart, 1986).

- Ethier then explores the outcome under non-integration when contract is constrained to call for state-invariant quality. In such case, the headquarters will only be able to demand the following state-independent per-unit transfers from the downstream firms:

  $$\begin{aligned} P &= Q - qw \\ P^* &= Q - qw^*. \end{aligned}$$

  This leaves the headquarters with

  $$p(R)(\mu P + (1-\mu) P^* - a_L Q w) + (1-p(R))(\mu P + (1-\mu) P^* - a_H Q w) - wR$$

[1] It is straightforward to show that this contract is incentive compatible.

or

$$Q\left[1-p(R)a_Lw-(1-p(R))a_Hw\right]-wR-q(\mu w+(1-\mu)w^*). \qquad (13.2)$$

- The headquarters will now choose R and Q to maximize (13.2). These choices again depend crucially on how large w is:

  1. If $w > 1/a_L > 1/a_H$, then it is clear that, as under integration, $Q = R = 0$.
  2. If $w < 1/a_H < 1/a_L$, $Q = Q_1$ and R is implicitly defined by $p'(R) = 1/Q_1(a_H - a_L)$, just as under integration.
  3. If $w \in (1/a_H, 1/a_L)$, the outcome is different than under integration:
     (a) If $w > 1/(p(R)a_L + (1-p(R))a_H)$, then $Q = R = 0$. This is identical to case 1 before, when $w > 1/a_L > 1/a_H$.
     (b) If $w < 1/(p(R)a_L + (1-p(R))a_H)$, then $Q = Q_1$ and R is implicitly defined by $p'(R) = 1/Q_1(a_H - a_L)$. This is identical to case 2 before, when $w < 1/a_H < 1/a_L$.

Integration vs. Non-Integration

- It follows from the analysis before that, for $w < w^*$, multinationals will only emerge in equilibrium whenever $w \in (1/a_H, 1/a_L)$. If the domestic wage is not in this range, then quality contingent contracts are not needed to bring about efficiency, and hence internalization is not needed (notice the implicit assumption that multinationals only emerge when they strictly dominate arm's length transactions).

- Another important point to notice is that under non-integration, the level of research R is always independent of wages, whereas we saw that R is decreasing in w precisely in the range in which multinationals will arise.

General Equilibrium

- Ethier embeds this simple firm behavior in a general equilbrium model with two sectors and two factors. A manufacturing sector employs labor to produce an endogenously determined measure of differentiated goods, whose demand and supply is as characterized above. Free entry into this sector ensures that all firms break even, so that (13.1) and/or (13.2) exactly equal zero in equilibrium. There is also a homogenous sectors that used labor and land to produce wheat.

- In equilibrium, the relative wage w/w^* is related to relative factor endowment differences across countries, as represented by their aggregate land-labor ratios. In particular, when the land-labor ratio abroad is much larger than at Home the wage at Home will tend to be low and an equilibrium with $w < 1/a_H < 1/a_L < w^*$ is more likely. In such case, quality-contingent contracts have no value and multinationals do not emerge.

- As relative factor endowments become more equal, the relative wage at Home will tend to increase thus yielding $w \in (1/a_H, 1/a_L)$. In such case, headquarters at Home will open subsidiaries in Foreign.

- Ethier also shows that as relative factor endowments converge even further, the model delivers factor price equalization, and provided that $w = w^* \in (1/a_H, 1/a_L)$, the model will feature two-way foreign direct investment.

- These predictions are clearly different (and arguably more realistic) than those emerging from Helpman's (1984) model of the multinational firm.

13.2 Ethier and Markusen (1996)

- Ethier and Markusen (1996) also tackle the important issue of internalization but emphasize instead the non-appropriable nature of knowledge. In particular, they develop a model in which, in servicing a foreign market, firms choose between exporting, opening a subsidiary (FDI), or licensing their technology to an independent firm. Exporting is costly because of the presence of transport costs, while foreign production entails a potential dissipation of knowledge and consequent loss of rents. Importantly, it is posited that the extent of dissipation under FDI is different than under licensing.

- We consider here a simplified partial-equilibrium version of the model, discussed in Markusen (1995), that focuses on the choice between FDI and licensing.

Set-up

- Consider a two-period model in which a domestic firm wishes to exploit a technology in a foreign market either through FDI (by setting up a subsidiary) or through licensing. Prohibitive transport costs make exporting unprofitable.
- If **FDI** is chosen, the home firm transfers the technology to the foreign subsidiary in the first period, and an enforceable contract is signed precluding the subsidiary or its workers from defecting and exploiting the technology independently in the second period. Let the total net rents associated with FDI be $2M - F$, where M are per-period rents and F is an upfront investment cost. It is assumed that the home (parent) firm is able to extract all this surplus through an ex-ante transfer.
- In the case of **licensing**, the home firm again transfers the technology in the first period, but in this case there is no available ex-ante contract that precludes defection in the second period. Defection can take one of two forms. On the one hand, the foreign producer can defect by opening a rival firm in the second period. On the other hand, the parent firm can defect by issuing a second license to another foreign firm in the second period. The licensing fees, and thus the division of the total rents, will then need to be specified in a way that ensures no defection. Let these total rents be given by $2R - F$, where to ensure a nontrivial tradeoff between FDI and licensing it is assumed that $R > M$. Let F be initially incurred by the home firm.
- The above assumed that the home foreign prefers to license the same firm in the two periods. Consider then a third option in which the license in the second period is issued to a different firm. In such case, the licensee will defect and the market structure will be one of duopoly in the second period. Let total rents in this case be given by $R + D - 2F$ and assume $R + D - 2F < 2M - F$.

- Assume that in the case of defection the defector needs to incur the upfront investment cost F in the second period and that an unforseen defection leaves the other firm with no time to react and produce a positive amount in that same period.

Equilibrium

- Because FDI is a secure option, the multinational firm will obtain net profits of $2M - F$ when exploiting the technology within firm boundaries.

- Consider next licensing. Let L_1 and L_2 be the licensing fees in periods one and two, respectively. Consider first an equilibrium with no defection. Notice that, in the second period, the original licensee can obtain a payoff of $R-F$ by defecting, while it gets $R - L_2$ if it does not split off. It thus follows that the license fee cannot be larger than F. Similarly, the parent firm can obtain $R-F$ by defecting, and L_2 by not defecting, and thus $L_2 \geq R - F$. In sum, no defection requires $R < 2F$ and the parent will optimally set $L_2 = F$.

- One can also show that when $R < 2F$, given that the equilibrium in the second period is one with no defection, the parent firm can extract all the surplus from the licensee by setting the first period license equal to $L_1 = 2R - F$. Notice that, overall, the licensee obtains $R - L_1 + R - L_2 = 0$, so its participation constraint is satisfied. On the other hand, the parent firm is left with $L_1 + L_2 - F = 2R - F$.

- Conversely, when $R > 2F$, the firms will anticipate defection in the second period and thus both firms will have an incentive to effect. Let the home firm keep onwership of the investment F in this case. Then, in the second period, the original licensee will obtain $D/2 - F$, while the second-period licensee gets $D/2$. Naturally, the home firm will in this case charge $L_2 = D/2$. In the first period, L_1 will be set as to make the first-period licensee's participation constraint just bind, which implies $L_1 = R + D/2 - F$. Overall, the parent firm is left with $L_1 + L_2 - F = R + D - F$.

- Given the assumptions on the size of the rents, this implies that FDI will dominate licensing when $R > 2F$, while licensing will be the preferred option when $R < 2F$.

The intuition is that defections tends to reduce the rents associated with licensing, and thus when defection is unavoidable, FDI becomes a more attractive option.

- Interpreting the case $F = 0$ as one of "pure" knowledge-capital technology (once learned, it requires no additional cost to exploit it), the model implies that FDI is more likely to emerge when the technology has the joint-input characteristic of knowledge-capital.

- Ethier and Markusen (1996) embed this simple model in a two-country general equilibrium model and relate the size of the rents to more fundamental parameters, such as factor prices and costs of production. They also incorporate costly exporting in the framework. Interestingly, they show that similarities in relative factor endowments may promote FDI over licensing, just as in Ethier (1986).

Chapter 14

The Transaction-Cost Approach in Industry Equilibrium: McLaren (2000) and Grossman and Helpman (2002)

- In the two models in Chapter 13, as well as in the bulk of the literature on the theory of the firm, a particular firm's integration decision is treated independently of the decision made by other firms in the same industry. In this chapter we will present two industry equilibrium models that feature interesting feedback mechanisms by which firms' decisions affect market conditions, thereby influencing other firm's decisions about organizational form.

14.1 Globalization and Vertical Structure: McLaren (2000)

- McLaren's (2000) model illustrates how a firm's decision to vertically integrate its supplier can exert a negative externality on the remaining non-integrated bilateral relationships by thinning the market for inputs and thus worsening opportunism problems in market transactions.

- His model features multiple equilibria, thus rationalizing the pervasiveness of different organizational forms (or industry systems) in ex-ante identical countries and industries. Furthermore, McLaren (2000) shows how trade opening, by thickening the market for inputs, may well lead to a worlwide move towards more disintegrated industrial systems, thus increasing world welfare and leading to gains from trade quite different from those emphasized in traditional trade theory.

- Let us discuss his framework in more detail.

Set-up

- Consider an industry composed of n downstream firms (DSF's hereafter) producing final goods and n upstream firms (USF's hereafter) producing specialized intermediate inputs. Entry of additional firms is prohibitively costly.

- The model has three stages: a Merger Stage, a Production Stage, and a Market Stage.

- Consider first the **Production Stage**:

  - Each of the n DSF's can reduce its fixed costs by using a specialized input, which is tailor-made for the firm by an USF using K units of labor.

  - Each DSF can use at most one input, and each USF can produce at most one input.

  - USF's design inputs following one of two strategies. Under the "maximal specialization" strategy, the input is perfectly specialized, thus leading to a reduction in the targeted DSF's costs equal to 1, but having no impact on the cost of alternative DSF's.

  - Under the strategy of "flexibility", the USF's input is still more valuable for the intended user, but it also serves to reduce costs of alternative DSF's.

  - Within the "flexibility strategy", the input can be "effective" with probability ρ or be a "dud" with probability $1 - \rho$.[1] An "effective" input lowers

[1] McLaren studies the case in which ρ is allowed to differ across pairs of firms. See his section IV for details.

the costs of the intended user by an amount $e < 1$, and lowers costs of alternative DSF's by $e' < e$. A "dud" lowers the costs of the intended user by an amount $d < 1$, and lowers costs of alternative DSF's by $d' < d < e'$.

  - The cost reduction to each DSF associated with each input is revealed at the end of the Production Stage.

- In the initial **Merger Stage**, the DSF's and USF's are numbered from 1 to n, and each DSF$_i$ is given the option of making a take-it-or-leave-it offer to USF$_i$.

  - If the offer is accepted, the two firms become integrated (IF$_i$) and USF$_i$ produces input i for DSF$_i$. It is assumed that in such case IF$_i$ incurs the sunk cost K and designs and produces the input using the expected-profit maximizing choice of technology. It is assumed, however, that integration entails governance costs, denoted by L.

  - If the offer in the merger stage is turned down, USF$_i$ may still produce a specialized input for DSF$_i$, but in such case the parties are not able to write an ex-ante contract specifying the purchase of a particular input for a certain price. Ex-post, once the input is produced, the unintegrated USF$_i$ will bring the input to the open market and sell it to the highest bidder (in the Market Stage), which in principle may or may not be DSF$_i$.

  - Because DSF$_i$ is allowed to make a take-it-or-leave-it offer, integration will occur if and only if the expected profit of IF$_i$ exceeds the sum of expected profits of the two unintegrated firms.

- Finally, in the **Market Stage**, DSF's place bids on the inputs produced by the different USF's, and USF$_i$ sells the input to the highest bidder, thus ending the game.

Ex-Post Price Determination

- Let us start from the Market Stage. Denote by b_{ij} the bid made by DSF$_i$ to USF$_j$, and by $P_j = \max_i \{b_{ij}\}$ the winning bid, which can also be called the vector of equilibrium prices. It is straightforward to show that the price of each input is determined by the runner-up bidder.

- McLaren's (2000) first result is:

  Proposition 1 *In any subgame-perfect equilibrium of the bidding game, no IF sells its input, and each independently produced input is sold to its originally intended user.*

  Intuitively, given the above assumptions, an input is always strictly more valuable to its intended user, and thus this particular DSF_i will outbid the rest.

- As a corollary, and given that integrated firms never sell their inputs in the open market, it follows that an integrated structure will always follow the strategy of maximal specialization (Proposition 2), and because that input will be useless to the remaining $n-1$ firms, the runner-up's bid will be zero, and so will the equilibrium price of any maximally specialized input (Proposition 3).[2] Hence, no unintegrated supplier will choose to produce a maximally specialized input. There is thus an exact correspondence between ownership structure and technology choice.

- A somewhat less straightforward result is that, focusing on lowest-price equilibria,

  Proposition 4 *The lowest-price equilibrium is well defined, and is as follows. If the inputs are either all duds or all effective, their prices are all zero. If there are at least one effective input and at least one dud, the price of each effective input is $e' - d > 0$, and the price of each dud is zero.*

  The intuition here is that an input will be sold at a positive price only if it has an *absolute* advantage over another input, so that the intended user of that "dominated" input is willing to bid a positive amount for the input. In equilibrium, the input will still go to the intended user, but the supplier will be able to extract more surplus through the runner-up bid.

[2] This requires focusing on "perfect" equilibria, which eliminate all weakly dominated strategies, and also requires picking the lowest-price equilibrium.

Industry Equilibrium

- Let $\mathcal{F}$ denote the set of USF's using flexible technology, which as we have seen coincides with the set of DSF's using independent suppliers. Also, let $N(\mathcal{F})$ denote the number of elements of $\mathcal{F}$.

- For any nonintegrated firm i, we can write the expected price of input i as a function μ of ρ and $N(\mathcal{F})$

  $$\mu(\rho, N(\mathcal{F})) = \rho\left[1 - \rho^{N(\mathcal{F})-1}\right](e' - d).$$

  This is simply the product of the price and the probability of that input being effective and at least one other input being a dud.

- It is then straightforward that

  Proposition 5 *The function μ is increasing in $N(\mathcal{F})$.*

  Intuitively, adding one additional pair of buyer and seller to the open market cannot lower the expected price paid for a given input, but it may well increase it if it generates an absolute advantage for that input. This is the sense in which the model features a feedback from market thickness to the integration decision.

- Moving back to the Production Stage, an unintegrated DSF$_i$ will only incur the sunk cost K provided that it expects to recoup it in the open market, that is, provided that $\mu(\rho, N(\mathcal{F})) \geq K$. Quite clearly, this can possibly hold only if $\rho(e' - d) > K$, a condition that is hereafter assumed. Furthermore, because of Proposition 5, it also follows that this inequality will hold whenever $N(\mathcal{F})$ exceeds some threshold $\overline{n}$, defined by $\overline{n} = \min\{m | \mu(\rho, N(\mathcal{F})) \geq K\}$

- Finally, let us move back to the Merger Stage. If DSF$_i$ and USF$_i$ integrate, then the net cost reduction will be equal to $1 - L - K$. On the other hand, the expected net cost reduction from an arm's-length relationship is $\rho e + (1 - \rho) d - K$. To make the problem interesting we need to assume

  $$L > 1 - \rho e - (1 - \rho) d,$$

so that vertical integration does not always dominate market transactions. In fact, under this assumption, if it is feasible (if $\mu(\rho, N(\mathcal{F})) \geq K$), an arm's-length arrangement will be the outcome of ex ante negotiations.

- It thus follows that, for any pair of firms i, if $\overline{n}-1$ firms are expected to outsource, i will also outsource. On the other hand, if less than $\overline{n}-1$ firms are expected to outsource, i will not outsource. This strategic complementarity in the integration decision delivers the following multiplicity result:

  Proposition 6 *In a small closed economy ($n < \overline{n}$), the only equilibrium is complete vertical integration. In a large economy ($n \geq \overline{n}$), there are two equilibria: Complete integration and universal use of independent suppliers.*

  Hence, the model is consistent with two identical countries (with $n \geq \overline{n}$) having completely different industrial systems.

Open-Economy Model

- The open-economy version of the model is a straightforward extension of the model above.

- Let there be two identical countries with n pairs of DSF and USF in each of them. Final goods are nontradable, but inputs can be traded at a cost of t.

- Following similar steps as above, it is straightforward to derive the following equilibrium prices in the Market Stage:

  - If $e' - d \leq t$, then in each economy the prices are the same as in the closed economy (no DSF finds it optimal to submit a positive bid for a foreign input).
  - If $e' - d > t$, then: (i) all duds have a price of zero; (ii) if all inputs in all countries are effective, all prices are zero; (iii) an effective input in a country with at least one dud has a price of $e' - d$; and (iv) an effective input in a country with only effective inputs has a price of $e' - d - t$ if there is at least one dud in the other country.

- The industry equilibrium is also analogous to the one above. Letting $N(\mathcal{F}')$ be the number of unintegrated pairs in the other country, the expected price paid to a USF is given by

$$\begin{aligned}\mu\left(\rho,t,N\left(\mathcal{F}\right),N\left(\mathcal{F}'\right)\right) &= \rho\left[1-\rho^{N(\mathcal{F})-1}\right]\left(e'-d\right)+\\ &+\rho^{N(\mathcal{F})}\left[1-\rho^{N(\mathcal{F}')-1}\right]\left(\max\left\{e'-d-t,0\right\}\right).\end{aligned}$$

  It is straightforward to see that μ is continuous and decreasing in t. Globalization (a fall in t) thus increases the expected price and alleviates the holdup problem faced by suppliers.

- Furthermore, denoting by medium-sized countries those satisfying $\overline{n}/2 < n < \overline{n}$, it is easy to see that:

  Proposition 7 *A sufficient globalization between medium-sized countries will make the more efficient arm's-length equilibrium possible in both economies. It would not be possible in either economy without globalization.*

- Finally, as a corollary of Proposition 6 above,

  Proposition 8 *Whatever the size of the economies concerned, if t is sufficiently low, any equilibrium will involve complete convergence of the vertical structure of the two economies.*

14.2 Integration vs. Outsourcing in Industry Equilibrium: Grossman and Helpman (2002)

- Grossman and Helpman (2002) also present an industry equilibrium model in which downstream firms (final-good producers) endogenously decide whether to integrate their suppliers or transact with them at arm's length.

- As in Williamson (1985) – and McLaren (2000) – they assume that integration is (exogenously) associated with relatively high governance costs associated with integration. The advantages of integration are that (i) it avoids the contractual-driven inefficiencies inherent in the holdup problem faced by suppliers in market transactions, and (ii) it saves on the search costs of finding a suitable independent supplier.

- For simplicity, I will first present a simplified variant of their model in which the search frictions are absent. As we will see, this still leads to a well defined tradeoff between integration and outsourcing.

The Model without Frictional Search

1. **Endowments and Preferences.** Consider a one-factor (labor), multi-sector closed economy, endowed with L units of labor. In each of J sectors, firms produce a continuum of varieties which are differentiated in the eyes of consumers. Preferences of the representative consumer are of the form:

$$u = \sum_{j=1}^{J} \mu_j \log \left[\int_0^{N_j} y_j(i)^{\alpha_j} di \right]^{1/\alpha_j},$$

with $\mu, \alpha \in (0,1)$, and $\sum_{j=1}^{J} \mu_j = 1$. As a result, demand for a variety i in sector j is given by

$$y_j(i) = A_j p_j(i)^{-1/(1-\alpha_j)}, \tag{14.1}$$

where

$$A_j = \frac{\mu_j E}{\int_0^{N_j} p_j(i)^{-\alpha_j/(1-\alpha_j)} di},$$

and E is aggregate spending.

2. **Technology.** On the supply side, goods are also differentiated in the eyes of producers. Each variety $y_j(i)$ requires a special and distinct intermediate input $x_j(i)$. The specialized intermediate input must be of high quality, otherwise output is zero. If the input is of high quality, production of the final good requires no further variable costs and $y_j(i) = x_j(i)$. An input tailored to a final-good variety i is useless for producing a variety $i' \neq i$.

Final goods may be produced by vertically integrated firms or by specialized producers that purchase their inputs at arm's length. A supplier that sells an input at arm's length is able to produce one unit of high-quality input with one unit of labor. An integrated supplier in industry j instead requires $\lambda_j \geq 1$ units of labor per unit of high-quality intermediate input. This higher variable costs reflect governance costs à la Williamson (1985). Low-quality intermediate inputs can be produced at a negligible cost.

There are also fixed costs of production which may vary by organization mode. Letting subscripts V and O denote vertical integration and outsourcing, respectively, these fixed costs are k_{jV} units of labor for an integrated pair in industry j and $k_{jO} = k_{js} + k_{jm}$ units of labor for a pair of stand-alone specialized final-good producer s and input manufacturer m. It is assumed throughout that $k_{jV} \geq k_{jO}$, but we will focus first in the case in which $k_{jV} = k_{jO}$.

3. **Contract Incompleteness.** An outside party cannot distinguish between a high-quality and a low-quality intermediate input. Hence, stand-alone input suppliers and stand-alone final goods producers cannot sign enforceable contracts specifying the purchase of a certain type of intermediate input for a certain price (remember that a low-quality input can be produced at a negligible cost). Ex-ante investments and sale revenues are not verifiable either.

   Since no enforceable contract will be signed ex-ante, the two firms will (costlessly) bargain over the surplus of the relationship after production. At this point, the quality of the input is observable and thus the costless bargaining will yield an ex-post efficient outcome. Assume that Generalized Nash Bargaining leaves the intermediate input producer with a fraction ω of the surplus.

   Following Williamson (1985), these contractual frictions are not present in integrated pairs, which are able to commit to an ex-ante choice of quality and level of production.

4. **Ex-ante Division of Surplus**. Before any investment is made, each final-good producer decides whether it wants to enter a given market, and if so, whether to

obtain the component from a vertically integrated supplier or from a stand-alone one. This decision is based on profit maximization.

Ex-ante, there is an infinitely elastic supply of potential input suppliers. Each final-good producer offers a contract that seeks to attract a supplier. The contract includes an up-front fee for participation in the relationship that has to be paid by the supplier (this fee can be positive or negative). The purpose of the fee is to secure the participation of the supplier in the relationship at minimum cost to the final-good producer.

The infinite supply elasticity implies that, in equilibrium, the supplier will be left with a net payoff equal to its ex-ante outside option, which for simplicity is set to zero.

Firm Behavior

- Let us focus on an industry j and variety i, and for simplicity let us drop these indices.

- A **vertically integrated** firm faces a marginal production cost of λ and a demand given by (14.1), where firms take A as given. It thus chooses a level of ouput y to maximize

  $$\pi = A^{1-\alpha} y^{\alpha} - \lambda y - k_V.$$

  This produces

  $$y_V = A \left(\alpha / \lambda\right)^{1/(1-\alpha)},$$

  with implied optimal price

  $$p_V = \lambda / \alpha$$

  and profits

  $$\pi_V = (1-\alpha) A \left(\alpha / \lambda\right)^{\alpha/(1-\alpha)} - k_V. \tag{14.2}$$

- Consider next a pair of stand-alone firms. In the ex-post bargaining, the surplus is given by sale revenues, or $py = A^{1-\alpha} y^{\alpha}$. On the other hand, both firm's outside options are zero. This is because the input is perfectly tailored to the final-good producer, and because it is assumed that the latter has no "time to react" and attract another supplier in case of contractual breach. In the Nash bargaining, the supplier will thus obtain a $(1-\omega)\,A^{1-\alpha}y^{\alpha}$ and will thus choose the amount of input produced to maximize

$$\pi = \omega A^{1-\alpha}x^{\alpha} - x - k_m,$$

which delivers

$$y_O = x_O = A\,(\alpha\omega)^{1/1-\alpha}\,, \tag{14.3}$$

and an implied price of the final good equal to

$$p_O = 1/\alpha\omega. \tag{14.4}$$

The diminished output ($\omega < 1$) and inflated price reflect the distorsionary impact of the incompleteness of contracts. Ex-ante, the final-good producer will set a transfer equal to

$$T = \omega\,(1-\alpha)\,A\,(\alpha\omega)^{\alpha/1-\alpha} - k_m,$$

and will be left with $(1-\omega)\,p_O y_O + T - k_s$, which simplifies to

$$\pi_O = (1-\alpha\omega)\,A\,(\alpha\omega)^{\alpha/1-\alpha} - k_O. \tag{14.5}$$

Choice of Organization

- The choice between integration and outsourcing is straightforward to analyze whenever $k_V = k_O$. In that case, we see that $\pi_O > \pi_V$ only if

$$\Theta\,(\omega,\alpha,\lambda) = \left(\frac{1-\alpha\omega}{1-\alpha}\right)(\lambda\omega)^{\alpha/1-\alpha} > 1.$$

It is straightforward to show that $\Theta\,(\cdot)$ is increasing in λ with $\lim_{\lambda\to 1}\Theta\,(\cdot) < 1$ and $\lim_{\lambda\to\infty}\Theta\,(\cdot) > 1$. Similarly, $\Theta\,(\cdot)$ is increasing in ω with $\lim_{\omega\to 0}\Theta\,(\cdot) = 0$ and $\lim_{\omega\to 1}\Theta\,(\cdot) > 1$. This implies that:

Proposition 16 *For a given ω and α, there exist a unique thresholds $\widehat{\lambda} > 1$ such* *that outsourcing dominates integration if* $\lambda > \widehat{\lambda}$*. For a given* λ *and* α*, there exist a unique thresholds* $\widehat{\omega} > 1$ *such that outsourcing dominates integration if* $\omega > \widehat{\omega}$*.*

- The first result is straightforward: for high enough governance costs, outsourcing dominates integration despite the contractual frictions associated with the former. The second result can be understood by noticing that as ω increases the hold up problem faced by suppliers diminishes and, in the limit $\omega \to 1$, the lack of contractibility is irrelevant because suppliers are, *ex-post*, full residual claimants.[3]

- Finally, differentiating with respect to α delivers:

$$\frac{\partial \Theta\left(\omega, \alpha, \lambda\right)}{\partial \alpha} = \frac{\left[\left(1-\omega\right)\left(1-\alpha\right) + \left(1-\alpha\omega\right)\ln\left(\lambda\omega\right)\right]\left(\lambda\omega\right)^{\alpha/(1-\alpha)}}{\left(1-\alpha\right)^{3}}.$$

  If $\lambda\omega > 1$, this derivative is positive, $\lim_{\alpha\to 0}\Theta\left(\cdot\right) = 1$ and $\lim_{\alpha\to 1}\Theta\left(\cdot\right) > 1$, so outsourcing dominates insourcing for all α. On the other hand, if $\lambda\omega < 1$, the relationship between $\Theta\left(\cdot\right)$ and α may be decreasing or $\cap$-shaped, $\lim_{\alpha\to 0}\Theta\left(\cdot\right) = 1$ and $\lim_{\alpha\to 1}\Theta\left(\cdot\right) = 0$. This implies that integration may dominate outsourcing for all α, but it could also be the case that this is only true for $\alpha > \widehat{\alpha}$. The model would then predict more vertical integration in more competitive industries. The reason is that the higher is α, the higher is the elasticity of profits with respect to the price charged for the final good and, when $\lambda\omega < 1$, integration is associated with a lower price than outsourcing.

- As we will discuss below, these comparative statics are similar *but not identical* to those in the original version of the model with search frictions.

- The above assumed that $k_V = k_O$. Now let $k_V > k_O$. In this case $\pi_O > \pi_V$ only if

$$A\alpha^{\alpha/(1-\alpha)}\left[\left(1-\alpha\right)\lambda^{-\alpha/(1-\alpha)} - \left(1-\alpha\omega\right)\omega^{\alpha/1-\alpha}\right] < k_V - k_O.$$

  Clearly, the inequality depends on the value of A, which can only be pinned down in industry equilibrium. This is the sense in which market conditions affect the integration decision by firms. So let us analyze the industry equilibrium.

[3] This result parallels the effect of asset specificity in the example developed in Lecture 10.

Industry Equilibrium

- In the industry equilibrium free entry ensures that firms break even. From equation (14.2), for integrating final-good producers to break even, A needs to equal

$$A_V = \frac{k_V}{(1-\alpha)(\alpha/\lambda)^{\alpha/1-\alpha}}.$$

On the other hand, from equation (14.5), for specialized final-good producers to break even, the demand level A needs to satisfy

$$A_O = \frac{k_O}{(1-\alpha\omega)(\alpha\omega)^{\alpha/1-\alpha}}.$$

- It is also easy to see that starting from an equilibrium with pervasive outsourcing, a deviating integrated firm's profits can be expressed as

$$\pi_V = k_V \left(\frac{A_O}{A_V} - 1 \right),$$

which can only be positive if $A_O > A_V$. Similarly, starting from an equilibrium with pervasive integration, a deviating specialized final-good producer's profits would be

$$\pi_O = k_O \left(\frac{A_V}{A_O} - 1 \right),$$

which are positive only if $A_V > A_O$.

- It thus follows that:

Proposition 17 *Generically, no industry has both vertically integrated and specialized producers ($A_O = A_V$ with probability zero). If $A_O > A_V$, the equilibrium is one with pervasive integration. If $A_O < A_V$, the equilibrium is one with pervasive outsourcing.*

- Furthermore, notice that

$$\frac{A_V}{A_O} = \left(\frac{1-\alpha\omega}{1-\alpha} \right) (\lambda\omega)^{\alpha/1-\alpha} \frac{k_V}{k_O},$$

which of course produces the same comparative statics as before, and adds the (intuitive) result that integration is more likely the lower is k_V/k_O.

The Model with Frictional Search

- Grossman and Helpman consider the above framework, but replace point 4 in the set up with the following:

4'. **Ex-ante Division of Surplus**. As in the previous version of the model, a vertically integrated pair enters the market jointly and maximizes joint profits. A way to interpret this assumption is by appealing to an ex-ante perfectly elastic supply of operators of integrated supplying firms and to the presence of an ex-ante lump-sum transfer, just as in the model without search. The case of pairs of specialized firms is different. Grossman and Helpman (2002) assume that each of these firms enter the market independently up to the point in which their expected profits are zero. Once s specialized final-good producers and m specialized manufacturers of inputs have entered, $n(s,m)$ pairs are formed, where $n(s,m) \leq \min\{s,m\}$. The remaining $s+m-n(s,m) \geq 0$ are forced to exit thus forfeiting the sunk cost of entry. It is assumed that $n(s,m)$ is increasing in both arguments and features constant returns to scale, so that defining $r = m/s$, we can express the probability of finding a match for a specialized final-good producer as $\eta(r) = n(s,m)/s$.

Notice that in this set-up, specialized final-good producers are unable to extract all surplus from specialized suppliers because the supply of suppliers is *not* perfectly elastic with an outside option of zero. The assumption that the mode of organization is chosen by the final-good producer to maximize its own profits is, however, maintained.

- Firm behavior for pairs of integrated firms is identical to that above and leads to expected profits for the final-good producer of

$$\pi_V = (1-\alpha) A (\alpha/\lambda)^{\alpha/(1-\alpha)} - k_V,$$

just as in equation (14.2).

- Within pairs of specialized firms, the ex-post division of surplus is as described before, so that the level of production and price of the final good is again given by equations (14.3) and (14.4). In this case, however, the final-good producer is left with expected profits of

$$\pi_s = \eta(r)(1-\omega)A(\alpha\omega)^{\alpha/1-\alpha} - k_s,$$

while the supplier obtains

$$\pi_m = \omega(1-\alpha)\frac{\eta(r)}{r}A(\alpha\omega)^{\alpha/1-\alpha} - k_m.$$

- Outsourcing is therefore now chosen whenever

$$A\alpha^{\alpha/(1-\alpha)}\left[(1-\alpha)\lambda^{-\alpha/(1-\alpha)} - \eta(r)(1-\omega)\omega^{\alpha/1-\alpha}\right] < k_V - k_s.$$

- An important point to notice is that with search frictions, even when $k_V = k_s$, **the integration decision will still depend on market conditions**, since the ratio $r = m/s$ is only pin down in the industry equilibrium.

- In the industry equilibrium specialized firms will enter up to the point to which $\pi_s = 0$ and $\pi_m = 0$, thus implying

$$r_O = \frac{\omega(1-\alpha)}{1-\omega}\frac{k_s}{k_m}. \tag{14.6}$$

Naturally, the relative number of input suppliers is decreasing in their relative fixed costs – k_m/k_s – and increasing in their relative *net* rents – $\omega(1-\alpha)/(1-\omega)$.

- We can thus define the demand levels that make integrating and non-integrating final-good producers break even as

$$A_V = \frac{k_V}{(1-\alpha)(\alpha/\lambda)^{\alpha/1-\alpha}}$$

and

$$A_O = \frac{k_m}{\omega(1-\alpha)(\alpha\omega)^{\alpha/1-\alpha}}\frac{r_O}{\eta(r_O)},$$

respectively, where r_O is given by (14.6). Grossman and Helpman next show that

Proposition 18 *Generically, no industry has both vertically integrated and specialized producers ($A_O = A_V$ with probability zero). If $A_O > A_V$, the unique equilibrium is one with pervasive integration. If $A_O < A_V$, the unique* stable *equilibrium is one with pervasive outsourcing.*

- Grossman and Helpman next analyze the determinants of the equilibrium mode of organization by studying the ratio

$$\frac{A_V}{A_O} = \omega\left(\lambda\omega\right)^{\alpha/1-\alpha}\frac{\eta\left(r_O\right)}{r_O}\frac{k_V}{k_m}.$$

The comparative statics with respect to λ and the fixed costs are completely analogous to those under the case with no search frictions. Moreover, with constant returns to scale in the matching function, $\eta\left(r_O\right)$ is independent of the size of the economy and thus the equilibrium mode of organization is independent of market size.

- The effect of α on the mode of organization has a similar flavor as before, depending crucially on whether $\lambda\omega$ is larger or smaller than one. Again this is explained by the effect of α on the elasticity of profits with respect to the price. It is no longer the case, however, that if $\lambda\omega > 1$, outsourcing necessarily dominates integration for all α. The reason is that α now also affects expected profits for suppliers and therefore r_O. In particular, a decrease in α will need to be compensated with a fall in $\eta\left(r_O\right)/r_O$ to maintain $\eta\left(r_O\right)$. This in turn will make outsourcing less attractive.

- The effect of ω is also richer than in the model with no search frictions. In particular, it is no longer the case that a higher ω unambiguously favors outsourcing. Intuitively, now ω not only affects incentives but also the ex-ante division of surplus. If the final-good producer were allowed to choose ω ex-ante, it would do so by trading the choice of a larger fraction of the revenue for a smaller revenue level. In general, this would produce $\omega^* \in (0,1)$. See Figure V in the paper and Antràs (2003) and Antràs and Helpman (2004) for a related discussion.

Extensions

- Grossman and Helpman next analyze a couple of interesting extensions of their model. First, they show that if the matching function features increasing returns to scale, the mode of organization may well depend on market size and, as in McLaren (2000), pervasive outsourcing is more likely to emerge the higher the market size. Notice, however, that the mechanism is somewhat different. In McLaren (2000), the thickness of the market for inputs affects the *ex-post* division of surplus by alleviating the holdup problem. The effect in Grossman and Helpman (2002) works instead through the *ex-ante* division of surplus, by increasing the probability of a match and therefore ex-ante expected profits.

- Second, Grossman and Helpman extend the model to allow for an endogenous specialization of inputs. This is also related to McLaren's choice of "maximal specialization" vs. "flexibility", but the modelling is quite different. See section VI of the paper for further details.

- The setup has also been extended in subsequent work by Grossman and Helpman (2003, 2004) to consider the possibility of partial contractibility and how the degree of contractibility affects both the location of production (in their REStud 2004 paper), as well as the mode of organization (in their JEEA 2003 paper). You are most encouraged to study these papers in detail.

Chapter 15

The Property-Rights Approach in International Trade (I): Antràs (2003a)

- In the models in Chapters 13 and 14 the costs of integration were treated as exogenous. Furthermore, it was assumed that the contractual frictions that plague the relationship between two nonintegrated firms disappeared when these firms integrated. As pointed out by Grossman and Hart (1986), this is not entirely satisfactory. After all, inside firms too agents are boundedly rational and opportunistic, and undertake important relationship-specific investments.

- In this chapter and the next, we will study theoretical frameworks that draw the boundaries of the multinational firm à la Grossman and Hart (1986).

- In this chapter, we will study Antràs (2003a), in which I unveiled some facts regarding the intrafirm component of trade and showed that these strong patterns can be rationalized combining elements of a Grossman-Hart-Moore view of the firm together with elements of a Helpman-Krugman view of international trade.

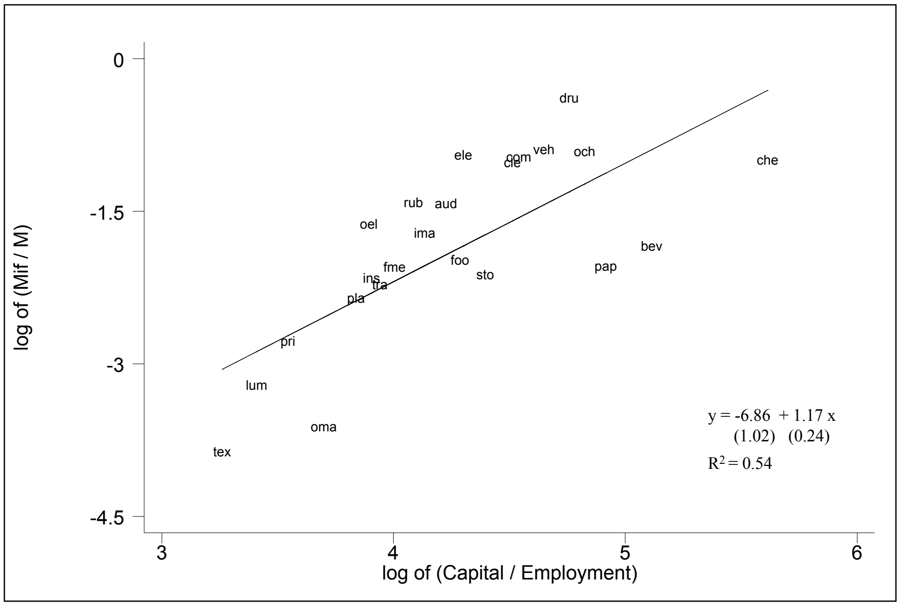
Introduction

- Roughly $\frac{1}{3}$ of world trade is intrafirm trade ($\frac{1}{3}$ of U.S. exports and more than 40% of U.S. imports). Furthermore, Figures 1 and 2 indicate that the intrafirm component of trade shows some strong patterns:

1. Fact 1: In a cross-section of industries, the share of intrafirm imports in total U.S. imports is larger the higher the capital intensity of the exporting industry (Figure 1). Firms in the U.S. import chemical products from affiliate parties, but import textiles from independent firms overseas.
2. Fact 2: In a cross-section of countries, the share of intrafirm imports in total U.S. imports is larger the higher the capital-labor ratio of the exporting country (Figure 2). Firms in the U.S. import from Switzerland within the boundaries of their firms, but import from Egypt at arm's length.

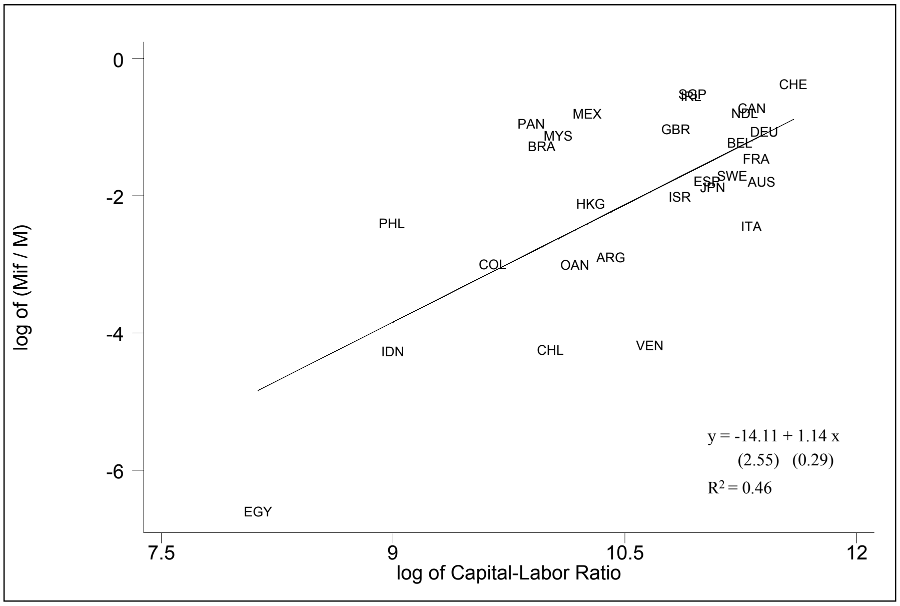
- These observations raise the following questions:
  - Why are capital-intensive goods transacted within firm boundaries while labor-intensive goods are traded mostly at arm's length?
  - Why is the share of intrafirm imports higher for capital-abundant countries?
  - Are these facts related?
- In Antràs (2003*a*), I suggest the following explanation for these facts:
  - I develop a property-rights model of the boundaries of the firm in which the endogenous benefits of integration outweigh its endogenous costs only in capital-intensive industries → *close to* Fact 1.
  - I then embed this framework in a general-equilibrium, factor-proportions model of international trade, with imperfect competition and product differentiation. In the general equilibrium, capital-abundant countries capture larger shares of a country's imports of capital-intensive goods.
  - Fact 2 follows from the interaction of transaction-cost minimization (Fact 1) and comparative advantage.

Figure 15.1: Share of Intrafirm U.S. Imports and Relative Factor Intensities



Notes: The Y-axis corresponds to the logarithm of the share of intrafirm imports in total U.S. imports for 23 manufacturing industries averaged over 4 years: 1987, 1989, 1992, 1994. The X-axis measures the average log of that industry's ratio of capital stock to total employment, using U.S. data. See Table A.1. for industry codes and Appendix A.4. for data sources.

Figure 15.2: Share of Intrafirm Imports and Relative Factor Endowments



Notes: The Y-axis corresponds to the logarithm of the share of intrafirm imports in total U.S. imports for 28 exporting countries in 1992. The X-axis measures the log of the exporting country's physical capital stock divided by its total number of workers. See Table A.2. for country codes and Appendix A.4. for details on data sources.

Heuristic Version of the Model

A. Grossman-Hart-Moore helps explain Fact 1

- A final-good producer needs to obtain a special and distinct intermediate input from a supplier. Production of the input requires certain noncontractible and relationship-specific investments in capital and labor.

- The final-good producer contributes to some of these investments but cost-sharing is relatively more important in capital investments. The empirical validity of this assumption is discussed in the introduction (see Table 1).

- The lack of ex-ante contracts implies that the bargaining over the terms of trade takes place after intermediate input has been produced and manufacturing costs are bygones. The combination of this ex-post bargaining and the lock-in effect stemming from the specificity of investments leads to a two-sided holdup problem which, in turn, results in underinvestment in both capital and labor.

- Ex-ante, there are two possible organizational forms: vertical integration or outsourcing. Ownership is defined as the entitlement of some residual rights of control. These residual rights translate into an outside option for the final-good that is higher under integration than under outsourcing.

- Inefficiency in labor investments is shown to be relatively higher under integration than under outsourcing; and conversely for capital. We will see that, ex-ante, this implies that firms will choose outsourcing only when the investment in labor is relatively important in production → *close to* Fact 1.

B. Helpman-Krugman and Fact 1 imply Fact 2

- This partial equilibrium is embedded in a general equilibrium setup with imperfect competition and product differentiation. Countries specialize in certain intermediate input varieties and export them worldwide. Capital-abundant countries tend to produce a larger share of capital-intensive varieties than labor-abundant countries. On the demand side, preferences are homothetic and identical everywhere.

- Under these assumptions, it is shown that the share of capital-intensive (and *thus* intrafirm) imports in total imports is then shown to be an increasing function of the capital-labor ratio of the exporting country (Romalis, 2002) → Fact 2.

The Closed-Economy Model

Set-up

- **Endowments and Preferences:** Consider a two-factor (K, L), two-sector (Y, Z) closed economy. K and L are inelastically supplied and freely mobile across sectors. In each sector, firms use K and L to produce a continuum of differentiated varieties. Preferences of the representative consumer are of the form:

  $$U = \left(\int_0^{n_Y} y(i)^\alpha di\right)^{\frac{\mu}{\alpha}} \left(\int_0^{n_Z} z(i)^\alpha di\right)^{\frac{1-\mu}{\alpha}}, \quad \mu, \alpha \in (0,1).$$

  Demand for final-good varieties is thus

  $$\begin{aligned} y(i) &= A_Y p_Y(i)^{-1/(1-\alpha)} \\ z\,(i) &= A_Z p_Z(i)^{-1/(1-\alpha)}. \end{aligned}$$

  Sale revenues are:

  $$\begin{aligned} R_Y(i) &= p_Y(i)y(i) = A_Y^{1-\alpha} y(i)^\alpha \\ R_Z(i) &= p_Z(i)z(i) = A_Z^{1-\alpha} z(i)^\alpha. \end{aligned}$$

- **Technology:** Each variety $y(i)$ requires a special and distinct intermediate input $x_Y(i)$ ($z(i)$ requires $x_Z(i)$). The input must be of high quality, otherwise output is zero. If the input is of high quality, production of the final good requires no further costs and $y(i) = x_Y(i)$, $z(i) = x_Z(i)$.

  Production of a high-quality intermediate input requires a combination of capital

(K_x) and labor (L_x).

$$x_k(i) = \left(\frac{K_x(i)}{\beta_k}\right)^{\beta_k} \left(\frac{L_x(i)}{1-\beta_k}\right)^{1-\beta_k},$$

for $k \in \{Y, Z\}$. Let $1 > \beta_Y > \beta_Z > 0$. Low-quality intermediate inputs can be produced at a negligible cost.

There are also fixed costs of production, which for simplicity we assume to have the same factor intensity as variable costs: $fr^{\beta_k}w^{1-\beta_k}$, $k \in \{Y, Z\}$.

- **Firm structure:** Before any production takes places, the final-good producer (F) decides whether it wants to enter a given market, and if so, whether to obtain the input from a vertically-integrated supplier (S) or from a stand-alone S. F chooses the mode of organization so as to maximize its ex-ante profits. It is assumed that, upon entry, S makes a lump-sum transfer $T_k(i)$ to F. $T_k(i)$ is such that S breaks even (this is equivalent to assuming an ex-ante competitive fringe of suppliers).

  The labor investment L_x is undertaken by S. The capital investment K_x is undertaken by F. These investments are incurred upon entry and are useless outside the relationship. This leads to Williamson's *fundamental transformation.*

- **Contract Incompleteness:** No outside party can distinguish between a high-quality and a low-quality intermediate input x_k, thus implying that F and S cannot sign **enforceable** quality-contingent contracts (the logic is the same as in Grossman and Helpman, 2002). Similarly, it is assumed that K_x, L_x, $R_Y(i)$, and $R_Z(i)$ are not verifiable either.

  Because no enforceable contract is signed ex-ante, F and S will bargain over the surplus of the relationship ex-post, when manufacturing costs are bygones. At this point, the quality of the input is observable and thus the costless bargaining will yield an ex-post efficient outcome. Assume that Generalized Nash Bargaining leaves the final-good producer with a fraction ϕ of the ex-post gains from trade.

  As in Grossman and Hart (1986), ownership will affect the distribution of ex-post surplus through its effect on each party's outside option. Because the input is completely specific to F, the outside option for S is zero regardless of ownership structure. If S is a stand-alone firm, F's outside option is also zero because a contractual breach leaves F with no "time to react" and attract another supplier. By integrating S, however, F obtains the residual rights over a fraction $\delta \in (0,1)$ of the amount of $x_k(i)$ produced, which translate into sale revenues of $\delta^{\alpha} R_k(i)$. Intuitively, in case of contractual breach, F can (i) fire the S manager, (ii) seize the input $x_k(i)$, which is sitting in the integrated suppliers's facility, and (iii) produce the final good, although at lower productivity as parameterized by δ.

Firm Behavior

- Notice that the payoffs in the Nash Bargaining are as follows:

	Final-good producer	Supplier
Non-Integration	$\phi R_k(i)$	$(1-\phi) R_k(i)$
Integration	$\overline{\phi} R_k(i)$	$\left(1-\overline{\phi}\right) R_k(i)$

where $\overline{\phi} = \delta^{\alpha} + \phi(1-\delta^{\alpha}) > \phi$. The investment K_x and L_x are set non-cooperatively to maximize these payoffs.

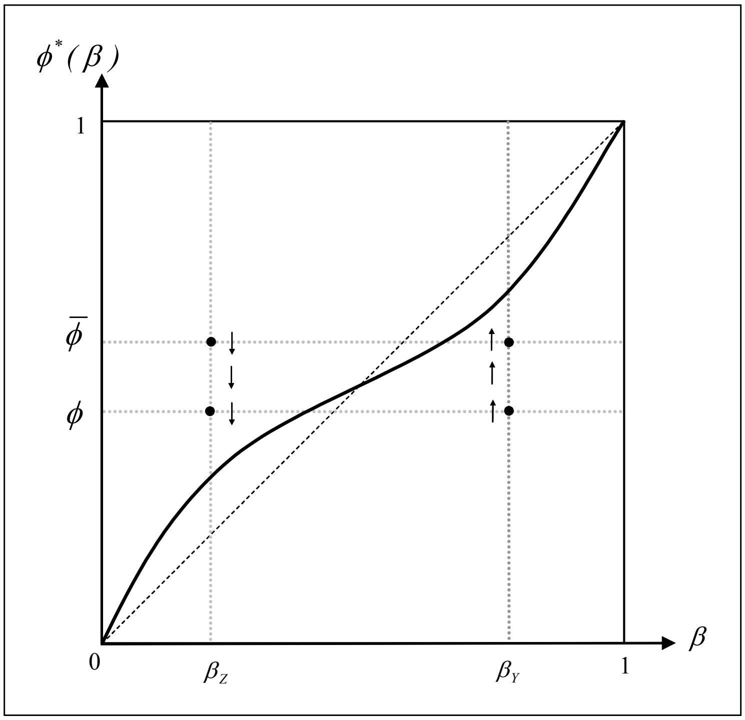
- Ex-ante, the choice between integration and outsourcing is equivalent to choosing between ϕ and $\overline{\phi}$ to maximize:

$$\begin{aligned}
\max_{\widetilde{\phi} \in \left\{\phi, \overline{\phi}\right\}} \quad & R_k\left(K_x\left(\widetilde{\phi}\right), L_x\left(\widetilde{\phi}\right)\right) - rK_x\left(\widetilde{\phi}\right) - wL_x\left(\widetilde{\phi}\right) - fr^{\beta_Y} w^{1-\beta_Y} \\
s.t. \quad & K_x\left(\widetilde{\phi}\right) = \arg\max_{K_x} \widetilde{\phi} R_k\left(K_x, L_x\left(\widetilde{\phi}\right)\right) - rK_x \\
& L_x\left(\widetilde{\phi}\right) = \arg\max_{L_x} \left(1-\widetilde{\phi}\right) R_k\left(K_x\left(\widetilde{\phi}\right), L_x\right) - wL_x
\end{aligned}$$

- Incomplete contracts leads to underinvestment in K_x and L_x. Crucially, underinvestment in L_x is relatively more severe under integration; underinvestment in K_x is relatively more severe under outsourcing (see Figure 4 in the paper).

Factor Intensity and Ownership Structure

Figure 15.3: Profit-Maximizing Ex-Post Division of Surplus



- Let $\Theta(\beta_k) \equiv \left(\pi^F_{k,V} + fr^{\beta_Y} w^{1-\beta_Y}\right) / \left(\pi^F_{k,O} + fr^{\beta_Y} w^{1-\beta_Y}\right)$ be operating profits under integration relative to outsourcing. Solving the problem above yields:

$$\Theta\left(\beta_k\right) = \left(1 + \frac{\alpha\left(1-\phi\right)\delta^{\alpha}(1-2\beta_k)}{1-\alpha(1-\beta_k)+\alpha\phi(1-2\beta_k)}\right)\left(1+\frac{\delta^{\alpha}}{\phi\left(1-\delta^{\alpha}\right)}\right)^{\frac{\alpha\beta_k}{1-\alpha}}\left(1-\delta^{\alpha}\right)^{\frac{\alpha}{1-\alpha}}.$$

Proposition 19 *There exists a unique $\widehat{\beta} \in (0,1)$ such that $\Theta(\widehat{\beta}) = 1$. Furthermore, for all $\beta < \widehat{\beta}$, $\Theta(\beta) < 1$, and for all $\beta > \widehat{\beta}$, $\Theta(\beta) > 1$.*

Hence, all firms with capital intensity below (above) a certain threshold $\widehat{\beta}$ choose to outsource (vertically-integrate) production of the intermediate input.

- To gain further intuition on this result, it is useful to consider the case in which F was free to choose the profit-maximizing ϕ^* (rather than $\underline{\phi}$ or $\overline{\phi}$). One can show that $\phi^*\left(\beta\right)$ satisfies $\phi^*\left(0\right) = 0$, $\phi^{*\prime}\left(\beta\right) > 0$, and $\phi^*\left(1\right) = 1$. This function is depicted in the Figure 15.3, where the arrows indicate the direction of increasing profits. It is clear that when β is high (low), integration dominates (is dominated by) outsourcing.

- It is important to emphasize the following points:

  1. $\Theta(\beta_Y)$ is independent of factor prices because of the Cobb-Douglas assumption in production. In general, the decision would depend on factor prices. The assumption is useful because it allows a sequential discussion of the equilibrium.

  2. Notice also that $\Theta\left(\beta_k\right)$ is not a function of the level of demand A_k and thus the integration decision is independent of the industry equilibrium. From our discussion of Grossman and Helpman (2002) in the previous chapter, it should be clear that the assumption that fixed costs are identical under integration and outsourcing is important in this. Nevertheless, it is straightforward to show that the model delivers the same link between capital-intensity and the integration decision even when $f_V > f_O$.

  3. Why is F providing K_x? Otherwise, S would choose K_x and L_x to maximize $\left(1-\widetilde{\phi}\right) R_k - rK_x - wL_x$.

     Lemma 20 *If $\overline{\phi} > \phi > 1/2$, final-good producers will always decide to provide the capital K_x required for production.*

     Key Point: the supplier is never given full control. There is an unmodelled non-contractible and inalienable investment by F that is indispensable for sale revenues to be positive. This also helps rationalize the assumption of the lack of forward integration

  4. How important is it that F provides K_x under non-integration? The result still holds true when $\phi < 1/2$, provided that $\bar{\phi} > 1 - \phi$.

Industry Equilibrium

- In the industry equilibrium, free entry implies that no firm makes positive profits, and n_k will adjust to ensure $\pi_k^F = 0$. Following the analysis of Grossman and Helpman (2002), one can show that:

Lemma 21 *A mixed equilibrium in industry $k \in \{Y, Z\}$ only exists in a knife-edge case, namely when $\beta_k = \widehat{\beta}$. An equilibrium with pervasive integration in industry k exists only if $\beta_k > \widehat{\beta}$. An equilibrium with pervasive outsourcing in industry k exists only if $\beta_k < \widehat{\beta}$.*

To make the model interesting we assume that $\beta_Y > \widehat{\beta} > \beta_Z$ (see also Figure 15.3). We thus have pervasive integration in the capital-intensive industry Y and pervasive outsourcing in labor-intensive industry Z.

General Equilibrium

- In the general equilibrium of the integrated economy, income equals spending,

$$E = rK + wL,$$

and the product, capital and labor markets clear:

$$\begin{aligned} \sum_{k \in \{Y,Z\}} n_k \left(K_{x,k} + K_{f,k}\right) &= K \\ \sum_{k \in \{Y,Z\}} n_k \left(L_{x,k} + L_{f,k}\right) &= L \end{aligned}$$

Plugging the equilibrium values of n_k, $K_{x,k}$, $K_{f,k}$, $L_{x,k}$, and $L_{f,k}$, we obtain the equilibrium wage-rental in the closed economy:

$$\frac{w}{r} = \frac{\sigma_L}{1 - \sigma_L} \frac{K}{L} = \frac{\mu(1 - \widetilde{\beta_Y}) + (1 - \mu)(1 - \widetilde{\beta_Z})}{\mu \widetilde{\beta_Y} + (1 - \mu)\widetilde{\beta_Z}} \frac{K}{L},$$

where the effective capital shares are:

$$\begin{aligned} \widetilde{\beta_Y} &= \beta_Y \Big(1 + \alpha \left(1 - \beta_Y\right) \left(2\bar{\phi} - 1\right)\Big) \\ \widetilde{\beta_Z} &= \beta_Z \left(1 + \alpha(1 - \beta_Z)(2\phi - 1)\right) \end{aligned}$$

Note that $\widetilde{\beta_Y} > \widetilde{\beta_Z}$, so that contract incompleteness does not create factor intensity reversals.

The Multi-Country Model

- Now suppose the world is divided in $J \geq 2$ countries, with country j receiving an endowment (K^j, L^j). Assume that preferences are identical in all J countries and that factors of production are internationally immobile. Furthermore, assume that for all $j \in J$, K^j/L^j is not "too different" from K/L (sufficient conditions are provided below), so that factor price equalization is attained and we can use the general equilibrium above to characterize the aggregate allocations in the world economy.

- Assume that intermediate inputs can be traded at zero cost, while final goods are nontradable so that each final-good producer (costless) sets J plants to service the J markets. On the other hand, because of the increasing returns to scale, intermediate inputs will be produced in only one country.

Pattern of Production

- The factor market clearing conditions in country $j \in J$ are now:

$$\begin{aligned}\sum_{k\in\{Y,Z\}} n_k^j\left(K_{x,k}^j + K_{f,k}^j\right) &= K^j \\ \sum_{k\in\{Y,Z\}} n_k^j\left(L_{x,k}^j + L_{f,k}^j\right) &= L^j\end{aligned}$$

- The fact that technology is the same in all J countries and that factor price equalization is attained implies that the investments $K_{x,k}^j$, $K_{f,k}^j$, $L_{x,k}^j$, and $L_{f,k}^j$ will be identical for all j. It thus follows that differences in production patterns will be channelled through n_Y^j and n_Z^j. In other words, factor market clearing is attained entirely through the extensive margin (relatively more varieties in the sector that uses intensively the abundant factor in country j). In particular,

  (Hecksher-Ohlin Theorem) *If country j is relatively capital-abundant (i.e. $K^j/L^j > K/L$), then $n_Y^j > s^j n_Y$ and $n_Z^j < s^j n_Z$, where s^j is j's share in world income, i.e*

$$s^j = \frac{rK^j + wL^j}{rK + wL}$$

- For the above allocation to be consistent with FPE, we require $n_Y^j > 0$ and $n_Z^j > 0$, or equivalently:

Assumption 3: $\frac{\widetilde{\beta_Y}\sigma_L}{\left(1-\widetilde{\beta_Y}\right)(1-\sigma_L)} > \frac{K^j/L^j}{K/L} > \frac{\widetilde{\beta_Z}\sigma_L}{\left(1-\widetilde{\beta_Z}\right)(1-\sigma_L)}$ for all $j \in J$.

Pattern of Trade

- Remember first that all world trade is in intermediate inputs x_Y and x_Z. A given country $N \in J$ will host $n_Y + n_Z$ producers of final-good varieties.
  - a measure n_Y^j will be importing from their *integrated* suppliers in every country $j \neq N$;
  - a measure n_Z^j will be importing from their *independent* suppliers in every country $j \neq N$.
- Each of these final-good producers in N will import a fraction s^N of world output of the corresponding variety. Assuming average cost transfer pricing ($p_{x_Y} = p_Y$ and $p_{x_Z} = p_Z$), we obtain

1. The volume of N's imports from S is

$$M^{N,S} = s^N \left(n_Y^S p_Y y + n_Z^S p_Z z\right) = s^N s^S \left(rK + wL\right)$$

2. The volume of N's intrafirm imports from S is

$$M_{i-f}^{N,S} = s^N n_Y^S p_Y y$$

3. The share of N's intrafirm imports from S is:

$$S_{i-f}^{N,S} = \frac{\left(\left(1-\widetilde{\beta_Z}\right)(1-\sigma_L)\frac{K^S}{L^S} - \widetilde{\beta_Z}\sigma_L\frac{K}{L}\right)}{\left(\widetilde{\beta_Y}-\widetilde{\beta_Z}\right)\left((1-\sigma_L)\frac{K^S}{L^S} + \sigma_L\frac{K}{L}\right)}$$

Figure 15.4: Volume of Intrafirm Imports

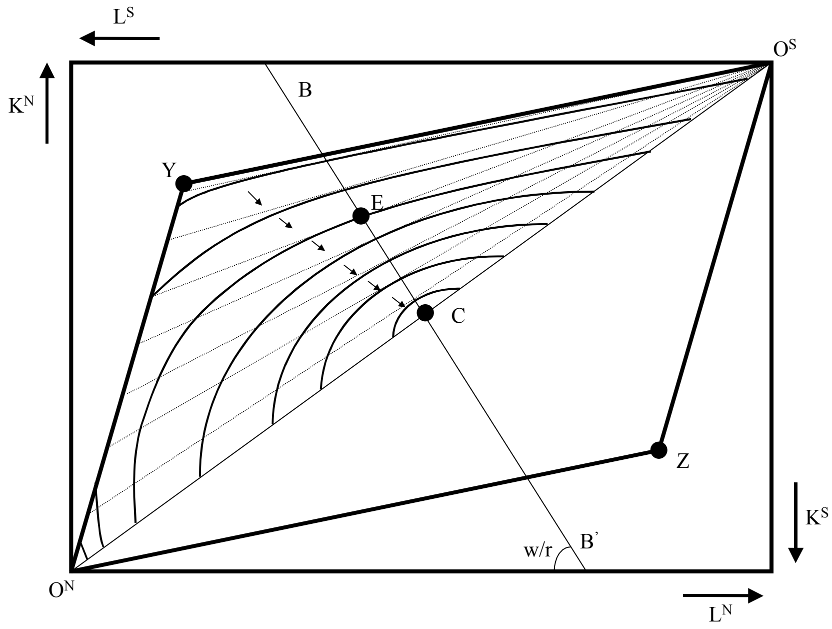
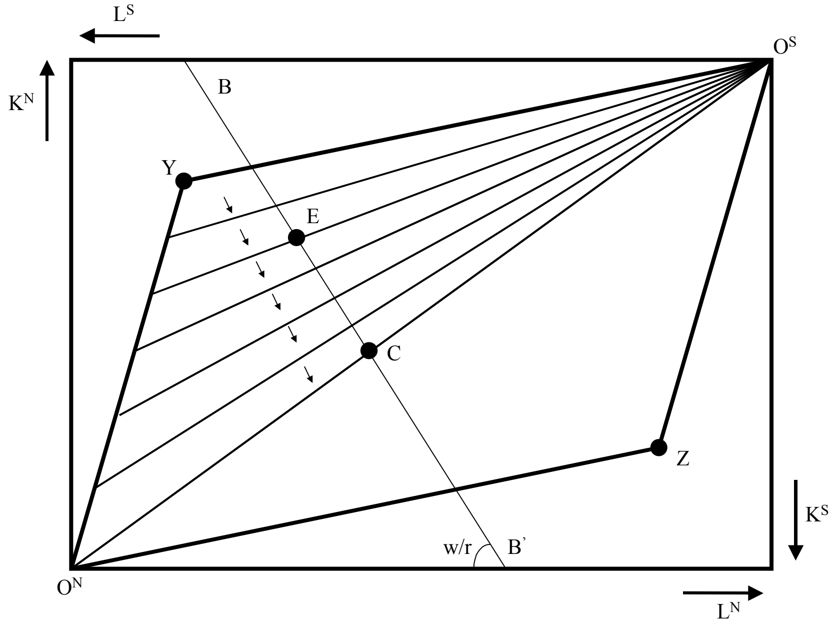


Figure 15.5: Share of Intrafirm Imports



Main Predictions

- *Let $N = USA$.*

  Lemma 22 *The volume $M_{i-f}^{USA,j}$ of U.S. intrafirm imports from country j is an increasing function of the capital-labor ratio K^j/L^j and the size s^j of the exporting country.*

  Proposition 23 *The share $S_{i-f}^{USA,j}$ of intrafirm imports in total U.S. imports from country j is an increasing function of the capital-labor ratio K^j/L^j of the exporting country and is independent of its size s^j.*

- Notice that this Proposition illustrates how in a world with international trade, the pattern of Figure 15.2 in the introduction is a direct implication of the pattern in Figure 15.1.

- Although for simplicity the model does not feature any FDI, it would be straightforward to extent the model to include it. The model would then predict that foreign direct flows should be heavily concentrated among capital-abundant, developed countries. This provides support for the view that international outsourcing to developing countries can be significant even when foreign direct flows are not (e.g. Feenstra and Hanson, 1996). In particular, the model developed above can help rationalize the recent surge in global production sharing (c.f. Feenstra, 1998 and references therein), and the lack of a parallel increase in foreign direct flows to developing countries (UNCTAD, 2001). For instance, an increase in the relative capital-labor ratio of developed countries, caused by trade integration with labor-abundant countries, could predict these trends.

Econometric Evidence

- The last section of the paper attempts to provide evidence that the patterns in Figures 15.1 and 15.2 are not driven by third omitted factors. The purpose is also to unveil additional factors affecting the relative prevalence of intrafirm trade.

Specification

- We first run $\ln\left(S_{i-f}^{USA,ROW}\right)_k = \theta_1 + \theta_2 \ln\left(K/L\right)_k + W_k'\theta_3 + \epsilon_k$, where we expect $\theta_2 > 0$. The model presented above actually predicts that the share should be 0 for industries with capital intensity β_k below a certain threshold $\widehat{\beta}$ and 1 for industries with $\beta_k > \widehat{\beta}$, a prediction that does not seem to be borne by the data. But Antràs (2003*a*) discusses how to smooth the prediction by appealing to a mismatch between the definition of an industry in the model and the classification used by the statistician (we discussed this already in Chapter 6).

- Next we run $\ln\left(S_{i-f}^{USA,j}\right) = \gamma_1 + \gamma_2 \ln\left(K^j/L^j\right) + \gamma_3 \ln\left(L^j\right) + W_j'\gamma_4 + \varepsilon_j$, where, from a log-linearization of the expression for $S_{i-f}^{N,S}$, we expect $\gamma_2 = \left(1-\sigma_L\right)\sigma_L / \left(1 - \sigma_L - \widetilde{\beta_Z}\right)$ and $\gamma_3 = 0$.

- Finally, we run $\ln\left(M_{i-f}^{USA,j}\right) = \omega_1 + \omega_2 \ln\left(K^j/L^j\right) + \omega_3 \ln\left(L^j\right) + W_j'\omega_4 + \varepsilon_j$, where, from a log-linearization of the expression for $M_{i-f}^{N,S}$, we expect $\omega_2 = \left(1-\sigma_L\right)\left(1-\widetilde{\beta_Z}\right) / \left(1 - \sigma_L - \widetilde{\beta_Z}\right) > \gamma_2$ and $\omega_3 = 1$.

Results

- These are the main results (regression below):

  1. In the cross-section of industries, capital-intensity and R&D intensity seem to be the major determinants of the decision to internalize imports. Other factors, such as human-capital intensity, have an insignificant effect on the share of intrafirm trade.

  2. In the cross-section of countries, the coefficient on capital abundance remains positive and significant after controlling for a whole set of country-specific variables such as size, human-capital abundance, corporate tax rates, and measures of institutions.

  3. Consistently with the theory, when the volume (and not the share) of intrafirm trade is chosen as the left-hand-side variable, the coefficient on capital abundance is higher $\omega_2 > \gamma_2$ and size seems to matter.

Table 4. Factor Intensity and the Share $S_{i-f}^{US,ROW}$

Dep. var. is	Random Effects Regressions					
$\ln\left(S_{i-f}^{US,ROW}\right)_m$	I	II	III	IV	V	VI
$\ln(K/L)_m$	0.947***	0.861***	0.780***	0.776***	0.703***	0.723***
	(0.187)	(0.190)	(0.160)	(0.162)	(0.249)	(0.253)
$\ln(H/L)_m$		0.369	-0.002	-0.038	-0.037	-0.081
		(0.213)	(0.188)	(0.200)	(0.206)	(0.221)
$\ln(R\&D/Sales)_m$			0.451***	0.470***	0.452***	0.421***
			(0.107)	(0.114)	(0.128)	(0.140)
$\ln(ADV/Sales)_m$				0.055	0.059	0.035
				(0.094)	(0.097)	(0.107)
$\ln(Scale)_m$					0.068	0.100
	99				(0.179)	(0.190)
$\ln(VAD/Sales)_m$						0.403
						(0.657)
R^2	0.50	0.55	0.72	0.73	0.73	0.73
No. of obs.	92	92	92	92	92	92

	Fixed Effects Regressions					
	I	II	III	IV	V	VI
$\ln(K/L)_m$	0.599**	0.610**	0.610**	0.610**	0.943**	1.058**
	(0.299)	(0.300)	(0.300)	(0.300)	(0.412)	(0.410)
p-value Wu-Hausman test	0.14	0.27	0.62	0.64	0.52	0.19

Note: Standard errors in parenthesis (*, **, and *** are 10, 5, and 1% significance levels)

Table 5. Factor Endowments and the Share $S_{i-f}^{US,j}$

Dep. var. is $\ln\left(S_{i-f}^{US,j}\right)$	I	II	III	IV	V	VI
$\ln(K/L)_j$	1.141^{***}	1.110^{***}	1.244^{***}	1.239^{***}	1.097^{**}	1.119^{**}
	(0.289)	(0.299)	(0.427)	(0.415)	(0.501)	(0.399)
$\ln(L)_j$		-0.133	-0.159	-0.158	-0.142	0.017
		(0.168)	(0.164)	(0.167)	(0.170)	(0.220)
$\ln(H/L)_j$			-1.024	-0.890	-1.273	-0.822
			(1.647)	(1.491)	(1.367)	(1.389)
$CorpTax_j$				-0.601	0.068	1.856
				(3.158)	(3.823)	(2.932)
$EconFreedom_j$					0.214	
					(0.213)	
$OpFDI_j$						-0.384^{*}
						(0.218)
$OpTrade_j$						0.292
						(0.273)
R^2	0.46	0.47	0.48	0.50	0.50	0.43
No. of obs.	28	28	28	28	28	26

Table 6. Factor Endowments and the volume $M_{i-f}^{US,j}$

Dep. var. is $\ln\left(M_{i-f}^{US,j}\right)$	I	II	III	IV	V	VI
$\ln(K/L)_j$	2.048^{***}	2.192^{***}	2.188^{***}	2.154^{***}	1.650^{**}	2.096^{***}
	(0.480)	(0.458)	(0.716)	(0.663)	(0.762)	(0.695)
$\ln(L)_j$		0.607^{**}	0.608^{**}	0.614^{**}	0.670^{**}	0.700
		(0.229)	(0.268)	(0.271)	(0.243)	(0.419)
$\ln(H/L)_j$			0.031	0.953	-0.406	0.708
			(3.289)	(3.316)	(2.992)	(3.052)
$CorpTax_j$				-4.135	-1.763	-0.647
				(5.294)	(5.955)	(5.295)
$EconFreedom_j$					0.795	
					(0.443)	
$OpFDI_j$						-1.006^{**}
						(0.474)
$OpTrade_j$						0.674
						(0.560)
R^2	0.44	0.52	0.52	0.53	0.60	0.49

Chapter 16

The Property-Rights Approach in International Trade (II): Antràs (2003b) and Antràs and Helpman (2004)

- In this chapter, we will discuss two further contributions incorporating the property-rights approach to international trade theory.
- We will first discuss Antràs (2003*b*), who studies the implications of incomplete contracting and the assignment of property rights for the emergence of product cycles and their organizational structure.
- We will then discuss the work of Antràs and Helpman (2004), who develop a theoretical model that combines the within-sectoral heterogeneity of Melitz (2003) with the structure of firms in Antràs (2003*a*). This allows them to study the impact of variations in productivity within sectors and of differences in technological and organizational characteristics across sectors on international trade, foreign direct investment, and the organizational choices of firms.

16.1 Incomplete Contracts and the Product Cycle: Antràs (2003*b*)

- As pointed out by Vernon (1966) in his classical "Product Cycle Hypothesis" paper, new goods are not only developed in high-wage countries, but they are also manufactured there for a while. Furthermore, Vernon emphasized the role of multinational firms in the eventual production transfer to less developed countries.

- There is by now substantial empirical evidence suggesting that indeed it takes time for low-wage countries to start producing relatively unstandardized goods. Interestingly, this empirical evidence also suggests that:

  - overseas assembly of new and unstandardized goods is kept within firm boundaries
  - arm's length production transfer (licensing, subcontracting) is more frequent for older goods and for goods with less product development requirements.

- The traditional theoretical literature on the product cycle either treats the emergence of product cycles as exogenous, as in Krugman (1979), or emphasizes the role of imitation ("it takes time to imitate"), as in Grossman and Helpman (1991a,b).

- Antràs (2003*b*) instead provides a theory of the product cycle that is more akin to Vernon's original one. In particular, the decision to shift production to the low-wage South will be a profit-maximizing one from the point of view of firms in the *industrialized* North. The time lag between the first appearance of the product and its manufacturing in the South will be explained by appealing to incomplete contracts *in international transactions.*

- Furthermore, when the model is extended to incorporate and endogenous choice of organizational structure, it is shown to immediately deliver the type of endogenous organizational cycles suggested by the empirical literature on the product cycle.

- We will also see that the model also delivers some interesting general equilibrium results that complement the important work of Krugman (1979) and Helpman (1993). For instance, at any point in time, the cross-sectional picture emerging from the model will be similar to Dornbusch-Fisher-Samuelson (1977), with the main difference being that comparative advantage will be *endogenous* and will depend on contractual parameters.

Partial Equilibrium Model

A. Endogenous Product Cycles

- Consider a world with two countries, the North and the South, and a single good y produced only with labor.

- **Preferences:** Demand for good y is:

$$y = \lambda p^{-1/(1-\alpha)}, \qquad 0 < \alpha < 1$$

- **Technology:** Production of y requires:

  - a special and distinct hi-tech input x_h (PD)
  - a special and distinct low-tech input x_l (M)
  - a fixed cost of f units of labor, wherever x_h is produced.

  Production of the final good requires no further costs and

$$y = \zeta_z x_h^{1-z} x_l^z, \qquad 0 \leq z \leq 1.$$

  There are two types of producers:

  - a stand-alone Research Center (R) controls the production of x_h (and y).
  - a stand-alone Manufacturing Plant (M) controls x_l (for now, we rule out vertical integration).

Both types of producers face a perfectly elastic supply of labor, with a wage that is assumed strictly higher in the North ($w^N > w^S$).

To produce one unit of x_h, a Northern R needs to hire one unit of labor. In the South this requires an infinite (or sufficiently high) amount of workers, thus implying that the South will have no R's. To produce one unit of x_l, both Northern and Southern M require one unit of labor.

Before x_h and x_l are produced, R decides whether it wants to produce y, and if so, whether to obtain x_l from a Northern M or a Southern one. The location of M is chosen by R to maximize its profits. As in Antràs (2003a), it is assumed that the ex-ante supply of M agents is infinitely elastic, so R is able to obtain all the surplus ex-ante through an upfront lump-sum fee. The choice of location will therefore be ex-ante efficient, in the sense the equilibrium location will maximize joint profits.

Both x_h and x_l are fully relationship–specific, so these inputs have zero value outside the relationship.

- **Contracting:** R and a Northern M are assumed to be able to sign enforceable contracts on purchases of intermediate inputs. Instead, contracts between R and a Southern M **cannot** be enforced. In the paper, I elaborate on this assumption (I allow parties to produce useless, bad-quality inputs at negligible cost and I assume that only when both inputs are produced in the same country can an outside party distinguish between a good-quality and a bad-quality intermediate input). It is also assumed that ex-ante labor investments and sale revenues are not verifiable either.

  The surplus is thus divided ex-post, and it is assumed that symmetric Nash Bargaining leaves R and M with $1/2$ of ex-post gains from trade.

Firm Behavior

- Because the analysis is similar to that in Antràs (2003a), I will skip most of the details.

- R chooses the location of M to maximize its profits. If it transacts with an M in the North, it obtains

$$\pi^N = \arg\max_{x_h, x_l} \left\{ S\left(x_h, x_l\right) - w^N x_h - w^N x_l - w^N f \right\},$$

because the parties are able to sign an enforceable ex-ante contract.

- If M is located in the South, it instead obtains

$$\pi^S = S\left(x_h^O, x_l^O\right) - w^N x_h^O - w^S x_l^O - w^N f$$

where

$$\begin{aligned} x_h^O &= \arg\max_{x_h} \frac{1}{2} S\left(x_h, x_l^O\right) - w^N x_h \\ x_l^O &= \arg\max_{x_l} \frac{1}{2} S\left(x_h^O, x_l\right) - w^S x_l \end{aligned}$$

The Equilibrium Choice

- It is straightforward to check that the low-tech produced will be produced in the South only if

$$A(z) \leq \omega \equiv w^N / w^S$$

where

$$A(z) \equiv \left(\frac{1-\alpha}{\left(1 - \frac{1}{2}\alpha\right) \left(\frac{1}{2}\right)^{\alpha/(1-\alpha)}} \right)^{(1-\alpha)/\alpha z}.$$

- Note that $A'(z) < 0$, $\lim_{z \to 0} A(z) = +\infty$ and $A\left(z\right) > 1$ for all $z \in [0,1]$. It thus follows that if $w^N = w^S$, x_l will not be produced in the South for any $z \in [0,1]$. In sum,

Lemma 1: *If $A(1) < \omega$, there exists a unique threshold $\bar{z} \in (0,1)$ such that the low-tech input is produced in the North if $z < \bar{z} \equiv A^{-1}(\omega)$, while it is produced in the South if $z > \bar{z} \equiv A^{-1}(\omega)$.*

- Intuitively, the benefits of Southern assembly are able to offset the distortions created by incomplete contracting only when the manufacturing stage is suffi-

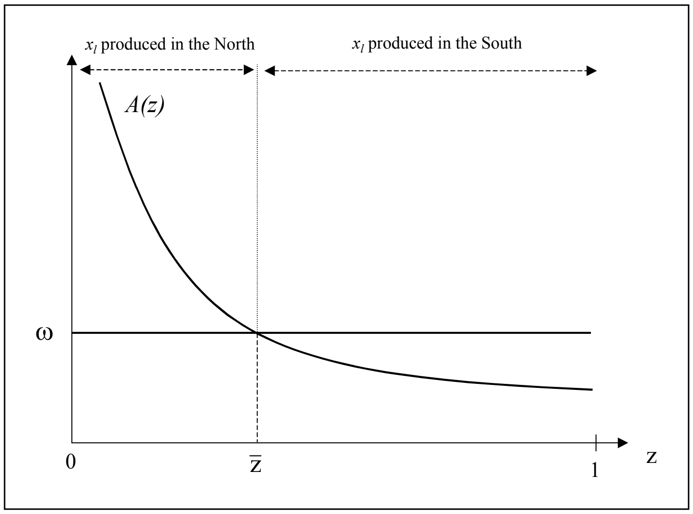


Figure 16.1: The Choice of Location

ciently important in production or when the wage in the South is sufficiently lower than that in the North.

Dynamics: The Product Cycle

- Let now time be continuous, indexed by t, with $t \in [0, \infty)$. Demand for y is assumed identical at each point in time. The parameters α and ω are also time-invariant. Furthermore, firm structure is such that reputational equilibria are not sustainable.

- Consider the following simple standardization process: $z(t) = h\,(t)$, with $h'(t) > 0$, $h\,(0) = 0$, and $\lim_{t\to\infty} h(t) = 1$.

- Intutively, most goods require a lot of R&D and product development in the early stages of their life cycle, while the mere assembling or manufacturing becomes a much more significant input in production as the good matures.

Proposition 24 *The model displays a product cycle. When the good is relatively new or unstandardized, i.e., $t \leq h^{-1}\,(\bar{z})$, the manufacturing stage of production takes place in the North. When the good is relatively mature or standardized, i.e., $t > h^{-1}\,(\bar{z})$, manufacturing is undertaken in the South.*

- For example, let $z(t) = h(t) = 1 - e^{-\frac{t}{\theta}}$, where $1/\theta$ is the rate of standardization. Then manufacturing is shifted at $\bar{t} = \theta \ln\left(\frac{1}{1-\bar{z}}\right)$. The faster the standardization, the earlier production transfer.

- Notice that with complete contracts, if $\omega > 1$, manufacturing is shifted to the South from period 0. If $\omega = 1$, location of manufacturing is indeterminate and product cycles emerge with probability zero. The presence of incomplete contracts is therefore *necessary* for a product cycle to arise.

B. Endogenous Organizational Cycles

- Consider next a simple extension of the above setup. In particular, let R choose the amount of control exterted on the production of the low-tech input. In particular, we now give R the option of vertically integrating his/her M.

- As in Grossman-Hart (1986), ownership will affect the distribution of ex-post surplus through its effect on each party's outside option. Since the x_l is specific to R, the outside option for M is always 0. If M is a stand-alone firm, the outside option for R is also 0. But by integrating M, R obtains the residual rights of control over x_l, so R can fire the M manager and still produce the final good. As in Antràs (2003a), it is assumed that such contractual breach is partially costly and only a fraction δ of output can be produced. For simplicity, assume: $\delta \leq \left(\frac{1}{2}\right)^{1/\alpha}$.

- Because contracts are complete when M is located in the North, we will not be able to pin down firm boundaries there. When M is located in the South matters are more interesting.

- A research center in the North has now three options:

  - Obtain x_l from a Northern M;
  - Obtain x_l from a stand-alone Southern M;
  - Obtain x_l from an integrated Southern M (i.e., become a multinational firm).

- The first two options yield profits π^N and π^S, as defined above. Assembly in the South by a Vertically-Integrated Manufacturing Plant leaves RC with

$$\pi_M^S = S\left(x_h^V, x_l^V\right) - w^N x_h^V - w^S x_l^V - w^N f$$

where

$$\begin{aligned} x_h^V &= \arg\max_{x_h} \bar{\phi} S\left(x_h, x_l^V\right) - w^N x_h \\ x_l^V &= \arg\max_{x_l} \left(1 - \bar{\phi}\right) S\left(x_h^V, x_l\right) - w^S x_l \end{aligned}$$

and $\bar{\phi} = \frac{1}{2}\left(1 + \delta^{\alpha}\right) > \frac{1}{2}$.

The Equilibrium Choice Revisited

- It is straightforward to show that there are now 3 thresholds implictly defined by:
  - $\bar{z}$ s.t. $\pi^N(z) > \pi^S(z)$ if and only if $z < \bar{z}$ (this was shown above)
  - $\bar{z}_{MN}$ s.t. $\pi^N(z) > \pi_M^S(z)$ if and only if $z < \bar{z}_{MN}$ (this can be shown analogously)
  - $\bar{z}_{MS}$ s.t. $\pi_M^S(z) > \pi^S(z)$ if and only if $z < \bar{z}_{MS}$.

- The existence of the third threshold is proved in Antràs (2003a) (see Chapter 15). Remember that, when z is low, product development is a relatively important input in production and the research center will want to integrate the transaction to have sufficient power inside the firm. When the good is standardized, the low-tech manufacturing input is a relatively more important input in production, and outsourcing will be chosen to make sure that the manager of M has the right incentives to generate the highest amount of sale revenues from the technology.

- Notice that:
  - For a sufficiently low z, the benefits from any type of Southern assembly will be low relative to the distortions from incomplete contracting, and x_l will be produced in the North.

  - For a sufficiently large z, a profit-maximizing research center will decide to outsource the manufacturing input to an independent manufacturing plant in the South.
  - Whether for intermediate values of z the research center becomes a multinational firm or not depends on parameter values. Multinational firms will emerge when $z \in [\bar{z}_{MN}, \bar{z}_{MS}]$, which requires $\bar{z}_{MN} < \bar{z}_{MS}$.

Dynamics: The Product Cycle

- From this discussion, in the dynamic version of the model, we obtain the following result:

Proposition 25 *The model displays a product cycle. If $\bar{z}_{MS} < \bar{z}_{MN}$, the product cycle is as before. If instead $\bar{z}_{MS} > \bar{z}_{MN}$, the following three-stage product cycle emerges:*

(i) When the good is relatively new, i.e., $t < h^{-1}(\bar{z}_{MN})$, the manufacturing stage of production takes place in the North.

(ii) For an intermediate maturity of the good, $h^{-1}(\bar{z}_{MN}) < t < h^{-1}(\bar{z}_{MS})$, manufacturing is shifted to the South but is undertaken within firm boundaries.

(iii) When the good is relatively standardized, i.e., $t > h^{-1}(\bar{z}_{MS})$, production is shifted to an unaffiliated party in the South.

- Notice that the same force that creates endogenous product cycles in the model is also instrumental in shaping the endogenous organizational cycles.
- Antràs (2003*b*) discusses time-series (Korean electronics industry) and cross-sectional evidence consistent with these predictions.

The General-Equilibrium Model

- Antràs (2003*b*) nexts embeds this partial-equilibrium model in a dynamic, general-equilibrium framework with varieties in different sectors standardizing at different rates. The setup is as follows:
- At each $t \in [0, \infty)$, the North is endowed with L^N units of labor; the South with L^S. At each $t \in [0, \infty)$, there exists a measure $N(t)$ of industries indexed by

j, each producing an endogenously determined measure $n_j(t)$ of differentiated goods. It is assumed that $N(t)$ grows at an exogenous rate g: $\dot{N}(t) = gN(t)$ and $N(0) = N_0 > 0$.

- The representative consumer in each country maximizes

$$U = \int_0^\infty e^{-\rho t} \int_0^{N(t)} \log \left(\int_0^{n_j(t)} y_j(i,t)^\alpha \, di \right)^{1/\alpha} dj dt,$$

  from which we derive the following demand for a variety i in industry j at time t:

$$y_j(i,t) = \lambda_j(t) p_j(i,t)^{-1/(1-\alpha)}.$$

- Production of $y_j(i,t)$ is as described in the partial-equilibrium model above. For simplicity, let us **initially rule out** vertical integration.

- Assume that all producers in a given industry share the same technology with a common time-varying elasticity $z_j(t-t_{0j})$, where t_{0j} is the date at which industry j appears. As before, we assume $z_j'(\cdot) > 0$, $z_j(t_{0j}) = 0$, and $\lim_{t-t_{0j}\to\infty} z_j(t-t_{0j}) = 1$. Notice that industries vary both in their "birth dates", as well as in the shape of their specific $z_j(\cdot)$ standardization processes. Firm structure is as above, with the additional feature that free entry at every period t ensures that the measure $n_j(t)$ adjusts so as to make $\pi = 0$.

- Because we abstract from the analysis of reputational equilibria, our assumptions allow us to focus on period-by-period analysis.

General Equilibrium

- Because ω and α are common across industries, so will $\bar{z}(t)$. All firms in all industries with $z_j(t-t_{0j}) < \bar{z}(t)$ will manufacture x_l in the North. Those with $z_j(t-t_{0j}) > \bar{z}(t)$ will do so in the South. The general equilibrium need only pin down $\omega(t)$ and $\bar{z}(t)$.

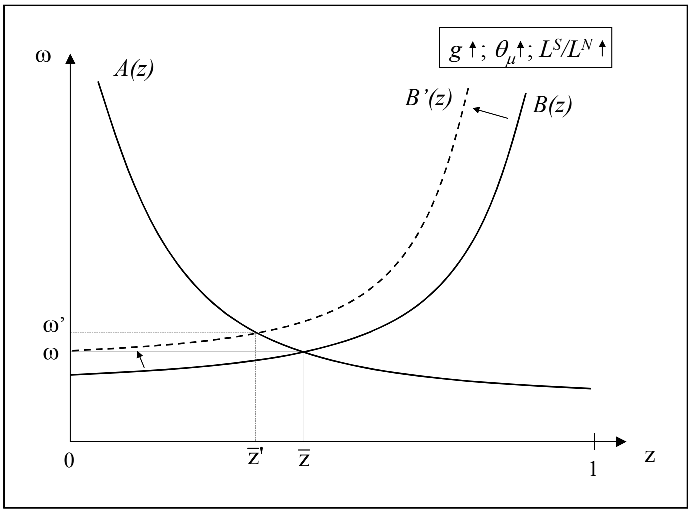


Figure 16.2: General Equilibrium

- This is achieved by noting that world income equals world spending

$$w^N(t)L^N + w^S(t)L^S = E(t),$$

  and by imposing labor market clearing, which yields a second equation (the first is $A(\bar{z}) = \omega$) linking $\bar{z}$ and ω.

$$\omega = B_t(\bar{z}) \equiv \frac{1 - \frac{1}{2}\alpha \int_{\bar{z}}^{1} z f_{z,t}(z)dz}{\frac{1}{2}\alpha \int_{\bar{z}}^{1} z f_{z,t}(z)dz} \frac{L^S}{L^N}$$

  Notice that the $B_t(\bar{z})$ schedule is increasing and depends on the distribution of z's in the world.

- This is depicted in Figure 16.2. In spite of heterogeneity in industry product-cycle dynamics, the cross-sectional picture is very similar to Dornbusch et al. (1977). But the $A(z)$ curve is here endogenous!

An Example

- Let $z_j(t - t_{0j}) = z_j(t, t_{0j}) = 1 - e^{-(t-t_{0j})/\theta_j}$ and assume θ_j is exponentially distributed with mean θ_μ, and is independent of t_{0j}. One can then show that the economy converges to a time-invariant c.d.f. for the z's given by:

$$F_z(z) = \frac{g\theta_\mu \ln\left(\frac{1}{1-z}\right)}{1 + g\theta_\mu \ln\left(\frac{1}{1-z}\right)}$$

Hence, in this case the general equilibrium values of $\bar{z}$ and ω are time-invariant. Furthermore, it is easy to see that:

Proposition 26 *Holding θ_j and t_{0j} constant, the relative wage in the North is higher and the shift to Southern assembly occur earlier: (i) the higher is g, (ii) the lower is $1/\theta_\mu$, (iii) the higher is L^S/L^N.*

- The effects on relative wages are analogous to those in Krugman (1979). Nevertheless, the model delivers implications for the timing of production transfer, which is exogenous in Krugman's model.

Welfare

- Antràs (2003*b*) then shows that:

Proposition 27 *Relative to a world with incomplete contracting, a shift to complete contracts unambiguously increases welfare in the South, while having an ambiguous effect on welfare in the North.*

- Three effects are at work (see the paper for details):
  - (i) the terms of trade (w^N/w^S) move in favor of the South;
  - (ii) production efficiency increases;
  - (iii) there is an ambiguous effect on the number of available varieties, but it is always outweighed by (i) and (ii).

- The result contrasts with Helpman's (1993) analysis of a tightening of intellectual property rights (IPRs) in a product cycle model in which imitation is the vehicle of production transfer. Intuitively, a tightening of IPRs moves terms of trade against the South and reduces production efficiency. Instead, in this model, an improvement in the contracting environment moves the terms of trade in favor of the South and increases production efficiency.

General Equilibrium with Multinational Firms

- Antràs (2003*b*) finally solves for the general equilibrium with multinational firms and derives a series of additional results.

Proposition 28 *Relative to a world with only arm's length transacting, allowing for intrafirm technology transfer by multinational firms weakly accelerates the transfer of production to the South (lowers $\widetilde{z}$), while having an ambiguous effect on the relative wage ω. Furthermore, provided that its effect on relative wages is small enough, allowing for intrafirm production transfer by multinational firms is welfare improving for both countries.*

- Intuitively, the introduction of multinational firms in a world with only arm's length transacting helps to alleviate the distortions generated by the incompleteness of contracts and faciliates an earlier transfer to the low-wage country.
- Finally, in our particular example it can be shown that:

Proposition 29 *The measure of product-development intensities for which multinationals exist, i.e.,* $\min\{\bar{z}_{MS} - \bar{z}_{MN}, 0\}$*, is non-decreasing in* g*,* θ_μ*, and* L^S/L^N*.*

- Hence, the same forces that, in view of Proposition 26, might explain a shortening of product cycles might also explain an increase in FDI to less developed countries.

16.2 Global Sourcing with Heterogenous Firms: Antràs and Helpman (2004)

- Antràs and Helpman (2004) incorporate intraindustry heterogeneity of the Melitz (2003) type in a property-rights model of the multinational firm. This is motivated by the substantial evidence of intraindustry heterogeneity in firms' participation in foreign trade and its relationship with firm characteristics, such as productivity.

- Their theoretical model delivers testable implications that go beyond the predictions derived in the work of Antràs (2003*a,b*). For instance, the model predicts that, in a cross-section of industries, the share of intrafirm imports of components in total imports of components should be higher in industries with higher productivity dispersion. They also study the effects of falling transport costs on the relative prevalence of intrafirm versus arm's-length foreign trade.

The Model

- **Environment and Preferences:** Consider a world with two countries, the North and the South, and a unique factor of production, labor. There is a representative consumer in each country with quasi-linear preferences:

  $$U = x_0 + \frac{1}{\mu}\sum_{j=1}^{J} X_j^{\mu},\ 0 < \mu < 1.$$

  where x_0 is consumption of a homogeneous good, X_j is an index of aggregate consumption in sector j, and μ is a parameter. Aggregate consumption in sector j is a CES function

  $$X_j = \left[\int x_j(i)^{\alpha} di\right]^{1/\alpha},\ \ 0 < \alpha < 1,$$

  of the consumption of different varieties $x_j(i)$, where the range of i will be endogenously determined. This specification leads to the following inverse demand function for each variety i in sector j:

  $$p_j\left(i\right) = X_j^{\mu-\alpha} x_j(i)^{\alpha-1}.$$

- **Technology:** Producers of differentiated goods face a perfectly elastic supply of labor. Let the wage in the North be strictly higher than that in the South ($w^N > w^S$). The market structure is one of monopolistic competition.

  - As in Melitz (2003), producers needs to incur sunk entry costs $w^N f_E$, after which they learn their productivity $\theta \sim G\left(\theta\right)$.

- As in Antràs (2003*a*), final-good production combines two specialized inputs according to the technology:

$$x_j(i) = \theta \left(\frac{h_j(i)}{\eta_j} \right)^{\eta_j} \left(\frac{m_j(i)}{1-\eta_j} \right)^{1-\eta_j}, \quad 0 < \eta_j < 1.$$

- h is controlled by a final-good producer (agent H), m is controlled by an operator of the production facility (agent M). This is true both when the input is produced in-house, as well as when it is purchased at arm's length.
- Sectors vary in their intensity of headquarter services η_j. Furthermore, within sectors, firms differ in productivity θ.
- Intermediates are produced using labor with a fixed coefficient. It is assumed that $h_j(i)$ is produced only in the North, which implies that the headquarters H are always located in the North. Productivity in the production of $m_j(i)$ is assumed identical in both countries.
- After observing θ, H decides whether to exit the market or start producing. In the latter case additional fixed cost of organizing production need to be incurred. It is assumed that these additional fixed cost are a function of the structure of ownership and the location of production. We assume that the fixed organizational costs are higher when M is located in the South regardless of ownership structure, because the fixed costs of search, monitoring, and communication are significantly higher in the foreign country. We also assume that, given the location of M, the fixed organizational costs of a V-firm are higher than the fixed organizational costs of an O-firm.
- In particular, if an *organizational form* is $k \in \{V, O\}$ and $\ell \in \{N, S\}$, these fixed costs are $w^N f_k^\ell$ and satisfy

$$f_V^S > f_O^S > f_V^N > f_O^N. \tag{16.1}$$

- **Contracting:** The setting is one of incomplete contracts. It is assumed that, regardless of the location of intermediate input production, parties cannot sign ex-ante enforceable contracts specifying the purchase of specialized intermediate inputs for a certain price. Furthermore, no contracts contingent on amount of labor hired or on sale revenues are available.

The surplus will thus be divided ex-post. It is assumed that H captures a fraction β of the ex-post gains from trade. As in Antràs (2003a,b), these gains from trade are also a function of the organization of production. To reiterate our discussions above, ex-post bargaining takes place both under outsourcing and under insourcing, but firm boundaries affect the threat points in the negotiations (Grossman and Hart, 1986). In particular, the outside option for H will be higher under integration. We will also assume that the better contracting environment in the North also (weakly) raises H's outside option in intrafirm domestic transactions (relative to transnational ones). In particular, the fraction of output that can be recovered in case of contractual breach in intrafirm transactions are δ^N and δ^S, with $\delta^N \geq \delta^S$. The outside option of H under outsourcing is zero regardless of the location of M. The outside option of M is zero regardless of ownership structure and location.

Following Antràs (2003a), the ex-post division of surplus is as follows: H gets a fraction β_k^ℓ of sale revenues, where β_k^ℓ is as follows:

	North	South
Non-Integration	$\beta_O^N = \beta$	$\beta_O^S = \beta$
Integration	$\beta_V^N = \left(\delta^N\right)^\alpha + \beta\left[1 - \left(\delta^N\right)^\alpha\right]$	$\beta_V^S = \left(\delta^S\right)^\alpha + \beta\left[1 - \left(\delta^S\right)^\alpha\right]$

Notice that

$$\beta_V^N \geq \beta_V^S > \beta_O^N = \beta_O^S = \beta.$$

- **Ex-ante Division of Surplus:** As in Antràs (2003a), it is assumed that the ex-ante supply of M agents is infinitely elastic, so H obtain all the surplus ex-ante and the H's profit-maximizing organizational mode will also maximize joint profits.

Equilibrium

- Let R be potential sales revenues. H then solves:

$$
\begin{aligned}
\max_{\beta_k^\ell \in \left\{\beta_V^N, \beta_V^S, \beta_O^N, \beta_O^S\right\}} \quad & \pi_k^\ell = \pi\left(h\left(\beta_k^\ell\right), m\left(\beta_k^\ell\right)\right) \\
s.t. \quad h\left(\beta_k^\ell\right) &= \arg\max_h \beta_k^\ell R\left(h, m\left(\beta_k^\ell\right)\right) - w^N h \\
m\left(\beta_k^\ell\right) &= \arg\max_m \left(1-\beta_k^\ell\right) R\left(h\left(\beta_k^\ell\right), m\right) - w^\ell m
\end{aligned}
$$

- This program simplifies to

$$
\max_{\beta_k^\ell \in \left\{\beta_V^N, \beta_V^S, \beta_O^N, \beta_O^S\right\}} \pi_k^\ell\left(\theta, X, \eta\right) = X^{(\mu-\alpha)/(1-\alpha)} \theta^{\alpha/(1-\alpha)} \psi_k^\ell\left(\eta\right) - w^N f_k^\ell \tag{16.2}
$$

where

$$
\psi_k^\ell\left(\eta\right) = \frac{1-\alpha\left[\beta_k^\ell \eta + \left(1-\beta_k^\ell\right)\left(1-\eta\right)\right]}{\left[\frac{1}{\alpha}\left(\frac{w^N}{\beta_k^\ell}\right)^\eta \left(\frac{w^\ell}{1-\beta_k^\ell}\right)^{1-\eta}\right]^{\alpha/(1-\alpha)}}.
$$

- By choosing k and ℓ, H is effectively choosing a triplet $\left(\beta_k^\ell, w^\ell, f_k^\ell\right)$. And:
  - π_k^ℓ is decreasing in w^ℓ and f_k^ℓ.
  - π_k^ℓ is largest when $\beta_k^\ell = \beta^*\left(\eta\right)$, with $\beta^{*\prime}\left(\eta\right) > 0$, $\beta^*\left(0\right) = 0$ and $\beta^*\left(1\right) = 1$ (see Figure 16.3). Intuitively, H wants to allocate relatively more power to the party undertaking a relatively more important investment in production.

Industry Equilibrium

- Upon observing its productivity level θ, a final-good producer H chooses the ownership structure and the location of manufacturing that maximizes (16.2), or exits the industry and forfeits the fixed cost of entry $w^N f_E$. It is clear from (16.2) that the latter outcome occurs whenever θ is below a threshold θ, denoted by $\underline{\theta} \in (0, \infty)$, at which the operating profits

$$
\pi\left(\theta, X, \eta\right) = \max_{k \in \{V,O\}, \ell \in \{N,S\}} \pi_k^\ell\left(\theta, X, \eta\right) \tag{16.3}
$$

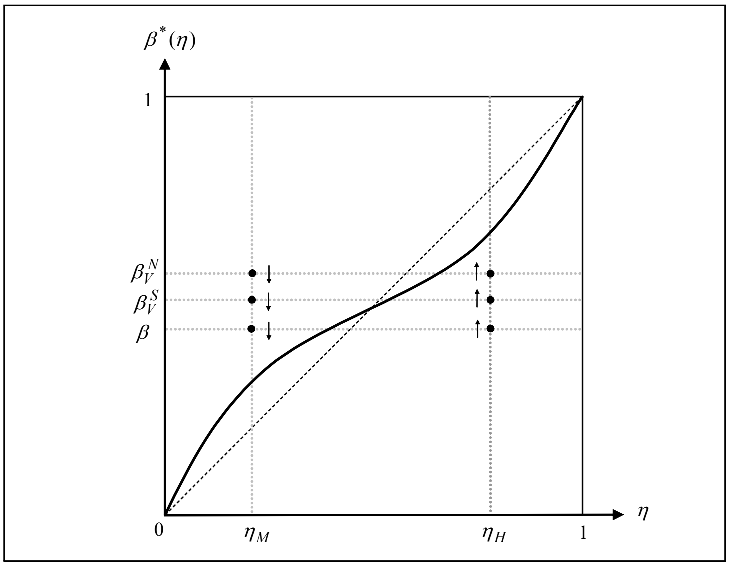


Figure 16.3: Profit-Maximizing Distribution of Surplus

equal zero. Namely, $\underline{\theta}$ is implicitly defined by

$$\pi\left(\underline{\theta}, X, \eta\right) = 0. \tag{16.4}$$

- Firms with $\theta \geq \underline{\theta}\left(X\right)$ stay in the industry and free-entry condition can be expressed as

  $$\int_{\underline{\theta}(X)}^{\infty} \pi\left(\theta, X, \eta\right) dG\left(\theta\right) = w^N f_E. \tag{16.5}$$

  This condition provides an implicit solution to the sector's real consumption index X, from which one can calculate all other variables of interest.

Relevant Trade-Offs

- The choice of an organizational form faces two types of tensions.
  - In terms of the location decision, variable costs are lower in the South, but fixed costs are higher there. As should be familiar from the work of Melitz (2003), it is clear in the present setup too, a firm's productivity θ will turn out to affect crucially the participation in international trade (e.g., the purchases of inputs from the South).

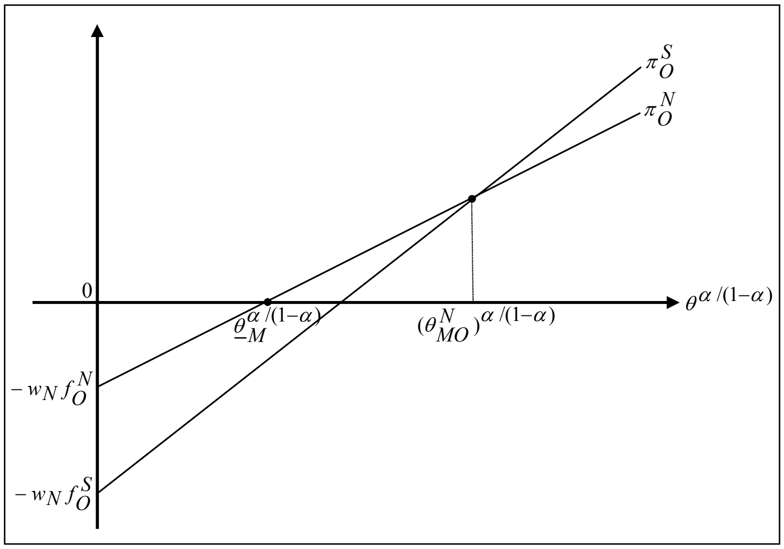


Figure 16.4: Equilibrium in the Component-Intensive Sector

- In terms of the integration decision, integration improves efficiency of variable production when the intensity of headquarter services is high, but involves higher fixed costs. This decision will thus crucially depend on η but also on θ.

- To simplify the discussion, we examine organizational forms in only two types of sectors:

1. A **Component-intensive sector** with $\eta < \beta^{*^{-1}}(\beta)$ and $w^N/w^S < \left(f_O^S/f_O^N\right)^{(1-\alpha)/\alpha(1-\eta)}$.

   - This implies $\psi_O^\ell(\eta) > \psi_V^\ell(\eta)$ for $\ell = N, S$, which together with (16.1), implies that any form of integration is dominated in equilibrium.
   - The equilibrium is depicted in Figure 16.4. Notice from equation (16.2) that, as in Melitz (2003) and Helpman et al. (2003), profits are linear in $\theta^{\alpha/(1-\alpha)}$. Thus the different profit curves are straight lines with an intercept equal to $-f_k^\ell$ and a slope proportional to ψ_k^ℓ.
   - Firms with productivity below $\underline{\theta}_M$ expect negative profits under all organizational forms. Therefore they exit the industry. Firms with productivity between $\underline{\theta}_M$ and θ_{MO}^N attain the highest profits by outsourcing in the North,

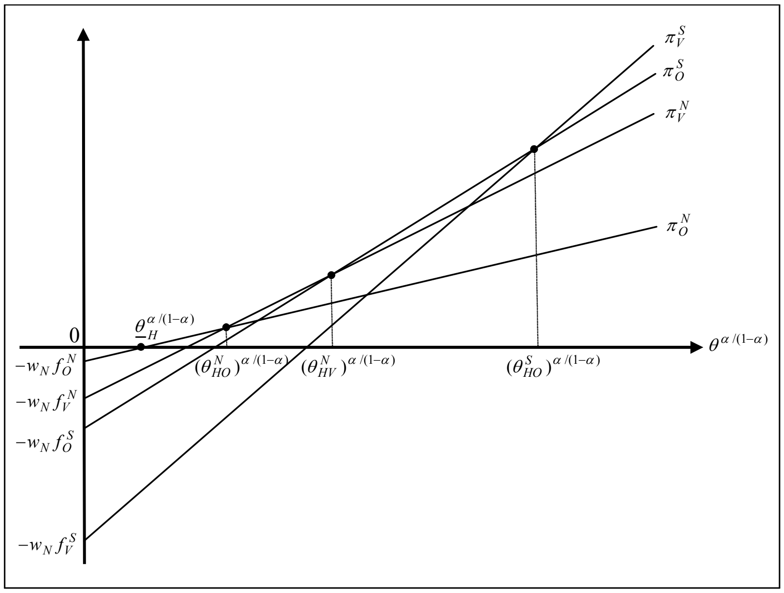


Figure 16.5: Equilibrium in the Headquarter-Intensive Sector

whereas firms with productivity above θ_{MO}^N attain the highest profits by outsourcing in the South.

- The cutoffs $\underline{\theta}_M$ and θ_{MO}^N are given by

$$\left.\begin{array}{l} \underline{\theta}_M = X^{(\alpha-\mu)/\alpha}\left[\frac{w^N f_O^N}{\psi_O^N(\eta)}\right]^{(1-\alpha)/\alpha}, \\ \theta_{MO}^N = X^{(\alpha-\mu)/\alpha}\left[\frac{w^N\left(f_O^S - f_O^N\right)}{\psi_O^S(\eta)-\psi_O^N(\eta)}\right]^{(1-\alpha)/\alpha}. \end{array}\right\} \qquad (16.6)$$

2. A **Heaquarter-intensive sector** with $\eta > \beta^{*^{-1}}\left(\beta_V^N\right)$, and $\left(w^N/w^S\right)^{1-\eta} > \phi\left(\beta_V^N, \eta\right)/\phi\left(\beta, \eta\right)$, where

$$\phi\left(\zeta,\eta\right) \equiv \left\{1-\alpha\left[\zeta\eta + (1-\zeta)(1-\eta)\right]\right\}^{(1-\alpha)/\alpha}\zeta^{\eta}\left(1-\zeta\right)^{1-\eta}.$$

- This implies the ranking of slopes

$$\psi_V^S(\eta) > \psi_O^S(\eta) > \psi_V^N(\eta) > \psi_O^N(\eta). \qquad (16.7)$$

which together with (16.1), implies the pattern of slopes and intercepts depicted in Figure 16.5.

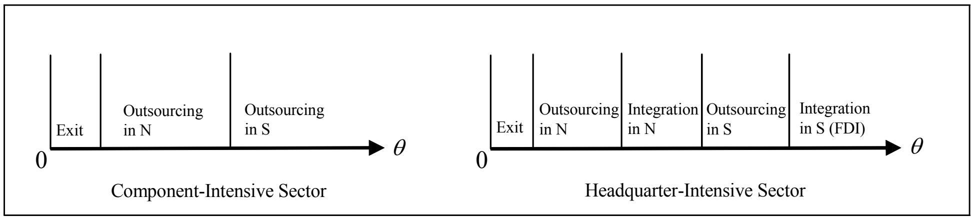


Figure 16.6: Organizational Forms

- In our benchmark case for headquarter-intensive sectors, all four organizational forms exist in equilibrium, with outsourcing and insourcing taking place in both countries. In particular, firms with productivity below $\underline{\theta}_H$ exit the industry, those with productivity between $\underline{\theta}_H$ and θ^N_{HO} outsource in the North, those with productivity between θ^N_{HO} and θ^N_{HV} integrate in the North, those with productivity between θ^N_{HV} and θ^S_{HO} outsource in the South, and those with productivity above θ^S_{HO} integrate in the South (engage in vertical FDI). These cutoffs are easily computed (see equation 15 in the paper).

Relative Prevalence

- These implied sortings are depicted in Figure 16.6. In the final section of the paper, we quantify the relative prevalence of the different organizational forms and how this prevalence varies across industries. Relative prevalence is measured by the share of products produced in various organizational forms (V or O, in N or S).

- This requires parameterizing the distribution of of θ. Following Helpman, Melitz and Yeaple (2003), we choose $G(\theta)$ to be a Pareto distribution with shape z, i.e.,

$$G(\theta) = 1 - \left(\frac{b}{\theta}\right)^z \text{ for } \theta \geq b > 0. \tag{16.8}$$

Remember that z is inversely related to the variance of the distribution.

- Denote by σ^ℓ_{MO} the fraction of active firms that outsource in country ℓ in the component-intensive sector. Then, as is clear from Figure 16.4,

$$\sigma_{MO}^{S}=\frac{1-G\left(\theta_{MO}^{N}\right)}{1-G\left(\underline{\theta}_{M}\right)}$$

and $\sigma_{MO}^{N}=1-\sigma_{MO}^{S}$. The Pareto distribution (16.8) then implies that $\sigma_{MO}^{S}=\left(\underline{\theta}_{M}/\theta_{MO}^{N}\right)^{z}$. Substituting (16.6) into this expression yields

$$\sigma_{MO}^{S}=\left[\frac{\psi_{O}^{S}(\eta)-\psi_{O}^{N}(\eta)}{\psi_{O}^{N}(\eta)}\ \frac{f_{O}^{N}}{f_{O}^{S}-f_{O}^{N}}\right]^{z(1-\alpha)/\alpha}. \tag{16.9}$$

- From (16.9) it follows that:
  - σ_{MO}^{S} is increasing in w^{N}/w^{S}, and decreasing in z and η. Hence, foreign outsourcing is more prevalent in countries with lower (efficiency-adjusted) wages and in industries with higher productivity dispersion and lower headquarter intensity.
  - Transport costs for components can easily be introduced in the model with their effect being analogous to a higher Southern wage w^{S}. Hence, foreign outsourcing increases when transport costs fall.

- The analysis of relative prevalence in the headquarter-intensive sector are analogous. The following are some of the results they derive:
  - A fall in the relative wage in the South or in trading costs, raise the share of imported inputs and also raise outsourcing relative to integration in every country. The paper discusses empirical evidence consistent with these trends.
  - In industries with more productivity dispersion (lower z), the share of imported inputs is higher and integration is higher relative to outsourcing in every country.
  - In sectors with higher headquarter intensity (higher η), the share of imported inputs is lower and integration is higher relative to outsourcing in every country. Thus, consistently with the findings of Antràs (2003*a*) that the

share of intra-firm imports in total U.S. imports is significantly higher, the higher the R&D intensity of the industry.

References

Aghion, Philippe and Jean Tirole (1997), "Formal and Real Authority in Organizations," The Journal of Political Economy, Vol. 105, No. 1, pp. 1-29.

Aitken, B. and A.Harrison (1999), "Do Domestic Firms Benefit from Foreign Direct Investment? Evidence from Panel Data," American Economic Review, 89(3), pp. 605-618.

Aitken, B., G. Hanson and A.E. Harrison (1997), "Spillovers, foreign investment, and export behavior", Journal of International Economics, 43(1), 103-132.

Antràs, Pol (2003a), "Firms, Contracts, and Trade Structure," Quarterly Journal of Economics, 118 (4), pp. 1375-1418.

Antràs, Pol (2003b), "Incomplete Contracts and the Product Cycle," NBER Working Paper 9945.

Antràs, Pol and Elhanan Helpman (2003), "Global Sourcing" NBER Working Paper 10082, forthcoming Journal of Political Economy.

Aw, B.Y., S. Chung and M.J. Roberts (2000), "Productivity and Turnover in the Export Market: Micro-level Evidence from the Republic of Korea and Taiwan (China)," World Bank Economic Review, 14(1), 65-90.

Bernard A. and J. B. Jensen (2004), "Why Some Firms Export", The Review of Economics and Statistics, forthcoming.

Bernard A., J. B. Jensen, and P. Schott (2003), "Falling Trade Costs, Heterogeneous Firms and Industry Dynamics," NBER Working Paper 9639.

Bernard, A.B. and J.B. Jensen (1999), "Exceptional Exporter Performance: Cause, Effect, or Both?" Journal of International Economics, 47(1), 1-25.

Bernard, A.B., J. Eaton, J.B. Jensen and S. Kortum (2003), "Plants and Productivity in International Trade", American Economic Review, Vol. 93, No. 4, September, pp. 1268-1290.

Brainard, S. Lael (1997), "An Empirical Assessment of the Proximity-Concentration Trade-off Between Multinational Sales and Trade," American Economic Review, 87:4, pp. 520-544.

Clerides, S., S. Lach and J. Tybout (1998), "Is learning by exporting important? Micro-dynamic Evidence from Colombia, Mexico, and Morocco", Quarterly Journal of Economics, 113 (3), 903-47.

Coase, Ronald H. (1937), "The Nature of the Firm," Economica, 4:16, pp. 386-405.

Das, M., M. Roberts and J. Tybout (2001), "Market Entry Costs, Producer Heterogeneity and Export Dynamics," NBER Working Paper 8629.

Dixit, A. (1989a), "Entry and Exit Decision under Uncertainty," Journal of Political Economy, 97(3), 620-638.

Dixit, A. (1989b), "Hysteresis, Import Penetration, and Exchange Rate Pass-Through," Quarterly Journal of Economics, 104(2), 205-228.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz (2003), "An Anatomy of International Trade: Evidence from French Firms," mimeo, NYU.

Eaton, Jonathan, Samuel Kortum, and Francis Kramarz (2003), "Dissecting Trade: Firms, Industries, and Export Destinations," forthcoming American Economic Review Papers and Proceedings.

Ethier, Wilfred J. (1986), "The Multinational Firm," Quarterly Journal of Economics, 101:4, pp. 805-833.

Ethier, Wilfred J. and James R. Markusen (1996), "Multinational Firms, Technology Diffusion and Trade," Journal of International Economics, 41:1, pp. 1-28.

Feenstra, Robert C. (1998), "Integration of Trade and Disintegration of Production in the Global Economy," Journal of Economic Perspectives, 12:4, 31-50.

Feenstra, Robert C. and Gordon H. Hanson (1996), "Globalization, Outsourcing, and Wage Inequality," American Economic Review, 86:2, pp. 240-245.

Grossman, G.M. and Helpman, E. (2002), "Integration vs. Outsourcing in Industry Equilibrium," Quarterly Journal of Economics 117 (1), 85-120.

Grossman, G.M. and Helpman, E. (2003), "Outsourcing in a Global Economy" mimeo Harvard University.

Grossman, Sanford J., and Oliver D. Hart (1986), "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration," Journal of Political Economy, 94:4, pp. 691-719.

Hanson G., Mataloni R. and M. Slaughter (2001), "Expansion Strategies of U.S. Multinational Firms," in Dani Rodrik and Susan Collins, eds., Brookings Trade Forum 2001, pp. 245-282.

Hart, Oliver (1995), Firms, Contracts, and Financial Structure, Oxford: Clarendon Press. Chapters 1-3.

Haskel, Jonathan E. and Sonia Pereira and Matthew Slaughter (2002), "Does Inward Foreign Direct Investment Boost the Productivity of Domestic Firms?," NBER Working Paper # 8724, January 2002.

Helpman, Elhanan (1984), "A Simple Theory of International Trade with Multinational Corporations", Journal of Political Economy, 92:3, pp. 451-471.

Helpman, Elhanan and Paul R. Krugman (1985), Market Structure and Foreign Trade, Cambridge, MA: MIT Press. Chapter 12.

Helpman, Elhanan, Marc J. Melitz, and Stephen R.Yeaple (2003), "Exports versus FDI with Heterogeneous Firms," American Economic Review, forthcoming.

Holmström Bengt and Paul Milgrom (1994), "The Firm as an Incentive System,", American Economic Review, 84:4, pp. 972-991.

Holmstrom, Bengt R. and Jean Tirole (1989), "The Theory of the Firm," In Handbook of Industrial Organization, Vol. 1, R. Schmalansee and R. Willing(Eds.), Elsevier Science Pub. Co., Amsterdam.

Hummels, David, Jun Ishii, and Kei-Mu Yi (2001), "The Nature and Growth of Vertical Specialization in World Trade," Journal of International Economics, 54:1, pp. 75-96.

Keller, Wolfgang, and Stephen Yeaple (2003), "Multinational Enterprises, International Trade, and Productivity Growth: Firm Level Evidence from the United States," mimeo UPenn.

Krugman, P. (1980), "Scale economies, product differentiation and the pattern of trade," American Economic Review, p. 950-959.

Marin, Dalia and Thierry Verdier (2003), "Globalization and the Empowerment of Talent," DELTA Mimeo 2003.

Markusen J. and A. Venables (1998): "Multinational Firms and the New Trade Theory", Journal of International Economics, 46(2), 183-203.

Markusen J.R. and A.J. Venables, (1999): "Foreign Direct Investment as a Catalyst for Industrial Development", European Economic Review, 43, 335-356.

Markusen, James R. (1984), "Multinationals, Multi-Plant Economies, and the Gains from Trade," Journal of International Economics, 16, pp. 205-226.

Markusen, James R. (1995), "The Boundaries of Multinational Enterprises and the Theory of International Trade," Journal of Economic Perspectives, 9:2, pp. 169-189.

Markusen, James R. (2002), Multinational Firms and the Theory of International Trade, Cambridge, MA: MIT Press.

Markusen, James R. and Anthony J. Venables (2000), "The Theory of Endowment, Intra-industry and Multi-national Trade," Journal of International Economics, 52, pp. 209-234.

McLaren, John (2000), "Globalization and Vertical Structure," American Economic Review, 90:5, pp. 1239-1254.

Melitz, Marc J. (2003), "The Impact of Trade on Intra-Industry Reallocations and Aggregate Industry Productivity," Econometrica, 71:6, pp. 1695-1725.

Melitz, Marc, and Gianmarco Ottaviano (2003), "Market Size, Trade and Productivity", working paper, Harvard University.

Pavcnik, N. (2002), "Trade Liberalization, Exit, and Productivity Improvements: Evidence from Chilean Plants", The Review of Economic Studies 69, January, pp. 245-76.

Puga, Diego, and Trefler, Daniel. "Knowledge Creation and Control in Organizations." Working Paper no. 9121. Cambridge: Mass,: NBER, August 2002.

Roberts M. and J. Tybout (1997), "The Decision to Export in Colombia: An Empirical Model of Entry with Sunk Costs," American Economic Review, 87(4), 545-564.

Rosen, Sherwin (1982), "Authority, Control, and the Distribution of Earnings," The Bell Journal of Economics, 13:2, pp. 311-323.

Tirole, Jean (1988), The Theory of Industrial Organization, Cambridge: MIT Press. Chapter 1.

Tybout, James (2001), "Plant- and Firm-level Evidence on the 'New' Trade Theories" (in E. Kwan Choi and James Harrigan, ed., Handbook of International Trade, Oxford: Basil-Blackwell, 2003, and NBER Working Paper No. 8418).

Williamson, Oliver E. (1985), The Economic Institutions of Capitalism, Free Press. Chapters 1-3.

Yeaple, Stephen (2003), "The Complex Integration Strategies of Multinationals and Cross Country Dependencies in the Structure of FDI," Journal of International Economics, 60, pp. 293-314.

Yeaple, Stephen (2003), "The Role of Skill Endowments in the Structure of U.S. Outward FDI," Review of Economics and Statistics, August, 85(3), pp. 726-734.