

Lecture 3:

Firm Heterogeneity in the Krugman Model

Instructor: THOMAS CHANEY
Econ 357 - **International Trade (Ph.D.)**

1 Introduction

In this lecture, we will start breaking up the original model of trade, and add new assumptions. For that lecture, we start from the Krugman (1980) model, and add firm level heterogeneity in terms of productivity. As we'll see when we look at the data, a major stylized fact is that in any industry, firms widely differ in terms of size and productivity. A corollary to that observation is that exporters are very different from the typical firm. They tend to be much larger, more productive, more capital intensive... We will try to integrate those features of the data in a simple model of international trade.

Melitz (2003) derives a simple model of industry equilibrium in an open economy with heterogeneous firms. For the trade part, it builds on the Krugman (1980) model; for the dynamic industry equilibrium part, it builds on Hopenhayn (1992). One of the predictions of the Melitz model is that opening up to trade will increase aggregate productivity. Chaney (2005) derives a simpler solution from a (simpler) model with heterogeneous firms, and gets predictions for the impact of trade barriers on trade flows in the presence of firm heterogeneity. First, trade barriers will have a larger impact when firms are heterogeneous. Second, the elasticity of trade flows with respect to trade barriers will not simply be the elasticity of substitution between goods as in Krugman (1980), it will actually be inversely related to this elasticity.

2 Melitz (2003)

Melitz adds firm level heterogeneity in productivity to the classical framework of Krugman (1980).

Firm heterogeneity assumptions:

- Firms differ in terms of their marginal productivity of labor (the only factor of production as in Krugman).
- The productivity of each firm is randomly drawn from some distribution, and firms do not know their productivity prior to starting production.

Additional assumption on trade barriers:

- In order to sell part of its output abroad, a firm has to pay not only a variable cost (iceberg type as in Krugman), but in addition, it has to pay a fixed cost.

These two assumptions give rise to a series of new features. More productive firms earn larger profits. Because of the presence of a fixed cost of entering foreign markets, only a subset of firms, those that are sufficiently productive, will be able to export. These firms benefit from the opening to trade (through an increased demand for their goods), whereas the other firms suffer from the additional competition from foreigners. The factors of production will therefore be partially reallocated towards the most productive firms, inducing an increase in aggregate productivity.

Let's dive into the formal model.

Autarky Equilibrium

Preferences and demand:

Many features are preserved from Krugman. Preferences are isoelastic, so that consumers solve the following maximization problem,

$$\begin{aligned} \max_{q(\omega)} U &\equiv \left(\int_{\Omega} q(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right)^{\frac{\sigma}{\sigma-1}} & (1) \\ \text{s.t.} \quad & \int_{\Omega} p(\omega) q(\omega) d\omega = R \end{aligned}$$

where R is aggregate expenditure, and Ω is the mass of available goods. This gives rise to the following iso-elastic demand functions,

$$q(\omega) = \left(\frac{p(\omega)}{P} \right)^{-\sigma} \frac{R}{P} \tag{2}$$

$$\text{with } P = \left(\int_{\Omega} p(\omega)^{1-\sigma} d\omega \right)^{\frac{1}{1-\sigma}} \tag{3}$$

Production and pricing:

As in Krugman, production takes place under increasing returns to scale. The difference is that *the marginal productivity of labor differs across firms*. Each firm draws a random productivity shock φ from some distribution G over \mathbb{R}^+ . Labor needed to produce q units for a firm with labor productivity φ is then,

$$l(q, \varphi) = f + \frac{q}{\varphi} \quad (4)$$

The fixed overhead cost f is paid each period. With iso-elastic demand functions, monopolist firms charge a constant mark-up over marginal cost. If we normalize the wage to 1, a firm with labor productivity φ will set a price,

$$p(\varphi) = \frac{\sigma}{(\sigma - 1)\varphi} \quad (5)$$

sells $r(\varphi) = p(\varphi)q(\varphi) = \left(\frac{\sigma-1}{\sigma}\varphi P\right)^{\sigma-1} R$ worth of output, and earns net profits $\pi(\varphi) = \left(\frac{\sigma-1}{\sigma}\varphi P\right)^{\sigma-1} \frac{R}{\sigma} - f$.

Since all firms with the same productivity φ will set the same price, sell the same quantities, and earn the same profits, we can relabel things, and look at the distribution of φ 's instead of the distribution of ω 's.

Convenient aggregation:

An equilibrium will be defined by a total mass of firms, M , and a distribution of productivities $\mu(\varphi)$ over \mathbb{R}^+ . It will be useful to introduce a special measure of aggregate productivity,

$$\tilde{\varphi} = \left(\int_0^\infty \varphi^{\sigma-1} \mu(\varphi) d\varphi \right)^{\frac{1}{\sigma-1}} \quad (6)$$

With that notation, we get the simple forms for the aggregate variables of the model,

$$P = M^{\frac{1}{1-\sigma}} p(\tilde{\varphi}) \quad (7)$$

$$R = Mr(\tilde{\varphi}) \quad (8)$$

$$\Pi = M\pi(\tilde{\varphi}) \quad (9)$$

Firms' entry and exit:

We now need to make a few additional assumptions about the model. Each period, some firms are started. To start production, an entrepreneur has to pay a fixed entry cost f^E . Once this cost is paid, the entrepreneur receives a productivity draw φ from a distribution G over \mathbb{R}^+ .

Each period, a firm may decide to stop production. If it doesn't, it may still die for some exogenous reasons. For simplicity, we assume that each period, surviving firms are hit by a random Poisson death shock that comes with probability δ . So each period, a fraction δ of all firms (and the same fraction whatever the productivity) dies. We will only consider stationary equilibria of this game.

Endogenous exit:

Once it has entered and received a productivity draw, a firm will stay in business only if it earns positive net profits each period. Since more productive firms earn higher profits, we can define a productivity cutoff of survival, $\bar{\varphi}$. Any firm with a productivity below $\bar{\varphi}$ immediately exits and never produces anything, and any firm with a productivity above $\bar{\varphi}$ stays in business until it exogenously dies. The productivity cutoff is defined by,

$$\pi(\bar{\varphi}) = 0 \tag{10}$$

In equilibrium, all firms above $\bar{\varphi}$ produce, and the distribution of productivity is given by,

$$\mu(\varphi) = \begin{cases} \frac{g(\varphi)}{1-G(\bar{\varphi})} & \text{if } \varphi \geq \bar{\varphi} \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

This defines the aggregate productivity as a function of the threshold $\bar{\varphi}$,

$$\tilde{\varphi}(\bar{\varphi}) = \left(\frac{1}{1-G(\bar{\varphi})} \int_{\bar{\varphi}}^{\infty} \varphi^{\sigma-1} g(\varphi) d\varphi \right)^{\frac{1}{\sigma-1}}$$

The average profit among surviving firms, that is the expected profits conditional on being about the cutoff $\bar{\varphi}$, is simply $\bar{\pi} = \pi(\tilde{\varphi}(\bar{\varphi}))$. We can rewrite the condition for zero profits, which Melitz nicely calls the "zero cutoff profits" (*ZCP*) condition, as,

$$\pi(\bar{\varphi}) = 0 \Leftrightarrow \bar{\pi} = f \left[\left(\frac{\tilde{\varphi}(\bar{\varphi})}{\bar{\varphi}} \right)^{\sigma-1} - 1 \right] \tag{ZCP}$$

Endogenous entry:

Prior to entry, potential entrants contemplate the expected profits they would generate if they enter (either nothing if the entrepreneur is not lucky enough to get $\varphi > \bar{\varphi}$, discounted [the Poisson death probability is the discount factor] sum of profits if it is sufficiently productive), and compare it to the cost of entry, f^E . In a stationary equilibrium, in each period, as long as it is alive, a firm earns the same profits each period.

Ex ante, the expected profits, conditional on being in business, is given by $\bar{\pi}$. The cost of entry is f^E . The net value of entering today ($t = 0$), given the constant probability of dying each period, is then,

$$v^E = E \left[\sum_{t=0}^{\infty} (1 - \delta)^t \pi(\varphi) - f^E \right] = \frac{1 - G(\bar{\varphi})}{\delta} \bar{\pi} - f^E \quad (12)$$

Free entry ensures that firms will enter until the net value of entering is driven down to zero. As in Krugman, as more firms enter, the market shares left shrink (some of those entrants are lucky enough to profitably produce, they get some market shares), until the expected sum of future profits equals the cost of entry. So we can state the free entry condition as follows,

$$v^E \leq 0 \quad (FE)$$

At each point in time, there cannot be a positive value of entering (otherwise, firms more firms would enter). Since we'll only look at stationary equilibria, we will only consider cases where this condition holds with equality. Note however that along some non stationary equilibria, or during the transition towards a new stationary equilibrium if the system is perturbed, we may observe some periods of time during which $v^E < 0$, and no firm enters.

General equilibrium:

The equilibrium of this economy is then simply given by two conditions, the zero cutoff profits condition, and the free entry condition, as well as the labor market clearing condition. Labor market clearing condition imposes that total expenditure on differentiated goods (expenditure by consumers) must equal the total revenue of consumers, L . We know that total expenditure is a function of the mass of firms, M , and the productivity cutoff $\bar{\varphi}$, $R = Mr(\bar{\varphi}) = M\sigma(\bar{\pi} + f)$.

$$\left\{ \begin{array}{ll} \pi(\bar{\varphi}) = 0 & \text{(Zero Cutoff Profits)} \\ v^E = 0 & \text{(Free Entry)} \\ R = L & \text{(Labor Market Clearing)} \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{ll} \bar{\pi} = f \left[\left(\frac{\bar{\varphi}(\bar{\varphi})}{\bar{\varphi}} \right)^{\sigma-1} - 1 \right] & (ZCP) \\ \bar{\pi} = \frac{\delta f^E}{1 - G(\bar{\varphi})} & (FE) \\ M = \frac{L}{\sigma(\bar{\pi} + f)} & (LMC) \end{array} \right.$$

The first two conditions give us a solution for $\bar{\pi}$ and $\bar{\varphi}$, which we plug into the last condition to get the total number of firms in the economy. From that, we can recover total output, aggregate prices, and the production of any individual firm. Melitz goes into the precise proof of existence and uniqueness of the equilibrium. For a generic distribution $G(\cdot)$, there need not be a unique solution. However, there will be a

unique solution if $\frac{\varphi g(\varphi)}{1-G(\varphi)}$ is increasing (which is satisfied by most common distributions).

Ownership structure:

I must make an important comment here: because we are in a stationary equilibrium, the ownership structure of the economy is irrelevant. More precisely. In each period, a fraction δ of all firms dies from an exogenous Poisson death shock (Hopenhayn derives a more plausible model where the probability of death varies across firms, so that the stationary distribution of productivity is not identical to the distribution at birth, as it is here in Melitz). The number of firms that die each period is $\frac{\delta M}{1-G(\bar{\varphi})}$. These firms will be replaced by new entrants each period. So in the stationary equilibrium, there will be $M^E = \frac{\delta M}{1-G(\bar{\varphi})}$ new entrants each period. Some labor has to be used to build those new firms, i.e. pay the sunk entry cost. The total cost of creating those new firms each period is $f^E M^E$. This investment must be somehow paid for by someone. It turns out that in the stationary equilibrium, the net profits earned by all the surviving firms exactly cover the fixed investment in each period: $f^E M^E = \Pi$.

One way to define the ownership structure in this economy is the following. All the workers in the economy own a share of a mutual fund. This mutual is made of all the firms in the economy. It redistributes dividends to the workers. The dividends of this mutual fund is the sum of all profits, minus what the mutual fund spends in investing in new firms. As long as it is profitable to create new firms, the mutual firm will provide some finance to entrepreneurs to create new firms, and that gives the free entry condition. In the stationary equilibrium, this mutual fund exactly breaks even: the profits that it collects from the firms it owns cover exactly the investment in new firms. And the expected net present value of these investments is exactly equal to the cost of these investments (the free entry condition), so that this mutual fund is not doing anything crazy with the workers' money.

Trade Equilibrium

We can now open trade between two identical economies. The symmetry assumption will simplify our life greatly, but it is not crucial. We'll see in the Chaney paper how a few simple assumptions allow us to actually get closed form solutions even in the case of many asymmetric countries.

As stated in the introduction, we add an extra assumption regarding trade barriers to the simple Krugman iceberg transportation cost framework. Namely, we assume that on top of the iceberg trade cost, firms

have to pay a fixed cost if they are to enter the foreign market. As for variable costs, and as in Krugman, a fraction $(\tau - 1)$ is lost in transit. For fixed costs, a firm has to pay each period a fixed cost (denominated in labor, with wages normalized to 1), f^X . Note that because we are looking at a steady state, it doesn't matter whether a firm has to pay a fixed cost once and for all, or a fixed cost each period it is operating (exporting). The equivalent once and for all fixed cost would be $F^X = \sum_{t=0}^{\infty} (1 - \delta)^t f^X = f^X / \delta$. Also note that a firm decides whether or not to enter the export market (and pay the fixed cost f^X) after its productivity shock is realized.

Export status:

More productive firms, if they do export, are able to charge a lower price, capture a larger market share, and generate larger profits. Depending on the assumption for the distribution of productivity shocks, the overhead cost of production, f , and the fixed entry cost into the export market, f^X , possibly a fraction of surviving firms only will export. A firm exports only if the net profits it generates from exporting are positive. As in Krugman, a firm will charge a price in the foreign market that is equal to a constant mark-up over its marginal cost of selling one unit in the foreign market. This price is just τ times the price charged for the domestic market: $p^x = \tau p$. We can then easily derive the net profits generated from exporting, $\pi^X(\varphi) = \left(\frac{\sigma-1}{\sigma} \frac{p}{\tau} P\right)^{\sigma-1} \frac{R}{\sigma} - f^X$. All firms with a productivity φ such that $\pi^X(\varphi) \geq 0$ export. This way, we can define a productivity cutoff for exporting, $\bar{\varphi}^X$. For a given firm with productivity φ , there is a simple relationship between sales/profits generated on the domestic market, and sales/profits generated in the foreign market. We use that to rewrite the zero profit cutoff condition for exports,

$$\pi^X(\bar{\varphi}^X) = 0 \Leftrightarrow \bar{\varphi}^X = \tau \left(\frac{f^X}{f}\right)^{\frac{1}{\sigma-1}} \bar{\varphi} \quad (ZCP^X)$$

We impose the simple restriction that $\tau \left(\frac{f^X}{f}\right)^{\frac{1}{\sigma-1}} > 1$ so that $\bar{\varphi}^X > \bar{\varphi}$ and not all firms export, which is consistent with the empirical evidence.

Open economy equilibrium:

We can now define all the conditions for the general equilibrium in the open economy case. The average profits that a firm earns, conditional on surviving, are the sum of profits earned domestically (π^D), and profits earned from exporting (π^X): $\bar{\pi} = \pi^D(\bar{\varphi}) + prob^X \pi^X(\bar{\varphi}(\bar{\varphi}^X))$, where $prob^X = \frac{1-G(\bar{\varphi}^X)}{1-G(\bar{\varphi})}$ is the probability of exporting, conditional on survival.

The stationary assumption implies that the labor market clearing condition is the same in the trade steady state as in the autarky steady state: $(LMC) \Leftrightarrow R = L$. The trade steady state is defined by the two zero cutoff profits conditions (which define $\bar{\varphi}$ and $\bar{\varphi}^X$), the free entry condition, and the labor market clearing condition,

$$\begin{aligned}\bar{\pi} &= \pi^D(\bar{\varphi}) + prob^X \pi^X(\bar{\varphi}^X) \\ \text{where } prob^X &= \frac{1 - G(\bar{\varphi}^X)}{1 - G(\bar{\varphi})}\end{aligned}$$

$$\begin{cases} \bar{\varphi}^X = \tau \left(\frac{f^X}{f} \right)^{\frac{1}{\sigma-1}} \bar{\varphi} & (ZCP^X) \\ \bar{\pi} = f \left(\left[\frac{\bar{\varphi}(\bar{\varphi})}{\bar{\varphi}} \right]^{\sigma-1} - 1 \right) + prob^X f^X \left(\left[\frac{\bar{\varphi}(\bar{\varphi}^X)}{\bar{\varphi}} \right]^{\sigma-1} - 1 \right) & (ZCP) \\ \bar{\pi} = \frac{\delta f^E}{1 - G(\bar{\varphi})} & (FE) \\ M = \frac{L}{\sigma(\bar{\pi} + f + prob^X f^X)} & (LMC) \end{cases}$$

Equilibrium properties, aggregate productivity and firm inequality:

1. The main finding put forward by Melitz is that trade opening will increase aggregate productivity in all trading economies. Formally, we can easily show that the productivity cutoff for survival $\bar{\varphi}$ goes up, so that the least productive firms disappear. At the same time, the most productive firms among the survivors (those with a productivity above $\bar{\varphi}^X$), export to the foreign market on top of selling for their domestic market, and therefore these firms employ disproportionately more labor than the less productive firms. Hence, the aggregate productivity in the economy is the average productivity of a better pool, with a larger weight on the most productive firms. This unambiguously leads to an increase in the aggregate productivity of the economy.

The reason for this increase in aggregate productivity is quite subtle though. Two forces are present. First, domestic firms now have to face the additional competition from the best foreign firms that export. This reduces the market share left for domestic firms, and drives down the profits of all firms. Because of the constant elasticity assumption, gross profits go down proportionally for all firms. This forces the least productive firms out of the market.

There is then a second channel that Melitz emphasizes. When the possibility of trade (at some cost) is opened up, there are additional profits to be expected by the most productive firms, those

firms that are productive enough to enter the foreign market. Two things. First, existing high productivity firms want to expand their scale of production in order to service the foreign market, and therefore they want to hire more workers. Second, new firms enter, attracted by the prospect of these higher profits. Those new firms also want to hire workers (both to cover the fixed cost and to produce). Those two things drive up the real wages, and forces the least productive firms to shut down. Melitz argues that only the second channel matters for the increase in aggregate productivity. It is actually not exactly correct¹. Indeed, a combination of each channel contributes to the increase in aggregate productivity.

2. The second important finding is that trade induces some reallocation of both market shares and profits among firms. More precisely, exporters' size increases, and non exporters' size shrinks. As for the net profits of firms, there is an increase in inequalities between firms: the least productive firms lose profits, and the most productive firms increase their profits.

Let's define by $r^A(\varphi)$ the sales of a firm with productivity φ in autarky, $r^D(\varphi)$ its sales for the domestic market in the trade equilibrium, and $r^X(\varphi)$ its sales to the export market in the trade equilibrium. The reallocation of market shares between firms (from non exporters towards exporters) is stated in the following set of inequalities,

$$r^D(\varphi) < r^A(\varphi) < r^X(\varphi), \quad \forall \varphi \geq \bar{\varphi} \quad (13)$$

The proof of these inequalities is not straightforward, and is presented in appendix E of Melitz's paper (the first inequality is easy to prove, the second one less so).

The corollary of this reallocation of market shares between firms is that the dispersion of profits will increase as well. Because of the entry of new foreign competitors, all firms lose some market shares domestically, and therefore their domestic profits shrink. However, for those firms that are able to export ($\varphi > \bar{\varphi}^X$), there is now an additional source of profits. There is a range of firms for which those profits are not enough to compensate the loss of "domestic"

¹If one were to solve for the transitional dynamics (as done in Chaney's "Productivity Overshooting" (2005), one would see that the initial impact of trade opening does drive up the aggregate productivity in the economy only because of the first channel: the entry of some foreign competitors forces the least productive firms out. In the long run however, the two channels are at play, so that the long run increase in aggregate productivity results from a combination of the two channels.

profits. There is however a productivity φ^\dagger above which the additional profits from exporting more than compensate for the loss of domestic profits. So some firms lose from opening to trade, and some firms gain. Maybe more important, the inequality in profits between firms increases when trade is opened up.

Export and productivity:

In the theoretical Melitz model, there is a simple and well defined concept of productivity for each firm, φ . It corresponds to the marginal productivity of labor in each firm. However, it is not obvious how one would match this theoretical concept of productivity to an actual empirical measure of productivity. Typically, when looking at actual firms, one computes the output per worker, and uses this as a proxy for labor productivity. Because of the CES preferences and monopolistic competition assumption, if one is to compute output per worker without taking into account the overhead costs of production (and the fixed costs of entering foreign markets), one would get a constant output per worker: $\frac{r(\varphi)}{q(\varphi)/\varphi} = \frac{\sigma}{\sigma-1}w$, $\forall \varphi$. This is because more productive firms (higher φ) produce larger quantities, sell at a lower price, and hire more workers, all the way until output (net of fixed costs) is equal to that of any other firm.

There are however fixed costs of production in this set-up, so that $\frac{r(\varphi)}{q(\varphi)/\varphi+f}$ is indeed increasing with φ . Since only the more productive firms export, if one were to measure productivity using only domestic operations, one would find that the measured productivity of exporters is indeed higher than the measured productivity of non exporters.

Unfortunately, this argument does not carry through once export operations are taken into account. One can easily prove that output per worker will be higher for exporters than it is for non exporters,

$$\frac{r^D(\varphi)}{q^D(\varphi)/\varphi+f} > \frac{r^D(\varphi) + r^X(\varphi)}{q^D(\varphi)/\varphi+f + q^X(\varphi)/\varphi+f^X} \quad (14)$$

if and only if $\tau \left(\frac{f^X}{f} \right)^{\frac{1}{\sigma-1}} > 1$

so that if a firm were to start exporting without changing its intrinsic productivity (nor what other firms are doing in the economy), its output per worker would decrease.

Measured productivity, $\frac{\text{output}}{\text{worker}}$ goes up with the conceptual measure of productivity φ , whether a firm exports or not. But an exporter that is just above the export cutoff would have a lower measured productivity than a non-exporter that is just below the export cutoff. The prediction

for the difference in measured productivity for average exporters versus non exporters is ambiguous. A more elaborate measure of productivity, that accounts for the fixed cost of production, or more generally, accounts for other factors of production, may not be subject to this critique.

Granted that productivity can be precisely estimated, the Melitz model then predicts that in the data, exporting will be systematically correlated with higher observed productivity. This is not due to a direct causality that allows exporters to acquire a better technology. The causality goes in the opposite direction. It is because a firm is more productive that it is able to export.

3 Chaney (2005)

Chaney (2005) expands the work of Melitz. One important finding in the Melitz model is that if firms are heterogeneous in terms of productivity, and if there are fixed costs associated with exporting (or more generally, if there are increasing returns to scale in the cost of exporting), there will be an endogenous selection of firms into the export market. A corollary to this statement is that as trade barriers move around (either fixed or variable cost of exporting), the set of firms that are able to overcome trade barriers is going to change. So changes in trade barriers, a trade liberalization among others, will not only change how much each firm exports, but it will also change the set of firms that exchange. There is both an intensive and an extensive margin of adjustment of trade flows to trade barriers.

This is unlike in the Krugman model. Krugman does find that when a country moves from autarky to some limited trade (limited by some trade barriers τ), only the extensive margin of trade varies. This is not an extremely informative statement. In a sense, one could say that when there is no trade, everything is as if trade barriers were infinite ($\tau = +\infty$), and domestic consumers do consume all foreign varieties, but zero quantity of each. More interestingly, in the Krugman model, when trade barriers move around (when τ moves around), the set of goods that consumers have access to does not change at all, they just consume more of the goods that become relatively cheaper. So in Krugman, there is only one margin of adjustment, the intensive margin: all firms always export, but they export more or less.

In the setting developed by Melitz, there will be an extra margin of adjustment. I develop a simplified version of the Melitz model (basically, I get rid of the free entry condition, and I remove the assumption regarding the dynamic part of the model, since we'll only be looking at steady state equilibria), and at the same time I add different sectors,

and different countries, possibly asymmetric.

The main findings of the paper are the following:

1. Once the extensive margin of trade is taken into account, trade barriers will have a larger impact on trade flows. When trade barriers go down (τ goes down), existing exporters, because they face a lower cost of exporting, are able to charge lower prices, and capture a larger market share. They increase their exports. This is the intensive margin of trade, exactly the same as in Krugman.

At the same time, it is now more profitable to export for any potential exporter. Some firms that were not productive enough, and therefore could not export, are now able to enter the export market. These firms export strictly positive quantities (and values), and they also contribute to increasing the aggregate volume of exports. This is the extensive margin of trade, that did not exist in Krugman.

So in total, since there is this additional margin of adjustment, trade will be more sensitive to trade barriers than in the Krugman model.

2. In models of trade with a representative firm (as in Anderson and van Wincoop), or with identical firms (as in Krugman), the elasticity of trade flows with respect to trade barriers will be high in sectors with a high elasticity of substitution. I predict that if firms are heterogeneous, the opposite is true: the elasticity of trade flows with respect to trade barriers will be higher in sectors with a **low** elasticity of substitution.

The reversal of the prediction comes precisely from the introduction of the extensive margin of trade. In sectors with a high elasticity of substitution, less productive firms are only able to capture a very small market share. When these less productive firms start exporting (following a reduction in trade barriers), they capture only small market shares, and do not increase aggregate exports a lot. In sectors with a low elasticity of substitution on the other hand, less productive firms still are able to capture a relatively large market share. When these firms start exporting, they capture a relatively large market share, and increase aggregate exports a lot.

Specific assumptions added to the Melitz model:

- There are potentially many (more than 2) countries, potentially asymmetric.
- There is no free entry, but the number of entrepreneurs getting a productivity draw is proportional to the size of the country.
- Productivity shocks are drawn from a Pareto distribution (power law distribution), which allows to solve all expressions in closed form.
- There are two sectors, as in the Helpman-Krugman extension of Krugman (1980) that we saw. Specifically, one sector is a homogenous good that is produced under constant returns to scale technology and is freely traded, and the other corresponds to a continuum of differentiated varieties, that are subject to both variable and fixed costs.

Set-up

Preferences:

I use the Melitz model, and add some minor modifications. First, there will be an additional sector, providing a homogenous good (i.e. non differentiated). This good, used as the numeraire, is produced under constant returns to scale, with one unit of output per worker, and is freely traded. Provided that this good is produced in every country, which can be insured as long as μ large enough, the wage will be constant and equal to 1 in every country. Consumers maximize,

$$U \equiv q_o^\mu \left(\int_{\Omega} q(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right)^{\frac{\sigma}{\sigma-1}(1-\mu)} \quad (15)$$

There are N potentially asymmetric countries, each with a size L_i . Trade barriers between countries are potentially asymmetric: τ_{ij} and f_{ij} are respectively the variable (iceberg) and fixed (in labor units) cost for a firm from country i selling in country j .

Pareto distributions:

The distribution of productivity is drawn from a Pareto distribution G with scaling parameter γ over $[1, +\infty)$,

$$P(\varphi > \Phi) = \Phi^{-\gamma} \quad (16)$$

with $\gamma > \sigma - 1$, $\gamma > 2$.

In each country n , there are L_n entrepreneurs who get a draw from this distribution and decide whether or not to enter. They don't have to pay a sunk cost to get access to this lottery as in Melitz.

The Pareto distribution was chosen essentially for two reasons:

- First, Pareto distributions have nice analytical properties. These distributions are stable to truncation from below: $P(\varphi > \Phi | \varphi > \bar{\varphi}) = \left(\frac{\Phi}{\bar{\varphi}}\right)^{-\gamma} = \frac{P(\varphi > \Phi)}{\bar{\varphi}^{-\gamma}}$. This property ensures that when we look at the endogenous selection into the export market, the subset of exporters is Pareto distributed with the same parameter as the distribution of domestic firms. Also useful is the fact that the density of a Pareto is a power function, i.e. the same type of function as the constant elasticity demand functions, which allows simple aggregation of demand across differentiated goods.
- Second, Pareto distributions, bundled with CES preferences and monopolistic competition, give predictions for the distribution of firm sizes (observable)

Ownership structure:

All the firms belong to a global mutual fund that collects all the profits from all the firms in the world. All the workers in the world own a single share of that fund. The fund redistributes those global profits (Π) in units of the numeraire good to the shareholders. So the aggregate income in country i is $(1 + \frac{\Pi}{L})L_i$ where L is the total world population.

Optimal pricing, demand and selection in markets:

Upon receiving its productivity draw, a firm decides which market to enter, and what price to set on each market so as to maximize profits. We know from the constant elasticity preferences and the monopolistic structure of the economy that if they are to enter a market, firms will charge a constant mark-up over marginal cost. So that total sales that a firm from country i with productivity φ would generate from selling in j is,

$$r_{ij}(\varphi) = \left(\frac{\sigma \tau_{ij}}{(\sigma - 1) \varphi P_j} \right)^{1-\sigma} \times \mu \left(1 + \frac{\Pi}{L} \right) L_j \quad (17)$$

A firm from country i will decide to enter the market in country j if and only if its profits there are positive. We can therefore define a zero cutoff profits condition for all pairs of countries. In country i , only the firms with a productivity above the threshold $\bar{\varphi}_{ij}$ will enter j 's market,

$$\pi_{ij}(\bar{\varphi}_{ij}) = 0 \Leftrightarrow \bar{\varphi}_{ij} = \lambda_1 f_{ij}^{\frac{1}{\sigma-1}} (P_j^{\sigma-1} L_j)^{\frac{-1}{\sigma-1}} \tau_{ij} \quad (ZCP_{ij})$$

with λ_1 a constant ².

General equilibrium

We can now determine the general equilibrium of this world equilibrium. To do so, we have to solve a fixed point problem. We know that all the firms from any country i with a productivity above $\bar{\varphi}_{ij}$ will enter into j 's market. The thresholds $\bar{\varphi}_{ij}$'s for the different countries i only depend on the price index in country j . This price index in turn depends on which firms enter into country j , that is what are the different thresholds $\bar{\varphi}_{ij}$'s. Thanks to the assumption we've made, because the wages are exogenously determined, and because the number of potential entrants in each country is also exogenously given, we do not have to solve simultaneously for the different price indices in every country, we can solve separately for the price index in each country (and the corresponding thresholds $\bar{\varphi}_{ij}$'s for exporting into j).

Equilibrium price indices:

With the assumption of Pareto distributed productivity shocks, we can get very simple closed form solutions for all variables,

$$\begin{aligned} P_j^{1-\sigma} &= \sum_{k=1}^N L_k \int_{\bar{\varphi}_{kj}}^{\infty} p_{kj}(\varphi)^{1-\sigma} dG(\varphi) \\ &= \sum_{k=1}^N L_k \int_{\bar{\varphi}_{kj}}^{\infty} \left(\frac{\sigma-1}{\sigma} \times \frac{\varphi}{\tau_{kj}} \right)^{\sigma-1} dG(\varphi) \\ &= \sum_{k=1}^N L_k \frac{\gamma}{\gamma - (\sigma-1)} \left(\frac{\sigma-1}{\sigma} \times \frac{1}{\tau_{kj}} \right)^{\sigma-1} \bar{\varphi}_{kj}^{\sigma-1-\gamma} \end{aligned}$$

Plugging back in the expression for the productivity thresholds, and rearranging, we get the equilibrium solution for the P_j 's,

$$P_j = \lambda_2 \times \left(\frac{L_j}{L} \right)^{\frac{1}{\gamma}} \times (L_j)^{-\frac{1}{\sigma-1}} \times \theta_j \quad (18)$$

$$\text{with } \theta_j^{-\gamma} \equiv \sum_{k=1}^N s_k \times \tau_{kj}^{-\gamma} \times f_{kj}^{-\left(\frac{\gamma}{\sigma-1}-1\right)} \quad (19)$$

² $\lambda_1 = \left(\frac{\sigma}{\mu} \right)^{\frac{1}{\sigma-1}} \left(\frac{\sigma}{\sigma-1} \right) \left(1 + \frac{\Pi}{L} \right)$. Note that there is a slight abuse of notations, as total world profits will be endogenously determined in equilibrium. However, firms as well as consumers take total world profits as a constant. I will solve for total profits in equilibrium.

λ_2 being a constant³. θ_j is an aggregate index of j 's remoteness from the rest of the world, which takes into account both the impact of variable and of fixed costs⁴.

Equilibrium exports and selection:

Having solved for equilibrium prices, we can now determine how much each firm is selling to each market,

$$r_{ij}(\varphi) = \begin{cases} \lambda_3 \times \left(\frac{L_j}{L}\right)^{\frac{\sigma-1}{\gamma}} \times \left(\frac{\theta_j}{\tau_{ij}}\right)^{\sigma-1} \times \varphi^{\sigma-1}, & \text{if } \varphi \geq \bar{\varphi}_{ij} \\ 0 & \text{, otherwise} \end{cases} \quad (20)$$

$$\bar{\varphi}_{ij} = \lambda_4 \times \left(\frac{L}{L_j}\right)^{\frac{1}{\gamma}} \times \left(\frac{\tau_{ij}}{\theta_j}\right) \times f_{ij}^{\frac{1}{\sigma-1}} \quad (21)$$

$$\Pi = \lambda_5 \times L \quad (22)$$

with λ_3 , λ_4 and λ_5 constants⁵. They are functions of fundamentals only: the size L_j , the trade barriers f_{ij} and τ_{ij} , and the measure of j 's remoteness from the rest of the world, θ_j . The behavior of each individual firm is identical to the behavior of firms in the Krugman model, but for the fact that only a subset of firms sells to each market. Among other, conditional on exporting to country j , the elasticity of a firm's export with respect to the variable trade barrier τ_{ij} is just given by the elasticity of substitution, $(\sigma - 1)$. If goods are very substitutable (σ high), a firm's exports will be very sensitive to changes in trade barriers.

Aggregate trade:

However, unlike in the Krugman model, there is endogenous selection into the export market, and not all firms do export to every country. As trade barriers move around, the set of exporters will change, and this additional margin of adjustment, the extensive margin, will radically change the behavior of aggregate trade. Total exports from i to j are,

$$X_{ij} = \lambda \times \frac{L_i L_j}{L} \times \left(\frac{\tau_{ij}}{\theta_j}\right)^{-\gamma} \times f_{ij}^{-\left(\frac{\gamma}{\sigma-1}-1\right)} \quad (23)$$

³ $\lambda_2 = \left(\frac{\gamma-(\sigma-1)}{\gamma}\right)^{1/\gamma} \left(\frac{\sigma}{\mu}\right)^{1/(\sigma-1)-1/\gamma} \left(\frac{\sigma}{\sigma-1}\right) \left(1 + \frac{\Pi}{L}\right)^{\frac{1}{\gamma}-\frac{1}{\sigma-1}}$. As for the definition of λ_1 , there is an abuse of notation in the sense that λ_2 depends on the equilibrium variable Π .

⁴ A simple way to interpret this aggregate index is to look at a symmetrical case: when $\tau_{kj} = \tau_j$ and $f_{ij} = f_j$ for all k 's, $\theta_j = f_j^{\frac{1}{\sigma-1}-\frac{1}{\gamma}} \times \tau_j$. In asymmetric cases, θ_j is a weighted average of bilateral trade barriers.

⁵ $\lambda_3 = \sigma \lambda_4^{1-\sigma}$, $\lambda_4 = \left(\frac{\sigma}{\mu} \times \frac{\gamma}{\gamma-(\sigma-1)} \times \frac{1}{1+\lambda_5}\right)^{\frac{1}{\gamma}}$, and $\lambda_5 = \frac{\mu(\sigma-1)}{\gamma\sigma-\mu(\sigma-1)}$.

with λ a constant⁶. This is to be compared to the predictions of an "equivalent" Krugman model, which would be,

$$\tilde{X}_{ij} = \tilde{\lambda} \times \frac{L_i L_j}{L} \times \left(\frac{\tau_{ij}}{\tilde{\theta}_j} \right)^{-(\sigma-1)} \quad (24)$$

Main differences between Krugman and a model with heterogeneous firms:

1. The elasticity of trade flows with respect to trade (variable) barriers is higher in the model with heterogeneous firms: $\gamma > \sigma - 1$. This is due to the additional margin of adjustment, the extensive margin of trade.
2. The elasticity of trade flows with respect to variable trade barriers, γ , does not depend on the elasticity of substitution σ anymore.
3. The elasticity of trade flows with respect to fixed costs, $(\frac{\gamma}{\sigma-1} - 1)$, is *negatively* related to the elasticity of substitution σ .
4. Both the elasticity of trade flows with respect to fixed costs and variable costs are higher in sectors where firms' productivity is less dispersed (γ large).

Intensive versus extensive margins

To understand these properties, and the differences between this model and the Krugman model, we must look at the relative importance of the intensive and the extensive margins of trade. Given that we have solved for everything in closed form, we can formally derive expressions for each of the margins of adjustment. We can differentiate the expression for aggregate exports, and see the impact of each type of trade barriers (variable and fixed) on each margin (intensive and extensive),

$$dX_{ij} = \underbrace{\left(\int_{\bar{\varphi}_{ij}}^{\infty} \frac{\partial r_{ij}(\varphi)}{\partial \tau_{ij}} dG(\varphi) \right)}_{\text{Intensive margin}} d\tau_{ij} - \underbrace{\left(r(\bar{\varphi}_{ij}) g(\bar{\varphi}_{ij}) \times \frac{\partial \bar{\varphi}_{ij}}{\partial \tau_{ij}} \right)}_{\text{Extensive margin}} d\tau_{ij} \\ + \underbrace{\left(\int_{\bar{\varphi}_{ij}}^{\infty} \frac{\partial r_{ij}(\varphi)}{\partial f_{ij}} dG(\varphi) \right)}_{\text{Intensive margin}} df_{ij} - \underbrace{\left(r(\bar{\varphi}_{ij}) g(\bar{\varphi}_{ij}) \times \frac{\partial \bar{\varphi}_{ij}}{\partial f_{ij}} \right)}_{\text{Extensive margin}} df_{ij}$$

⁶ $\lambda = (1 + \lambda_5) \times \mu..$

Solving for each expression, and expressing each margin in elasticities, we get the following predictions,

$$\zeta \equiv -\frac{d \ln X_{ij}}{d \ln \tau_{ij}} = \underbrace{(\sigma - 1)}_{\substack{\text{Intensive margin} \\ \text{Elasticity}}} + \underbrace{(\gamma - (\sigma - 1))}_{\substack{\text{Extensive margin} \\ \text{Elasticity}}} = \gamma \quad (25)$$

$$\xi \equiv -\frac{d \ln X_{ij}}{d \ln f_{ij}} = \underbrace{0}_{\substack{\text{Intensive margin} \\ \text{Elasticity}}} + \underbrace{\left(\frac{\gamma}{\sigma - 1} - 1\right)}_{\substack{\text{Extensive margin} \\ \text{Elasticity}}} = \frac{\gamma}{\sigma - 1} - 1 \quad (26)$$

$$\Rightarrow \frac{\partial \zeta}{\partial \sigma} = 0 \text{ and } \frac{\partial \xi}{\partial \sigma} < 0$$

Main properties of each margin of adjustment:

- When the variable trade barriers move (τ_{ij}), the intensive margin of trade responds in exactly the same way as in Krugman. The elasticity of exports by existing exporters with respect to τ is $(\sigma - 1)$, as in Krugman. Each firm is facing a constant elasticity residual demand, and therefore, when goods are very substitutable (σ high), the export of each individual exporter is very sensitive to the trade barriers she faces.
- When variable trade barriers move, the extensive margin on the other hand behave quite differently than in the Krugman model. As goods become more substitutable (σ gets higher), the market share of the least productive firms, compared to the more productive firms, shrinks. In a sense, increasing σ makes the market more competitive, and in such a competitive market, a small difference in productivity translates into large differences in market shares. So in equilibrium, the marginal exporters, which are the least productive exporters, export very small quantities compared to the existing exporters. When variable trade barriers go down, some of these less productive firms start exporting, but they have a negligible impact on aggregate trade is σ is high.
- So the intensive margin is more sensitive to trade barriers when σ is high, and the extensive margin is *less* sensitive when σ is high. With the specific functional form chosen here (Pareto distributed productivity shocks), those two effects exactly cancel out each other, so that at aggregate, the elasticity of trade flows with respect to trade barriers does not depend at all on the elasticity of substitution σ . Note that this results is similar to the finding in Eaton and Kortum. The reasoning in their Ricardian model is

somehow similar, and σ drops out for some specific distribution of productivity shocks.

- The response of trade to changes in the fixed cost are somehow similar. First, note that since this is a fixed cost, changing the fixed cost is not going to change the behavior of existing exporters (to a first order approximation). Those exporters have already decided to enter, and they have optimally chosen how much to export. Changing the fixed cost at the margin will not change their decision. All the adjustment will come from changes in the extensive margin. As far as the extensive margin is concerned, its sensitivity to fixed trade barriers can be understood in the same way as for the sensitivity to variable trade barriers. All comes from the equilibrium effect that in high σ sectors, the less productive firms have a very small market share compared to the more productive firms.
- Also note that when the dispersion of productivity among firms is small (γ large), there are few high productivity firms, but a large mass of less productive firms. So when trade barriers move around, there is a large number of potential candidates for exporting. The extensive margin is therefore more sensitive to changes in trade barriers in sectors where firms' productivity is less dispersed.
- Note that some of the results are functional form specific. It turns out however that the assumed functional form fits both micro data and aggregate data quite well.

3.1 Empirical estimation

I test the main predictions of the model using sectoral data on bilateral trade flows.

Data used:

- Data on sectoral bilateral trade flows.
- Measures of trade barriers (distance, common language, contiguity, as well as freight and tariffs).
- Estimates of sector-level demand elasticities. These estimates have been computed by Broda and Weinstein (2006).

To construct those estimate, Broda and Weinstein use data on prices and quantities at a highly disaggregated level over time.

Table 1: Firm heterogeneity distorts gravity.

Note: Dependent variable, log of exports from country i to country j in sector h in 1996. All regressions include sector dummies, origin country and destination country dummies. Observations are clustered within country pairs. Robust standard errors are given in parentheses. Significant at the 1% (***) , 5%(**), 10% level (*). Source: 1996 bilateral trade flows, Feenstra (2000); firm heterogeneity, Compustat, rank-size scaling coefficient of sales in 1996; data are aggregated over 35 BEA sectors; countries with a GDP/capita lower than \$3000 (in PPP) or a population smaller than 1 million have been ignored.

These are prices and quantities of goods imported by the US from many countries. They estimate separately demand and supply elasticities for each sector (conceptually, we want to look at demand elasticities, not supply elasticities). Most importantly, for each sector, they have many different subsectors. It is not as good as firm level data, which would give estimate of the demand elasticity that each firm is facing, but it's as good as it gets. So in the end, the elasticities they estimate are really elasticities of substitution between different varieties of the same good.

My model predicts that using some measure of the sensitivity of trade flows to trade barriers, as has been done extensively in the literature, would not give estimates of the elasticity of substitution, but would actually give estimates that are inversely related to the elasticity of substitution.

- Estimates of the dispersion of productivity within sectors. To compute those estimates, I estimate the parameters driving the whole distribution of firm sizes in different sectors, which directly give an estimate for $\frac{\gamma}{\sigma-1}$.

I find strong support for the predictions of the model. First, in sectors where firms are very dispersed ($\frac{\gamma}{\sigma-1}$ small), trade barriers have a mild impact on aggregate trade flows. Second, in sectors where goods are very substitutable (σ high), trade barriers have a mild impact on trade flows. These results are displayed on tables (1) and (2) respectively.

This evidence is also consistent with the work of Jim Rauch (1999). He find that in sectors where goods are arguably more homogenous

Table 2: Market structure distorts gravity.

Note: Dependent variable, log of exports from country i to country j in sector h in 1997. All regressions include sector dummies, origin country and destination country dummies. Observations are clustered within country pairs. Robust standard errors are given in parentheses. Significant at the 1% (***) , 5%(**) , 10% level (*). Source: 1997 bilateral trade flows, Feenstra (2000); elasticities of substitution, Broda and Weinstein (2004), 1980-1997 estimates; data are aggregated at the 3-digit SITC level; countries with a GDP/capita lower than \$3000 (in PPP) or a population smaller than 1 million have been ignored.

(goods that are traded on organized markets, or goods for which there is a price quote), trade barriers, and distance among others, have a lower impact on trade flows. This is consistent with my model. There is also more specific evidence on the relative importance of the intensive and the extensive margins of trade in different sectors, that seem to be systematically related to the degree of product differentiation. This evidence has been collected on firm level data on exports of French firms by Pamina Koenig (2006).