Firms in Trade and Macro - Lecture Notes*

Alessandro Ferrari†

†University of Zurich

April 2024

*Elie Gerschel provided excellent teaching assistance in the preparation of these notes.

Contents

Foreword

These notes are meant as an introduction to topics on firms in macro and trade. The key idea is to study how firms' behaviour impacts the economy within and across borders.

I start by reviewing a basic framework to think about an optimizing firm making choices in competitive markets.

This workhorse model allows us to then derive a way to think about the data and to measure policy-relevant quantities like misallocation. Within this setting, I then provide a brief introduction to the problem of estimating production functions and markups.

In the core of these notes, we cover models of firms' decisions about production and trade and conclude with a section on imperfect competition, granular firms and production networks.

Throughout the lecture notes, there are digressions. Some of them are important for putting what we discuss into practice or to provide a little more context to some results I derive. This is to say they are digressions but sometimes they are where the interesting stuff lies.

If you find any mistake in these notes please let me know at `alessandro.ferrari@econ.uzh.ch`.

1 Basics of Firms

We start by reviewing the most basic framework of firm choices. The model starts with a number of assumptions that we will relax throughout these notes as we progress.

1.1 A simple framework

Assume there is a continuum of identical firms, selling the same homogeneous good. Markets are competitive both on the output and inputs side. We can represent the production problem by introducing the notion of a production function. In general, this is a map f that takes a set of inputs $X = \{x_1, x_2, \ldots, x_N\}$ and turns it into output $Y = f(X)$. Y is output or value added. It is important to realize that Y is only true value added if $Y \notin X$. Later on in these notes, this will not be the case anymore, and we will have to be careful about what is value added (GDP) and how it differs from Y.

To make this more concrete, we restrict X to consist only of labour L. The firm's problem then boils down to how much to produce and sell, which can be restated as how much labour to employ.

As we assumed that markets are competitive, the firm takes all prices as given. In particular, denote p, the price of the output good and w, the wage (price of labour). Further, we normalize $p = 1$ so that w is also the real wage or the price of labour in terms of units of the output good.[1]

Let us also assume that the firm cannot store the output good and that it can produce without limits at a given point in time. The firm's objective is assumed to be to maximise its profits, which are given by

$$\max_{Y,L} \pi = pY - wL \quad st \tag{1}$$

$$Y = f(L). \tag{2}$$

Thanks to our normalisation and the definition of the production function, we can rewrite this as

$$\max_{L} \pi = f(L) - wL. \tag{3}$$

Taking the first order condition with respect to labour, we obtain

$$f'(L) = w, \tag{4}$$

where $f'(L)$ denotes the partial derivative of $f(L)$ with respect to L.

[1] Recall that by Walras' Law, an economy with N markets has one redundant market-clearing condition. This implies that we can only pin down $N - 1$ relative prices. Hence, our normalization of one of the prices to 1.

In economics, $f'(L)$ is the marginal product of labour (MPL), and the first order condition states that this needs to be equal to w, which is the marginal cost of production.

This is a case in which the firm is maximising profits by imposing that the $MPL = w$. This will, in general, be true in our models throughout the course. However, this setting is special in two ways which make this condition even more important than usual. First, labour is the only input, so picking labour is equivalent to picking output. Second, all markets are competitive, which implies that prices are given, and therefore, the marginal product is equal to the marginal revenue product.

To see how these matter, consider a production function $Y = f(L, K)$, where r denotes the price of capital. Using the same procedure, we get

$$MPL = f_L(L, K) = w, \tag{5}$$

$$MPK = f_K(L, K) = r, \tag{6}$$

where f_i denotes the partial derivative with respect to input i. The combination of these two conditions now gives us the production plan. This can be seen as the firm picking i) with what optimal mix of L and K; ii) how much output to produce. In this order, the firm asks itself what is the optimal mix of inputs that minimizes the marginal cost. In other words, it might be able to produce one unit of output with infinitely many combinations of L, K, but they all imply a different cost of production. In the first step, the firm minimizes the cost of producing a single unit. In the second step, the firm takes the optimal input mix and decides how much to produce. We will come back to under which conditions we can think of these steps as sequential or separate.

Next, we start thinking about the competitive output market assumption. We will discuss this in detail further down the road, but for now, suppose that the firm choice of how much to sell changes the price, i.e. it is not price-taker on the output market. We write this formally as

$$\max_{Y,L} \pi = p(Y)Y - c(Y) \tag{7}$$

where $c(Y)$ denotes the cost function. The first order condition now is

$$p_Y(Y)Y + p(Y) - c_Y(Y) = 0. \tag{8}$$

The second term on the left is familiar, it tells us that producing and selling one more unit increases our profits by $p(Y)$. The third term states that producing one more unit increases our costs and therefore decreases our profits by $c_Y(Y)$. So absent the first term, we will have the familiar $p = MC$. The first term instead says that the firm will take into account that its choices move the price. Furthermore, note that in any downward-sloping demand system, $p_Y < 0$, so

the price is larger than the marginal cost. We can rewrite the last equation as

$$p_Y(Y)Y + p(Y) = c_Y(Y), \tag{9}$$

where the left-hand side is called the Marginal Revenue Product (MRP). In the previous example, we had that the firm was equating the marginal product to the marginal cost because the price was fixed and normalized to 1.

1.2 Returns to Scale

So far, we have not assumed anything about the production function, other than, implicitly, that it is differentiable. We now impose additional structure to study how the properties of this function impact firm choices.

The notion of returns to scale is possibly the most important one when it comes to production functions. They answer the question on how the technology behaves depending on whether we are producing small or large quantities. As before, let us start in a general way with the following production function: $Y = f(x_1, \ldots, x_N)$.

We state that f has constant returns to scale if $f(\chi x_1, \ldots, \chi x_N) = \chi f(x_1, \ldots, x_N) = \chi Y$ for $\chi > 0$. In words, this states that if we increase all inputs by a factor χ, then output will increase by a factor χ. Similarly, we say that it has increasing returns to scale if $f(\chi x_1, \ldots, \chi x_N) > \chi f(x_1, \ldots, x_N) = \chi Y$ and decreasing returns to scale if $f(\chi x_1, \ldots, \chi x_N) < \chi f(x_1, \ldots, x_N) = \chi Y$.

Going back to our firm problem, we can immediately ask what returns to scale do to a firm's choice. Let's start with the case of constant returns to scale. Denote p_i the price of input i, then the first order conditions imply

$$f_{x_i} = p_i, \quad \forall i. \tag{10}$$

We can now invoke the Euler theorem upon realizing that returns to scale are the economics version of what we call the degree of homogeneity in maths. The Euler theorem states that, for a homogeneous of degree χ function, the following holds

$$\sum_i^N x_i f_{x_i} = \chi f(x_1, \ldots, x_N). \tag{11}$$

We now note that CRS is the case where $\chi = 1$, while DRS is $\chi < 1$ and IRS is $\chi > 1$. For CRS, this allows us to write

$$f(x_1, \ldots, x_N) = \sum_i^N x_i f_{x_i} = \sum_i^N p_i x_i. \tag{12}$$

Note that the rightmost formulation is the total cost bill of the firm. With this production function, the profits of a firm are given by $\pi = f(x_1, \ldots, x_N) - \sum_i^N p_i x_i$. The left-hand side is the total value of output (recall the price is 1) and, therefore, the firm's revenues. As total revenues are equal to total costs, the firm makes zero profits. If instead, the firm has a DRS production function, then we have that $\sum_i^N p_i x_i = \chi f(x_1, \ldots, x_N)$, which in turn implies that $\pi = (1 - \chi) f(x_1, \ldots, x_N) > 0$. We can generalise this by noting that a firm with returns to scale $\chi < 1$ will have a profit rate of $1 - \chi$.

To study what returns to scale do to cost functions, we introduce the Cobb-Douglas production function. In particular, we assume that output is given by

$$Y = AK^{\alpha} L^{\beta}, \tag{13}$$

where A is called Total Factor Productivity (TFP), and we will come back to it extensively. For now, assume $A = 1$. It is immediate to show that this function is homogeneous of degree $\alpha + \beta$.

Solving the optimal problem of the firm, we obtain

$$rK = \alpha Y \tag{14}$$

$$wL = \beta Y. \tag{15}$$

Solving for the inputs and plugging back into the production function

$$K = \left(\frac{\beta}{w} \frac{r}{\alpha} \right)^{\frac{-\beta}{\alpha+\beta}} Y^{\frac{1}{\alpha+\beta}} \tag{16}$$

$$L = \left(\frac{\alpha}{r} \frac{w}{\beta} \right)^{\frac{-\alpha}{\alpha+\beta}} Y^{\frac{1}{\alpha+\beta}}. \tag{17}$$

Recalling that the total cost function is given by

$$TC(Y, r, w) = r \left(\frac{\beta}{w} \frac{r}{\alpha} \right)^{\frac{-\beta}{\alpha+\beta}} Y^{\frac{1}{\alpha+\beta}} + w \left(\frac{\alpha}{r} \frac{w}{\beta} \right)^{\frac{-\alpha}{\alpha+\beta}} Y^{\frac{1}{\alpha+\beta}}. \tag{18}$$

It is immediate to verify that the marginal cost c_Y is equal to the average cost TC/Y if and only if $\alpha + \beta = 1$.

1.3 Constant Elasticity Demand and Monopolistic Competition

So far, we have worked out an example in which a firm is taking prices as given on both input and output markets. We have also briefly discussed how the problem changes when we let the firm output choices affect the market price of its good. In this subsection, we work from there to build first a general pricing rule and then a specific one after we settle on a demand curve.

From equation 9 and assuming a constant marginal cost $c_Y(Y) = c$, we obtain

$$p_Y(Y)Y + p(Y) = c. \tag{19}$$

At this point, we can introduce the familiar notion of price elasticity of demand ε_p. First, call $Y(p)$ the demand function ($p(Y)$ is the inverse demand). Then define[2]

$$\varepsilon_p \equiv \frac{-Y_p(p)p}{Y(p)}. \tag{20}$$

For convenience, we focus here on cases where $\varepsilon_p > 1$. As a consequence $p_Y(Y) = -\frac{p(Y)}{Y\varepsilon_p}$, which implies

$$p(Y) = \frac{\varepsilon_p}{\varepsilon_p - 1}c > c. \tag{21}$$

So firms that can affect the price of their goods choose a price above marginal cost. The ratio between price and marginal cost, called the markup, is a function only of the elasticity of demand. In this simple setting, we can note a few interesting observations. If the elasticity of demand is just a constant number, then so is the markup (more on this soon). If the elasticity of demand goes down with a firm size then, larger firms will face less elastic demands and optimally charge higher markups and, therefore, higher prices. This optimal pricing rule holds independent of the demand a firm faces. We can now proceed to study a specific case called the Constant Elasticity of Substitution (CES) demand system.

This preference system, introduce by Dixit and Stiglitz (1977), has the property that the ensuing demand is iso-elastic. Formally, assume that the consumer in our economy maximises utility, in the form of consumption, which is an aggregate of many different consumption goods

$$C = \left(\int_i c_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}, \tag{22}$$

where different goods are indexed by i and $\sigma > 1$ is the elasticity of substitution between different types of goods. Suppose also that the consumer has the following budget constraint

$$\int_i p_i c_i di = I, \tag{23}$$

where I is some exogenous income they have and p_i is the market price of good i. We proceed by asking how a consumer would split the income to maximise utility. Formally, taking the first

[2]The negative sign in the definition of the price elasticity of demand implies that we are defining a positive number (most of the time). That is easier to work with.

order condition with respect to a generic variety c_i

$$\frac{\sigma-1}{\sigma}c_i^{\frac{\sigma-1}{\sigma}-1}\frac{\sigma}{\sigma-1}\left(\int_i c_i^{\frac{\sigma-1}{\sigma}}di\right)^{\frac{\sigma}{\sigma-1}-1}-\lambda p_i=0, \tag{24}$$

where λ is the Lagrange multiplier associated to the budget constraint. We can immediately take the price to the RHS and divide by the same condition for variety j to obtain

$$c_i=\left(\frac{p_i}{p_j}\right)^{-\sigma}c_j. \tag{25}$$

We can now define the ideal price index $P\equiv\left(\int_i p_i^{1-\sigma}di\right)^{\frac{1}{1-\sigma}}$. This price index is designed so that $\int_i p_i c_i di = PC$ at the optimal choice of the consumer.[3] Multiply both sides of equation (25) by p_i and integrate over i to obtain

$$PC=P^{1-\sigma}p_j^{\sigma}c_j, \tag{26}$$

Which, finally, implies

$$c_j=\left(\frac{p_j}{P}\right)^{-\sigma}C. \tag{27}$$

This is the demand for variety j as a function of all other prices and the total level of consumption. This demand function has intuitive and sensible properties: it decreases in the price of the good and increases in the price of other goods, through P. It also increases homothetically when total consumption increases. Finally, as the name might give away, its price elasticity is constant, and in particular, it is equal to σ. Without further solving, we can then conclude that a firm facing this type of demand for its own variety will have an optimal price

$$p=\frac{\sigma}{\sigma-1}c. \tag{28}$$

In other words, the markup is just a constant number. Intuitively, when $\sigma\to\infty$ the markup $\mu\to 1$. This is a case where goods are perfect substitutes, and so there is little product differentiation and market power coming from it. The opposite case is when $\sigma\to 1$ and therefore $\mu\to\infty$. Note

[3] To obtain this note that, at the optimum, $c_i = p_i^{-\sigma}p_j^{\sigma}c_j$ and we want $\int_i p_i c_i di = PC$. Then $\int_i p_i p_i^{-\sigma}p_j^{\sigma}c_j di = P\left(\int_i c_i^{\frac{\sigma-1}{\sigma}}di\right)^{\frac{\sigma}{\sigma-1}}$, by the definition of the aggregator C in eq. 22. Using the optimality condition $p_j^{\sigma}c_j\int_i p_i^{1-\sigma}di = P\left(\int_i\left(p_i^{-\sigma}p_j^{\sigma}c_j\right)^{\frac{\sigma-1}{\sigma}}di\right)^{\frac{\sigma}{\sigma-1}}$. Simplifying we obtain $p_j^{\sigma}c_j\int_i p_i^{1-\sigma}di = Pp_j^{\sigma}c_j\left(\int_i p_i^{1-\sigma}di\right)^{\frac{\sigma}{\sigma-1}}$. Canceling out $p_j^{\sigma}c_j$ from both sides, $\int_i p_i^{1-\sigma}di = P\left(\int_i p_i^{1-\sigma}di\right)^{\frac{\sigma}{\sigma-1}}$, which yields the desired price index $P=\left(\int_i p_i^{1-\sigma}di\right)^{\frac{1}{1-\sigma}}$.

that there is one key unstated assumption behind this result. If we were to derive this optimal pricing rule, formally, we would have to ask ourselves whether we want to allow p_j to affect P. We wrote down this model in the context of a continuum of varieties, as shown by the integral, rather than a sum, so it is somewhat natural to think that $\partial P/\partial p_j = 0$. This is equivalent to assuming that firms are each a monopolist on their own variety, but they do not have any aggregate effect. This set of assumptions falls under the name of monopolistic competition. We will possibly discuss alternatives to this setting later in the course, for example, oligopolistic competition models, in which this needs not to hold. The combination of CES preferences and monopolistic competition is by far the most used in international trade and is often used in modern macro.

This concludes the preliminaries we needed to cover some ground in thinking about data through models and how to measure theoretical objects. In the next section, we go back to some of the facts we discussed in the first lectures and dig deeper into how those empirical observations came about.

2 Measurement[4]

In this section, we take the framework described so far and use it as a measurement device. In particular we will make structural assumptions which allow us to identify important parameters in the data. We begin by studying production function estimation, which leads us to estimate firm productivity. With estimates of a firm's production function, we can then study the allocation of an economy and ask whether we could engineer better ones in terms of how we use resources. We conclude by discussing the problem of estimating markups and how this impacts the empirical observations of the first class.

2.1 Production Function Estimation

Suppose a competitive firm i produces with the following production function

$$Y_i = A_i K_i^{\alpha} L_i^{\beta}. \tag{29}$$

Suppose that we have data on Y_i, K_i and L_i. Can we recover the key parameters of interest A_i, α and β? We can start by taking logs

$$y_i = \omega_i + \alpha k_i + \beta l_i, \tag{30}$$

where smaller case letters represent the log of capital letters and $\omega_i = \log A_i$. If we have many firms, under the assumption that α and β are common for all these firms, we can estimate this via OLS. In principle, we could simply run our OLS regression, and get $\hat{\alpha}, \hat{\beta}$. The first issue there is that ω_i would just be in the error term. Unfortunately, that's only the beginning of this mess. When ω_i is in the error term, if the choices of capital and labour depend in any way on ω_i, our OLS estimates $\hat{\alpha}, \hat{\beta}$ will be biased. Given their definitions, this is equivalent to asking whether choices of capital and labour depend on productivity. We know that is the case from our first order conditions in the previous section. TFP determines marginal products of inputs and, therefore, how much of them the firm uses.

In summary, it is immediate that OLS is a bad idea as it will deliver biased estimates for our production function parameters.

2.1.1 Fixed Effects and First-Differences

If we have access to panel data in which we can observe the same firm repeatedly and we are willing to make the following unrealistic assumption $\omega_{it} = \omega_i, \forall t$, then maybe there is hope.

[4]This section is partially based on Daron Acemoglu's and Florian Ederer's notes on production function and markups estimation.

We will not be able to estimate TFP, but we can get a α and β. Start by first-differencing log output in consecutive periods

$$y_{it} - y_{it-1} = \alpha(k_{it} - k_{it-1}) + \beta(l_{it} - l_{it-1}) + \epsilon_{it}. \tag{31}$$

This will consistently estimate α and β as the productivity component is differenced out. The catch is that if the assumption $\omega_{it} = \omega_i$, $\forall t$ is not true, then we are back to a biased estimate as changes in productivity are related to changes in factor usage.

In a similar vein, consider the following firm fixed effects model of the production function

$$y_{it} = \omega_{it} + \alpha k_{it} + \beta l_{it} + \theta_i + \epsilon_{it}, \tag{32}$$

where θ_i is a set of firm fixed effects. In this setting, θ_i will absorb all firm-specific time-invariant variation. Under the assumption that productivity does not change across periods, we then have that $\hat{\theta}_i$ will be equal to ω_i. Note, however, that this is only true if productivity is the only firm-specific time-invariant element in the production function, if there are others we do not observe, then θ_i will be a function of productivity and other determinants. In either case, if we believe productivity to be time-invariant, we have removed the endogeneity problem since we are controlling for productivity directly. Hence we can unbiasedly estimate α and β.

These two approaches are fundamentally very similar. The time-invariance assumption implies that productivity can be differenced away. Also, recall that fixed effects models, or within estimators, are the same as demeaning, so you can think of the second approach as subtracting the mean of log output for firm i across all the sample periods. You would estimate the following equation:

$$\tilde{y}_{it} = \alpha \tilde{k}_{it} + \beta \tilde{l}_{it} + u_{it}, \tag{33}$$

with $\tilde{x}_{it} = x_{it} - \frac{1}{T}\sum_t x_{it}$ for $x = y, k, l$.

2.1.2 Using FOCs

An alternative approach is to use the first order conditions directly. We know from our previous discussion that a firm in a competitive environment (both inputs and output) will choose

$$\frac{\partial Y}{\partial K}\frac{K}{Y} = \frac{rK}{pY} \tag{34}$$

$$\frac{\partial Y}{\partial L}\frac{L}{Y} = \frac{wL}{pY}. \tag{35}$$

These quantities are often observable. In balance sheets, we typically know the wage bill wL, some form of capital stock K and interest rate payments r and revenues pY. When we impose our Cobb-Douglas assumption, the LHS are just β and α. So potentially, without any estimation, we could get at the key parameters of interest. We cannot quite get at A directly but if we believe our estimates of α and β and we observe input quantities and prices separately (namely r and K rather than capital expenditure rK and w and L rather than the wage bill wL) we can use the FOC

$$p\hat{\alpha}AK^{\hat{\alpha}-1}L^{\hat{\beta}} = r \tag{36}$$

or the labour equivalent to get the product pA, which we call TFPR as it is based on revenues. In the rare event that we observe output Y and price p separately, rather than revenues pY, then we can compute A, which we call TFPQ as it is based on output quantity.

This approach comes at two key costs, one is an inherent limit of using the FOCs, and the other is quite general to most production function estimation problems. Implicitly, using this approach requires that firms make static choices. In other words, they need to only care about this period. In a world with dynamic decisions, adjustment costs and labour market frictions, our FOCs are misspecified, and therefore we cannot identify our parameters of interest.

The second, broader issue, is the one of assuming perfect competition. If firms face downward sloping demand, then they will have

$$\frac{\partial Y}{\partial L}\frac{L}{Y} = \mu\frac{wL}{pY} \tag{37}$$

$$\frac{\partial Y}{\partial K}\frac{K}{Y} = \mu\frac{rK}{pY}. \tag{38}$$

Where $\mu = p/mc$ is the markup. In this environment, we will not be able to get the actual production function parameters α and β since we will always observe them jointly with the markup. What we can observe is α/β since, in the ratio, the markup cancels out. To be able to separately identify them, we would need to know the markup itself, which would require us to know the elasticity of demand the firm faces.

2.1.3 Instrumental Variable

A standard approach to endogeneity issues is to find an instrumental variable. We would like something that affects output but only through input choices and is not correlated with ω_i or the error term.

If we live in the world of competitive markets, then theory tells us that output and input prices are potentially good candidates. Take input prices, for example. They clearly move the input choices but do not affect output directly, only through the choice of inputs themselves. It is

trickier to think about when they would actually be uncorrelated with ω_i. For example, suppose a firm operates in an imperfectly competitive output market. Then p_i is definitely related to ω_i as more productive firms will produce more, which will drive down the price. Suppose instead that input markets are not competitive. You could have that more productive firms have more market power vs workers and therefore face effectively lower wages, invalidating w as an instrument. Another non-trivial issue is whether, when we observe two firms paying different wages, we can infer anything about their productivity, since they might just be employing a different kind of workers without us (the econometricians) observing it in the data.

2.1.4 "Modern" Approaches to Production Function Estimation

We conclude with the two approaches that represent the basics of how we estimate production functions these days. These are two papers by Olley and Pakes (1996) and Levinsohn and Petrin (2003). I list in the end refinements and extensions to these methods in more recent years, which I will not cover.

Olley & Pakes (1996), Econometrica

Suppose again that we have data on many firms for many periods within a sector so that, in principle, we could run the following regression

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \omega_{it} + \varepsilon_{it}. \tag{39}$$

We know that k_{it}, l_{it} are correlated with ω_{it}, and therefore we cannot just run OLS. The big idea in this approach is to note that all of the endogeneity is coming from the fact that ω_{it} is unobserved by the econometrician while it is observed by the firm. Note that i) if we, the econometricians, knew ω_{it} we would use it in the regression and identify β_1, β_2; also, if the firm did not know ω_{it} when choosing l_{it} and k_{it} we would not have an endogeneity problem.

What Olley and Pakes do to solve this problem is to make an assumption about how productivity moves over time. In particular

$$\omega_{it} = f(\omega_{it-1}) + \xi_{it}. \tag{40}$$

Where we build $f(\cdot)$ so that $\mathbb{E}[\xi_{it}|\omega_{it-1}] = 0$. This is a first order Markov process since ω_{it} only depends on itself at time $t-1$. Next, we assume that labour l can be adjusted frictionlessly, while capital K follows the law of motion

$$K_{it} = \delta K_{it-1} + i_{it-1}, \tag{41}$$

where δ is the fraction of yesterday's capital which has not depreciated and can still be used

today, and i is the investment made yesterday that turns into capital today. So the trick here is to say that since the capital of today is decided yesterday somehow, it is directly related to productivity yesterday. This is exactly the variation we will leverage to get at our parameters of interest. Towards this, let's rewrite the investment policy implicitly as

$$i_{it} = g_t(k_{it}, \omega_{it}). \tag{42}$$

Note that we are allowing the function g to change over time but not across firms. For example, if wages are high this period, g might vary, but it needs to do so for all firms in the same way. Otherwise, we lost our identifying variation. Next, we invert this to write productivity as

$$\omega_{it} = g_t^{-1}(k_{it}, i_{it}). \tag{43}$$

Technically we made just made two important assumptions here: i) strict monotonicity of g so that the function is invertible; ii) that there is nothing else we do not observe in the investment equation and varies across firms. Effectively ii) is saying that two firms whose investment is different but have the same capital can only differ in productivity, not in the input prices they face, the demand elasticity they face and so on and so forth. ii) is a very strong assumption, but we cannot proceed without it in this setting.

From here on out, it is pretty straightforward, we have ω_{it} as a function of observables. We need to choose some g_t function. We can do this non-parametrically as a higher order polynomial of i_{it} and k_{it}. For example, using order 2, we can estimate

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \gamma_{0t} + \gamma_{1t} k_{it} + \gamma_{2t} i_{it} + \gamma_{3t} k_{it}^2 + \gamma_{4t} i_{it}^2 + \gamma_{5t} k_{it} i_{it} + \varepsilon_{it}. \tag{44}$$

The key thing here is that our endogeneity problem is no more since we have expressed ω_{it} as a function of observables, and it is not in the error term anymore. In an ideal world, we are done. Unfortunately, we cannot separately identify β_1 from γ_{1t} since they are perfectly collinear.

Note, however, that there is no labour in the polynomial g_t function, so β_2 is well identified. So what we can do is estimate 44. We will get weird stuff for capital, but we will have the right estimate of $\hat{\beta}_2$.

With $\hat{\beta}_2$ in the bag, we can start working our (long) way towards β_1. From our regression with the polynomial expansion, we got

$$y_{it} = \beta_2 l_{it} + \underbrace{\beta_0 + \beta_1 k_{it} + g_t^{-1}(k_{it}, i_{it})}_{\Phi_{it}} + \varepsilon_{it} \tag{45}$$

Where

$$\Phi_{it} = \chi_{0t} + \chi_{1t}k_{it} + \gamma_{2t}i_{it} + \gamma_{3t}k_{it}^2 + \gamma_{4t}i_{it}^2 + \gamma_5 k_{it}i_{it}. \tag{46}$$

We can now make use of our assumed law of motion for ω_{it},

$$\hat{\omega}_{it} = f(\hat{\omega}_{it-1}) + \xi_{it}, \tag{47}$$

and combine it with the observation that

$$g_t^{-1}(k_{it}, i_{it}) = \hat{\Phi} - \beta_1 k_{it} = \hat{\omega}_{it}, \tag{48}$$

where I am including the β_0 inside ω_{it} since it is common to all firms. Plugging these terms in the law of motion yields

$$\hat{\Phi}_{it} - \beta_1 k_{it} = f(\hat{\Phi}_{it-1} - \beta_1 k_{it-1}) + \xi_{it}. \tag{49}$$

Rearranging

$$\hat{\Phi}_{it} = \beta_1 k_{it} + f(\hat{\Phi}_{it-1} - \beta_1 k_{it-1}) + \xi_{it}. \tag{50}$$

We use the same trick as before and use a polynomial expansion of Φ and k. Note, however, that they are at time $t-1$, so now we can identify β_1.

Finally, with $\hat{\beta}_1$, we can immediately recover $\hat{\omega}_{it}$ as $\hat{\Phi}_{it} - \hat{\beta}_1 k_{it}$. We made it to consistently estimating β_1, β_2 and the productivity ω_{it}.

Levinsohn & Petrin (2003), Review of Economic Studies

The Olley & Pakes approach is the big starting point of this literature. We will not cover the numerous available improvements (listed below), but it is worth discussing one important extension.

Levinsohn and Petrin (2003) observe an important limitation of the OP approach: they use investment as the key instrument, but most firms do not invest every year. In their original dataset of Chilean firms, half of the firms have 0 investment. Unless we believe that all these firms have the same productivity, we have violated strict monotonicity. We can relax it to weak monotonicity (there is a threshold $\underline{\omega}$ below which a firm does not invest), but it comes at additional costs in terms of throwing away observations. So either way, we are not in great shape since we cannot use OP for those firms, but clearly, excluding them will introduce large selection bias.

Levinsohn and Petrin make a further observation: all firms use material inputs. Think of

electricity or wood and metal for Ikea furniture. So we write a production function which is the same Cobb-Douglas as before but now has an additional input

$$y_{it} = \beta_0 + \beta_1 k_{it} + \beta_2 l_{it} + \beta_3 m_{it} + \omega_{it} + \varepsilon_{it}. \tag{51}$$

We can then think of the firm's optimal material input choice as we were thinking about investment before, so we write $m_{it} = g_t(k_{it}, \omega_{it})$. We then proceed as in OP. First, we estimate $\hat{\beta}_2$ by using a polynomial expansion of the g_t^{-1} function (with the caveat that now we cannot identify β_1 nor β_3). Next, we use the same approach for the second stage to recover $\hat{\beta}_1$, $\hat{\beta}_3$ and finally $\hat{\omega}_{it}$.

2.1.5 Discussion

So far, we addressed the fundamental identification problem of our output elasticities of interest. Let's say we call it a win because we have some methods that, provided our specified production function, under the stated assumptions, deliver us α and β.

How much can we trust our estimates? There are two well-known issues in the literature, which we briefly mentioned so far, and it is worth expanding on. The first big issue in this whole literature is effectively an omitted price bias. The variation we use in our estimates is typically across firms within a sector. Effectively we are assuming that the production function elasticities are the same across firms in the same sector, so Pepsi and Coca-Cola have the same production function by assumption. Now assume for some reason that Coca-Cola has more market power than Pepsi, so it charges a higher markup. Since we want to be careful, suppose we have a deflator P at the sector level so that we turn our revenue data into quantity using this price index. Clearly, by aggregation, we deflated too much Pepsi and too little Coca-Cola since $p^P < P < p^{CC}$. So, everything else equal, it is going to look like Coca-Cola is more productive than Pepsi because they make more "deflated revenues" for every unit of inputs. In an ideal world, we have information on prices and quantities, and we can solve this problem by deflating appropriately.

The second problem worth mentioning is one of multiple products. Suppose Coca-Cola produces two different goods, Coke and Fanta. Typically we do not observe data separately at the product level, so we have to assume that Coke and Fanta have the same production function. This might be true for these very similar products, but it is definitely not the case for Ikea's food vs Ikea's sofas...

Even worse, suppose that we do have information at the product level, and suppose that we have information about prices. We know exactly everything we would like to know about the Pepsi and Coca-Cola inputs, output and so on. Unfortunately, we do not know anything about quality. Maybe Coca-Cola has a higher price because it is just better than Pepsi, not because it has a higher market power. It turns out that sometimes using quantity data instead of revenues can make matters worse since prices are a measure of quality which we would not observe

anyway. See De Loecker et al. (2016) for a discussion on this.

Last, honourable mention in the list of issues to think about, there is attrition. Firms enter and exit all the time, as we saw in the first class. These are not random choices and are probably very related to productivity itself. So possibly all these estimates suffer from attrition bias.

If you want to know more about this literature, these are 3 important innovations over the methods we discussed so far: Doraszelski and Jaumandreu (2013), Ackerberg et al. (2015),Gandhi et al. (2020). Finally, if you want to know a lot more, De Loecker and Syverson (2021) provide an excellent review article of all these themes and the ones we are about to cover.

2.2 Markups Estimation

There are many reasons why we may want to work through all the pains of getting the production function estimation right. First, we know that productivity itself is an important and interesting economic object. Second, related to our discussion on recent trends, getting the production function right is one of two possible alternative routes to estimating markups. The alternative we will not cover has to do with estimating the demand directly, since once we have demand, we have the price elasticity and, therefore, the markup.

The production function approach to markup estimation leverages a condition we have already derived:

$$\frac{\partial Y}{\partial L}\frac{L}{Y} = \mu \frac{wL}{pY} \tag{52}$$

or its equivalent for capital. We have already mentioned that, under Cobb-Douglas, $\frac{\partial Y}{\partial L}\frac{L}{Y} = \beta$ and $\frac{\partial Y}{\partial K}\frac{K}{Y} = \alpha$ so if we have these elasticities correctly estimated from our production function since we typically observe wage bill and revenues we can immediately compute the markup by inverting equation (52).

Unfortunately, this is a special case. Suppose we have a production function in which we have a generic variable input V_{it} with a price p_{it}^V and capital. Let's also assume that there is some fixed cost f_{it}, which is a cost that does not scale with production but reduces profits. The cost minimization problem of the firm implies

$$\mathcal{L}(V_{it}, K_{it}, \lambda_{it}) = p_{it}^V V_{it} + r_{it} K_{it} + f_{it} - \lambda_{it}(f(V_{it}, K_{it}) - \bar{Y}). \tag{53}$$

Optimizing with respect to V, we obtain

$$p_{it}^V = \frac{\partial f(\cdot)}{\partial V}\lambda_{it}. \tag{54}$$

Denote $\theta_{it}^V = \frac{\partial Y}{\partial V}\frac{V}{Y}$ the output elasticity of V. Multiplying on both sides by V, one then gets

$$p_{it}^V V_{it} = \lambda_{it}\theta_{it}^V Y_{it}. \tag{55}$$

Recall that λ_{it} tells us how much the cost bill increases when we tighten the constraint by 1, which is equivalent to saying by how much does cost go up when we increase output by 1. That is the definition of marginal cost, so by rearranging this equation and multiplying both sides by the price P_{it}, we get

$$\frac{P_{it}}{\lambda_{it}} = \theta_{it}^V \frac{Y_{it}P_{it}}{p_{it}^V V_{it}} \tag{56}$$

Recalling that the markup μ_{it} is defined as $\frac{P_{it}}{MC_{it}}$ where MC_{it} is the marginal cost, we conclude that

$$\mu_{it} = \theta_{it}^V \frac{Y_{it}P_{it}}{p_{it}^V V_{it}}. \tag{57}$$

This is known as the ratio estimator, proposed by Hall (1989) and then De Loecker and Warzynski (2012).

So we are back at the conclusion we got when we had labour in the production function as the only variable input. If we could observe all variable inputs, revenues and output elasticities, we would be done. The latter we get by estimating the production function, revenues are observed, so we are left with variable inputs. Unfortunately, what constitutes a variable input V rather than a fixed cost f is very much up for discussion. Formally the f are costs that do not move with output. The cost of opening a business is definitely a fixed cost. Electricity or physical input (wood in Ikea tables) are definitely variable inputs, but what about advertising? management wage bill? building rents? Those are in a very grey area.

The markup estimation literature is, to some extent, plagued by the same problem as the production function estimation literature and possibly some additional ones. As a simple first step, as Doraszelski and Jaumandreu (2019) show, if there is a markup, then the production function estimation should account for it to get the elasticity in the first place. Otherwise, we are estimating something inconsistent between the two steps.

A clear example of this is what we get if we try to estimate markups based on revenue data. Namely, we do not observe prices separately. In what follows, we briefly go through the critique by Bond et al. (2021), which questions the validity of the ratio estimator and then discuss what we can still learn about markups through the insights of De Ridder et al. (2021).

The bulk of the criticism moved by Bond et al. (2021) is based on the idea that, when we only observe revenues, the ratio estimator in equation (57) is uninformative about markups. To see their argument let's first define a couple of important quantities. We have already introduced

the output elasticity $\theta^V \equiv \frac{\partial Y}{\partial V}\frac{V}{Y}$. Recalling that revenues are given by PY, we can define the revenue elasticity to variable input as $\theta^{RV} \equiv \frac{\partial PY}{\partial V}\frac{V}{PY}$ and finally let's bring back the price elasticity of demand $\varepsilon^P \equiv -\frac{\partial Y}{\partial P}\frac{P}{Y}$. Breaking up the derivative of the sum, we can write the revenue elasticity as

$$\theta^{RV} = \left(\frac{\partial P}{\partial V}Y + \frac{\partial Y}{\partial V}P\right)\frac{V}{PY}, \tag{58}$$

multiplying and dividing the first term by $\partial Y/Y$, we get

$$\theta^{RV} = \frac{\partial P}{\partial Y}\frac{\partial Y}{\partial V}\frac{YV}{PY} + \frac{\partial Y}{\partial V}\frac{V}{Y} = \tag{59}$$

$$= (-1/\varepsilon^P + 1)\theta^V = \theta^V\frac{\varepsilon^P - 1}{\varepsilon^P}. \tag{60}$$

So if we now estimate our markup $\hat{\mu}$ using the revenue elasticity, we will get

$$\hat{\mu} = \frac{\varepsilon^P - 1}{\varepsilon^P}\mu, \tag{61}$$

but we know from the previous section that $\mu = \varepsilon^P/(\varepsilon^P - 1)$ so we get $\hat{\mu} = 1$ independently of what the true markup is.

Note, however, that such an issue only arises if we estimate our output elasticity by regressing revenues on input quantities. This is something that we never do. We either do revenues on input expenditures or, if we are lucky to have the data, output quantity on input quantity.

In a recent paper De Ridder et al. (2021) show two important practical results: i) while the Bond et al. (2021) critique is correct on average, a revenue-based estimated markup distribution has approximately the correct dispersion; ii) even if the average revenue markup is uninformative, if you are willing to assume that the production function parameters are time-invariant, you can still use revenue-based markups to look at trends. In summary, if you have revenue data, you will not get the average right, but you can still get a lot of useful information.

Digression on Variable vs Fixed Costs We spent an inordinate amount of time by now discussing variable costs. We also talked about how useful the notion of fixed costs can be to generate inaction or sorting. We have introduced the two as obviously different things because there is something quite intuitive about a cost that moves 1 to 1 with an input and one that does not move at all.

In the real world, these lines are way more blurred, and sometimes they just do not exist depending on what the unit of observation of the time period of our data is. A good example in this setting is a manager. Suppose we need a manager for every 10 employees. Is the manager's wage a variable or a fixed cost? Well, it does move with output but not 1 to 1. It's locally

fixed between the 1st and the 9th employee. Whether we think of managers or advertisement expenditure as fixed or variable costs turns out to be consequential when we want to empirically measure the quantities we write down in our model. A practical example is the one of markups. We will discuss later how there is a giant literature about the rise of markups and market power. Now a markup in our model is the ratio between prices and marginal costs. When we go to the balance sheet data, we often do not observe prices or marginal costs. What we have to do is to decide which balance sheet item can be considered marginal cost. This choice turns out to matter a whole lot to figure out if markups increased over time or not. For a debate on this, see De Loecker, Eeckhout and Unger (2020), and Traina (2018). Figure 1, from Traina (2018) shows exactly this point.

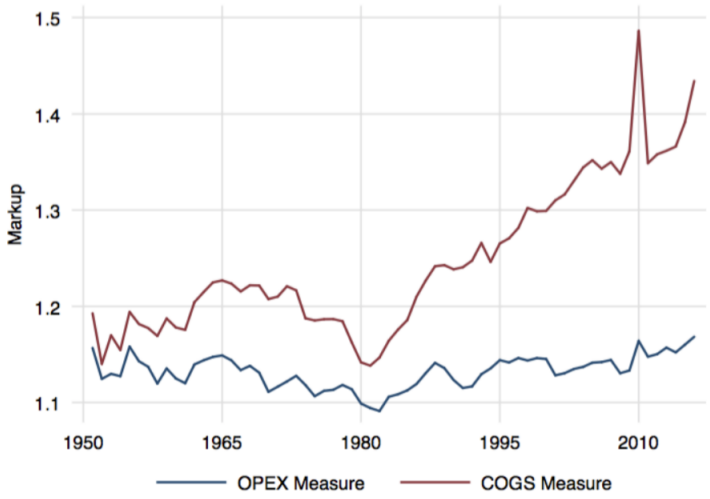


Figure 1: Markups Trends from Traina (2018).

A more nuanced approach to this is to think about costs as living on a continuum that goes from fully variable to fully fixed depending on if and, if so, how they move with output. And indeed, from a very highly theoretical level, we do not need to make such a distinction, particularly if it turns out to be problematic when we try to look at the real world.

End of Digression

Markup Trends We are now in the position to think about what is driving the observed increase in markups (conditional on it being well-measured). Consider the result from before

$$\mu_{it} = \theta_{it}^{V} \frac{Y_{it} P_{it}}{p_{it}^{V} V_{it}}. \tag{62}$$

Now we want to think about the economy. The first non-trivial question is how to aggregate if there are multiple firms selling different quantities. In general, we want to write

$$\bar{\mu}_t = \sum_i m_{it}\mu_{it}. \tag{63}$$

The key question is what to use as weight m_{it}. An obvious candidate is to use sales shares y_{it}/y_t, with $y_t \equiv \sum_i y_{it}$. This is what De Loecker et al. (2020) use. If this is the measure we run with, then if we see the markup moving over time, it can be because of 3 possible sources:

1. the output elasticity θ^V_{it} changed;
2. the cost share (and thus its inverse $\frac{Y_{it}P_{it}}{p^V_{it}V_{it}}$) changed;
3. the aggregation weight m_{it} changed.

One way to answer the question is to produce counterfactual markups series. For example, we could check how close we get to the observed μ_t series if we build

$$\tilde{\mu}_t = \sum_i m_{it}\bar{\mu}_i, \tag{64}$$

which is the counterfactual markup series in which only reallocation is active. Similarly, by turning each of the three on, one at a time, we can assess how much of the observed trend we can explain.

As a side note on aggregation, observe the following. Call an aggregate ratio $Y/X = \sum_i y_i / \sum_i x_i$. The following holds

$$\frac{Y}{X} = \sum_i \frac{x_i}{X}\frac{y_i}{x_i}, \tag{65}$$

$$\frac{Y}{X} \neq \sum_i \frac{y_i}{Y}\frac{y_i}{x_i}. \tag{66}$$

So when we think about markups as revenues/costs, we should use cost shares, rather than revenue (sales) shares to do the aggregation. It turns out that this actually matters for the markup trends, as pointed out by Edmond et al. (2018).

Another standard diagnostic is the following decomposition (which applies to any weighted average). Supposing there is no entry nor exit, starting from equation (63), we can write

$$\Delta\bar{\mu}_t = \underbrace{\sum_{i\in N_t} m_{it-1}\Delta\mu_{it}}_{\Delta \text{ within}} + \underbrace{\sum_{i\in N_t} \Delta m_{it}\mu_{it-1}}_{\Delta \text{ between}} + \underbrace{\sum_{i\in N_t} \Delta m_{it}\Delta\mu_{it}}_{\Delta \text{ cross term}} \tag{67}$$

When there is entry or exit, the formula has to be amended. We now introduce a term for entrants and a term for exiters, and we need to look at demeaned variables.

$$\begin{aligned}\Delta\bar{\mu}_t = &\underbrace{\sum_{i\in N_{t-1}, i\in N_t} m_{it-1}\Delta\mu_{it}}_{\Delta \text{ within}} + \underbrace{\sum_{i\in N_{t-1}, i\in N_t} \Delta m_{it}\tilde{\mu}_{it-1}}_{\Delta \text{ between}} + \underbrace{\sum_{i\in N_{t-1}, i\in N_t} \Delta m_{it}\Delta\mu_{it}}_{\Delta \text{ cross term}} \\ &+ \underbrace{\sum_{i\in N_t, i\notin N_{t-1}} m_{it}\hat{\mu}_{it}}_{\Delta \text{ entry}} - \underbrace{\sum_{i\notin N_t, i\in N_{t-1}} m_{i,t-1}\tilde{\mu}_{i,t-1}}_{\Delta \text{ exit}}\end{aligned} \tag{68}$$

where we have to define $\tilde{\mu}_{it} = \mu_{it} - \bar{\mu}_t$ and $\hat{\mu}_{it} = \mu_{it} - \bar{\mu}_{t-1}$. De Loecker et al. (2020) carry out this exercise and conclude that the bulk of the change in markup is driven by reallocation effects, namely the second and third terms. This is particularly true from the '90s onwards, while in the '80s also within, net entry played a role.

2.3 Misallocation - Hsieh and Klenow (2009)[5]

In this section, we briefly discuss how to use theory to measure misallocation. Economists are inherently interested in the problem of the allocation of resources and its efficiency. For example, we often note that, when comparing developed and developing countries, the difference between the most productive firms is not particularly large. What makes up for the bulk of the aggregate productivity differences is, for example, the size of these firms. If very productive firms are very small and unproductive firms are large, then resources are poorly allocated, and we could reshuffle them and end up making more goods.

The key question is then how to measure the degree of misallocation and how much it may contribute to the aggregate productivity differences, which, together with relative factor abundance/scarcity, make up for a lot of the difference in the degrees of development of countries.

To work towards this goal, we can start by writing down what the efficient allocation would look like. Consider the problem of the firm with CRS Cobb-Douglas $y_i = A_i K_i^{\alpha} L_i^{1-\alpha}$ maximizing profits

$$\pi_i = p_i y_i - wL_i - rK_i. \tag{69}$$

Suppose further that the firm faces a downward sloping demand derived from CES preferences. We know (see Appendix 5.1) that, denoting c the marginal cost for $A_i = 1$, the price is given by

$$p_i = \frac{\sigma}{\sigma - 1}\frac{c}{A_i}. \tag{70}$$

[5]This model is based on Chris Edmond's lecture slides.

And we know from the Cobb-Douglas cost minimization problem that this is

$$p_i = \frac{\sigma}{\sigma - 1} \left(\frac{r}{\alpha}\right)^{\alpha} \left(\frac{w}{1-\alpha}\right)^{1-\alpha} A_i^{-1}. \tag{71}$$

We can then derive the optimal capital-labour ratio

$$\frac{K_i}{L_i} = \frac{\alpha}{1-\alpha} \frac{w}{r}. \tag{72}$$

This property derived from the Cobb-Douglas production function remains true even in the presence of a markup since the latter only changes the size of the firm, not the optimal input mix. Note that this would not be the case if we introduced mark-downs on specific inputs rather than an output markup.

Recalling that demand for a given good i is given by

$$y_i = p_i^{-\sigma} P^{\sigma} Y, \tag{73}$$

using the price equation, we can write that

$$y_i \propto A_i^{\sigma}. \tag{74}$$

Meaning that output is proportional to firm productivity with curvature given by σ.

Lastly, recall that by the firm's optimization problem, we have that

$$MRPK_i = r, \tag{75}$$

$$MRPL_i = w, \quad \forall i. \tag{76}$$

Namely, firms choose inputs so that the marginal revenue products are equal to input prices, and therefore the marginal profit is zero. This holds for all firms, provided that they all face the same input prices r and w. As a direct consequence, we should observe no dispersion in measured MRPK and MRPL in the data, if we had an efficient economy.

The ultimate goal is to figure out what the measured aggregate productivity of this economy is. Towards this, define aggregate productivity as the productivity of the aggregate production

$$Y = AK^{\alpha} L^{1-\alpha}, \tag{77}$$

where $K = \sum_i K_i$ and $L = \sum_i L_i$. We know from the firm's first order condition that

$$rK_i = \frac{\alpha c}{A_i} y_i, \tag{78}$$

$$wL_i = \frac{(1-\alpha)c}{A_i} y_i. \tag{79}$$

where $\frac{c}{A_i} y_i$ is the total cost. It follows that aggregate capital and labour, as given by

$$K = \sum_i K_i = \frac{\alpha c}{r} \sum_i \frac{y_i}{A_i}, \tag{80}$$

$$L = \sum_i L_i = \frac{(1-\alpha)c}{w} \sum_i \frac{y_i}{A_i} \tag{81}$$

We can plug these into the aggregate production function

$$A = Y K^{-\alpha} L^{\alpha-1} \tag{82}$$

$$= Y \left(\frac{\alpha c}{r} \sum_i \frac{y_i}{A_i} \right)^{-\alpha} \left(\frac{(1-\alpha)c}{w} \sum_i \frac{y_i}{A_i} \right)^{\alpha-1} \tag{83}$$

$$= Y \left(\frac{\alpha}{r} \right)^{-\alpha} \left(\frac{1-\alpha}{w} \right)^{\alpha-1} c^{-1} \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{84}$$

$$= Y \left(\frac{\alpha}{r} \right)^{-\alpha} \left(\frac{1-\alpha}{w} \right)^{\alpha-1} \left(\frac{r}{\alpha} \right)^{-\alpha} \left(\frac{w}{1-\alpha} \right)^{\alpha-1} \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{85}$$

$$= Y \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{86}$$

Inverting this condition, we obtain

$$A^{-1} = Y^{-1} \sum_i \frac{y_i}{A_i}, \tag{87}$$

using the demand to solve for the relative output y_i / Y

$$A^{-1} = \sum_i \left(\frac{p_i}{P} \right)^{-\sigma} A_i^{-1}. \tag{88}$$

We know what both p_i and P are, and we can take the ratio

$$\frac{p_i}{P} = \frac{\frac{\sigma}{\sigma-1}\frac{c}{A_i}}{\left(\sum_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{89}$$

$$= \frac{\frac{\sigma}{\sigma-1}\frac{c}{A_i}}{\left(\sum_j \left(\frac{\sigma}{\sigma-1}\frac{c}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{90}$$

$$= \frac{\sigma}{\sigma-1}\frac{c}{A_i}\frac{\sigma-1}{\sigma c}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{91}$$

$$= \frac{1}{A_i}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}}. \tag{92}$$

We can plug this into 88 and obtain

$$A^{-1} = \sum_i A_i^{\sigma-1}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{-\sigma}{\sigma-1}}, \tag{93}$$

which implies

$$A = \left(\sum_i A_i^{\sigma-1}\right)^{-1}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{\sigma}{\sigma-1}} \tag{94}$$

$$= \left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}}. \tag{95}$$

That is, the aggregate productivity of the economy is a power mean of firm-level productivities. Note that if all firms have the same productivity $\bar{A}$, then this collapses to $A = \bar{A}$. This is a direct consequence of constant returns to scale and downward sloping demand. Firms will optimally have different sizes since they have different productivities, so if we take away this heterogeneity, the economy collapses to a representative firm with productivity $\bar{A}$.

We have now characterized the efficient benchmark and can introduce inefficiencies.

Consider now an economy with "wedges" which can potentially create misallocation. In particular, suppose firms' profits are given by

$$\pi_i = (1-\tau_i^Y)p_i(y_i)y_i - wL_i - (1+\tau_i^K)rK_i, \tag{96}$$

where τ^Y distorts output, thereby increasing the marginal product of both capital and labour, and τ^K distorts capital, thereby increasing the marginal product of capital relative to labour. An

example of the latter are financial constraints, where the firms would like to borrow more to achieve their optimal capital-to-labour ratio but cannot because they cannot borrow.

To make our life easier, denote $\tilde{r}_i = r(1+\tau_i^k)$ and $\tilde{p}_i = p(1-\tau_i^Y)$. Then the problem becomes

$$\pi_i = \tilde{p}_i(y_i)y_i - wL_i - \tilde{r}_i K_i. \tag{97}$$

The FOCs yield

$$\tilde{r}_i = \lambda_i \alpha A_i K_i^{\alpha-1} L_i^{1-\alpha} \tag{98}$$

$$w = \lambda_i (1-\alpha) A_i K_i^{\alpha} L_i^{-\alpha}. \tag{99}$$

Solving for the input mix and output decisions, we obtain that the optimal capital-to-labour ratio is given by

$$\frac{K_i}{L_i} = \frac{\alpha}{1-\alpha}\frac{w}{r}\frac{1}{1+\tau_i^K}, \tag{100}$$

so it is distorted by a factor $(1+\tau_i^K)^{-1}$, relative to the efficient level. Similarly, it is easy to show that

$$MRPL_i = w\frac{1}{1-\tau_i^Y}, \tag{101}$$

$$MRPK_i = r\frac{1+\tau_i^K}{1-\tau_i^Y}. \tag{102}$$

So both marginal revenue products will be distorted relative to the efficient level. As a direct consequence, we will have a non-degenerate distribution of marginal revenue products as long as the wedges are not identical across firms. We could then use data to estimate the production functions and marginal revenue products to check how much dispersion we have in a given sector.

Before solving further, think about what these wedges are doing to the economy. The output wedge operates in a way that is potentially indistinguishable from productivity A_i. It implies that for a given level of input usage, less output is obtained, so a high wedge is the same as a lower productivity. Effectively it will imply that some firms are smaller than they should be. The capital wedge instead operates by distorting the relative marginal product of capital and labour. Therefore, firms will use a suboptimal capital per-worker ratio. The intuition is that for a given level of MPK, the cost of capital increases due to the wedge. Therefore, capital will be underutilized in equilibrium.

We can now solve the model further to recover aggregate productivity as we did before.

Solving the firm's problem, we obtain the optimal price

$$p_i = \frac{\sigma}{\sigma - 1} \frac{c}{A_i} \frac{(1 + \tau_i^K)^\alpha}{1 - \tau_i^Y}. \tag{103}$$

To obtain this result start from the cost minimization problem

$$\min_{K_i, L_i} wL_i + (1 + \tau_i^K) rK_i + \lambda(\bar{y} - A_i K_i^\alpha L_i^{1-\alpha}), \tag{104}$$

recalling that λ is the marginal cost of the firm. Next, you can maximize the profits in 96 using the marginal cost you just derived to obtain the optimal price in 104. An alternative way to see this is to think about τ_i^Y as a tax, then the firm will set the *after-tax* price as the markup over the marginal cost

$$(1 - \tau_i^Y) p_i = \frac{\sigma}{\sigma - 1} \lambda. \tag{105}$$

To find the exact price, we need to solve for λ. We can do this by combining the FOCs of the cost minimization and the production function. From the FOCs we have

$$K_i = \frac{\lambda_i \alpha y_i}{\tilde{r}_i}, \tag{106}$$

$$L_i = \frac{\lambda_i (1 - \alpha) y_i}{w}. \tag{107}$$

Plugging into the production function

$$y_i = A_i \left(\frac{\lambda_i \alpha y_i}{\tilde{r}_i} \right)^\alpha \left(\frac{\lambda_i (1 - \alpha) y_i}{w} \right)^{1-\alpha}, \tag{108}$$

which implies

$$\lambda_i^{-1} = A_i \left(\frac{\alpha}{\tilde{r}_i} \right)^\alpha \left(\frac{(1 - \alpha)}{w} \right)^{1-\alpha}. \tag{109}$$

Therefore the marginal cost is given by

$$\lambda_i = \frac{\tilde{c}_i}{A_i} = A_i^{-1} \left(\frac{\alpha}{\tilde{r}_i} \right)^{-\alpha} \left(\frac{(1 - \alpha)}{w} \right)^{\alpha - 1}. \tag{110}$$

Note that now the marginal cost is also distorted since the optimal input mix is distorted. As before, we can define $c = \left(\frac{\alpha}{r}\right)^{-\alpha} \left(\frac{(1-\alpha)}{w}\right)^{\alpha-1}$, which implies that the effective marginal cost of firm i is $cA_i^{-1}(1 + \tau_i^K)^\alpha$. Plugging this into 105, we obtain 103.

We can now retrace our steps from the efficient model. First, recall that

$$y_i = p_i^{-\sigma} P^{\sigma} Y, \tag{111}$$

which, using the price, implies

$$y_i = \left(\frac{\sigma}{\sigma - 1} c\right)^{-\sigma} A_i^{\sigma} \left(\frac{1 - \tau_i^Y}{(1 + \tau_i^K)^{\alpha}}\right)^{\sigma}. \tag{112}$$

Note that, while in the undistorted economy we had $y_i \propto A_i^{\sigma}$, here

$$y_i \propto A_i^{\sigma} \left(\frac{1 - \tau_i^Y}{(1 + \tau_i^K)^{\alpha}}\right)^{\sigma}. \tag{113}$$

Next, we can characterize the distortion in factor shares. From the cost minimization problem we have

$$(1 + \tau_i^K) r K_i = \alpha \frac{c}{A_i} (1 + \tau_i^K)^{\alpha} y_i, \tag{114}$$

$$w L_i = (1 - \alpha) \frac{c}{A_i} (1 + \tau_i^K)^{\alpha} y_i. \tag{115}$$

Therefore the effective capital and labour shares are

$$\Theta_K = \frac{rK}{PY} = \frac{\alpha}{\frac{\sigma}{\sigma - 1}} \sum_i \frac{1 - \tau_i^Y}{1 + \tau_i^K} \frac{p_i y_i}{PY} \tag{116}$$

$$\Theta_L = \frac{wL}{PY} = \frac{1 - \alpha}{\frac{\sigma}{\sigma - 1}} \sum_i (1 - \tau_i^Y) \frac{p_i y_i}{PY}. \tag{117}$$

Note that we can also use these shares and the expenditures derived above to get

$$PY = \left(\frac{r}{\Theta_K}\right)^{\alpha} \left(\frac{w}{\Theta_L}\right)^{1 - \alpha} K^{\alpha} L^{1 - \alpha}. \tag{118}$$

Finally, to characterize aggregate productivity, we write again the aggregate production function

$$A = Y K^{-\alpha} L^{\alpha - 1}, \tag{119}$$

We can now define revenue TFP of firm i, called TFPR, in this economy as

$$p_i A_i = \frac{\sigma}{\sigma - 1} c \frac{(1 + \tau_i^K)^{\alpha}}{1 - \tau_i^Y}. \tag{120}$$

Note that the term using the definition of the aggregate production function, we can write

$$PA = \left(\frac{r}{\Theta_K}\right)^{\alpha}\left(\frac{w}{\Theta_L}\right)^{1-\alpha} \equiv TFPR. \tag{121}$$

This is the aggregate TFPR since it maps input usage in the production function to aggregate revenues PY. Further note that it is very similar to the cost index we derived multiple times, but now instead of the α, $1-\alpha$ coefficient, it includes the effective factor shares Θ_K, Θ_L which account for potential misallocation induced by the wedges τ^Y, τ^K.

From here, we can work towards isolating A by noting that $A = TFPR/P$. From the definition of the price index

$$P = \left(\sum_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}} \tag{122}$$

$$= \left(\sum_j \left(\frac{TFPR_j}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}, \tag{123}$$

where we used the definition of the firm-specific TFPR, $TFPR_i = p_i A_i$. We can then plug this into the price index and obtain

$$A = TFPR\left(\sum_j \left(\frac{TFPR_j}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{\sigma-1}} \tag{124}$$

$$= \left(\sum_j \left(A_j \frac{TFPR}{TFPR_j}\right)^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{125}$$

So the aggregate productivity A in this economy is not the same as in the efficient economy. If $TFPR = TFPR_j$, $\forall j$ then 125 collapses to 95. Namely, if there is no dispersion in $TFPR_j$, the economy is back at its efficient benchmark. In this economy, the dispersion in $TFPR_j$ is solely driven by the wedges, and it will lower productivity. To see this formally, you can note that in a two-firm economy, having $TFPR/TFPR_j = \{.9, 1.1\}$ will imply a lower A than having $TFPR/TFPR_j = \{1, 1\}$ due to the concavity of the aggregator.

Note the key assumptions required for this approach to work: we need firms with the same production function, and we need to have integrated capital and labour markets. Also, note that it does not really matter that we assumed a distortion on capital but not in labour. We could have done the opposite and gotten the same results. What matters is that there is both something distorting the size of the firm and something that distorts the optimal input mix.

We conclude the discussion on measuring misallocation by thinking through the mapping

between the model and the data. It is useful to remember that, in this economy, the market share of firm i is defined as $s_i = \frac{p_i y_i}{PY} = (y_i/Y)^{\frac{\sigma-1}{\sigma}}$, see Appendix 5.1 for the derivation.

We would like to measure TFPQ at the firm level: A_i. Suppose we have external measures of α and σ and we observe revenues $p_i y_i$, the number of workers L_i, the wage bill wL_i and the capital stock K_i, as typical in firm/plant-level data. We can write the inverted production function as

$$A_i = y_i K_i^{-\alpha} L_i^{\alpha-1}, \tag{126}$$

we can invert market shares to obtain output as a function of revenues and aggregates: $y_i = \left(\frac{p_i y_i}{PY}\right)^{\frac{\sigma}{\sigma-1}} Y$ and plug it in

$$A_i = K_i^{-\alpha} L_i^{\alpha-1} \left(\frac{p_i y_i}{PY}\right)^{\frac{\sigma}{\sigma-1}} Y \tag{127}$$

$$= K_i^{-\alpha} L_i^{\alpha-1} (p_i y_i)^{\frac{\sigma}{\sigma-1}} \chi, \tag{128}$$

where $\chi := (PY)^{\frac{\sigma}{1-\sigma}} Y$ is just a function of aggregates. From 128, we can estimate TFPQ at the firm level, since everything on the RHS is either data, parameters, or common factors across firms. Measuring TFPR is even easier since

$$TFPR_i = (p_i y_i) K_i^{-\alpha} L_i^{\alpha-1}. \tag{129}$$

From here we can directly compare the distribution of revenue and quantity productivities and estimate counterfactuals.

2.4 Concentration

We have now established that there are a lot of potential pitfalls in trying to figure out the markup distribution of an economy. The reason why the recent (more macro than IO) literature has taken this approach based on production function estimation is that estimating demand directly is an even harder problem. It is difficult to figure out what is the relevant market, if not markets, a firm is competing in. This fundamental issue is also at the heart of the problems in measuring market power from concentration measures.

We proceed in steps by first introducing the most common measure of concentration, and then we discuss when and how it is informative.

By far, the most common measure used to think about market power is the Herfindhal Hirschman Index (HHI). Formally, suppose you have N firms, denoted by the index i, define

$$HHI = \sum_i s_i^2, \tag{130}$$

where $s_i = y_i / \sum_i y_i$ is firm i's market share. This measure is largely used, particularly by Antitrust authorities, because it is very easy to compute. For example, the US Department of Justice states that *"the agencies generally [] consider markets in which the HHI is in excess of 2,500 points to be highly concentrated"* [6]. The fundamental question is whether it is informative about market power.

Before diving into counterexamples, note the properties of HHI: if we live in a world with a monopolist, then $HHI = 1$. In perfect competition, $HHI = 0$. If two identical firms compete in a Cournot duopoly game, $HHI = 1/2$. Now take the monopolistic competition model with CES demand we studied earlier. We know firms will sell according to

$$q_i = Y p_i^{-\sigma} P^{\sigma}. \tag{131}$$

rewriting this in revenues (multiply by p_i on both sides) and turning it into a market share (dividing by PY):

$$s_i = p_i^{1-\sigma} P^{\sigma-1}. \tag{132}$$

Next, recall that the price is given by the constant markup $\sigma/(\sigma-1)$ and their marginal cost c_i and that $P = \left(\sum_i p_i^{1-\sigma} di\right)^{1/(1-\sigma)}$, where I am using the discrete version of the price index discussed before. Combining these, we obtain

$$s_i = \frac{c_i^{1-\sigma}}{\sum_i c_i^{1-\sigma}}. \tag{133}$$

Hence

$$HHI = \sum_i s_i^2 = \left(\sum_i c_i^{1-\sigma}\right)^{-2} \sum_i c_i^{2-2\sigma}. \tag{134}$$

Note that markups do not enter in this measure. So the HHI in this setting is not informative about the degree of market power, just about productivity or marginal cost distributions.

The HHI becomes particularly problematic when we want to think about dynamics. The problem with looking at a plot of HHI over time and inferring trends about market power is that we need to hold the definition of market constant. This is particularly troublesome when we think about the recent decades and how globalization has changed the relevant definition of market. For example, think of your purchase behaviour over the last couple of years. Presumably, a large share of your expenditure is not local at all.

Lastly, note the following problem related to data availability. Suppose that we have scattered

[6] Oftentimes HHI is reported on a scale up to 10,000, meaning that a 90% market share is reported as $s_i = 90$.

missing data on the sales of some firms in our industry of interest. This will bias our HHI estimate. For example, if we only have data on the largest firms (typically the case), then $\widehat{HHI} < HHI$. An alternative measure which does not suffer from this issue is using concentration measures directly. For example, $C(n) = \sum_{j=1}^{n} s_j$, where we rank firms by their market share, is the cumulative market share of the n largest firms.

In summary, HHI needs to be handled with care, as a measure of market power because lots of (heroic) assumptions are required for HHI to be informative on the competition structure. Nonetheless, further down the road, we will cover a competition model in which HHI is exactly the right metric to think about market power and how to aggregate an economy.

3 Heterogeneous Firms

In this section, we start to develop richer models of firm choices. In particular, we make two key extensions to the model discussed in Section 1. First, we allow for firm heterogeneity so that we will be able to generate a sensible firm distribution. Second, we allow for firm entry and exit. This is motivated by the observation that there is a high (but declining in the US) turnover rate of firms, meaning high entry rate and high exit rate. Our previous model cannot account for this under the assumption of a representative firm. Next, we discuss how the same framework, with small tweaks, can be used to think about trade and, in particular why some firms trade while others do not and if that selection is random.

3.1 Firm Dynamics - Hopenhayn (1991)[7]

Take the following simplified version of the Hopenhayn model. Firms are identified by a productivity draw z when they are born, and this remains the same forever. Assume that the production function is given by $y = F(n) = zn^{\alpha}$ with $\alpha \in (0,1)$. Firms use labour n paying a wage w and have to pay a per-period fixed cost c_f to produce. The firm profits can be written as

$$\pi(z,p) = pzn^{\alpha} - wn - c_f. \tag{135}$$

We can use the wage as the numeraire (which means we will have to solve for the output price as the key aggregate object). First, we can ask how much does a firm with productivity z want to produce, for a fixed p, and get the first order condition

$$p\alpha z n^{\alpha-1} = 1, \tag{136}$$

which implies that the optimal amount of labour hired is

$$n^{\star} = (z\alpha p)^{\frac{1}{1-\alpha}} \tag{137}$$

and, substituting into the production function, the optimal size of the firm (in terms of output) is

$$y^{\star} = z(z\alpha p)^{\frac{\alpha}{1-\alpha}}. \tag{138}$$

With this solution, we can compute the firm profits at the optimum. The assumption that there is a per period cost of production implies that some firms may find it optimal to exit the market because their profit can turn negative, depending on changes in the aggregate state,

[7] This model is based on Chris Edmond's lecture slides.

as indexed by p. We can then impose the zero profit condition and pin down the exit cutoff. We can conjecture the existence and uniqueness of a $z^\star(p)$ such that firms with lower productivity $z \leq z^\star(p)$ would choose to exit the market. The reason we can conjecture this cutoff rule is that profits are strictly increasing in z, and if production is zero then profits are equal to $-c_f$. First, profits are given by

$$\begin{aligned}\pi(z,p) &= pz(z\alpha p)^{\frac{\alpha}{1-\alpha}} - (z\alpha p)^{\frac{1}{1-\alpha}} - c_f \\ &= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}(pz)^{\frac{1}{1-\alpha}} - c_f. \end{aligned} \tag{139}$$

The zero profit condition implies that the indifferent firm $z^\star$ will have $\pi(z^\star(p), p) = 0$. This is the first of two equilibrium conditions. At this point, we assume that there is a large mass of potential entrants. They can choose to enter and pay an entry cost, c_e, before knowing with what level of productivity they will enter. We also assume that they draw from a distribution $g(z)$. The goal of this is to eventually write the value for an entrant, impose free entry and find the equilibrium level of the price $p^\star$, which, together with $z^\star$, completes the description of the equilibrium. In equilibrium, because the aggregate state is constant (no price changes), firms do not exit, and there is no entry either because potential entrants are indifferent between entering or not.

The value for an incumbent firm with productivity z when the economy is forever on the equilibrium is given by

$$v(z,p) = \max\left\{0, \sum_{t=0}^{\infty} \beta^t \pi(z,p)\right\} = \max\left\{0, \frac{\pi(z,p)}{1-\beta}\right\}. \tag{140}$$

Since firms can choose to exit, if the present discounted value of profit is negative, they will leave the market and choose 0. Note that we can replace the sum of future profits with its annuity value since there is no residual uncertainty. It follows that a potential entrant will face a value which is the expectation over all possible z realizations of $v(z,p)$, hence

$$v^E(p) = \beta \int v(z,p) g(z) dz. \tag{141}$$

Note that we implicitly assumed that if a firm enters today, they start producing tomorrow. You can see this from the discount factor β in the equation. We can now impose a free entry condition. This condition states that since there is a large mass of firms waiting outside, if there are positive profits to be made, some firms will enter. This will happen until there are no more profits left. Technically, free entry implies that $v^E(p) = c_e$, which pins down $p^\star$. Practically we

can plug the definition of profits into the free entry condition and work it out.

$$c_e = \beta \int v(z,p) g(z) dz. \tag{142}$$

$$= \beta \int \max \left\{ 0, \frac{\pi(z,p)}{1-\beta} \right\} g(z) dz. \tag{143}$$

Now consider what happens to a firm that decides to pay the c_e entry cost but then draws a low productivity. We know that a firm needs to draw a $z > z^\star(p)$ to obtain positive profits. Any firm entering and drawing $z < z^\star(p)$ would then find it optimal to exit immediately since zero profits are better than losses. Then we can break the integral at $z^\star(p)$ and obtain

$$c_e = \beta \int_0^{z^\star(p)} \max \left\{ 0, \frac{\pi(z,p)}{1-\beta} \right\} g(z) dz + \beta \int_{z^\star(p)}^{\infty} \max \left\{ 0, \frac{\pi(z,p)}{1-\beta} \right\} g(z) dz. \tag{144}$$

We know that at $z < z^\star(p)$, $\pi(z,p) < 0$ while at $z > z^\star(p)$, $\pi(z,p) > 0$, then we can get rid of the max operators

$$c_e = \beta \int_0^{z^\star(p)} 0 g(z) dz + \beta \int_{z^\star(p)}^{\infty} \frac{\pi(z,p)}{1-\beta} g(z) dz. \tag{145}$$

Clearly, the first term is just zero, so, by plugging in equation (139) and rearranging, we obtain

$$(1-\beta) c_e = \beta \int_{z^\star(p)}^{\infty} ((1-\alpha) \alpha^{\frac{\alpha}{1-\alpha}} (pz)^{\frac{1}{1-\alpha}} - c_f) g(z) dz. \tag{146}$$

This tells us that firms need to be able to cover the entry cost in expectation. Since they can exit immediately upon seeing their productivity, they do not have to cover the production cost c_f too. If they were forced to pay c_f, then the expected value of entering would need to be large enough to pay both c_e and c_f. From equation (139) and the zero profit condition $\pi(z^\star, p^\star) = 0$ we can obtain

$$p^\star = c_f^{1-\alpha} (1-\alpha)^{\alpha-1} \alpha^{-\alpha} z^{\star -1}. \tag{147}$$

Plugging this solution into equation (146), we get

$$(1-\beta) \frac{c_e}{c_f} = \beta \int_{z^\star(p)}^{\infty} \left[\left(\frac{z}{z^\star} \right)^{\frac{1}{1-\alpha}} - 1 \right] g(z) dz. \tag{148}$$

Equation (148) provides the value of the exit cutoff. Plugging this into equation (147), we obtain the equilibrium price, which completes the characterization of the equilibrium. We immediately note that increasing the cost of entry will generate less entry and therefore, less productive incumbents will be able to stick around ($z^\star \downarrow$). As a consequence, the price will increase as the

average productivity in the economy declined ($p^\star \uparrow$). An increase in the per-period production cost will increase the exit cutoff, this however has ambiguous effects on the equilibrium price.

This model is a powerful starting point. We can characterize everything elegantly, and it is straightforward to extend in many interesting directions. An example of this is Melitz (2003). The paper does two main things: i) takes the Hopenhayn (1992) model and replaces the perfect competition assumption with monopolistic competition in differentiated varieties; ii) adapts the setting to talk about international trade. In general, these models are quite flexible and are also amenable to thinking about misallocation and reallocation.

Moreover, note an important element of this model. Most of the action comes from the existence of fixed costs. Without a fixed cost we would have that firms stay in forever. If firms stay in forever, we need zero entry to have a stationary distribution. Indeed the trick that Melitz (2003) uses to get around this is to assume that an exogenous fraction of firms δ disappears every period. In such a setting, then the stationary distribution just requires that entry replenishes the disappearing firms.

3.2 Heterogeneous Firms in an Open Economy - Melitz (2003)

We now turn to the model designed by Melitz (2003) to think about how firms operate in international environments. In particular, the model will be able to account for the non-random selection of firms into trade. The setting is one of consumers with CES utilities and heterogeneous firms.

Formally, consider a representative agent with utility

$$U = Q = \left[\int_{\omega \in \Omega} q(\omega)^{\frac{\sigma-1}{\sigma}} d\omega \right]^{\frac{\sigma}{\sigma-1}}. \tag{149}$$

We know that this implies demand for each variety given by

$$q(\omega) = Q p(\omega)^{-\sigma} P^{\sigma} \tag{150}$$

and expenditure $r(\omega) = p(\omega)q(\omega)$ given by

$$r(\omega) = I p(\omega)^{1-\sigma} P^{\sigma-1}. \tag{151}$$

where I is total income and the price index $P = \left[\int_{\omega \in \Omega} p(\omega)^{1-\sigma} d\omega \right]^{\frac{1}{1-\sigma}}$.

Next, assume that there is a unit mass of workers so that the total labour supply $L = 1$ and also normalize the wage $w = 1$ as the numeraire.

Firms are heterogeneous in their productivity φ and operate in monopolistic competition.

They use a production function

$$q = \varphi(l - f), \tag{152}$$

where l is their labour used and f is a fixed cost of production. Their profit maximization under CES implies that the optimal price is

$$p(\varphi) = \frac{\sigma}{\sigma - 1}\frac{1}{\varphi}. \tag{153}$$

The profits are then given by

$$\pi = r(\varphi) - l(\varphi), \tag{154}$$

where $l(\varphi)$ is labour used by a firm with productivity φ. From the expenditure function (equation 151), we know that

$$r(\varphi) = I\left(\frac{\sigma}{\sigma - 1}\varphi^{-1}\right)^{1-\sigma} P^{\sigma-1}. \tag{155}$$

which implies that profits are

$$\begin{aligned}
\pi(\varphi) &= r(\varphi) - l(\varphi), \\
&= r(\varphi) - \frac{q(\phi)}{\phi} - f, \\
&= r(\varphi) - \frac{p(\phi)q(\phi)}{p(\phi)\phi} - f, \\
&= r(\varphi) - \frac{r(\varphi)}{\frac{\sigma}{\sigma-1}} - f, \\
\pi(\varphi) &= \frac{r(\varphi)}{\sigma} - f.
\end{aligned} \tag{156}$$

Note the following features of what we derived so far. First, if we were to measure revenues per worker, we would estimate

$$\frac{r(\varphi)}{l(\varphi)} = \frac{\sigma}{\sigma - 1}[1 - \frac{f}{l(\varphi)}]. \tag{157}$$

Which is increasing in φ but includes a markup and a term accounting for the fixed cost of production f. Next, note that if we compare two firms with productivities φ_1 and φ_2 we obtain

$$\frac{q(\varphi_1)}{q(\varphi_2)} = \left(\frac{\varphi_1}{\varphi_2}\right)^{\sigma}, \tag{158}$$

meaning that a firm with higher productivity, say $\varphi_1 > \varphi_2$, will produce more but not proportionally so due to the imperfect substitutability as measured by σ. We know from before that more productive firms charge lower prices, so it is not obvious that they would have larger revenues. However,

$$\frac{r(\varphi_1)}{r(\varphi_2)} = \left(\frac{\varphi_1}{\varphi_2}\right)^{\sigma-1}, \tag{159}$$

namely, they have higher revenues since $\sigma > 1$, but the difference between the firms in revenues is smaller than in quantities. This is reasonable since the price difference is working against the quantity one.

We are now in the position to try to aggregate up the firm-level outcomes. In particular, denote $\mu(\varphi)$ the probability density function of firms with productivity φ conditional on their survival. We also know that if a firm has a productivity φ, it will have a price given by equation (153). Hence we can rewrite the price index P defined above as

$$P = \left(\int_0^\infty p(\varphi)^{1-\sigma} M\mu(\varphi)d\varphi\right)^{\frac{1}{1-\sigma}}, \tag{160}$$

where M is the mass of surviving firms. Using our pricing solution, we obtain

$$P = M^{\frac{1}{1-\sigma}} \frac{\sigma}{\sigma-1} \left(\int_0^\infty \varphi^{\sigma-1}\mu(\varphi)d\varphi\right)^{\frac{1}{1-\sigma}}. \tag{161}$$

For convenience, let's call $\tilde{\varphi} = \left(\int_0^\infty \varphi^{\sigma-1}\mu(\varphi)d\varphi\right)^{\frac{1}{\sigma-1}}$ which is a power average of prices charged by individual firms. We can do the same for other aggregates and obtain the total expenditure $R = Mr(\tilde{\varphi})$, the total profit $\Pi = M\pi(\tilde{\varphi})$ and the aggregate quantity $Q = M^{\frac{\sigma}{\sigma-1}} q(\tilde{\varphi})$. You can obtain this by the same procedure, namely compute the index as the integral of the individual firms with the appropriate pdf, and express quantities as a function of $\tilde{\varphi}$. For example, in the case of total expenditure:

$$\begin{aligned} R &= \int_0^\infty r(\varphi) M\mu(\varphi)d\varphi && (162)\\ &= \int_0^\infty r(\tilde{\varphi}) \left(\frac{\varphi}{\tilde{\varphi}}\right)^{\sigma-1} M\mu(\varphi)d\varphi && (163)\\ &= Mr(\tilde{\varphi}) \frac{1}{\tilde{\varphi}^{\sigma-1}} \int_0^\infty \varphi^{\sigma-1}\mu(\varphi)d\varphi && (164)\\ &= Mr(\tilde{\varphi}). && (165) \end{aligned}$$

The important result here is that we managed to obtain all our aggregate quantities of interest as a function of a weighted average productivity. Note for example that if all firms

had productivity $\varphi = \varphi^*$ then the average productivity would be $\tilde{\varphi}(\varphi^*) = \varphi^*$, and the average profits $\bar{\pi} = \pi(\tilde{\varphi}(\varphi^*)) = \pi(\varphi^*) = 0$, per definition.

Clearly, we are not done since this average is an endogenous object which depends on the firms' survival. To further solve this, we now use the same assumptions we used in Hopenhayn about entry but change exit slightly. Suppose that there is a large pool of potential entrants who can pay a cost of entry f_e and draw a productivity level from a cdf $G(\varphi)$. Active firms die with probability δ, which is exogenous. So, differently from Hopenhayn, we will not have endogenous exit since there is no per-period cost of production, but we can sustain a stable (stationary) distribution of firms by assuming that they randomly exit.

Like in the Hopenhayn model, we assume that firms maximize the present discounted value of profits

$$v(\varphi) = \max\{0, \sum_t (1-\delta)^t \pi(\varphi)\} = \max\left\{0, \frac{\pi(\varphi)}{\delta}\right\}, \tag{166}$$

where the firm is discounting at rate δ since with probability δ it will exogenously exit in every period.

Given our assumptions on the profit function, we can conjecture a rule such that there exists a cutoff productivity level φ^* such that below that firms do not find profitable to enter. In other words $\varphi^* = \inf\left\{\varphi \geq 0 : \frac{\pi(\varphi)}{\delta} \geq 0\right\}$ so that

$$\pi(\varphi^*) = 0. \tag{167}$$

Note that our assumptions on the profit function also ensure that this cutoff rule is unique.

We can now define the four objects of interest as a function of this cutoff and then eventually solve for the cutoff itself. For example, we can write the pdf of firm productivity as

$$\mu(\varphi) = \begin{cases} \frac{g(\varphi)}{1-G(\varphi^*)} & \varphi \geq \varphi^*, \\ 0 & \varphi < \varphi^*. \end{cases} \tag{168}$$

With this definition, we can compute our power average productivity $\tilde{\varphi}$ as

$$\begin{aligned} \tilde{\varphi} &= \left(\int_0^\infty \varphi^{\sigma-1} \mu(\varphi) d\varphi\right)^{\frac{1}{\sigma-1}} \\ &= \left(\frac{1}{1-G(\varphi^*)} \int_{\varphi^*}^\infty \varphi^{\sigma-1} g(\varphi) d\varphi\right)^{\frac{1}{\sigma-1}}. \end{aligned} \tag{169}$$

This characterizes the average productivity in the economy as a function of the entry cutoff. Working towards the solution of the cutoff itself, note that if we define the average profit in

the economy as $\bar{\pi} = \Pi/M$ then free entry, which implies that expected profits are equal to the entry cost, are given by

$$0 \times G(\varphi^*) + \frac{\bar{\pi}}{\delta}(1 - G(\varphi^*)) = f_e. \tag{170}$$

Namely, with probability $G(\varphi^*)$, the entrant draws a productivity below the cutoff, so it will now find it profitable to stay. With complement probability, it will get a high enough productivity draw to remain in the market and obtain the profit flow. It follows that

$$\bar{\pi} = \frac{\delta f_e}{1 - G(\varphi^*)}. \tag{171}$$

This is known as the free entry condition (FE). It can be interpreted as follows: holding constant the cost of entering, if firm death is more likely, firms will need a higher average profits flow to compensate the cost of entering. In other words, the firm in expectation will be active $1/\delta$ periods, so they need enough profits to cover the entry cost in $1/\delta$ periods.

From the previous derivations, we know that the average level of profits is given by the profits associated with the appropriately averaged productivity so that

$$\bar{\pi} = \pi(\tilde{\varphi}(\varphi^*)) = \frac{r(\tilde{\varphi}(\varphi^*))}{\sigma} - f = f\left[\frac{r(\tilde{\varphi}(\varphi^*))}{\sigma f} - 1\right]. \tag{172}$$

We also know that by definition of the cutoff φ^* we have that $\pi(\varphi^*) = 0 \Leftrightarrow r(\varphi^*) = \sigma f$, following equation (156). Finally, we can use our result on ratios of revenues to write

$$\bar{\pi} = f\left[\frac{r(\tilde{\varphi}(\varphi^*))}{r(\varphi^*)} - 1\right] = f\left[\left(\frac{\tilde{\varphi}(\varphi^*)}{\varphi^*}\right)^{\sigma-1} - 1\right]. \tag{173}$$

This is the zero cutoff profit (ZCP) in this economy. Jointly with the free entry condition derived above, it defines a system of two equations in two unknowns: φ^* and $\bar{\pi}$ that uniquely pin down the equilibrium in the economy. It is easy to show that uniqueness is implied by the fact that the free entry condition is strictly increasing in φ. We will show next that the zero profit condition is constant in φ in the Pareto case (and it is decreasing for many other distribution choices). Once we know φ^*, we can immediately recover $\tilde{\varphi}$. The rest of the economy can be solved for by imposing market clearing conditions. In particular, total income is given by Lw, which is equal to L as the wage is the numeraire. This needs to be equal to total expenditure $R = M\bar{r}$.

Note the following, we can write the ZCP as $\bar{\pi} = k(\varphi^*)f$, where $k(\varphi^*) = \left(\frac{\tilde{\varphi}(\varphi^*)}{\varphi^*}\right)^{\sigma-1} - 1$. The FE also gives us $\bar{\pi} = \frac{\delta f_e}{1-G(\varphi^*))}$. So we can plot this in the φ, π space: This tells us that there is a unique solution of the system of equations to get φ^* and $\bar{\pi}$. Next, we solve this formally.

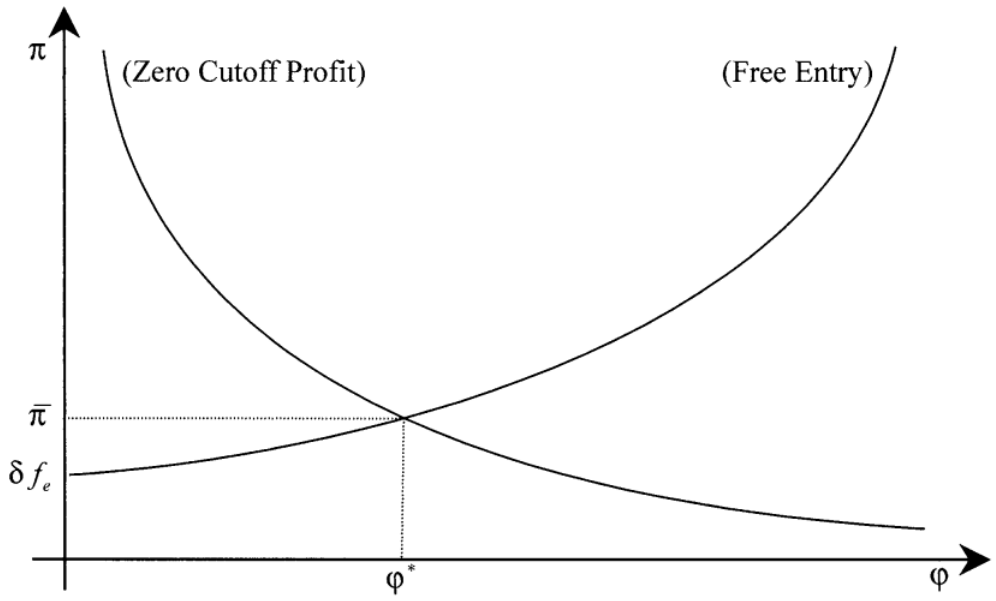


Figure 2: ZCP and FE conditions

We can start by using equation 156 and noting that profits can be written as

$$\pi(\varphi) = \left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1}\frac{R}{\sigma}\varphi^{\sigma-1} - f. \tag{174}$$

In words, it tells us that the profits of a firm with productivity φ depend on aggregate quantities and a measure of effective productivity $\varphi^{\sigma-1}$. This is represented in Figure 3.[8] Using this in the free entry condition, we can write

$$\left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1}\frac{R}{\sigma}\int_{\varphi^*}\varphi^{\sigma-1}g(\varphi)d\varphi - f\int_{\varphi^*}g(\varphi)d\varphi = \delta f_e. \tag{175}$$

At this point, we can use the zero profit cutoff condition, namely, use $\pi(\varphi^*) = 0$ to rewrite 174

$$\left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1}\frac{R}{\sigma}\varphi^{*\sigma-1} = f. \tag{176}$$

We can now divide 175 by 176 and obtain

$$\int_{\varphi^*}\left[\left(\frac{\varphi}{\varphi^*}\right)^{\sigma-1} - 1\right]g(\varphi)d\varphi = \frac{\delta f_e}{f}. \tag{177}$$

[8] I am denoting $B \equiv \left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1}\frac{R}{\sigma}$ which are just parameters and aggregate variables, so it is a constant in the $\varphi^{\sigma-1}, \pi(\varphi)$ space.

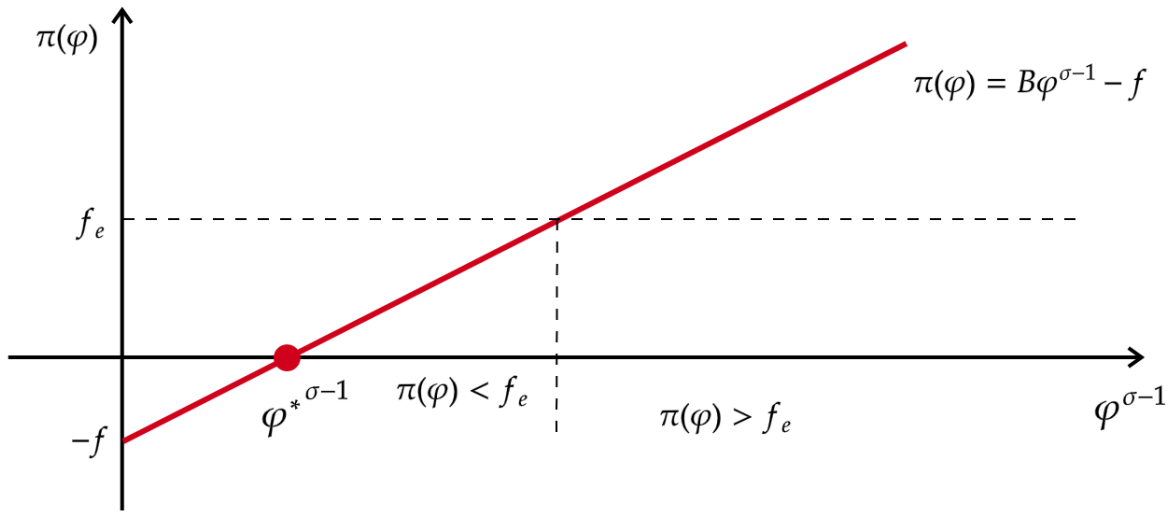


Figure 3: Closed Economy Equilibrium. Note that the x-axis is $\varphi^{\sigma-1}$ so that everything is linear.

This implicitly defines the cutoff productivity φ^*. However, we cannot proceed further without making an assumption on $g(\varphi)$. The most convenient realistic assumption we can make is that productivity follows a Pareto distribution. Namely, $G(\varphi) = 1 - \left(\frac{\varphi}{\underline{\varphi}}\right)^{-\theta}$, where $\underline{\varphi}$ is the lower bound of the support of the distribution, while the upper bound is ∞, and $\theta > \sigma - 1$ governs the shape. There are two main reasons why we make this specific assumption: i) the empirical distribution of firm size is not far from Pareto, particularly in the right tail; ii) Pareto distributions have the property that if we truncate them, they remain Pareto with the same parameter governing their shape. Note, importantly, that this assumption restricts the results significantly. In particular, you can show that under Pareto distribution, the ZCP is flat, rather than decreasing, as in Figure 2. Under this assumption we have that $g(\varphi) = \theta \underline{\varphi}^{\theta} \varphi^{-\theta-1}$. We can

plug this into the implicit definition of our cutoff and obtain

$$\begin{aligned}
\frac{\delta f_e}{f} &= \int_{\varphi^*}^{\infty} \left[\left(\frac{\varphi}{\varphi^*} \right)^{\sigma-1} - 1 \right] \theta \underline{\varphi}^{\theta} \varphi^{-\theta-1} d\varphi \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \int_{\varphi^*}^{\infty} \left[\varphi^{\sigma-1} - \varphi^{*\sigma-1} \right] \varphi^{-\theta-1} d\varphi \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[\int_{\varphi^*}^{\infty} \varphi^{\sigma-1} \varphi^{-\theta-1} d\varphi - \int_{\varphi^*}^{\infty} \varphi^{*\sigma-1} \varphi^{-\theta-1} d\varphi \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[\int_{\varphi^*}^{\infty} \varphi^{\sigma-\theta-2} d\varphi - \varphi^{*\sigma-1} \int_{\varphi^*}^{\infty} \varphi^{-\theta-1} d\varphi \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[\frac{\varphi^{\sigma-\theta-1}}{\sigma-\theta-1} \Big|_{\varphi^*}^{\infty} + \frac{\varphi^{*\sigma-1}}{\theta} \varphi^{-\theta} \Big|_{\varphi^*}^{\infty} \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[\frac{\infty^{\sigma-\theta-1} - \varphi^{*\sigma-\theta-1}}{\sigma-\theta-1} + \frac{\varphi^{*\sigma-1}}{\theta} (\infty^{-\theta} - \varphi^{*-\theta}) \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[-\frac{\varphi^{*\sigma-\theta-1}}{\sigma-\theta-1} - \frac{\varphi^{*\sigma-1}}{\theta} \varphi^{*-\theta} \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*1-\sigma} \left[\frac{-\varphi^{*\sigma-\theta-1}}{\sigma-\theta-1} - \frac{\varphi^{*\sigma-\theta-1}}{\theta} \right] \\
&= \theta \underline{\varphi}^{\theta} \varphi^{*-\theta} \left[-\frac{1}{\sigma-\theta-1} - \frac{1}{\theta} \right] \\
&= \underline{\varphi}^{\theta} \varphi^{*-\theta} \frac{\sigma-1}{\theta-(\sigma-1)}.
\end{aligned} \tag{178}$$

Hence

$$\varphi^* = \left(\frac{f}{\delta f_e} \frac{\sigma-1}{\theta-(\sigma-1)} \right)^{\frac{1}{\theta}} \underline{\varphi}. \tag{179}$$

Which is the closed-form solution for the cutoff. You can immediately note the following properties: i) if the cost of entry increases, the minimum productivity to operate decreases; ii) if the cost of operating f increases, the minimum productivity to survive goes up. From here, we could go back and get all our aggregate objects of interest in closed form.

So far, we operated in a closed economy environment, and the equilibrium was such that there was positive selection in being a firm, namely unproductive firms would prefer to stay out of the market. The next step is to introduce the international dimension of the problem. In particular, assume that there are n identical countries indexed by c so that $w_c = 1$ in all countries.

Exporting goods implies an iceberg trade cost $\tau \geq 1$. In other words, if an agent ships τ units of a good, only 1 unit reaches the destination. Furthermore, there is a fixed cost f_x that

firms need to pay to export every period.

We assume that firms can price-discriminate across markets. Much like before, firms will choose the monopolistic competition price of a constant markup over marginal cost so that domestically they will charge

$$p_d(\varphi) = \frac{\sigma}{\sigma - 1}\frac{1}{\varphi}, \tag{180}$$

while abroad, they will price at

$$p_x(\varphi) = \frac{\sigma}{\sigma - 1}\frac{\tau}{\varphi}. \tag{181}$$

Notice that the trade cost enters like a scalar increase in the marginal cost. Similarly to what we did before for the closed economy, we can then express domestic and export revenues as

$$r_d(\varphi) = R_d \left[P_d \frac{\sigma - 1}{\sigma}\varphi\right]^{\sigma-1}, \tag{182}$$

$$r_x(\varphi) = R_x \tau^{1-\sigma} \left[P_x \frac{\sigma - 1}{\sigma}\varphi\right]^{\sigma-1}. \tag{183}$$

Since all countries are symmetric, it must be that $P_d = P_x = P$ and $R_d = R_x = R = I$. Given the symmetry, we immediately recover that

$$r_x(\varphi) = \tau^{1-\sigma} r_d(\varphi). \tag{184}$$

We can write domestic and export profits as

$$\pi_d(\varphi) = \frac{r_d(\varphi)}{\sigma} - f, \tag{185}$$

$$\pi_x(\varphi) = \frac{r_x(\varphi)}{\sigma} - f_x. \tag{186}$$

We can now think about how firms select into foreign markets. Recall that the expected value of a firm with productivity φ is given by

$$v(\varphi) = \max\left\{0, \frac{\pi(\varphi)}{\delta}\right\}, \tag{187}$$

where the profits now take the form

$$\pi(\varphi) = \pi_d(\varphi) + \max\{n\pi_x(\varphi), 0\}. \tag{188}$$

Like in the closed economy, we can define a cutoff level of productivity under which firms will

not find it profitable to be active:

$$\varphi^* = \inf\left\{\varphi \geq 0 : \frac{\pi(\varphi)}{\delta} > 0\right\}. \tag{189}$$

But we can now complement this with an additional cutoff level above which the firm finds it optimal to export

$$\varphi_x^* = \inf\left\{\varphi \geq 0 : \frac{\pi_x(\varphi)}{\delta} > 0\right\}. \tag{190}$$

If we assume that $\tau^{\sigma-1} f_x > f$ we know that $\varphi_x^* > \varphi^*$. We can see this by comparing

$$\pi_d(\varphi) = \left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1} \frac{R}{\sigma}\varphi^{\sigma-1} - f, \tag{191}$$

$$\pi_x(\varphi) = \left(\frac{\sigma-1}{\sigma}P\right)^{\sigma-1} \frac{R}{\sigma}\left(\frac{\varphi}{\tau}\right)^{\sigma-1} - f_x. \tag{192}$$

Using the definition of the cutoffs, $\pi_d(\varphi^*) = 0$ and $\pi_x(\varphi_x^*) = 0$ we obtain

$$\varphi_x^* = \tau\left(\frac{f_x}{f}\right)^{\frac{1}{\sigma-1}} \varphi^*. \tag{193}$$

Under this assumption, there is a natural ordering of firms into inactivity, domestic and foreign activity. Firms below φ^* will be inactive, firms between φ^* and φ_x^* will operate only in the domestic market, and firms with productivity $\varphi > \varphi_x^*$ will sell both domestically and abroad. Note that this selection also implies that exporting firms will be larger than domestic firms. The intuition behind this latter result is that exporting implies paying an additional fixed cost. Since all firms charge the same markup, the only way exporting firms can recover their costs is by selling more units and therefore being larger. In the open economy, aggregate productivity is given by an average of the productivity of domestic and exporting firms. In particular

$$\bar{\varphi}_t = \left[\frac{1}{M_t}\left[M\tilde{\varphi}^{\sigma-1} + nM_x\left(\frac{\tilde{\varphi}_x}{\tau}\right)^{\sigma-1}\right]\right]^{\frac{1}{\sigma-1}}, \tag{194}$$

Where $M_t \equiv M + nM_x$, $M_x = \frac{1-G(\varphi_x^*)}{1-G(\varphi^*)}M$ and the productivity averages are given by

$$\tilde{\varphi} = \left[\frac{1}{1-G(\varphi^*)}\int_{\varphi^*} \varphi^{\sigma-1} g(\varphi) d\varphi\right]^{\frac{1}{\sigma-1}}, \tag{195}$$

and

$$\tilde{\varphi}_x = \left[\frac{1}{1 - G(\varphi_x^*)} \int_{\varphi_x^*} \varphi^{\sigma-1} g(\varphi) d\varphi \right]^{\frac{1}{\sigma-1}}. \tag{196}$$

These are, respectively, the average productivity across all firms and the average productivity across exporting firms. As in the closed economy, we can compute aggregate variables as a function of aggregate productivity

$$P = M_t^{\frac{1}{1-\sigma}} \frac{\sigma}{\sigma - 1} \bar{\varphi}^{-1}, \tag{197}$$

$$R = M_t r(\bar{\varphi}), \tag{198}$$

$$\Pi = M_t \pi(\bar{\varphi}), \tag{199}$$

$$Q = M_t^{\frac{\sigma}{\sigma-1}} q(\bar{\varphi}). \tag{200}$$

Note that we are heavily relying on the symmetry assumption here. In particular, the welfare of domestic consumers should depend on imported goods and, therefore, the productivity of firms producing the imported varieties. Under symmetry, the exporters from country A to country B are identical to the exporters from B to A. Therefore, we can write domestic welfare as a function of the productivity of firms exporting.

We can then introduce the free entry condition, which is unchanged,

$$\bar{\pi} = \frac{\delta f_e}{1 - G(\varphi^*)}, \tag{201}$$

with the difference that now the average profits include profits from exporting

$$\bar{\pi} = \pi_d(\bar{\varphi}) + n \frac{1 - G(\varphi_x^*)}{1 - G(\varphi^*)} \pi_x(\tilde{\varphi}_x), \tag{202}$$

where $\frac{1-G(\varphi_x^*)}{1-G(\varphi^*)}$ is the probability of exporting, conditional of successful entry. We can use again the definitions of the cutoff productivities

$$r_d(\varphi^*) = \sigma f, \tag{203}$$

$$r_x(\varphi_x^*) = \sigma f_x, \tag{204}$$

and obtain

$$\frac{r_x(\varphi_x^*)}{r_d(\varphi^*)} = \frac{f_x}{f} \Rightarrow \varphi_x^* = \varphi^* \tau \left(\frac{f_x}{f} \right)^{\frac{1}{\sigma-1}}. \tag{205}$$

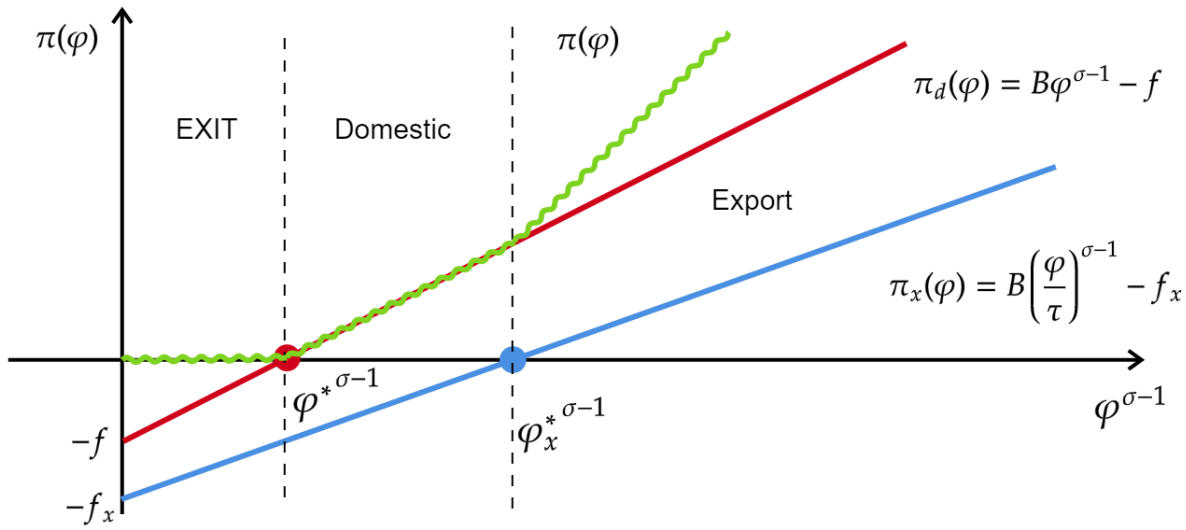


Figure 4: Open Economy Equilibrium

using the ZPC

$$\bar{\pi} = f\left[\left(\frac{\bar{\varphi}(\varphi^*)}{\varphi^*}\right)^{\sigma-1} - 1\right] + nf_x\frac{1-G(\varphi_x^*)}{1-G(\varphi^*)}\left[\left(\frac{\tilde{\varphi}_x(\varphi^*)}{\varphi_x^*(\varphi^*)}\right)^{\sigma-1} - 1\right]. \tag{206}$$

From this, we conclude that when the economy opens to trade the cutoff for entering firms moves rightwards. The intuition is that while profits increase because exporters gain from selling to other destinations, there is higher competition from foreign exporters. For domestic firms, only the latter effect operates. Opening to trade, therefore, implies that the worst firms exit as they suffer the pressure from foreign exporters, thereby increasing the average productivity of the economy. This is a selection effect which also implies a reallocation of market shares towards more productive firms. Consumers unambiguously gain because the higher competition implies a decline in their price index and because they have access to more varieties.

This model provides strong predictions on the sorting of firms. First, it tells us that the observed productivity distribution is truncated below, suggesting that we should be worried about selection issues. Next, it predicts that firms engaging in export should be more productive than firms operating only domestically. This prediction is, by construction, very stark: it states that if we split our data based on export status, there should be no overlap between their productivity distributions. Note that in this model, by the symmetry assumptions, if a firm export to one country, it will export to all. This, in turn, means that we cannot think about whether a firm in the data that exports to N countries should be more or less productive than a firm exporting to M countries.

4 Large Firms and Networks in Macro

In this section, we analyse the role of large firms in the economy. We start by stating providing an important benchmark result called the Hulten theorem. This result formalizes how idiosyncratic shocks transmit to aggregate quantities in efficient economies. Then we discuss three distinct elements of recent advances in the field of firms in macro that break that result. We start by discussing the results in Gabaix (2011). In particular, we study a setting in which, as firms are *large*, idiosyncratic shocks turn into aggregate ones. We then move to look at production networks. In this context, we show that Gabaix's notion of *large* can actually just mean that a firm or a sector is *central* in the network. Finally, we discuss the role of imperfect competition and market power, looking at the recent literature on oligopoly in macro.

4.1 Hulten Theorem[9]

We are interested in understanding how idiosyncratic shocks to firms turn into aggregate fluctuations. We start we a pretty general result called Hulten's theorem. We derive it in a multi-sector economy, where one sector can use other sectors' products as intermediate inputs, together with labor. All markets are perfectly competitive. The economy has N sectors, and the representative household provides an exogenous quantity of labor (many of these features can be generalized). The household derives utility from the consumption of the N types of goods. Solving the maximization problem:

$$\max_{\{c_i\}} U(c_1, \ldots, c_N), \quad \text{st} \tag{207}$$

$$\sum_i p_i c_i \leq w\bar{l} + \sum_i \pi_i \tag{208}$$

where p_i is the price of good i, w is the wage, and π_i is profits from sector i. The first-order conditions of the corresponding Lagrangian yield that:

$$u_{c_i} = \lambda p_i \quad \forall i, \tag{209}$$

And the Lagrange multiplier is equal to $\lambda = \frac{\partial L}{\partial I} = \frac{1}{P^*}$, where I is the household' exogenous income and P^* is the price index of the economy. If we choose the price index to be the numeraire, we can set $\lambda = 1$, so that $u_{c_i} = p_i$, $\forall i$.

Let's now look at the firm side. Firms are assumed to have a Hicksian productivity parameter A_i, so that their production function is

$$y_i = A_i F_i(l_i, x_{1i}, \ldots, x_{Ni}), \quad \forall i \tag{210}$$

[9]This presentation follows Baqaee and Farhi (2019) Appendix A and the notes of Toshihiko Mukoyama.

where i is labor input at sector i and x_{ji} is the quantity of product j used in sector i. The corresponding profit is

$$\pi_i = p_i y_i - w l_i - \sum_j p_j x_{ji}, \quad \forall i. \tag{211}$$

In such an economy, some of the output is used as input by other firms. Therefore, GDP is not the sum of output but the sum of value added, which is equivalent to the sum of consumption

$$Y = \sum_i p_i c_i. \tag{212}$$

Finally, the market clearing conditions are

$$L = \sum_i l_i \tag{213}$$

$$y_i = c_i + \sum_j p_j x_{ij}. \tag{214}$$

From the definition of GDP, we can replace prices by marginal utilities and obtain

$$Y = \sum_i u_{c_i} c_i = U(c_1, \ldots, c_N) \tag{215}$$

where the last equality is obtained from assuming U to be homogeneous of degree 1[10].

The planner problem in such an economy is

$$\max_{\{c_i\},\{l_i\},\{x_{ij}\}} U(c_1, \ldots, c_N), \quad \text{st} \tag{216}$$

$$c_i + \sum_j x_{ij} \leq A_i F_i(l_i, x_{1i}, \ldots, x_{Ni}), \quad \forall i \qquad (\mu_i)$$

$$\sum_i l_i \leq L, \qquad (\nu)$$

Given the setup, the First Welfare Theorem holds, meaning that the competitive equilibrium is Pareto-efficient and corresponds to the solution of the planner's problem. The Lagrangian is

$$\mathcal{L} = U + \sum_i \mu_i \big(A_i F_i(l_i, x_{1i}, \ldots, x_{Ni}) - c_i - \sum_j x_{ij} \big) + \nu \big(L - \sum_i l_i \big) \tag{217}$$

[10]This convenient assumption implies that preferences are homothetic: whatever the income level of the household, the proportion spent on goods from sector i is always the same.

with the following first-order conditions:

$$\frac{\partial U}{\partial c_i} = \mu_i \tag{218}$$

$$\frac{\partial U}{\partial A_i} = \mu_i F_i(l_i, x_{1i}, \ldots, x_{Ni}) = \mu_i \frac{y_i}{A_i}. \tag{219}$$

Combining the last two conditions, we have

$$\frac{\partial U}{\partial A_i} = \mu_i \frac{y_i}{A_i} = \frac{\partial U}{\partial c_i}\frac{y_i}{A_i} = c_i' \frac{y_i}{A_i} = \frac{p_i y_i}{A_i}. \tag{220}$$

where this equality comes from the fact that we are operating in an efficient economy where the Welfare Theorems hold and, therefore, the consumption allocations will be the same in the competitive equilibrium and planner solution. If in the competitive equilibrium $c_i' = p_i$ and in the planner solution $c_i' = \mu_i$, given that c_i' is the same in the two, it follows that $p_i = \mu_i$. Therefore,

$$\frac{\partial Y}{\partial A_i} = \frac{\partial U}{\partial A_i} = \frac{p_i y_i}{A_i} \tag{221}$$

which we can rewrite as

$$\frac{\partial Y}{Y} = \underbrace{\frac{p_i y_i}{Y}}_{\text{Domar weight}=\frac{sales_i}{GDP}} \frac{\partial A_i}{A_i}. \tag{222}$$

This last equation means that productivity shocks to one sector affect aggregate output proportionally to the share of that sector in the economy, as measured by the Domar weights, which represent a variation of sales share (because the denominator is GDP and not the sum of all sales, their sum may not be equal to one).

Lucas (1977) Irrelevance of Micro Shocks Consider an economy where the Hulten theorem holds. Suppose that firms have a common variance of shocks σ. Then, the standard deviation of GDP growth is given by

$$\sigma_{GDP} = (Var(d \log Y))^{\frac{1}{2}} = \left(\sum_i s_i^2\right)^{\frac{1}{2}} \sigma. \tag{223}$$

If there is a large number of firms N of similar sizes, the term $\left(\sum_i s_i^2\right)^{\frac{1}{2}}$ is of order $1/N$. Lucas (1977) argues that, given the large number of firms in the U.S., idiosyncratic shocks vanish extremely quickly by a diversification argument. For every firm that gets a positive shock, there

will be a firm that gets an equal but negative shock, and the two wash out. Lucas (1977)'s conclusion is that if we want to understand aggregate fluctuations, we do not need to think about idiosyncratic shocks, all that actually matters is aggregate shocks. See also Dupor (1999) for the same type of argument in multi-sector models.

In what follows, we ask how robust is this result. First, we study Gabaix (2011), who argues that under fat-tailed first size distribution Lucas (1977)'s argument breaks.

4.2 Large Firms and Granularity[11]

Throughout the course, we looked at economies where firms were small. The definition of small was that no individual firm's choice or shock were able to move aggregate quantities on its own. This notion comes from the assumption of a continuum of firms. The underlying statistical result is basically the law of large numbers. A key deviation from this result is formalized in Gabaix (2011). I would suggest reading the paper as it is beautiful and extremely clear (the proofs are quite hard). The fundamental idea is quite simple: some economies are dominated by extremely large firms. The paper provides the example of Nokia in the 2000s. Nokia's sales represented 26% of Finnish GDP. Our models so far cannot account for the fact that if Nokia has a good year so does Finland.[12] To formalize this notion, we start with an extremely simple islands economy.

We study a setting in which the economy has N firms. Production is given by an endowment, and firm i produces s_{it} of the good. A firm's stochastic endowment has a growth rate of $\sigma_i \epsilon_{i,t+1}$, where σ_i is the volatility scalar, and ϵ is distributed according to $F(0,1)$. GDP is simply given by

$$y_t = \sum_{i=1}^{N} s_{it}, \tag{224}$$

which immediately implies that GDP growth is

$$\frac{\Delta y_{t+1}}{y_t} = \frac{1}{y_t} \sum_{i=1}^{N} \Delta s_{it+1} = \sum_{i=1}^{N} \sigma_i \frac{s_{it}}{y_t} \epsilon_{it+1}. \tag{225}$$

[11] For a deeper discussion on this, see the review article by Gabaix (2016) on the role of Power Laws in economics.

[12] Di Giovanni and Levchenko (2012) find "In Korea, the 10 biggest business groups account for 54% of GDP and 51% of total exports. . . . The largest one, Samsung, is responsible for 23% of exports and 14% of GDP".

As shocks are iid, the variance of GDP growth is $\sigma_y = \left(\text{var}\frac{\Delta y_{t+1}}{y_t}\right)^{1/2}$. Hence

$$\sigma_y = \left(\sum_{i=1}^{N} \sigma_i^2 \left(\frac{s_{it}}{y_t}\right)^2\right)^{1/2}. \tag{226}$$

Note that if firms have symmetric volatilities $\sigma_i = \sigma$, then

$$\sigma_y = \sigma h, \tag{227}$$

with $h = \left(\sum_{i=1}^{N} \left(\frac{s_{it}}{y_t}\right)^2\right)^{1/2}$ being the Herfindhal index of the economy.

From here we can start deriving results building up to how idiosyncratic shocks affect aggregates. First, suppose firms are equally sized $1/N$ and have symmetric volatilities. Then

$$\sigma_y = \frac{\sigma}{\sqrt{N}}. \tag{228}$$

This is the standard result that follows from the law of large numbers. Gabaix goes on by saying that he estimates that $\sigma \approx 12\%$ for US firms and that there are $N = 10^6$ firms in the US. Then

$$\hat{\sigma}_y \approx 0.012\% \text{ per year}. \tag{229}$$

The aggregate measured volatility of US GDP is approximately 1%, so we are off by orders of magnitude. More generally, the paper shows that even if the firms are not equally sized, as long as their size is drawn from a finite variance distribution, GDP volatility will have a $\sqrt{N}$ scaling.

Gabaix then proceeds to study what happens to this result when we allow firm size to be drawn from a power law. He shows that depending on how thick the tail of the power law is, we get different convergence rates. In particular, suppose size s is such that $P(s > x) = ax^{-\zeta}$, for $x > a^{1/\zeta}$ and $\zeta \geq 1$, then, as $N \to \infty$,

$$\begin{aligned} \sigma_y &\sim \frac{\nu_\zeta}{\ln N} && \zeta = 1 \\ \sigma_y &\sim \frac{\nu_\zeta}{N^{1-1/\zeta}} && 1 < \zeta < 2 \\ \sigma_y &\sim \frac{\nu_\zeta}{N^{1/2}} && \zeta \geq 2, \end{aligned} \tag{230}$$

where ν_ζ is a random variable whose distribution is independent of N and σ. The proof of this result and its technical aspects are beyond what we want to understand here. What is of interest is what that means for the economy. First, note that $\zeta = 1$ is a so-called Zipf distribution. This distribution has a very fat tail, which means that there are extremely large firms in this economy. Gabaix tells us that if this is the case, then idiosyncratic shocks decay at a much slower rate. So

if we take the numbers from before: $N = 10^6$, we draw from a Zipf distribution, and we get a median Herfindhal of $h = 12\%$, with $\sigma \approx 12\%$ we get $\sigma_y \approx 1.4\%$. So now we are much closer to the observed volatility without using any aggregate shock. The paper then considers many extensions, including one in which the volatility of a firm depends on its size. The last part of the paper looks at the data and uses the model-based decomposition to claim that about 1/3 of the US GDP volatility can be explained by shocks to the top 100 US firms.

Measuring Granularity To analyse the problem empirically, we can extend our basic island endowment economy to a simple production structure. Suppose firms produce using $y_{it} = e^{z_{it}} l_{it}$ where z_{it} is the firm productivity. This implies that $z_{it} = \ln \frac{y_{it}}{l_{it}}$, in words, it is log output per worker. Assume further that productivity moves over time such that $g_{it} = z_{it} - z_{it-1}$. We can think of g as the change in a firm's productivity. Part of this change might be predictable based on observed characteristics, for example, if a firm invested a lot in R&D, it might have a higher productivity growth in the future. To get at the unpredictable part, we can postulate an empirical model $g_{it} = \beta' X_{it} + \epsilon_{it}$. In this case, ϵ_{it} is the *shock*. We can estimate the regression and compute the shock as $\hat{\epsilon}_{it} = g_{it} - \hat{g}_{it}$. From here, we can build the *granular residual*

$$\Gamma_t \equiv \sum_{i=1}^{K} \frac{s_{it-1}}{y_{t-1}} \hat{\epsilon}_{it}, \tag{231}$$

for the biggest K firms in the economy. A special case for example is using $X_i t = \bar{g}_t = \frac{1}{Q} \sum_{i=1}^{Q} g_{it}$, namely the average of the Q largest firms growth rates.[13] In this case, we would then have

$$\Gamma_t \equiv \sum_{i=1}^{K} \frac{s_{it-1}}{y_{t-1}} (g_{it} - \bar{g}_t). \tag{232}$$

Using this formulation Gabaix (2011) shows that the granular residual for $K = 100$ explains between 25 and 30% of GDP growth and the Solow Residual of the US economy.

Without going into further detail about the topic of granularity, note that this paper sparked a large stream of research. Some of the most recent contributions also noted that, for example, as trade is mostly carried out by large firms, foreign shocks tend to be granular in nature for domestic economies (see Di Giovanni, Levchenko and Mejean, 2020). Along similar lines di Giovanni and Levchenko (2009) use a combination of a Melitz and Gabaix model to show that small open countries are much more volatile than larger countries as trade makes the firms bigger and, therefore the economy more granular. Carvalho and Gabaix (2013) show that the granular hypothesis, together with long-run changes in the sectoral composition of the US economy, can explain the Great Moderation and the subsequent rise in GDP volatility. In particular, they argue this is driven by the secular decline of manufacturing and the rise of

[13] We could also restrict these to be in the same sector as the firm itself.

the financial sector. Finally, Burstein et al. (2019) use a granular oligopoly model to show that if business cycle is driven by shocks to the largest firms, then the cyclicality of markups depends on whether we study it at the firm, sector or aggregate level.

Digression on Power Laws, Properties and Genesis[14] The main result of Gabaix (2011) relies on the notion that firm size is distributed according to a Power Law. To better understand this, we start by defining this class of distributions. Importantly, these are empirical regularity of many interesting economic phenomena such as the distribution of city size, the distribution of firm size, and the distributions of income and wealth in the population.

Define a power law as a relation of the type $Y = aX^{\beta}$. A common way to detect power laws in the data is to estimate

$$\log(Rank_X) = \log a - \beta \log X \tag{233}$$

Where $Rank_y$ is the rank of observation Y in terms of order statistics (the unit with the highest value of Y has rank 1). This relationship gives us the power law tail index $\hat{\beta}$. This regression, when applied to the size of city gives a $\hat{\beta} = 1.03$, while for firm size $\hat{\beta} = 1.06$. For wealth $\hat{\beta} \approx 1.5$ and for income between 1.5 and 3.

A natural question is how to rationalize the existence of power law distributions in the data. It turns out that to generate this kind of empirical observation, we need something like proportional random growth. This model is one in which all firms have the same expected growth rate and std. dev of growth rate. Absent additional elements, the firm size distribution explodes (variance grows unbounded), but if we add frictions so that a stationary distribution exists, then this distribution is a power law. This result tells us that we can obtain power laws from the repeated assignment of random growth shocks, but, importantly, the exponent need not be 1. To obtain Zipf's Law, we need something bounding the system. For example, suppose that we are interested in a proportional random growth model of city size. If the total size of the population to be assigned to cities is fixed, then the exponent in the power law of city size goes to 1.

While this explanation is successful at generating the empirical distribution, it is clearly unsatisfactory as a *theory*. More interesting theories underlying power law limit distributions have to do with matching and supermodular assignment. As in the previous digression on assortative matching, think of a problem with two-sided heterogeneity. For example, think about firms and managers. Under positive assortative matching, the best manager is matched with the best firm. This assignment process generates a complementarity between the qualities of the two parties, such that if we look at the effect of increasing a firm's productivity, it will be more than 1-to-1 as it will also improve the quality of the manager they attract. The

[14] This digression is partially based on the review paper by Gabaix (2016).

limit distribution of this problem (provided something that bounds it, to ensure existence and finiteness of the limit distribution) is a power law. Similarly, richer cities attract highly educated individuals. There are many examples of these supermodular assignment problems, the basic theory of which is laid out in Rosen (1981).

End of Digression

4.3 Network Economies

Gabaix (2011) is titled "The Granular Origins of Aggregate Fluctuations". A year later, Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012)'s "The Network Origins of Aggregate Fluctuations" came out. This equally seminal paper argues that, while Gabaix uses a fat-tailed distribution of size as the underlying force in his idiosyncratic to aggregate mechanism, an alternative explanation (or a source of size in Gabaix's economy) can be found in input-output linkages.

In this section, we first go through the model, we characterize a fundamental benchmark result about aggregation called the Hulten Theorem and discuss how input-output links can generate granular economies. Finally, we look at volatility and weakest-link or O-ring production approaches.

4.3.1 Cobb-Douglas Network Economies

We start from the workhorse production network in macro model. Note that the exact formulation of this model is a mix of Acemoglu, Akcigit and Kerr (2016) and Carvalho and Tahbaz-Salehi (2019).[15] There is a representative consumer with 1 unit of labour, supplied inelastically. The consumer aggregates goods by

$$u(\underline{c}) = \sum_{i=1}^{n} \beta_i \ln(c_i/\beta_i). \tag{234}$$

Note that these preferences are equivalent to a Cobb-Douglas aggregator over consumption goods. We also normalize $\sum_i \beta_i = 1$. Competitive firms produce goods used both for consumption and as intermediates by other sectors through the CRS production function

$$y_i = z_i \Gamma_i l_i^{\alpha_i} \prod_{j=1}^{n} x_{ji}^{a_{ji}}, \tag{235}$$

where l_i is labour used by firm i, $\alpha_i \in (0,1)$ is the labour share and x_{ji} is output of firm j used to produce output of firm i. Γ_i is a normalization constant and z_i is a productivity shock iid across firms with $\epsilon_i \equiv \ln z_i \sim F_i$. For simplicity we assume CRS, namely $\alpha_i + \sum_j a_{ji} = 1$, $\forall i$. We call the

[15] These are to some extent, review articles which are a very good starting point to think about networks in macro, in particular, Carvalho and Tahbaz-Salehi (2019).

collection $\mathcal{A}$ of a_{ji} the input-output matrix of the economy. Further, call $d_i \equiv \sum_j a_{ij}$ the outdegree of firm i. Intuitively this measures how important firm i is in the production of all other sectors j, but only directly. Importantly this is different from the indegree $in_i = \sum_j a_{ji} = 1 - \alpha_i$ (which measures the reliance of firm i on other sector's production and always lies between 0 and 1). Clearly, we are going to study changes in z_i and how they propagate in the economy.

With this in mind, market clearing takes the form

$$y_i = c_i + \sum_j x_{ij}. \tag{236}$$

We start by noting that the minimization of expenditure for firms implies that $p_j x_{ji} = a_{ji} p_i y_i$ and $wl_i = \alpha_i p_i y_i$. Further, by a similar argument on the optimization of consumers we get that, calling E the total expenditure on consumption, the expenditure on good i is given by $p_i c_i = \beta_i E$. Using these results in the market clearing condition from equation (236) (after we multiply both sides by p_i) we get

$$p_i y_i = \beta_i E + p_i \sum_j a_{ij} \frac{p_j y_j}{p_i} = \beta_i E + \sum_j a_{ij} p_j y_j. \tag{237}$$

Solving the model it is also possible to show that, denoting $\lambda_i = \frac{p_i y_i}{GDP}$ firm i's sales share or Domar weight are equal to

$$\lambda_i = \beta_i + \sum_{j=1}^{n} a_{ij} \lambda_j. \tag{238}$$

This first important result states that a firm's sector share over GDP is given by its importance in consumer baskets plus its customers' sales share weighted by the appropriate connection a_{ij}. Upon noting that we have as many such conditions as firms in the economy we can denote Λ the vector of Domar weights and B the vector of βs. Writing this system of equation in matrix form we get $\Lambda = B + \mathcal{A}\Lambda$, which we can solve to get

$$\Lambda = [I - \mathcal{A}]^{-1} B. \tag{239}$$

We can now introduce another important piece of notation: denote $L = [I - \mathcal{A}]^{-1}$ the Leontief inverse. Elements ℓ_{ij} of this matrix measure how important a firm i is for a firm j through both direct and indirect connections. It is trivial to show that in this economy the spectral radius of

L is strictly less than 1 (because of DRS on reproducible inputs),[16] therefore we can express it as

$$L = [I - \mathcal{A}]^{-1} = \sum_{j=0}^{\infty} \mathcal{A}^j. \tag{240}$$

This representation is insightful because it tells us that the first-degree connections are given by the matrix $\mathcal{A}$, the second degree by $\mathcal{A}^2$ and so on. Once we sum over them all, we get the total importance of a firm in the economy.

This allows us to solve the Domar weights λs as

$$\lambda_i = \sum_{j=1}^{n} \ell_{ij} \beta_j. \tag{241}$$

Careful that on the RHS we have the element of the Leontief inverse. What this tells us is that a firm's sales share depends on how important this firm is for other firms, weighted by how important these firms are for consumers.

What we have so far is a relationship between sales shares and primitives given by the Input-Output matrix and the consumer weights. Our goal however is to study the effect of productivity shocks on GDP. To do so we need to solve for prices so that we can get real output from the sales we have already figured out. To this end, start with the firm's first order conditions and the production function. Using our results on x_{ji} and l_i into the production function we obtain

$$y_i = z_i \Gamma_i \left(\frac{\alpha p_i y_i}{w} \right)^{\alpha_i} \prod_j \left(\frac{a_{ji} p_i y_i}{p_j} \right)^{a_{ji}}. \tag{242}$$

From here we note that, since $\alpha_i + \sum_j a_{ji} = 1$, output y_i cancels out. This is not surprising since this firm has CRS technology, which implies that its size is indeterminate from the supply side. Solving for the price we get

$$p_i^{-1} = z_i \Gamma_i \left(\frac{\alpha_i}{w} \right)^{\alpha_i} \prod_j \left(\frac{a_{ji}}{p_j} \right)^{a_{ji}}. \tag{243}$$

This condition tells us that the price of each firm is a function of its productivity and the appropriately weighted prices of its suppliers (which enter as part of the marginal cost). Again this is not surprising, these are competitive firms, so they price at marginal cost. What we have on the RHS is nothing but the optimal marginal cost, where optimal is to be understood as the marginal cost associated with the optimal input mix from the expenditure minimization

[16] This is the matrix version of the condition on the convergence of geometric series with roots in the unit circle you are probably familiar with.

problem.

We can go further to solve this problem. First, we take logs of equation (243). Note here the convenience of the Cobb-Douglas assumption: as the production function and therefore the price is multiplicative, it becomes additive in logs. This allows us to solve it as a linear system again. First, note that in logs

$$\ln p_i = -\epsilon_i + \alpha_i \ln w + \sum_{j=1}^{n} a_{ji} p_j - \ln \Gamma_i - \Delta_i, \tag{244}$$

where I have collected into $\Delta_i = \alpha_i \ln \alpha_i + \sum_{j=1}^{n} a_{ji} \ln a_{ji}$. These are just constants that will not change as we shock the economy. At this point we pick $\Gamma_i = exp(-\Delta_i)$, which is just a free constant, to get rid of these parameters. Then we get

$$\ln p_i = -\epsilon_i + \alpha_i \ln w + \sum_{j=1}^{n} a_{ji} p_j \tag{245}$$

We can use again the fact that $\alpha_i + \sum_{j=1}^{n} a_{ji} = 1$ to get

$$\ln(p_i/w) = -\epsilon_i + \sum_{j=1}^{n} a_{ji} \ln(p_j/w). \tag{246}$$

This is promising because we can define some vector of relative prices to the wage $\hat{P}$ of which p_i/w are elements and solve the linear system

$$\hat{P} = -\epsilon + A'\hat{P} \tag{247}$$

to get

$$\hat{P} = -[I - A']^{-1}\epsilon. \tag{248}$$

Note two things: first, we have successfully solved for relative prices as a function of primitives; secondly, this condition tells us that relative prices are nothing but appropriately weighted productivities. This comes through by thinking that if some firm j becomes more productive, some other firm i will reduce its price if j is a direct or indirect supplier of i. This occurs because as j becomes more productive, the marginal cost of i decreases and, by marginal cost pricing, so does its price.

We are almost there. We still want to figure out GDP. To that end, note that this new result allows us to write $\ln(p_i/w) = -\sum_{j=1}^{n} \epsilon_j \ell_{ji}$, using the definition of the Leontief Inverse. Then note that GDP in this economy is by definition equal to the wage. This is trivially true upon noting that profits are zero and therefore all income (value added) is paid out to workers. As the

population is normalized to 1, $GDP = w$. It follows that we can rewrite $p_i/GDP = -\sum_j \epsilon_j \ell_{ji}$. Recall that from before we had $p_i y_i/GDP = \sum_{j=1}^n \ell_{ij}\beta_j$, taking logs

$$\ln y_i + \ln(p_i/GDP) = \ln(\sum_{j=1}^{n} \ell_{ij}\beta_j). \tag{249}$$

Note that the right-hand side is made only of parameters. Hence we can ignore it when computing the economy's response to productivity shocks. We now combine this with $\ln(p_i/GDP)$ to get

$$\ln y_i = \sum_j \epsilon_j \ell_{ji} + C \tag{250}$$

where C collects these constants we do not care about. With this in mind, we can now show the main result of the competitive equilibrium. We study a change in productivity of all firms, denoted $d \ln z$. The effect of this productivity shock on the output of firm i is

$$d \ln y_i = d \ln z_i + \sum_{j=1}^{n} (\ell_{ji} - \mathbb{1}_{i=j}) d \ln z_j. \tag{251}$$

First, note that the first term gives us the direct effect of a productivity change of the firm itself. The summation represents instead network effects. It is trivial to show that if a firm does not use any input from other firms, then $\ell = 0$ and the second term vanishes.
The result can be reformulated as:

$$\frac{d \ln y_i}{d \ln z_j} = \ell_{ji}. \tag{252}$$

The response of a firm i to a productivity shock of some other firm j only depends on their direct or indirect connections from j to i, as summarized by ℓ_{ji}. Note that in this economy, productivity shocks only travel downstream (you can also show that demand shocks only travel upstream). This is a result of the Cobb-Douglas assumption and the fact that expenditure shares are constant.

We have now studied the individual firm response to a productivity shock anywhere in the network. We now want to get GDP and study the aggregate impact of idiosyncratic shocks. Start from the optimal relative price in logs

$$\ln(p_i/w) = -\sum_{j=1}^{n} \epsilon_j \ell_{ji}, \tag{253}$$

we can multiply by β_i and sum over i.

$$\sum_i \beta_i \ln p_i - \underbrace{\sum_i \beta_i \ln w}_{=1} = -\sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_j \ell_{ji} \beta_i, \tag{254}$$

Note that the first term on the left-hand side is the log of the price index $P = \prod_i p_i^{\beta_i}$, which is the optimal price index for consumers and which we now use as a numeraire and normalize to 1. This immediately implies that the first term becomes zero as it is the log of 1. By changing the sign and switching the indexes, we get

$$\ln GDP = \sum_{i=1}^{n} \epsilon_i \underbrace{\sum_{j=1}^{n} \ell_{ij} \beta_j}_{\lambda_i}. \tag{255}$$

What this tells us is that the importance of a firm in the economy is given by its position in the network. Here the sales distribution comes from the optimal choice of firms on how to source their inputs and by the input-output structure of the economy.

Note that the weight of the firm is still given by its sales share λ_i. This is also the statement of the Hulten Theorem. Formally the theorem states that

$$\frac{d \ln GDP}{d \ln z_i} = \lambda_i. \tag{256}$$

Recent contributions in this literature show that the Hulten Theorem actually holds exactly only in the Cobb-Douglas case. The reason behind this is that with Cobb-Douglas aggregators the sales shares distribution is constant. With other aggregators, for example, CES, Baqaee and Farhi (2019) show that the theorem only holds as a first order approximation, while when we go to higher orders substitution patterns kick in and the sales shares distribution moves with the idiosyncratic shocks. Also, note that the property that productivity shocks only travel downstream (and demand shocks upstream) is only true in the Cobb-Douglas case.

From here we can go back to discussing aggregate volatility of this economy based on what the input-output structure looks like. Intuitively, we will be able to get back the same logic we had for Gabaix (2011) into this setting.

4.3.2 Volatility and Granularity in Network Economies

The economy derived so far has the same aggregate properties as the one in Gabaix (2011). We could ignore the existence of Input-Output network and just use the Hulten theorem as a starting point. Importanlty, however, this economy has a lot more to say about micro moments

such as the comovement between industries.

We can now go back to discussing the volatility and granularity argument in the context of network economies. The key starting point is the definition of log GDP as sales share weighted sum of shocks. From there, assuming $\alpha_i = \alpha \, \forall i$ for simplicity, we immediately get

$$\sigma_{\ln GDP} = \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{1/2} \sigma, \tag{257}$$

where σ is the volatility of the shocks in logs. This is exactly what we had in the basic version of Gabaix (2011). We can now go further from here. Start by noting that $\sum_i \lambda_i = 1/\alpha$, which is intuitive, as this is the ratio between sales and value added. Then we can rewrite the equation above as

$$\sigma_{\ln GDP} = \sigma \left(\sum_{i=1}^{n} \lambda_i^2 \pm \frac{1}{n} \left(\sum_{i=1}^{n} \lambda_i \right)^2 \right)^{1/2} \tag{258}$$

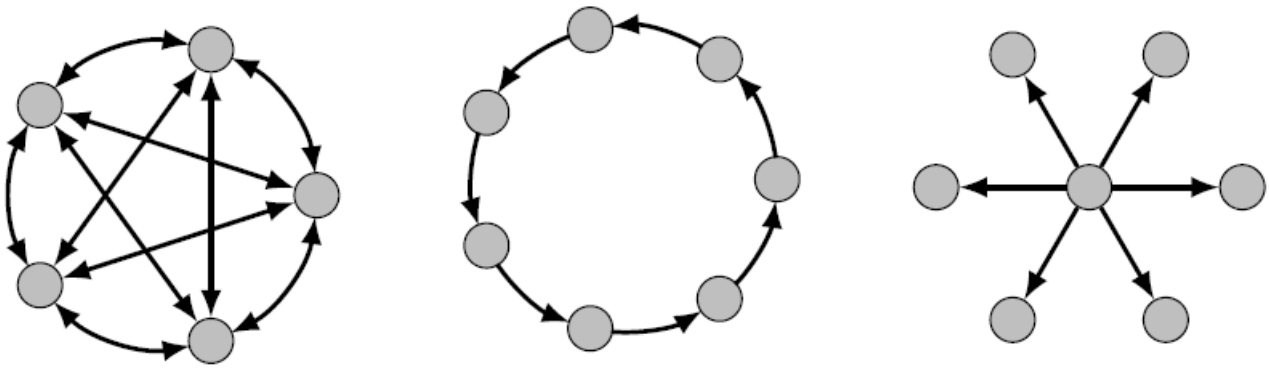
$$= \sigma \left(\frac{1}{n\alpha^2} + n \text{var}(\underline{\lambda}) \right)^{1/2} \tag{259}$$

$$= \frac{\sigma}{\alpha\sqrt{n}} \left(1 + n^2 \alpha^2 \text{var}(\underline{\lambda}) \right)^{1/2}. \tag{260}$$

From this formulation, we note a number of important insights. First, if the Domar weights are symmetric (i.e., $\text{var}(\underline{\lambda}) = 0$), aggregate volatility simply scales with $\sqrt{n}$. Second, aggregate volatility depends on the heterogeneity of the Domar weights distribution. Further, you can show that if the Domar weights follow a Pareto distribution with exponent $\gamma \in (1, 2)$, then aggregate volatility is proportional to $n^{1/\gamma - 1}$ as $n \to \infty$, which is exactly the same result as in Gabaix (2011).

To better understand the intuition, consider the three network economies in Figure 5. Economy (a) is a fully connected graph where all firms are connected to all other firms. The density of this network (the percentage of connections realized out of potential connections) is 1. Since all the connections have equal weight, all firms have the same Domar weight and the variance of λ is equal to 0. This economy has little scope for idiosyncratic shocks to turn into aggregate fluctuations since, for each firm which receives a large negative shocks there is bound to be another firm which receives the identical positive shock. Since all firms have the same weight these shocks wash out. In economy (b) we have a similar structure but the network is not fully connected. Nonetheless, since all connections have the same weight, shocks again was out quickly as we increase the number of edges in the graph. Finally, consider economy (c). This is the most asymmetric network possible since there is only one firm with high centrality and all others are peripheral. In this economy, shocks to peripheral nodes will wash out, however,

shocks to the central firm will not be counterbalanced by any other firm. The variance of λ in this economy is the highest possible and therefore so is the variance of GDP growth.



(a) $\text{var}(\underline{\lambda}) = 0$
$\sigma_{\ln GDP} \sim 1/\sqrt{n}$

(b) $\text{var}(\underline{\lambda}) = 0$
$\sigma_{\ln GDP} \sim 1/\sqrt{n}$

(c) $\text{var}(\underline{\lambda}) \sim n$
$\sigma_{\ln GDP} \sim 1$

Figure 5: Network Economies

As a final remark, note the following comparative statics. Consider two input-requirement matrices $\mathcal{A}$ and $\tilde{\mathcal{A}}$ of dimension n, such that $\lambda_i = \tilde{\lambda}_i, \forall i$ and $\alpha_i = \tilde{\alpha}_i = \alpha, \forall i$. Namely, the economies described by these input requirement matrices have the same distribution of Domar weights, and all firms have the same output labour elasticity. We say that $\tilde{\mathcal{A}}$ is more connected than $\mathcal{A}$ if $\tilde{a}_{ij} = \gamma a_{ij} + (1-\gamma)\frac{1-\alpha}{n}, \ \forall i, j$ for some $\gamma \in (0,1)$. In words, this says that the elements of $\tilde{\mathcal{A}}$ are more evenly distributed than those of $\mathcal{A}$. We say that $\tilde{\mathcal{A}}$ is more interconnected than $\mathcal{A}$. Then, we can show (see Carvalho and Tahbaz-Salehi, 2019) that a) the average pairwise correlation of (log) outputs is higher in the more interconnected economy, and b) the industries in the less interconnected economy are more volatile. In other words, a) says that the more interconnected economy features higher comovement between industries, while b) states that in the less interconnected economy, there is less diversification of shocks and, therefore, industries are more volatile.

4.3.3 Measuring Production Networks

Through the lens of the model in this section, we know how theory suggests we should look at the production network empirically. In practice, many countries now provide Input-Output data at some level of aggregation. A comprehensive example of this is the World Input-Output Database, (see Timmer et al., 2015), which includes I-O data for 44 countries over 15 years. This data typically comes in the following format:

The World Input-Output Table represents a world economy with J countries and S industries per country. The $(S \times J)$ by $(S \times J)$ matrix whose entries are denoted by Z represents flows of output used by other industries as intermediate inputs. Specifically, Z_{ij}^{rs} denotes the value of

Figure 6: World Input Output Table

			Input use & value added							Final use			Total use
			Country 1			⋯	Country J			Country 1	⋯	Country J	
			Industry 1	⋯	Industry S	⋯	Industry 1	⋯	Industry S				
Intermediate inputs supplied	Country 1	Industry 1	Z_{11}^{11}	⋯	Z_{11}^{1S}	⋯	Z_{1J}^{11}	⋯	Z_{1J}^{1S}	F_{11}^{1}	⋯	F_{1J}^{1}	Y_1^1
		⋯	⋯	Z_{11}^{rs}	⋯	⋯	⋯	Z_{1J}^{rs}	⋯	⋯	⋯	⋯	⋯
		Industry S	Z_{11}^{S1}	⋯	Z_{11}^{SS}	⋯	Z_{1J}^{S1}	⋯	Z_{1J}^{SS}	F_{11}^{S}	⋯	F_{1J}^{S}	Y_1^S
	⋯	⋯	⋯	⋯	⋯	Z_{ij}^{rs}	⋯	⋯	⋯	⋯	F_{ij}^{r}	⋯	Y_i^r
	Country J	Industry 1	Z_{J1}^{11}	⋯	Z_{J1}^{1S}	⋯	Z_{JJ}^{11}	⋯	Z_{JJ}^{1S}	F_{J1}^{1}	⋯	F_{JJ}^{1}	Y_J^1
		⋯	⋯	Z_{J1}^{rs}	⋯	⋯	⋯	Z_{JJ}^{rs}	⋯	⋯	⋯	⋯	⋯
		Industry S	Z_{J1}^{S1}	⋯	Z_{J1}^{SS}	⋯	Z_{JJ}^{S1}	⋯	Z_{JJ}^{SS}	F_{J1}^{S}	⋯	F_{JJ}^{S}	Y_J^S
Value added			VA_1^1	⋯	VA_1^S	VA_j^s	VA_J^1	⋯	VA_J^S				
Gross output			Y_1^1	⋯	Y_1^S	Y_j^s	Y_J^1	⋯	Y_J^S	⋯			

output of industry r in country i used as intermediate input by industry s in country j. In addition to the square matrix of input use, the table provides the flows of output used for final consumption. These are denoted by F_{ij}^r, representing the value of output of industry r in country i consumed by households, government and non-profit organizations in country j. Following the literature, I denote $F_i^r = \sum_j F_{ij}^r$, namely the value of output of sector r in country i consumed in any country in the world. By the definition of output, all rows sum to the total production of an industry. Finally, the table provides a row vector of value added for every industry, which implies that columns, too, sum to sectoral output.

From this matrix, we can start computing some of the objects we specified in the theory. First, our Cobb-Douglas assumption implies that $a_{ji}p_iy_i = p_jx_{ji}$. The data in the I-O table is already in sales terms, so it includes the prices. So we know the value of output of an industry, denote Y_i^r in the table, and the value of sales from each other industry denoted Z_{sr}^{ji}. We can immediately compute $a_{ji}^{sr} = Z_{sr}^{ji}/Y_i^r, \forall r, s, i, j$. In turn, this implies that we now know the whole input requirement matrix $\mathcal{A}$. We can then compute a number of important measures of the network of specific industries.

A first important metric to understand how the network is shaped is given by its density. The density of the network "counts" how many connections are active. This is tantamount to calculating

$$density = \frac{1}{(S \times J)^2} \sum_{ijrs} \mathbb{1}\{a_{ij}^{rs} > 0\} \tag{261}$$

A high density tells us that the network has most potential links active, while a low density (high sparsity) says the network has few active links. Note, importantly, that the density of a network will significantly depend on the level of aggregation. For example, firm-level networks are typically very sparse since firms are typically connected to a very small subset of other firms. On the other hand, the I-O network of "macro-sectors" features virtually no empty cells.

To further study what a production network looks like, we can compute the distribution of degrees and centrality. First, define the indegree and outdegree as

$$indegree_j^s = \sum_i \sum_r a_{ij}^{rs}, \tag{262}$$

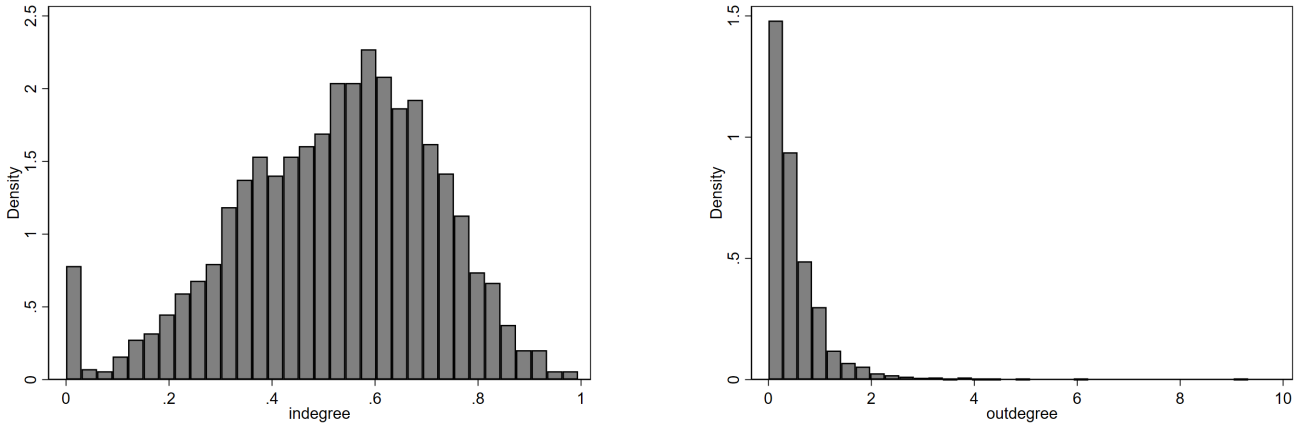
$$outdegree_i^r = \sum_j \sum_s a_{ij}^{rs}. \tag{263}$$

The indegree measures the fraction of gross output attributed to inputs (note that $indegree_j^s = 1 - va_j^s$ where va_j^s is the value added share). In the language of our model, since labour is the only primary factor, the indegree is always equal to $1 - \alpha_i$ for a sector i, namely the complement of the labour share.

The weighted outdegree is defined as the sum over all using industries of the fraction of gross output of industry s in country j that can be attributed to industry r in country i. This measure ranges between 0, if the sector does not supply any inputs to other industries, and $S * J$, which is the total number of industries in the economy if industry r in country i is the sole supplier of all industries.

The distribution of these measures tells us a lot about how the network is structured. In particular, the outdegree distribution tells us whether there are some sectors that are important in the economy because they serve a large fraction of others sectors' output.

Figure 7: Degree Distributions



(a) Indegree (b) Outdegree

Note: The figure shows the distributions of the indegree and outdegree across all sectors and years in the WIOD data.

Similarly, we can measure the Katz-Bonacich centrality as

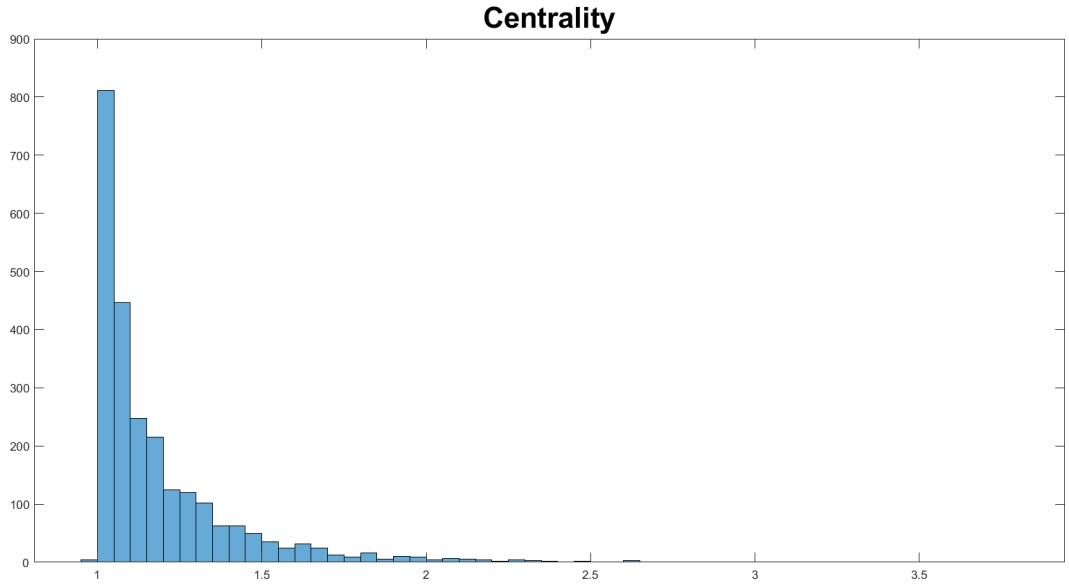
$$KB = [I - (\mathbf{1} - \underline{\alpha}) \otimes A]^{-1}\mathbf{1}, \tag{264}$$

where $\underline{\alpha}$ is the row vector of labour shares like in our model, and $\otimes$ represents the element by element multiplication. This measure tells us whether some sectors serve a large number of other sectors in the economy both directly and indirectly. The second part is best understood upon noting that this measure is the matrix limit of the following recursion:

$$KB_i = \sum_j (1 - \alpha_j) a_{ij} KB_j + 1, \tag{265}$$

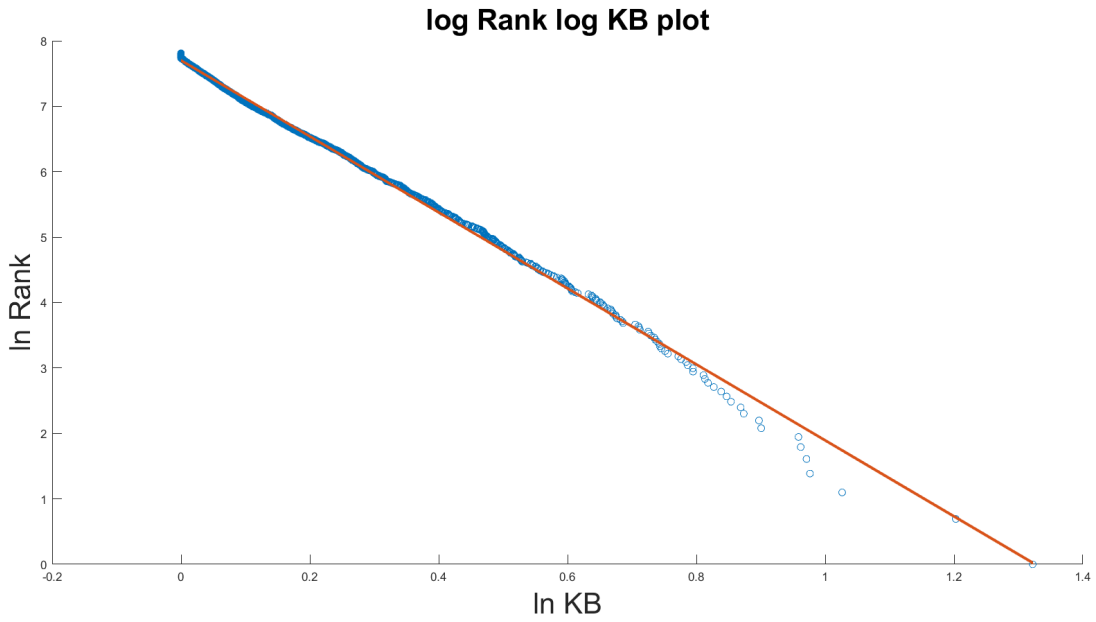
namely, the centrality of sector i depends on the centrality of the sectors it is connected with, weighted by the intensity of their connection. In the WIOD data for the year 2000, the distribution of centrality looks as follows (Figure 8). Interestingly we can plot the log-rank log variable

Figure 8: Distribution of KB Centrality



scatter plot to obtain the graph in Figure 9. The observation that this plot is almost linear would suggest that the centrality measure is likely to come from a power law. Estimating this relationship we find an estimate for the shape parameter of the power law of 5.8. We can also ask how does the geography of the network look like in terms of distance from consumption. For example, we could have that some sectors only produce intermediate inputs while others only produce final goods. It could also be that all sectors produce both intermediate and final goods. To formalize this concept, we can compute the measure of upstreamness. Using our notation, we can define it as Antràs, Chor, Fally and Hillberry (2012), who characterize upstreamness based on the number of steps that exist between production and consumption. The index is constructed by assigning value 1 to the share of sales directly sold to final consumers, value 2 to the share sold to consumers after it was used as an intermediate good by another industry,

Figure 9: Log Rank - Log KB



Note: The graph shows the scatter plot of log rank over log KB. The red line is a linear fit.

and so on. Formally:

$$U_i^r = 1 \times \frac{F_i^r}{Y_i^r} + 2 \times \frac{\sum_{s=1}^{S}\sum_{j=1}^{J} a_{ij}^{rs} F_j^s}{Y_i^r} + 3 \times \frac{\sum_{s=1}^{S}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{k=1}^{K} a_{ij}^{rs} a_{jk}^{st} F_k^t}{Y_i^r} + ... \tag{266}$$

where F_i^r is the value of output of sector r in country i consumed anywhere in the world and Y_i^r is the total value of output of sector r in country i. a_{ij}^{rs} is dollar amount of output of sector r from country i needed to produce one dollar of output of sector s in country j, defined as $a_{ij}^{rs} = Z_{ij}^{rs}/Y_j^s$. This formulation of the measure is effectively a weighted average of distance, where the weights are the distance-specific share of sales and final consumption.

Provided that $\sum_i \sum_r a_{ij}^{rs} < 1$, which is a natural assumption given the definition of a_{ij}^{rs} as input requirement, this measure can be computed by rewriting it in matrix form: [17]

$$U = \hat{Y}^{-1}[I - \mathcal{A}]^{-2}F, \tag{267}$$

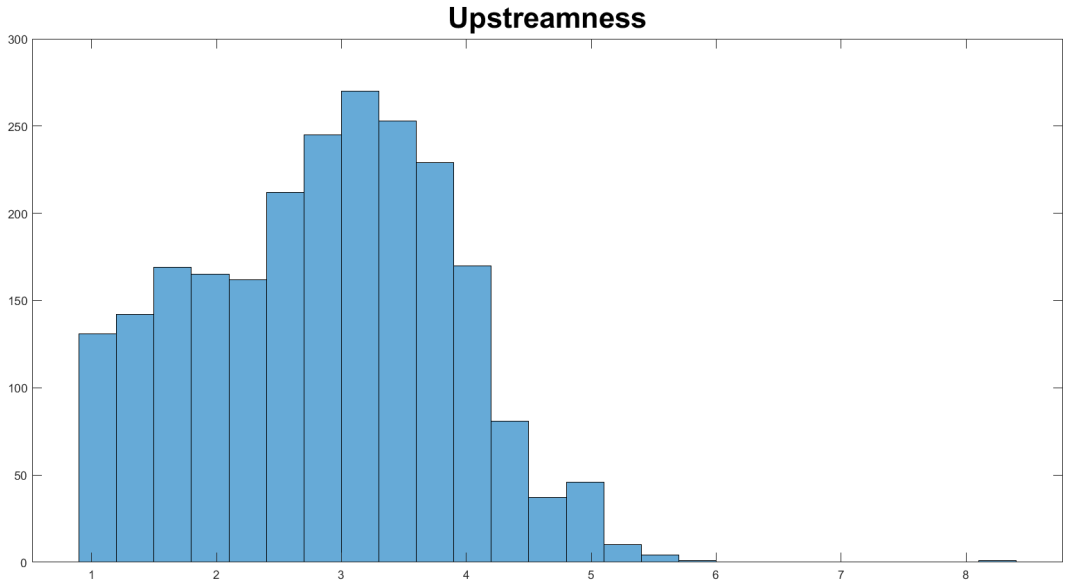
where U is a $(J \times S)$ by 1 vector whose entries are the upstreamness measures of every industry in every country. $\hat{Y}$ denotes the $(J \times S)$ by $(J \times S)$ diagonal matrix whose diagonal entries are the output values of all industries. The term $[I - \mathcal{A}]^{-2}$ is the power of the Leontief inverse, in which $\mathcal{A}$ is the $(J \times S)$ by $(J \times S)$ matrix whose entries are all a_{ij}^{rs} and finally the vector F

[17] For this not to be true some industry would need to have negative value added since $\sum_i \sum_r a_{ij}^{rs} > 1 \Leftrightarrow \sum_i \sum_r Z_{ij}^{rs}/Y_j^s > 1$, meaning that the sum of all inputs used by industry s in country j is larger than the value of its total output.

is an $(J \times S)$ by 1 whose entries are the values of the part of industry output that is directly consumed. Equation 266 shows the value of upstreamness of a specific industry r in country i can only be 1 if all its output is sold to final consumers directly. Formally, this is occurs if and only if $Z_{ij}^{rs} = 0, \forall s, j$, which immediately implies that $a_{ij}^{rs} = 0, \forall s, j$.

The distribution of upstreamness in the 2000 WIOD data can be seen in Figure 10.

Figure 10: Upstreamness Distribution



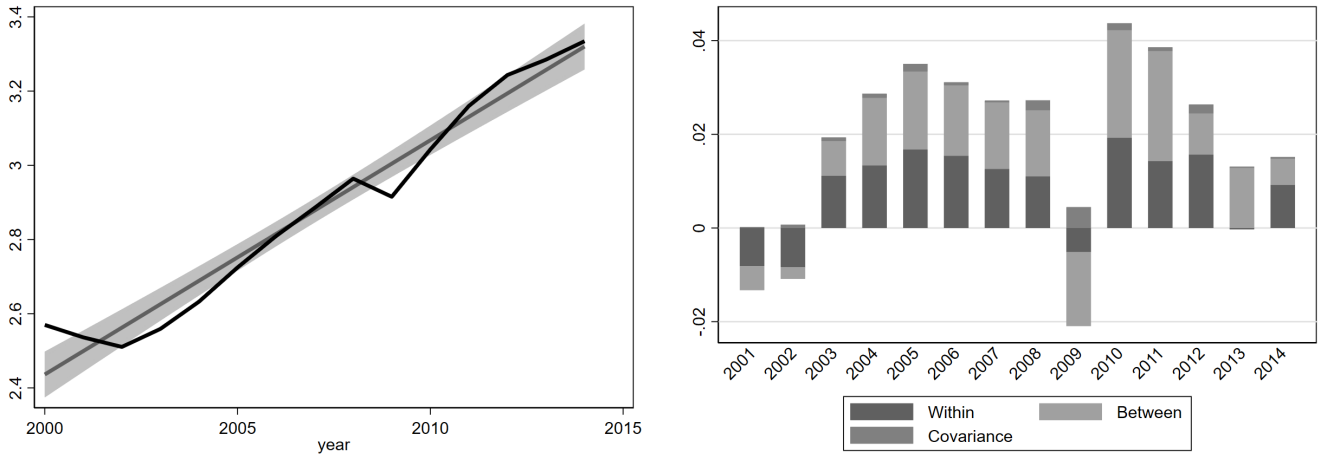
In summary, these observations suggest that most industries are not very central and neither far nor close to consumers. Few industries, on the other hand, have an outsized importance in the economy. As the model discussed above highlights, these are the industries that can drive the fluctuations of an economy.

These observations become ever more important when we notice that the global economy has been undergoing significant changes in the way goods are made. In particular, two observations are worth noticing:

Production chains have increased in length In the past few decades, modes of production have changed markedly. As highlighted by the World Bank Development Report 2020, a growing share of production now occurs in many stages and crosses borders multiple times before reaching consumers. Figure 11 shows, on average, how many production steps a good undergoes before it is finally consumed. This measure steadily increased in the period covered by the WIOD dataset, from 2.6 in 2000 to 3.3 in 2014. This change is driven in equal measure by an increase in importance of already long chains (between component) and the increase in length of important chains (within component). As mentioned in the discussion of the existing

literature, a salient feature of current models of production networks is that, as the source of an external shock increases in distance from the production chain, the shock itself dissipates in intensity. Taken together with the increasing length of production chains, this feature would imply that the world is becoming more resilient to demand shocks.

Figure 11: Upstreamness Dynamics



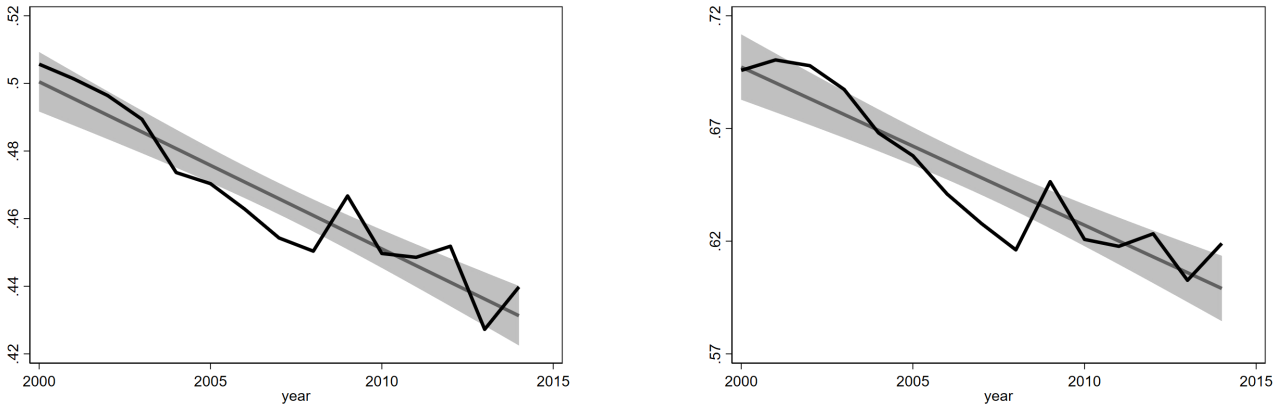
(a) Upstreamness Dynamics

(b) Dynamics Decomposition

Note: The figure shows the dynamics of the weighted upstreamness measure computed as $U_t = \frac{\sum_i \sum_r y^r_{it} U^r_{it}}{\sum_i \sum_r y^r_{it}}$. The left panel shows the average over time and it includes the estimated linear trend and the 95% confidence interval around the estimate. The right panel shows the decomposition of these changes into the stacked contributions (in levels) of the different components of the changes in the weighted average upstreamness measure. The components are given by $\Delta U_t = \sum_i \sum_r \underbrace{\Delta U^r_{it} w^r_{it-1}}_{Within} + \underbrace{U^r_{it-1} \Delta w^r_{it}}_{Between} + \underbrace{\Delta U^r_{it} \Delta w^r_{it}}_{Covariance}$.

The spatial distribution of sales has become less concentrated A second element related to the increasing role of international linkages in production is that of diversification. Using the WIOD data, it is possible to construct sales shares of each industry which account for intermediate linkages. In summary, this measure represents how much of a given industry's output is eventually consumed in a given destination, whether it is sold directly or indirectly. Figure 12 shows the trend in the Herfindahl-Hirschman Index of these sales shares. The key observation is that the HHI has been significantly declining over the period 2000-2014, whether I use a simple or a sales-weighted average. Quantitatively the unweighted HHI went from .51 in 2000 to .44 in 2014, while the sales-weighted HHI went from .7 to .6 in the same period. This observation would suggest that as industries are now exposed to a wider array of destinations, they should be less exposed to idiosyncratic shocks. In turn, this should reduce output volatility.

Figure 12: Herfindahl Index of Sales Shares



(a) Simple Average HHI

(b) Weighted Average HHI

Note: The figure shows the behaviour of the Herfindahl Index of destination shares over time. The Herfindahl Index is computed at the industry level as $HHI_t^r = \sum_j \xi_{ij}^{r\,2}$. The left panel shows the simple average across industry, i.e. $HHI_t = R^{-1} \sum_r HHI_t^r$. The right panel shows the weighted average using industry shares as weights: $HHI_t = \sum_r \frac{Y_t^r}{Y_t} HHI_t^r$. The plots include the estimated linear trend and the 95% confidence interval around the estimate.

4.4 Imperfect Competition and Market Power

So far we have always worked with firms operating in a competitive environment or in monopolistic competition. The convenient aspect of this clearly extreme assumption is that often firms were kind of dummies that we used to supply goods but they did not have much of a strategic role. In this section, we depart from this assumption. We have already studied simple economies with monopolistic competition. We now go further and allow for a continuum of industries with a finite number of firms in them and study oligopoly. In this case, firms will play a game inside the industry but still be small in the economy (because of the continuum of industries). The goal is always the study of what this more reasonable firm behaviour implies for the aggregate economy.

To build towards a more general model consider the simple case of a firm that has two products i and j. The two goods are produced independently (meaning that there is no technology spillover between them) with constant returns to scale. On the demand side, they are potentially related, meaning that they could be substitutes or complements. Denote the objective function of the firm

$$\tilde{\pi} = \pi_i + \pi_j = p_i(q_i, q_j)q_i + p_j(q_i, q_j)q_j - c_i q_i - c_j q_j, \tag{268}$$

where the inverse demands depend on both quantities because the goods may be not independent. As before, denote $\epsilon_i^D \equiv -\frac{\partial q_i}{\partial p_i}\frac{p_i}{q_i}$ but now also introduce $\epsilon_{ji}^P \equiv \frac{\partial q_i}{\partial p_j}\frac{p_j}{q_i}$ which is the cross-price

elasticity. If this is positive, then an increase in the price of good j implies an increase in demand for good i. Namely, the two goods are substitutes. The first order condition with respect to the quantity of q_i reads

$$\frac{\partial p_i}{\partial q_i} q_i + p_i + \frac{\partial p_j}{\partial q_i} q_j = c_i. \tag{269}$$

Following the same steps as in Section 1.3, we can rewrite this in terms of elasticities and obtain

$$c_i = -\frac{p_i}{\epsilon_i^D} + p_i + \frac{q_j p_j}{q_i \epsilon_{ji}^P} \tag{270}$$

$$= -\frac{p_i}{\epsilon_i^D} + p_i + p_i \frac{q_j p_j}{p_i q_i \epsilon_{ji}^P} \tag{271}$$

$$= p_i \left[1 - \frac{1}{\epsilon_i^D} + \frac{r_j}{r_i \epsilon_{ji}^P} \right] \tag{272}$$

Where the last term is given by the definition of revenues as $r_i = p_i q_i$. It is immediate that the price is given by

$$p_i = c_i \left[\frac{\epsilon_i^D - 1}{\epsilon_i^D} + \frac{r_j}{r_i \epsilon_{ji}^P} \right]^{-1} \tag{273}$$

Note that if the two goods are independent and the cross-price elasticity is zero, then we have the same pricing rule as in Section 1.3 with a markup $\mu_i = \frac{\epsilon_i^D}{\epsilon_i^D - 1}$. Consider the case of two goods which are substitutes. The higher the substitution, the higher ϵ_{ji}^P, and the higher the markup. The intuition behind this result is that if two goods are substitutes, then a price increase of one good generates a demand increase of the other which the firm now internalizes and profits from. Conversely, for complement goods, the price increase has a negative effect on the other good, which the firm wants to avoid by charging a lower markup.

This simple case is the scenario in which a monopolist owns two related products. Similar results can be derived in less stark cases, such as common ownership or cartels. Suppose that firm i cares about the profits of firms j because they either collude or i owns a fraction κ_{ij} of firm j. We can then write i's objective function as

$$\tilde{\pi}_i = \pi_i + \sum_{j \neq i} \kappa_{ij} \pi_j. \tag{274}$$

We can then obtain a similar result on firm i's pricing as

$$p_i = c_i \left[\frac{\epsilon_i^D - 1}{\epsilon_i^D} + \sum_{j \neq i} \kappa_{ij} \frac{r_j}{r_i \epsilon_{ji}^P} \right]^{-1}. \tag{275}$$

This is a somewhat general result from which it is hard to make much further progress, particularly in terms of aggregating up to the economy level. To do so, we now specify preferences further so that everything can be aggregated based on the model in Atkeson and Burstein (2008).

Suppose the economy is populated by a very large number of industries I making differentiated products. These products are aggregated into a final consumption good (the numeraire) according to

$$Y = \left(\sum_I y_I^\rho \right)^{\frac{1}{\rho}}, \tag{276}$$

with $\rho \in (0,1)$. Each industry is then populated by a finite number of firms N_I producing differentiated goods, which are aggregated into industry output according to

$$y_I = \left(\sum_i^{N_I} y_i^\eta \right)^{\frac{1}{\eta}}, \tag{277}$$

with $0 < \rho < \eta \leq 1$.Meaning that goods are more substitutable within than across sectors. This nested demand system implies that the inverse demand faced by firm i is given by

$$p_i = \left(\frac{Y}{y_I} \right)^{1-\rho} \left(\frac{y_I}{y_i} \right)^{1-\eta}, \tag{278}$$

where I imposed that the numeraire $P = 1$, which is the price index associated to Y[18].

We then assume that firms compete in a Cournot quantity game producing with CRS production out of labour $y_i = \varphi_i l_i$. Cournot competition implies that firms maximize profits using quantities while taking as given the quantity choices of other firms, formally

$$\max_{y_i} p_i y_i - \frac{w}{\varphi_i} y_i \quad \text{st} \quad p_i = \left(\frac{Y}{y_I} \right)^{1-\rho} \left(\frac{y_I}{y_i} \right)^{1-\eta} \tag{279}$$

$$y_I = \left(\sum_j^{N_I} y_j^\eta \right)^{\frac{1}{\eta}} \tag{280}$$

$$\{y_{-i}\}. \tag{281}$$

[18]The final good producer maximizes profits $\pi = Y - \sum_I \sum_i^{N_I} p_i y_i$. The first-order condition with respect to y_i is $y_I^{\rho-1} Y^{1-\rho} y_i^{\eta-1} y_I^{1-\eta} = p_i$.

Note a number of elements. First, I substituted in the production function so that we can optimize only over output. Next, since the firm has market power, it internalizes the inverse demand and the effect that choosing output has on it. Third, since there are few firms in the sector, it internalizes that changing y_i affects y_I, namely the size of the sector itself. Lastly, since there are many industries (possibly a continuum), it does not internalize the effect its choices have on Y, the size of the economy.

We can then proceed by plugging in the constraints and optimizing over y_i to obtain

$$\frac{w}{\varphi_i} = p_i + p_i' y_i \tag{282}$$

$$= p_i + y_i Y^{1-\rho} \left[\eta y_i^{\eta-1} \frac{\rho-\eta}{\eta} \left(\sum_j y_j^\eta \right)^{\frac{\rho-\eta}{\eta}-1} y_i^{\eta-1} + (\eta-1) y_i^{\eta-2} \left(\sum_j y_j^\eta \right)^{\frac{\rho-\eta}{\eta}} \right] \tag{283}$$

$$= p_i + p_i y_i \left[\eta \frac{\rho-\eta}{\eta} \frac{y_i^{\eta-1}}{\sum_j y_j^\eta} + (\eta-1) y_i^{-1} \right] \tag{284}$$

$$= p_i + p_i \left[(\rho-\eta) \frac{y_i^\eta}{\sum_j y_j^\eta} + \eta - 1 \right] \tag{285}$$

$$= p_i \left[\eta + (\rho-\eta) \frac{y_i^\eta}{\sum_j y_j^\eta} \right], \tag{286}$$

Hence

$$p_i = \frac{w}{\varphi_i} \left[\eta + (\rho-\eta) \frac{y_i^\eta}{\sum_j y_j^\eta} \right]^{-1}. \tag{287}$$

We can make further progress by noting that the market share is defined in sales terms so that

$$s_i = \frac{p_i y_i}{\sum_j p_j y_j} \tag{288}$$

$$= \frac{y_i \left(\frac{Y}{y_I} \right)^{1-\rho} \left(\frac{y_I}{y_i} \right)^{1-\eta}}{\sum_j \left(\frac{Y}{y_I} \right)^{1-\rho} \left(\frac{y_I}{y_j} \right)^{1-\eta} y_j} \tag{289}$$

$$= \frac{y_i^\eta}{\sum_j y_j^\eta}, \tag{290}$$

and therefore

$$p_i = \frac{w}{\varphi_i} \left[\eta + (\rho-\eta) s_i \right]^{-1}. \tag{291}$$

Recall that $\rho < \eta$ so that the markup is increasing in the firm's market share. We can think of this as averaging between the cross-industry and within industry elasticities via the market share. If the firm is a monopolist, ie $s_i = 1$, the markup is just $1/\rho$, which is the optimal markup of a firm competing with other industries. If the firm instead has no market share, we are back in monopolistic competition so that the markup is $1/\eta$.

This formulation of the problem is convenient because it allows us to easily aggregate upward to the industry and economy-wide level. To see that, note that the firm's markup is

$$\mu_i = \frac{p_i}{\lambda_i} \tag{292}$$

by definition, where λ_i is the marginal cost. Next, we can define an industry level markup as

$$\mu_I = \frac{p_I}{\lambda_I}, \tag{293}$$

where λ_I is the output-share weighted marginal cost of firms. Note, importantly, that it has to be weighted through output rather than sales shares because the share of a firm in terms of cost is given by how much it produces rather than how it sells. The latter is distorted by the heterogeneous markup. We can then rewrite this as

$$\mu_I = \frac{p_I}{\sum_i \lambda_i \frac{y_i}{y_I}} = \left(\sum_i \lambda_i \frac{y_i}{p_I y_I} \right)^{-1} \tag{294}$$

$$= \left(\sum_i \frac{\lambda_i}{p_i} \frac{p_i y_i}{p_I y_I} \right)^{-1} = \left(\sum_i \frac{\lambda_i}{\mu_i \lambda_i} s_i \right)^{-1} = \left(\sum_i \mu_i^{-1} s_i \right)^{-1}. \tag{295}$$

Using the firm's markup, we obtain

$$\mu_I = \left(\sum_i (\eta + (\rho - \eta) s_i) s_i \right)^{-1} = \left(\eta \sum_i s_i + (\rho - \eta) \sum_i s_i^2 \right)^{-1} \tag{296}$$

$$= \left(\eta + (\rho - \eta) HHI_I \right)^{-1}, \tag{297}$$

where I have used the definition $HHI_I = \sum_i s_i^2$. So the industry markup is a function of elasticities and industry concentration. Note that often concentration is not a good metric for market power (think of the CES+monopolistic competition model, where HHI measures productivity dispersion since markups are the same for all firms).

Similarly, we can aggregate further up at the economy level with exactly the same steps. Define $\lambda = \mathbb{C}$ the marginal cost of the economy as an output-weighted average of marginal

costs of industries, then the economy-wide markup is

$$\mu = \frac{P}{\lambda} = \left(\sum_I \lambda_I \frac{y_I}{PY} \right)^{-1} = \left(\sum_I \frac{\lambda_I}{p_I} \frac{p_I y_I}{PY} \right)^{-1} = \left(\sum_I \mu_I^{-1} s_i \right)^{-1}. \tag{298}$$

Using the industry-level markup derived above

$$\mu = \left(\eta \sum_I s_I + (\rho - \eta) \sum_I s_I HHI_I \right)^{-1} = \left(\eta + (\rho - \eta) \sum_I s_I HHI_I \right)^{-1}. \tag{299}$$

This result, which is identical to the findings of Burstein et al. (2019), highlights two important relationships. First, industries with higher concentration have larger markups. When highly concentrated industries have large shares in the economy (large s_I), the economy's average markup is also high.

Digression on Large Firms in Large Sectors We have gone through some models in which firms are "large" in their sector but not in the economy. There is, however, a very recent push for considering firms that are actually "large" in the economy. A plea to this comes from a recent column by Xavier Vives.[19] To some extent, this push is affine to our previous discussion on granularity. We spent decades writing models in which firms were small, then we made firms large but without them knowing. What I mean by this is that most of our models consider firms who, even if they have granular potential, behave as price-takers in the economy. For example, there is a recent literature about how firms have monopsonistic power over workers. In an economy in which labour is mobile, this immediately requires that firms are "large" in the economy. In Azar and Vives (2021), they build a model in which oligopolistic firms know that they are large in the economy and take it into account when deciding how much to produce and how much labour to hire.

End of Digression

[19] https://voxeu.org/article/taking-oligopoly-seriously-macroeconomics.

References

Acemoglu, Daron, Ufuk Akcigit, and William Kerr, "Networks and the macroeconomy: An empirical exploration," *Nber macroeconomics annual*, 2016, *30* (1), 273–335.

__ , Vasco M Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi, "The network origins of aggregate fluctuations," *Econometrica*, 2012, *80* (5), 1977–2016.

Ackerberg, Daniel A, Kevin Caves, and Garth Frazer, "Identification properties of recent production function estimators," *Econometrica*, 2015, *83* (6), 2411–2451.

Antràs, Pol, Davin Chor, Thibault Fally, and Russell Hillberry, "Measuring the Upstreamness of Production and Trade Flows," *American Economic Review*, May 2012, *102* (3), 412–416.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, December 2008, *98* (5), 1998–2031.

Azar, José and Xavier Vives, "General equilibrium oligopoly and ownership structure," *Econometrica*, 2021, *89* (3), 999–1048.

Baqaee, David and Emmanuel Farhi, "The macroeconomic impact of microeconomic shocks: beyond Hulten's Theorem," *Econometrica*, 2019, *87* (4), 1155–1203.

Bond, Steve, Arshia Hashemi, Greg Kaplan, and Piotr Zoch, "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data," *Journal of Monetary Economics*, 2021.

Burstein, Ariel, Basile Grassi, and Vasco Carvalho, "Bottom-Up Markup Fluctuations," Working Paper 2019.

Carvalho, Vasco and Xavier Gabaix, "The great diversification and its undoing," *American Economic Review*, 2013, *103* (5), 1697–1727.

Carvalho, Vasco M and Alireza Tahbaz-Salehi, "Production networks: A primer," *Annual Review of Economics*, 2019, *11*, 635–663.

di Giovanni, Julian and Andrei A Levchenko, "Trade openness and volatility," *The Review of Economics and Statistics*, 2009, *91* (3), 558–585.

Dixit, Avinash K and Joseph E Stiglitz, "Monopolistic competition and optimum product diversity," *The American economic review*, 1977, *67* (3), 297–308.

Doraszelski, Ulrich and Jordi Jaumandreu, "R&D and productivity: Estimating endogenous productivity," *Review of Economic Studies*, 2013, *80* (4), 1338–1383.

__ and __ , "Using cost minimization to estimate markups," 2019.

Dupor, Bill, "Aggregation and irrelevance in multi-sector models," *Journal of Monetary Economics*, 1999, *43* (2), 391–409.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "How costly are markups?," Technical Report, National Bureau of Economic Research 2018.

Gabaix, Xavier, "The granular origins of aggregate fluctuations," *Econometrica*, 2011, *79* (3), 733–772.

__ , "Power laws in economics: An introduction," *Journal of Economic Perspectives*, 2016, *30* (1), 185–206.

Gandhi, Amit, Salvador Navarro, and David A Rivers, "On the identification of gross output production functions," *Journal of Political Economy*, 2020, *128* (8), 2973–3016.

Giovanni, Julian Di and Andrei A Levchenko, "Country size, international trade, and aggregate fluctuations in granular economies," *Journal of Political Economy*, 2012, *120* (6), 1083–1132.

__ , __ , and Isabelle Mejean, "Foreign shocks as granular fluctuations," Technical Report, National Bureau of Economic Research 2020.

Hall, Robert E, "Invariance properties of Solow's productivity residual," 1989.

Hopenhayn, Hugo A, "Entry, exit, and firm dynamics in long run equilibrium," *Econometrica: Journal of the Econometric Society*, 1992, pp. 1127–1150.

Hsieh, Chang-Tai and Peter J Klenow, "Misallocation and manufacturing TFP in China and India," *The Quarterly journal of economics*, 2009, *124* (4), 1403–1448.

Levinsohn, James and Amil Petrin, "Estimating production functions using inputs to control for unobservables," *The review of economic studies*, 2003, *70* (2), 317–341.

Loecker, Jan De and Chad Syverson, "An industrial organization perspective on productivity," in "Handbook of Industrial Organization," Vol. 4, Elsevier, 2021, pp. 141–223.

__ **and Frederic Warzynski**, "Markups and firm-level export status," *American economic review*, 2012, *102* (6), 2437–71.

__ **, Jan Eeckhout, and Gabriel Unger**, "The rise of market power and the macroeconomic implications," *The Quarterly Journal of Economics*, 2020, *135* (2), 561–644.

__ **, Pinelopi K Goldberg, Amit K Khandelwal, and Nina Pavcnik**, "Prices, markups, and trade reform," *Econometrica*, 2016, *84* (2), 445–510.

Lucas, Robert E, "Understanding business cycles," *Carnegie-Rochester Conference Series on Public Policy*, 1977, pp. 7–29.

Melitz, Marc J, "The impact of trade on intra-industry reallocations and aggregate industry productivity," *econometrica*, 2003, *71* (6), 1695–1725.

Olley, Steven and Ariel Pakes, "The dynamics of productivity in the telecommunications equipment industry," *Econometrica*, 1996, (6), 1263–1297.

Ridder, Maarten De, Basile Grassi, Giovanni Morzenti et al., "The Hitchhiker's Guide to Markup Estimation," Technical Report 2021.

Rosen, Sherwin, "The economics of superstars," *The American economic review*, 1981, *71* (5), 845–858.

Timmer, Marcel P, Erik Dietzenbacher, Bart Los, Robert Stehrer, and Gaaitzen J De Vries, "An illustrated user guide to the world input–output database: the case of global automotive production," *Review of International Economics*, 2015, *23* (3), 575–605.

Traina, James, "Is aggregate market power increasing? production trends using financial statements," *Production Trends Using Financial Statements (February 8, 2018)*, 2018.

World Bank, *World Development Report 2020* number 32437. In 'World Bank Publications.', The World Bank, 12-2019 2020.

5 Appendix

5.1 CES + Monopolistic Competition

assume that the consumer in our economy maximises utility, in the form of consumption, which is an aggregate of many different consumption goods

$$C = \left(\int_i c_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}, \tag{300}$$

where different goods are indexed by i and $\sigma > 1$ is the elasticity of substitution between different types of goods. Suppose also that the consumer has the following budget constraint

$$\int_i p_i c_i di = I, \tag{301}$$

where I is some exogenous income they have and p_i is the market price of good i. We proceed by asking how a consumer would split the income to maximise utility. Formally, taking the first order condition with respect to a generic variety c_i

$$\frac{\sigma-1}{\sigma} c_i^{\frac{\sigma-1}{\sigma}-1} \frac{\sigma}{\sigma-1} \left(\int_i c_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}-1} - \lambda p_i = 0, \tag{302}$$

where λ is the Lagrange multiplier associated to the budget constraint. We can immediately take the price to the RHS and divide by the same condition for variety j to obtain

$$c_i = \left(\frac{p_i}{p_j} \right)^{-\sigma} c_j. \tag{303}$$

We can now define the ideal price index $P \equiv \left(\int_i p_i^{1-\sigma} di \right)^{\frac{1}{1-\sigma}}$. This price index is designed so that $\int_i p_i c_i di = PC$ at the optimal choice of the consumer.[20] Multiply both sides of equation (303) by p_i and integrate over i to obtain

$$PC = P^{1-\sigma} p_j^{\sigma} c_j, \tag{304}$$

[20] To obtain this note that, at the optimum, $c_i = p_i^{-\sigma} p_j^{\sigma} c_j$ and we want $\int_i p_i c_i di = PC$. Then $\int_i p_i p_i^{-\sigma} p_j^{\sigma} c_j di = P \left(\int_i c_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}$, by the definition of the aggregator C in eq. 300. Using the optimality condition $p_j^{\sigma} c_j \int_i p_i^{1-\sigma} di = P \left(\int_i \left(p_i^{-\sigma} p_j^{\sigma} c_j \right)^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}$. Simplifying we obtain $p_j^{\sigma} c_j \int_i p_i^{1-\sigma} di = P p_j^{\sigma} c_j \left(\int_i p_i^{1-\sigma} di \right)^{\frac{\sigma}{\sigma-1}}$. Canceling out $p_j^{\sigma} c_j$ from both sides, $\int_i p_i^{1-\sigma} di = P \left(\int_i p_i^{1-\sigma} di \right)^{\frac{\sigma}{\sigma-1}}$, which yields the desired price index $P = \left(\int_i p_i^{1-\sigma} di \right)^{\frac{1}{1-\sigma}}$.

Which, finally, implies

$$c_j = \left(\frac{p_j}{P}\right)^{-\sigma} C. \tag{305}$$

This is the demand for variety j as a function of all other prices and the total level of consumption. This demand function has intuitive and sensible properties: it decreases in the price of the good and increases in the price of other goods, through P. It also increases homothetically when total consumption increases. Finally, as the name might give away, its price elasticity is constant, and in particular, it is equal to σ. Without further solving, we can then conclude that a firm facing this type of demand for its own variety will have an optimal price

$$p = \frac{\sigma}{\sigma - 1} c. \tag{306}$$

In other words, the markup is just a constant number. Intuitively, when $\sigma \to \infty$ the markup $\mu \to 1$. This is a case where goods are perfect substitutes, and so there is little product differentiation and market power coming from it. The opposite case is when $\sigma \to 1$ and therefore $\mu \to \infty$. Note that there is one key unstated assumption behind this result. If we were to derive this optimal pricing rule, formally, we would have to ask ourselves whether we want to allow p_j to affect P. We wrote down this model in the context of a continuum of varieties, as shown by the integral, rather than a sum, so it is somewhat natural to think that $\partial P / \partial p_j = 0$. This is equivalent to assuming that firms are each a monopolist on their own variety, but they do not have any aggregate effect. This set of assumptions falls under the name of monopolistic competition. Section 4.4 shows a counterexample. The combination of CES preferences and monopolistic competition is by far the most used in international trade and is often used in modern macro.

Note two final remarks on the CES + monopolistic competition model. First, there is no cross-sectional misallocation of resources. This is evident in Hsieh and Klenow (2009) model solution, which implies that, given the symmetric markup rule, the wedge is identical across firms and, therefore, relative sizes are undistorted.[21] Second, note that, while there is dispersion in market shares, this is no indication of market power. To see start by the definition of market

[21] Recall, however, that the total size of the economy is distorted. In particular, the economy is too small by a factor $(\sigma - 1)/sigma$.

share:

$$s_i = \frac{p_i y_i}{PY} = \frac{y_i y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y\left(\int_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{307}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y\left(\int_j (y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} P)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{308}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y Y^{\frac{1}{\sigma}} P\left(\int_j (y_i^{-\frac{1}{\sigma}})^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{309}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}}}{Y\left(\int_j y_i^{\frac{\sigma-1}{\sigma}}\right)^{\frac{1}{1-\sigma}}} \tag{310}$$

$$= y_i^{\frac{\sigma-1}{\sigma}} Y^{-}1\left(\int_j y_i^{\frac{\sigma-1}{\sigma}}\right)^{\frac{1}{\sigma-1}} \tag{311}$$

$$= y_i^{\frac{\sigma-1}{\sigma}} Y^{-}1 Y^{\frac{1}{\sigma}} = \left(\frac{y_i}{Y}\right)^{\frac{\sigma-1}{\sigma}}. \tag{312}$$

Or, alternatively, by using the demand (rather than the inverse demand):

$$s_i = \frac{p_i y_i}{PY} = \frac{p_i p_i^{-\sigma} P^{\sigma} Y}{PY} \tag{313}$$

$$= p_i^{1-\sigma} P^{\sigma-1} = p_i^{1-\sigma}\left(\int_j p_j^{1-\sigma}\right)^{-1} \tag{314}$$

$$= \left(\frac{\sigma}{\sigma-1}\frac{c}{A_i}\right)^{1-\sigma}\left(\int_j \left(\frac{\sigma}{\sigma-1}\frac{c}{A_j}\right)^{\sigma-1}\right)^{-1} \tag{315}$$

$$= \frac{A_i^{\sigma-1}}{\int_j A_j^{\sigma-1}}. \tag{316}$$

In words, market shares are only driven by technology differences.