Macro III - Lecture Notes*

Alessandro Ferrari

University of Zurich

September 2024

*These notes are partially based on notes by Arpad Abraham, Chris Edmond, Nir Jaimovich, Philipp Kircher, Ramon Marimon, Alireza Tahbaz-Salehi, and Johannes Wieland. Lorenzo Pesaresi and Jan Ringling provided excellent teaching assistance in the production of these notes. The feedback from past students has been vital in improving these notes. Rest assured that all mistakes are mine and mine alone.

Contents

Economic activity undergoes periods of expansions and contractions. Business cycle research aims to understand the underlying reasons and whether policy can partially undo such fluctuations.

These lecture notes are meant as an introduction to the workhorse RBC model and some of its extensions. They are by no means meant to be a comprehensive discussion of the RBC literature and they are clearly heavily skewed toward my own interest and research.

I start by analyzing the workhorse stochastic growth model and its most basic extensions. This is followed by a long detour on search theory. I then go back to the simple RBC model to study how this can be extended with the recent advances in the literature of firms in macro. In particular, the notes cover some concepts on granular economies, production networks, and market power in a macro context.

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. Others discuss more recent advances in the literature. This is to say they are digressions, but sometimes they are where the interesting stuff lies. The first time you read the notes, it is probably a good idea to skip the digressions and go back to them once you are more familiar with the basic material.

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

1 Real Business Cycle Model[1]

This section starts by discussing risk in an endowment economy. To keep things simple we work in a setting with no aggregate uncertainty. This model represents a useful benchmark in which to think about Pareto optimal levels of risk sharing. We then proceed to introduce the same notion of risk in the Neoclassic Growth Model you are familiar with. This setting is at the core of modern macro. We conclude the section by discussing simple frictions in this setting.

1.1 Risky Endowment Economy

We start with some notation. Denote $s_t \in S_t$ an event (realization) in the event space S_t. Denote $s^t = \{s_0, ..., s_t\}$ a sequence (history) of realizations. Denote $\pi(s_t), \pi(s^t) \in [0, 1]$ the probabilities of an event and of an event history, respectively. Clearly, $\sum_{s_t} \pi(s_t) = 1$. Denote $\pi(s_t|s_{t-1})$ a conditional probability. $\pi(s_t) = \pi(s_t|s_{t-1})$ iff $s_t \perp s_{t-1}$. If shocks are iid then $\pi(s_t) = \pi(s^t)$. If shocks follow a Markov chain $\pi(s_t|s_{t-1}) = \pi(s_t|s^{t-1})$.

With this notation, we can start describing the economy. We assume that there is a unit measure of agents, each denoted by i. Every agent has an asset (Lucas tree) that bears fruit $y_t^i(s^t)$ every period.

Two important benchmark cases are:

1. No Aggregate Risk: $\sum_i y_t^i(s^t) = \bar{y}$, but $y_t^i(s^t) \neq y_t^j(s^t)$, $i \neq j$.
2. No Idiosyncratic Risk: $y_t^i(s^t) = y_t^j(s^t)$, $\forall i, j$.

We proceed by introducing preferences and the Planner problem and then come back to why these are important cases.

We assume agents have preferences over the consumption good

$$U(c^i) = \sum_{t=0}^{\infty} \sum_{s^t \in S^{t+1}} \beta^t \pi(s^t) u(c_t^i(s^t)) = \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t^i). \tag{1}$$

In this notation, $\beta \in (0, 1)$ is the subjective discount factor (very important when you try to solve things numerically). This parameter measures the degree of impatience of agents. Further, preferences here are time-separable and linear in probabilities. To make the problem "nice" we assume that the instantaneous utility function $u(\cdot)$ is twice continuously differentiable, increasing and concave ($u' > 0, u'' < 0$). We also assume that they satisfy Inada conditions: $\lim_{c \to 0^+} u' = \infty$, $\lim_{c \to \infty} u' = 0$.

[1] This section is inspired by Arpad Abraham and Nir Jaimovich's lecture notes on real business cycle.

Digression on Preferences in RBC The two most common preferences in RBC models are the general constant relative risk aversion (CRRA) and its special case of log utility. Formally, for some $\gamma > 1$ being the coefficient of relative risk aversion (curvature)

$$u(c) = \frac{c^{1-\gamma} - 1}{1 - \gamma}. \tag{2}$$

These preferences satisfy all the conditions before and also $u''' > 0$ (known as prudence). If we take the limit of $\gamma \to 1$ we obtain $u(c) = log(c)$. In the log utility special case income and substitution effects on savings offset each other. In this setting, the income effect is pushing for lower savings as the agent is richer today and wants to consume more. The substitution effect is increasing savings as the agent wants to transfer resources from today to tomorrow. In the log case, these cancel out and we have a constant savings rate.

End of Digression

We conclude the description of the economy by introducing the resource constraint. This intuitively states that every period's consumption must at most equal output. Formally, an allocation is feasible if

$$\sum_i c_t^i(s^t) \leq \sum_i y_t^i(s^t) \quad \forall t, s^t. \tag{3}$$

1.1.1 Planner Allocation

Before diving into the actual allocation we need to define the benchmark of the efficient allocation.

Definition 1 (Efficient Allocation)**.** *An allocation $\{c^i\}_i$ is Pareto efficient if the allocation is feasible and there is no other feasible allocation $\{\tilde{c}^i\}_i$ such that $U(\{\tilde{c}^i\}_i) > U(\{c^i\}_i)$.*

We now consider a Planner maximising a welfare function W. To define the welfare function we need to establish a set of Pareto weights $\{\mu_i\}_i$, with $\mu_i > 0$, $\forall i$ and $\sum_i \mu_i = 1$. These weights represent how important each agent's utility is to the planner. Further note that their scale is irrelevant because they will only be used as a splitting rule. The welfare function can then be characterized as

$$W = \sum_i \mu_i U(c^i). \tag{4}$$

The planner then maximizes the welfare function (4) subject to the sequence of resource constraints (3). The planner's Lagrangian is

$$\mathcal{L} = \sum_{t=0}^{\infty} \sum_{s^t \in S^{t+1}} \sum_i \mu_i \beta^t \pi(s^t) u(c_t^i(s^t)) + \lambda_t(s^t) \sum_i y_t^i(s^t) - c_t^i(s^t), \tag{5}$$

with $\lambda_t(s^t)$ the Lagrange multiplier of the resource constraint for s^t. First, note that non-satiation immediately implies that the resource constraint must hold with equality. This, in turn, (loosely) implies that the multipliers must be strictly positive.[2] Solving the first order conditions of the problem we get a set of equations

$$\mu_i \beta^t u'(c_t^i(s^t))\pi(s^t) = \lambda_t(s^t). \tag{6}$$

One useful trick to solve for an allocation is always to figure out the splitting rule for any level of binding resource constraint (endowment) by dividing the FOCs for two agents i, j

$$\frac{\mu_j}{\mu_i} = \frac{u'(c_t^i(s^t))}{u'(c_t^j(s^t))}. \tag{7}$$

This tells us that the relative marginal utilities of agents are equal to the relative Pareto weights. Using strict monotonicity (hence invertibility) of u' we can write

$$\sum_i c_t^i(s^t) = \sum_i u'^{-1}\left(\frac{\mu_j}{\mu_i} u'(c_t^j(s^t))\right) = \sum_i y_t^i(s^t). \tag{8}$$

This condition, together with the set of relative marginal utilities, defines the allocation of the economy.

1.1.2 Discussion

Now that we have the allocation we can look at its properties. A few observations:

- if the Planner cares equally about agents, i.e. $\mu_i = \mu_j$, $\forall i, j$ then all the marginal utilities are equalized (see equation 7).
- if the aggregate endowment is constant (no aggregate risk), then the RHS of equation (8) is constant, therefore so is the LHS. Agents' consumption is constant. This is a property of full risk sharing (or full insurance) economy: if there is no aggregate risk, then the marginal utilities and therefore consumption streams are constant. There is no risk at the individual level.
- In an economy with aggregate risk (RHS of 8 varies over time) agents split the endowment according to the Pareto Efficient sharing rule. As the total endowment fluctuates the marginal utilities (and therefore consumption) of all agents perfectly comove.
- The Planner allocation is not history dependent: agents' consumption only depends on the current state of the economy and the realization of the aggregate endowment.

[2] Recall that the resource constraint multipliers represent the marginal welfare effect of having a larger endowment. Because the welfare function does not have bliss points the multipliers must be strictly positive.

This economy is a useful benchmark for a number of reasons. First, it establishes what full risk sharing looks like so that we can think of deviations from it. For example, it tells us that if we were to go to the data and observed consumption and idiosyncratic endowments we could estimate

$$\Delta log c_t^i = \alpha + \beta \Delta log y_t^i + \gamma \Delta log Y_t + \epsilon_t^i. \tag{9}$$

If we were in a full risk sharing economy we should observe $\hat{\beta} = 0$ as consumption should not move with private endowments. In this context actually, β is a measure of how much risk is shared among agents.

Digression on Efficiency The allocation we just solved for through the Planner problem is clearly Pareto efficient. We could also have solved a decentralized equilibrium in which agents would trade endowments every period. As usual, we need to fix some set of assets to study efficiency in competitive equilibria. Suppose that the state space is finite-dimensional, then, if we have an independent asset space of equal cardinality we know that agents can trade claims such that the competitive equilibrium will be Pareto efficient. In this sense, we need asset market completeness. Suppose otherwise that for a state space of cardinality $S > 1$ the economy only has one asset. It is immediate that agents cannot trade in a state-contingent way and therefore marginal utilities will not be equalized.

End of Digression

1.2 Stochastic Neo-Classic Growth Model[3]

In this section, we introduce stochasticity in the production economy of the neo-classic growth model. The goal is to study how the economy responds to changes in productivity. We can use a lot of the heavy lifting from the endowment economy and just work out how to introduce production and capital. We start by assuming that there is a representative competitive firm producing according to the following technology

$$y_t = A_t F(k_t, n_t) \tag{10}$$

Where A_t is a Hicks-neutral productivity shifter and F has the usual properties: $F_i' > 0$, $F_{ii}'' < 0$, $\lim_{i \to 0} F' = \infty$, $\lim_{i \to \infty} F' = 0$ for $i = k, n$.

You are familiar with the deterministic version of this model from Q1. So for now we focus on the novelty and assume that households have preferences over the consumption good and leisure

[3]For a textbook analysis of the RBC model see L. Ljungqvist and T.J. Sargent (2012): Recursive Macroeconomic Theory: Chapter 12 or J. Adda and R. Cooper (2003): Dynamic Economics: Chapter 5.

and that the planner maximizes welfare subject to the following constraints

$$l_t(s^t) + n_t(s^t) = 1 \quad \forall s^t, t \qquad \text{(Labour Constraint)}$$
$$c_t(s^t) + i_t(s^t) = A_t F(k_t(s^{t-1}, n_t(s^t)) \quad \forall s^t, t \qquad \text{(RC)}$$
$$k_{t+1}(s^t) = (1-\delta)k_t(s^{t-1}) + i_t(s^t). \qquad \text{(LOM)}$$

The first constraint states that agents, who have an endowment of time equal to 1, can either work or enjoy leisure. The second constraint states that whatever is produced in a given period (RHS) needs to be either consumed or invested in the capital stock. The last constraint describes the law of motion of the capital stock. In particular, a fraction δ of the current capital is burnt every period and can be replenished through investment. Note that I have not imposed that $i_t \geq 0$, this implies that capital "can be eaten" or is reversible.

1.2.1 Planner Problem

We proceed as in the previous section and study the Planner solution. We can write the Lagrangian (after substituting in some constraints) where we take the starting capital k_0 as given

$$\mathcal{L} = \sum_t \sum_{s^t} \beta^t \pi(s^t) \left\{ u(c_t(s^t), 1 - n_t(s^t)) + \right.$$
$$\left. + \lambda_t(s^t)[A_t(s^t)F(k_t(s^{t-1}), n_t(s^t)) + (1-\delta)k_t(s^{t-1}) - c_t(s^t) - k_{t+1}(s^t)] \right\}. \qquad (11)$$

From here we can take FOCs and get a number of important insights.

$$u_c(s^t) = \lambda_t(s^t). \qquad (12)$$

First, the FOC on consumption tells us that the marginal effect of relaxing the resource constraint is equal to the marginal utility that the agent derives from consumption.

$$u_l(s^t) = u_c(s^t) A_t(s^t) F_n(s^t). \qquad (13)$$

The FOC on labour and the one on consumption tell us that the marginal utility of leisure (hence the marginal disutility from labour) must be equal to the utility value of the additional output we produce.

$$u_c(s^t) = \beta \sum_{s^{t+1}|s^t} \pi_{t+1}(s^{t+1}|s^t) u_c(s^{t+1})[(1-\delta) + A_{t+1} F_k(s^{t+1}))]. \qquad (14)$$

Lastly, the FOCs on capital and consumption give us the Euler Equation. This tells us that the marginal cost of saving, i.e. the marginal utility of consumption, must be equal to the expected

value of investment. The latter is given by the expected utility value of an additional unit of capital tomorrow, which is the utility value of additional output produced through the undepreciated capital. In other words, saving today means consuming less, hence the "cost" u_c. Saving one unit today means that we will have extra production tomorrow, in particular the expected value of $A_{t+1}F_k(s^{t+1})$. Further, that extra capital will be brought forward in the future at rate $1-\delta$. All these benefits are weighted by the appropriate marginal utility.[4]

These three conditions, together with the resource constraint, pin down the allocation.

Digression on Solving the RBC Numerically The formulation above is not exactly the easiest to work with when we want to solve the problem numerically. To do so we start by assuming a Markov chain structure of productivity. From the previous section this means that $\pi(s_t|s^{t-1}) = \pi(s_t|s_{t-1})$. Denote Π the transition matrix, where $\Pi = \{\pi\}$.

We can rewrite the problem in a recursive way substituting in some of the constraints

$$V(K,A) = \max_{N,K'} u(AF(K,N) + (1-\delta)K - K', 1-N) + \beta \sum_{A'|A} \Pi_{A'|A} V(K', A'). \tag{15}$$

In practice, we will use the fact that the value function in this problem is a contraction and therefore it has a unique fixed point (by Banach Fixed Point Theorem or Contraction Mapping Theorem). This means that we can define a guess V_0, write the problem as $V_1 = u(c^*) + \beta\Pi V_0$, and then replace V_0 with V_1 until they converge.

1.2.2 Competitive Equilibrium

We move on to study the decentralized version of this economy. We start by assuming that there is a representative firm hiring labour and renting capital from the household and producing the consumption good. Denote the prices of labour and capital $w_t(s^t)$ and $r_t(s^t)$ respectively. The firm solves a static optimization problem (does this matter in this context?). The household maximises expected utility subject to a budget constraint

$$C_t(s^t) + K_{t+1}(s^t) \leq (1 - \delta + r_t(s^t))K_t(s^{t-1}) + w_t(s^t)N_t(s^t), \quad \forall t, s^t. \tag{16}$$

Firms solve their static problem, maximising over the quantity of output sold (which in this case is equivalent to maximising over inputs). Importantly they take factor and output prices as given.

$$\max_{K_t^d(s^t), N_t^d(s^t)} A_t(s^t)F(K_t^d(s^t), N_t^d(s^t)) - w_t(s^t)N_t^d(s^t) - r_t(s^t)K_t^d(s^t). \tag{17}$$

[4] So far I have assumed that the technology transforming savings into future capital is efficient. Suppose instead that there were imperfectly competitive intermediaries turning savings into capital stock. How would the Euler equation change?

Where the superscripts d denote demanded quantities.

Definition 2 (Competitive Equilibrium)**.** *In this economy, a competitive equilibrium is a sequence of factor prices* $\{w_t(s^t), r_t(s^t)\}$*, household policies* $\{C_t(s^t), L_t(s^t), N_t(s^t), K_{t+1}(s^t)\}$ *and firm policies* $\{K_t^d(s^t), N_t^d(s^t)\}$ *such that*

1. *Taking the initial level of capital, factors, and output prices as given, households optimize;*
2. *Taking factors and output prices as given, firms optimize;*
3. *Goods, Capital, and labour markets clear;*
4. *The time constraint is satisfied.*

As before, we can write the Lagrangian substituting in some of the constraints and study the first order conditions.

$$\mathcal{L} = \sum_t \sum_s^t \left\{ \beta^t \pi(s^t) u(C_t(s^t), 1 - N_t(s^t)) + \lambda_t(s^t)[(1 - \delta + r_t(s^t))K_t(s^{t-1}) + w_t(s^t)N_t(s^t) - C_t(s^t) - K_{t+1}(s^t)] \right\}. \tag{18}$$

From the first order conditions,

$$\beta^t \pi(s^t) u_C(C_t(s^t), 1 - N_t(s^t)) = \lambda_t(s^t). \tag{19}$$

This condition states that the marginal utility value of relaxing the budget constraint is the marginal utility of consumption

$$u_L(C_t(s^t), 1 - N_t(s^t)) = w_t(s^t) u_C(C_t(s^t), 1 - N_t(s^t)). \tag{20}$$

From the labour and consumption FOCs, we get that the marginal disutility of labour (LHS) must be equal to the utility value of supplying an additional unit of labour, i.e. the utility value of the market wage (RHS).

$$u_C(C_t(s^t), 1 - N_t(s^t)) = \sum_{s^{t+1}|s^t} \beta \pi(s^{t+1}|s^t) u_C \left(C_{t+1}(s^{t+1}), 1 - N_{t+1}(s^t + 1) \right) [r_{t+1}(s^{t+1}) + 1 - \delta]. \tag{21}$$

Lastly, from the choice of future capital, we get again the Euler equation but now we compare the disutility from saving (lower consumption today) to the benefit of an additional unit of capital tomorrow, which is given by the expected rental rate and bringing the capital forward in the future (up to depreciation).

We now have to combine these conditions with the firm's FOCs, which are given by

$$A_t(s^t)F_N(K_t^d(s^t), N_t^d(s^t)) = w_t(s^t) \tag{22}$$
$$A_t(s^t)F_K(K_t^d(s^t), N_t^d(s^t)) = r_t(s^t) \tag{23}$$

These conditions state that at the firm's optimum the wage must be equal to the marginal product of labour and the rental rate must be equal to the marginal product of capital.

It is immediate to note that once we impose factor market clearings and combine the firm's FOCs with the household ones we get the FOCs describing the Planner problem. Since time and resource constraints are the same between the decentralized and Planner problems we conclude the allocations are the same. This is an application of the First Welfare Theorem.

Digression on Returns to Scale, Homogeneity and the Euler Theorem Using the firm first order, imposing market clearings so that $K^d = K$ and $N^d = N$ we can immediately note that the firm makes zero profits. To see this, note that the FOCs state (simplifying the notation and normalizing the output price to 1) and generalizing to any multiple input, constant returns to scale production function $Y = Af(x_1, ..., x_N)$

$$Af_{x_i} = p_i, \quad \forall i. \tag{24}$$

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

$$\sum_i x_i f_{x_i} = \chi f(x_1, ..., x_N). \tag{25}$$

We now note that CRS is the case where $\chi = 1$, which allows us to write

$$f(x_1, ..., x_N) = A\sum_i x_i f_{x_i} = \sum_i p_i x_i. \tag{26}$$

Note that the rightmost formulation is the total cost bill of the firm. 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. We can generalize this by noting that a firm with returns to scale $\chi < 1$ will have a profit rate of $1 - \chi$.

End of Digression

Digression on Capital Ownership So far we have made the assumption that households own the capital stock. Suppose otherwise that firms own capital. Instead, households have an asset

they use to save. This implies that firms are also solving a dynamic problem. We start with the household's problem, whose budget constraint is now (omitting history dependence)

$$C_t + \Delta B_{t+1} \leq w_t N_t + r_{t-1} B_t \tag{27}$$

Where we assume that the return on the bonds is predetermined. The firm's dynamic problem can now be written as a maximization of the value of the firm

$$\max_{N_t, I_t, D_{t+1}, K_{t+1}} V_0 = \mathbb{E}_0 \sum_{t=0}^{\infty} M_{0,t}(A_t F(K_t, N_t) - w_t N_t - I_t + D_{t+1} - (1 + r_{t-1}) D_t), \tag{28}$$

subject to the capital law of motion

$$K_{t+1} = I_t + (1 - \delta) K_t. \tag{29}$$

An important element of the maximization problem is how firms discount the future. Here firms discount at rate $M_{0,t}$. We assume that this is the stochastic discount factor of the households, defined as

$$M_{0,t}(s^t) = \beta^t \frac{u'\left(C_t(s^t)\right)}{u'(C_0)} \tag{30}$$

This is the current utility value of an extra unit of consumption at time t and state of the world s^t. We assume firms discount at this rate because they are ultimately owned by the household.

Once we make this assumption it is trivial to get the same first order condition we had in the case in which the household itself owned the capital. This tells us that in this environment it is ultimately irrelevant (and therefore indeterminate) whether firms finance their production capital directly or through equity claims or bonds. The allocation remains the same. This is a version of the Modigliani-Miller theorem.

End of Digression

Digression on Big K vs Little k In the competitive equilibrium with a representative firm/household there is a fundamental inconsistency between price takers and price makers. When you think about it the household is the only supplier of factors and the firm is the only supplier of the consumption good. Yet they all take prices as given. This is because we think of the economy as populated by a continuum of atomistic firms/households that we assume cannot collude. In particular, an atomistic household maximises "little k" without knowing that, by aggregation, that also coincides with "big K". Similarly for the firm. In practice, this is a convenient trick and one that also helps a lot in solving these models numerically. In particular, it allows us to guess some

price, have agents optimize given that price, check ex-post that it clears the market. Rinse and repeat until it does.

End of Digression

1.3 Business Cycle in RBC Models[5]

In the model, we laid out so far the only source of stochasticity is the productivity process. This is very common in macro models and should not be taken literally. First, it should be noted that in this context TFP is defined residually. If we were to take, for example, the log of a Cobb-Douglas production function we would be able to estimate our model as

$$y_t = a_t + \alpha k_t + \beta l_t + \epsilon_t \tag{31}$$

Where $a_t = \ln A_t$. Also, note that in the A process the mean is irrelevant. We can just normalize and the whole economy will just be rescaled. Another clear oddity of this approach is the interpretation of negative TFP shocks. If we were to interpret them literally we would have to think about a case in which a technology disappeared from our recipes set. We, however, tend to think about the technology set as something that is weakly expanding.

On the other hand, this approach is extremely convenient. We get to use a shock which is sort of a placeholder for many things we think are going on along the cycle. We can get fancier and more sophisticated, and we know how to model expectations driven cycles, uncertainty driven cycles, and so on. The TFP approach is just a starting point and one we rely on for its parsimony. So in summary what's a TFP shock? We don't know, nothing in particular, but possibly many things. We often refer to TFP as a "measure of our ignorance". Welcome to modern macro.

With this in mind, we can ask how our economy responds to a change in TFP. Suppose TFP follows an AR(1) process

$$\ln A_t = \rho \ln A_{t-1} + \epsilon_t,\ \epsilon_t \sim_{iid} N(0, \sigma^2). \tag{32}$$

Note that if we were doing this in levels we would need to have a drift, otherwise, this process is centered around zero, which would mean that we get zero output (recall A enters multiplicatively in the production function). This process is stationary and has the Markov property.

1.3.1 Comparative Statics in PE

Suppose the economy is in steady state and productivity suddenly increases.[6] This shock reverts to the mean according to its persistence ρ.

[5] This section is partially based on the Handbook chapter by King and Rebelo (1999).

[6] This is called an MIT shock. This is opposed to a case in which agents know that with some probability a shock will occur.

To establish the partial equilibrium response of consumption and investment, note that the Euler equation reads

$$u_C = \beta \mathbb{E}_t u'_C [MPK' + 1 - \delta]. \tag{33}$$

and that the budget constraint implies

$$C = AF(K, N) + (1 - \delta)K - K', \tag{34}$$

where the prime notation denotes objects in the next period.

Think of the LHS and RHS of the Euler equation 33 in the space of K'. First note that the LHS is increasing in K' since to obtain a higher K' agents have to increase investment, which, ceteris paribus, implies lower consumption. Since $u_{CC} < 0$, then u_C is decreasing in K'. In other words, the marginal cost of investment increases in the level of investment and therefore capital tomorrow.

Similarly, the RHS is decreasing in K'. This is driven by two forces: first, a higher level of capital tomorrow implies, all else equal, a higher level of output and therefore consumption tomorrow. Since u_{CC} is negative, then at a higher level of consumption C' corresponds to a lower marginal utility of consumption u'_C. Furthermore since $F_{KK} < 0$, a higher level of capital tomorrow implies a lower MPK'. This is plotted in Figure 1.

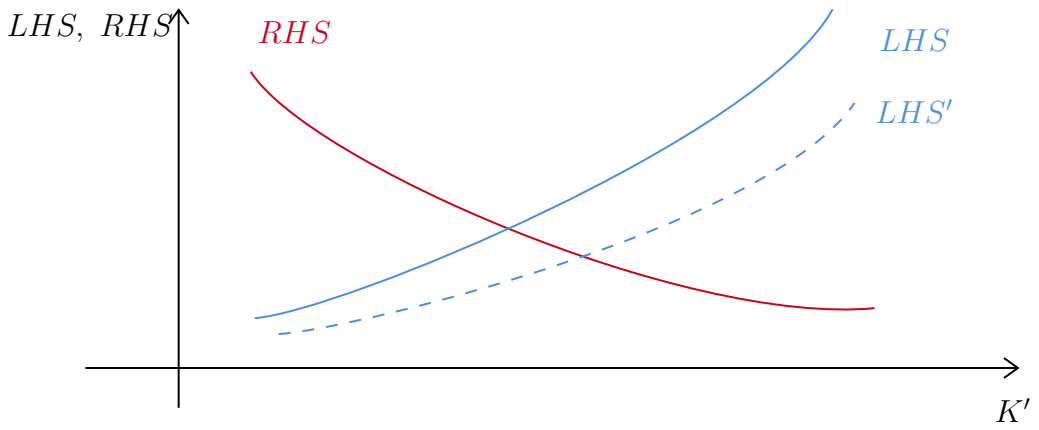


Figure 1: Consumption and Investment

Consider now an increase in A. Mechanically, fixing factors, a higher productivity implies a higher level of consumption, for any level of K'. As consumption increases, the marginal utility of consumption shifts downward. In partial equilibrium (disregarding movements in the RHS), consumption and investment increase, generating a positive comovement between A, Y, C and I.

Consider the changes in the intratemporal choice, governed by

$$\frac{u_N}{u_C} = w = AF_N \tag{35}$$

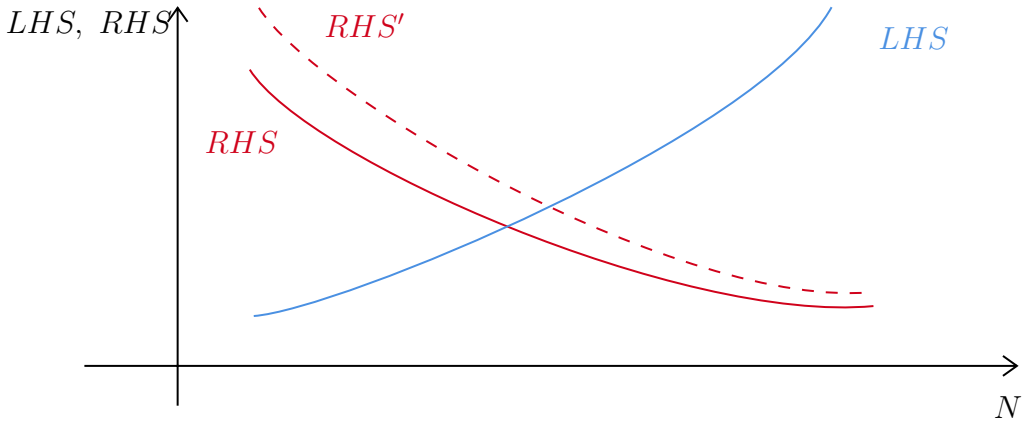


Figure 2: Labour - No income effect

First, as preferences are concave in leisure, they are convex in labour supplied. In the N space, the LHS is increasing since $u_{NN} > 0$, and to higher labour corresponds higher production and consumption, so the denominator also decreases via $u_{CC} < 0$. Note that in GE, as the wage increases, the consumer is richer and might decide to reduce labour supply if the income effect dominates, this would flip the slope of the LHS. Assuming that substitution effects dominate, the LHS is increasing in N. The RHS instead is decreasing in N since $F_{NN} < 0$, so at higher N, F_N declines. This is plotted in figure 2. Consider now an increase in A. At higher productivity, firms demand more labour since the marginal product of labour increases. This shifts up the RHS curve. Absent income effects, the LHS does not move. If income effects are present, the LHS will shift as well. Absent income effects, the economy in the new equilibrium features higher levels of hours worked. As a consequence, N also positively comoves with the other aggregates.

The first, mechanical effect is that, given the amount of current capital, output increases because the technology improved. Furthermore, the marginal product of labour increases, so firms will demand more of it, and this, in turn, will lead to an increase in the wage and an increase in the equilibrium hours worked.

Recall that consumption smoothing motives implies that agents do not "eat" fully the higher productivity benefits but save part of the extra output to increase the capital stock.

1.3.2 Quantitative Performance[7]

The model has sensible qualitative predictions. To assess the quantitative properties, we have to first figure out how to discipline it quantitatively.

There are generally two ways (and convex combinations thereof) of taking a model to the data. The first one is, assuming that you have outside estimated parameters, to use those parameters directly. For example, suppose that in an experiment, you measured the rate of time preference β. Brushing aside obvious and important questions on extrapolating from your experimental pool to

[7]For a more systematic discussion of the quantification of RBC models, please refer to Adda and Cooper (2003).

the whole economy, you can, in principle, use the empirically estimated value directly.

When this option is not viable for all (or any) of the parameters, the model needs to be calibrated. This effectively means using indirect inference to back out values of the parameters. What follows is a number of sparse observations about the process of calibrating a model. After these, I will use the structure of the basic RBC model to map parameters to data moments.

First, note the use of the word "moments". This largely refers to empirical observations that we can measure in the data and define in the model. The goal is to have the model and data objects match as close as possible.

Any serious discussion of the process of calibration has to start with the Sonnenschein, Mantel, Debreu, Mas-Colell theorem.[8] This, loosely, states that, given a set of prices and allocations, there exists an economy with some number of consumers for which these prices and allocations constitute an equilibrium. This is effectively a theorem on observational equivalence. We then need to impose restrictions on the primitives (preferences, technology) and parameters. The process of calibration is an attempt to do exactly that, by restricting the parameters so that the economy displays a set of properties.

Note also that, unless your model features functional form assumptions that simply nest others (eg. assuming CES nests Cobb-Douglas, but not vice versa), you cannot typically test functional form assumptions: they are the equivalent of identifying assumptions in the empirical work.

Concretely, think of your model G as a map between a set of parameters θ and moments of the generated data $\mathcal{M}$, such that

$$\mathcal{M} = G(\theta). \tag{36}$$

Suppose further that you observe the empirical counterparts of some of the moments $\tilde{\mathcal{M}}$. If your model is invertible, then we can back out the parameters of interest as

$$\hat{\theta} = G^{-1}(\tilde{\mathcal{M}}). \tag{37}$$

In general, the exact inversion of the model is hard as few models allow us to write down the mapping between moments and all parameters in closed form. An alternative to this is the minimization of some norm between the generated and the actual data. For example, denote $\{x_t(\theta_0)\}_{t=1}^T$ a T-period sequence of actual data, where θ_0 is the true value of the parameters, which is unobserved. Call $\{x_t^s(\theta)\}_{t=1}^T$ a T-period sequence of simulated data where $s = 1 \dots S$ is the index of each simulation sequence. Denote $\mu(x_t)$ a statistic (moment) of the actual data. We can write

[8] See Chapter 17.E of Mas-Colell et al. (1995).

down an estimator

$$\hat{\theta}_{S,T}(W) = \arg\min_{\theta} \left[\sum_{t=1}^{T} \mu(x_t) - \frac{1}{S} \sum_{s=1}^{S} \mu\left(x_t^s(\theta)\right) \right]' W_T^{-1} \left[\sum_{t=1}^{T} \mu(x_t) - \frac{1}{S} \sum_{s=1}^{S} \mu\left(x_t^s(\theta)\right) \right]. \tag{38}$$

This is the estimator of the Simulated Method of Moments (SMM) with a weighting variance W_T. Specifically, this estimator is the equivalent of the Generalized Method of Moments but applied to data simulated through the model.

A natural question at this point is the one of choosing the moments vector and how to weigh them. In the simplest case where the system is exactly identified (i.e., length of θ and μ are the same), then no weighting choice is necessary. More in general, when the size of μ is (much) larger than that of θ, you can either target an overidentified system or choose some moments to be *untargeted*. The underlying idea is that your model can have out-of-sample power if, conditional on matching the set of targeted moments, it can also predict some untargeted ones.

Beware, in general, about the "correlation/collinearity" of targeted and untargeted moments. When these are too close to one another, the whole exercise of looking at untargeted moments is just window-dressing.

Sometimes we extend the logic of the SMM to other statistics. We generally call this method *indirect inference*. Suppose we know from the data the coefficient of the regression between an outcome Y and dependent variable X to be β. We can run the same regression on the simulated data for each $T-$period sequence s. Define the likelihood of the auxiliary model (the regression) as $\phi(x_t, \beta)$ such that:

$$\hat{\beta}_T = \arg\max \prod_{t=1}^{T} \phi(x_t, \beta). \tag{39}$$

Define the same problem on simulated data sequence such that:

$$\hat{\beta}_{s,T} = \arg\max \prod_{t=1}^{T} \phi(x_t^s(\theta), \beta). \tag{40}$$

Then the indirect inference estimator is

$$\hat{\theta}_{ST} = \arg\min_{\theta} [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)]' \Omega_T [\hat{\beta}_T - \hat{\beta}_{ST}(\theta)], \tag{41}$$

where Ω_T is a positive definite weighting matrix.

The last important word of caution is about measurement in the context of models. You should also try to ask yourself whether the moment in the data and in the model are really the same thing. The typical example is the comparison between economic and accounting profits. Similarly, when using aggregate data in the national accounting system, one should wonder if the national

accountants creating the real data aggregate information the same way as the national accountants in the model do.

Calibrating the RBC model Practically this requires choosing the parameters in the model. Suppose we assume preferences are such that the per-period utility is $u_t = \ln C_t + \psi ln L_t$, that the production function is a constant returns Cobb-Douglas $Y_t = A_t K_t^{\alpha} N_t^{1-\alpha}$ and $N_t + L_t \leq 1$. In this case, we have the following parameters to pick: β, δ, α, ψ, ρ, σ.

The first identifying assumption is that the data is in steady-state so that we can use the steady-state conditions on the model to back out the parameters of interest. We normalize A in steady state to 1. First, note that the Euler equation in steady-state boils down to $\beta = (R^{\star} + 1 - \delta)^{-1}$, defining the real interest rate as $r = R - \delta$ we obtain $\beta = (1 + r^{\star})^{-1}$, where $\star$ denotes steady-state quantities. The real interest rate is something we can observe in the data (with assumptions) and allows us to back out $\beta = (1 + \tilde{r})$, where I use $\sim$ to denote empirical objects.

Next, note that the Cobb-Douglas assumption on the production function implies that the firm solving $\max_{K,N} \pi = K^{\alpha} N^{1-\alpha} - wN - RK$ yields the first order conditions $\frac{wN}{Y} = 1 - \alpha$ and $\frac{RK}{Y} = \alpha$. These are the total payments to labour and capital divided by the value of output (recall the price of output is the numeraire). They are called the capital (KS) and labour share (LS) and they are available in most national account data (easier to measure the labour share than the capital share). So we can compute $\alpha = 1 - \tilde{LS}$.

Turning to the depreciation rate, we can manipulate the model starting from the law of motion of capital $K' = (1 - \delta)K + I$. In steady-state, this collapses to $I^{\star}/K^{\star} = \delta$. While I/K is not as easily measured (hard to compute what the stock of capital K is), we are pretty good at measuring flows, for example, I/Y, the investment rate of the economy. With a trivial manipulation, we can get

$$\frac{I^{\star}}{Y^{\star}} = \frac{\frac{I^{\star}}{K^{\star}}}{\frac{Y^{\star}}{K^{\star}}} = \frac{\delta}{\left(\frac{K^{\star}}{N^{\star}}\right)^{\alpha - 1}}. \tag{42}$$

From the Euler equation we note that $u_C / u'_C = \beta[\alpha A (K/N)^{\alpha - 1} + 1 - \delta]$. In steady state $u_C = u'_C$ and $A = 1$ so that $(K/N)^{\alpha - 1} = \frac{\beta^{-1} - 1 + \delta}{\alpha}$. Plugging in we obtain

$$\frac{I^{\star}}{Y^{\star}} = \frac{\delta \alpha}{\beta^{-1} - 1 + \delta}, \tag{43}$$

which, given α and β allows us to pin down δ as a function of the investment rate in the data $\tilde{I/Y}$.

To back out ψ, note that from the intratemporal first order condition $u_L = u_C F_N$ which, given our functional form assumptions, we get $C^{-1} \frac{Y}{N}(1 - \alpha) = \frac{\psi}{1 - N}$. Suppose we want to have a model in which on average agents work 8 hours a day, namely 1/3 of their available time budget, which we assumed to be 1. Then we can set $N^{\star} = 1/3$ and manipulate the FOC to compute ψ. First,

recall from before that we know that $K^\star/N^\star = \left(\frac{\beta^{-1}-1+\delta}{\alpha}\right)^{\frac{1}{\alpha-1}}$. We can then write

$$\psi = \frac{1-N}{C}\frac{Y}{N}(1-\alpha) = \frac{1-N^\star}{N^\star}\frac{K^{\star\alpha}N^{\star 1-\alpha}(1-\alpha)}{C^\star}. \tag{44}$$

From the budget constraint, we know that $C^\star = Y^\star - I^\star = Y^\star - \delta K^\star \Rightarrow C^\star/N^\star = (K^\star/N^\star)^\alpha - \delta K^\star/N^\star$ so that

$$\psi = \frac{1-N^\star}{N^\star}\frac{(K^\star/N^\star)^\alpha(1-\alpha)}{(K^\star/N^\star)^\alpha - \delta K^\star/N^\star} \tag{45}$$

We have already established parameter values for $K^\star/N^\star$, $N^\star$, α and δ. Plugging them all in allows us to pin down a value of ψ.

The model is now fully calibrated in steady-state. If we want to have business cycle we go back to our assumption on the law of motion for lnA_t which implies that we need to back out ρ and σ. To estimate them, King and Rebelo (1999) propose a two-step procedure. First, note that we can obtain de-trended versions of output Y_t, hours N_t and capital K_t for a number of years. The production function assumption implies that in logs

$$\ln Y_t = \ln A_t + \alpha \ln K_t + (1-\alpha)\ln N_t. \tag{46}$$

Since we have already computed our value of α we can invert it and back out $\ln A_t$ for each year in our sample. This gives us a vector of T estimated Solow residuals. We can then estimate the regression

$$\ln A_t = \beta \ln A_{t-1} + \epsilon_t \tag{47}$$

and obtain $\rho = \hat{\beta}$ and the variance of the residual $\hat{\sigma}^2$. This completes the parametrization of the model.

From here, we can simulate the business cycle performance of the data by obtaining a sequence of generated shocks that follows 47 by discretizing the process into a Markov Chain. With the sequence of shocks and the agents' policy functions for C_t, N_t, K_{t+1}, Y_t we obtain the generated data of the model economy, which we compare the moments of the actual data.

This simple model, compared to the data, does surprisingly well. Table 1 provides the moments of the simple model compared to the data.[9]

There are a couple of well-known issues: first consumption is way too smooth. The agents have incentives to smooth consumption a lot as, for example, we do not distinguish between durables and non-durables. Secondly, hours worked are also not moving much relative to the data. Lastly, the model predicts a very strong correlation of macro variables with output, i.e. we get too much

[9]This table is from Nir Jaimovich's notes on RBC.

	Data		Model	
	σ_x/σ_y	$\rho_{y,x}$	σ_x/σ_y	$\rho_{y,x}$
Y	1	1	1	1
I	4.6	.91	4.4	.98
C	.82	.87	.29	.61
N	1.29	.87	.57	.96
A	.82	.77	.59	1

Table 1: Moments

comovement.

A lot of the coming extensions, frictions and complications are effectively aimed at either improving the fit of the model to the data or getting more bang for the buck, meaning getting the model to endogenously amplify the shocks we feed.

1.4 Extensions, Frictions and an Accounting Framework in RBC

In this section, we briefly go through some extensions that can help get the model closer to the data and some of the most common frictions we introduced in the literature.

1.4.1 Extensions

Indivisible Labour[10] One of the model's shortcomings was the inability to match the volatility of hours worked. In particular, the model implies a volatility that is counterfactually low, relative to the data. So far, we have modeled a household whose only labour supply margin is the intensive one. To make progress in both quantitative performance and realism, consider the following extension, based on the models in Hansen (1985) and Rogerson (1988). Maximizing with respect to labour subject to the standard budget constraint yields

$$u_l = w_t \lambda_t, \tag{48}$$

with λ being the multiplier on the budget constraint. Define the Frisch elasticity of labour supply as the elasticity of labour N_t with respect to the wage rate w_t, holding constant the marginal utility of wealth λ_t. This is a key parameter in determining the responsiveness of labour to TFP shocks in the RBC model. Formally this is

$$\left.\frac{\partial \ln N_t}{\partial \ln w_t}\right|_{\lambda_t} \tag{49}$$

[10] This is inspired by lecture notes by Nir Jaimovich and Eric Sims.

For example, suppose the household has preferences

$$u(C_t, 1 - N_t) = \ln C_t + \psi \ln(1 - N_t). \tag{50}$$

The FOC is

$$\frac{\psi}{1 - N_t} = w_t \lambda_t, \tag{51}$$

hence the Frisch elasticity is given by

$$\frac{\partial N_t}{\partial w_t} \frac{w_t}{N_t} = \frac{\psi}{\lambda_t w_t} \frac{1}{N_t} = \frac{1 - N_t}{N_t} > 0. \tag{52}$$

Consider instead these CRRA within period preferences of the household

$$u(C_t, N_t) = \ln C_t - \psi \frac{N_t^{1+\theta}}{1 + \theta}. \tag{53}$$

Then labour supply is implicitly defined by

$$w_t \lambda_t = \psi N_t^{\theta} \tag{54}$$

and the Frisch elasticity is given by

$$\left. \frac{\partial \ln N_t}{\partial \ln w_t} \right|_{\lambda_t} = \frac{1}{\theta}. \tag{55}$$

It is immediate to see that the highest elasticity is associated with $\theta \to 0$, which corresponds to the linear preferences case. So, in general, a model in which aggregate labour supply is given by linear preferences, the Frisch elasticity will be infinitely elastic. We build such a model by introducing labour indivisibilities. Consider a setting in which a large household is composed by individuals who either work $\bar{N}$ or 0 hours. A fraction τ_t of the household works $\bar{N}$ and the complement 0. The within-period preferences are then given by

$$u_t(C_t, \tau_t) = \ln C_t - \tau_t \psi \frac{\bar{N}^{1+\theta}}{1 + \theta} - (1 - \tau_t) \psi \frac{0^{1+\theta}}{1 + \theta}. \tag{56}$$

The labour income is then given by $w_t \tau_t \bar{N}$, and τ_t is a choice variable. You can think of τ as the outcome of a lottery such that in expectation, the household works $N_t = \tau_t \bar{N}$. Substituting in the utility and relabeling constants, we get

$$u_t(C_t, N_t) = \ln C_t - B N_t, \tag{57}$$

where $B = \psi \frac{\bar{N}^\theta}{1+\theta}$ is a constant. Note that in the preferences, we have a positive θ, so to some extent, at the single individual level, there is an upward-sloping supply, but at the household level, the elasticity is infinite and equivalent to the $\theta \to 0$ case above. Changing the basic RBC model along these lines allows us to have hours responding much more than before.

Digression on Aggregation The result that we can decouple the individual and the aggregate labor supply elasticity is an example of what aggregation can do. Next we can discuss three counterintuitive results of aggregation.

Sonnenschein-Mantel-Debreu (Anything Goes) Theorem Let the excess demand function $z(p)$ for $p \in \mathbb{R}^L_{++}$ satisfy the following properties:

1. z is continuous.
2. z is homogeneous of degree zero.
3. $p\dot{z}(p) = 0$ for all p. (Walras' Law)

Fix any $z: P_\varepsilon \to \mathbb{R}$ for z satisfying these properties, where

$$P_\varepsilon = \left\{p \in \mathbb{R}^L_{++} | p_k/p_l \geq \varepsilon > 0 \ \forall k, l\right\}$$

Then there exists a private ownership pure exchange economy (with L consumers) whose excess demand function coincides with z on P_ε.

In summary, we can "fit" any possible excess demand function, by choosing some specific structure for our economy. Without knowledge about the agents in our economy, any equilibrium outcomes and comparative statics (if they can be generated by an excess demand function satisfying above's assumption) are possible. However, once we have specific knowledge about agents' preferences, we can narrow down the set of possible outcomes.

Houthakker (1955) Consider a continuum of profit-maximizing firms, each with its own production function. Suppose individual production functions are Leontief, e.g., for firm j's production function, a_{1j} units of x_1 and a_{2j} units of x_2 are needed to produce one unit of output y_j. Input requirements a_1, $a_2, \ldots$ are heterogeneous across firms. If the input requirement distribution follows a generalised Pareto:

$$\varphi(a_1, a_2) = A a_1^{\alpha_1 - 1} a_2^{\alpha_2 - 1} \ (\alpha_1 \geq 1, \alpha_2 \geq 1),$$

then the aggregate production function has the Cobb-Douglas form:

$$Y = C X_1^{\gamma_1} X_2^{\gamma_2},$$

where C is constant and γ_1 and γ_2 are simple functions of a_1, a_2, etc.

We conclude that despite having extreme complementaries between inputs at the firm level, the aggregate production function exhibits significantly larger substitutability.

Boehm and Pandalai-Nayar (2022) Consider a competitive aggregating firm and monopolistically competitive intermediate goods firms (with measure 1) within an industry. The competitive aggregating firm uses a CES aggregator:

$$Y_t = \left(\int_0^1 \omega_{lt}^{\frac{1}{\theta}} y_{lt}^{\frac{\theta-1}{\theta}} \right)^{\frac{\theta}{\theta-1}}$$

where shocks ω are drawn i.i.d. from distribution G with unit mean and finite variance.

Intermediate producers' capacity can limit production in the short run. Denoting q_{lt} the firm's production capacity that's predetermined within the period, the production function is

$$y_{lt} = q_{lt} \min\{v_{lt}, 1\}.$$

Hence, q_{lt} denotes the production capacity since $y_{lt} \leq q_{lt}$. Production capacity is of the form $q_{lt} = z_t k_{lt}^{\alpha}$, where k_{lt} develops following the standard law of motion for capital.

It can then be shown that the industry's price index (its inverse supply curve) is given by

$$\ln P_t^Y = \mathcal{M}(\ln u_t) + \ln(mc_t),$$

where mc_t is the industry's marginal cost and $\mathcal{M}$ is the industry's log average markup, that is solely an increasing function of the industry's utilization rate and no other variables. The shape of $\mathcal{M}$ is a priori ambiguous, depending on G, but for all sensible parameterizations, $\mathcal{M}$ is a convex function of capacity utilization. In summary, we can obtain convex aggregate supply curves from firms whose supply is either horizontal (because of perfect competition) or vertical (whenever the constraint binds).

End of Digression

Capital Utilization We start by adding the possibility that firms use their plants at less than 100%. In the data, utilization is strongly procyclical and volatile. To introduce it, we make the following extensions to the baseline model. This is a version of the mode in Burnside and Eichenbaum (1996). First, the production function is

$$Y_t = A_t (u_t K_t)^{\alpha} L_t^{1-\alpha}, \tag{58}$$

where u_t is a choice variable for the firm. High utilization comes at a higher depreciation cost, with the law of motion being

$$K_{t+1} = (1 - \delta(u_t))K_t + I_t \tag{59}$$

and $\delta(u_t)$ has $\delta' > 0$.

Intuitively this extension amplifies the business cycle response to a given TFP shock. The firm now increases utilization after a positive TFP shock. In particular, the firm utilization FOC is

$$\alpha \frac{Y_t}{u_t} = \delta'(u_t) K_t. \tag{60}$$

From this condition, it is immediate that u_t is increasing in Y_t, hence in A_t. To see the increase in amplification we can parametrize $\delta(u_t) = \frac{1}{\gamma} u_t^{\gamma}$, with $\gamma > 1 > \alpha$. We use the firm's FOC and substitute it in the production function:

$$Y_t(u_t^*) = \alpha^{\frac{\alpha}{\gamma-\alpha}} A_t^{\frac{\gamma}{\gamma-\alpha}} K_t^{\frac{\alpha(\gamma-1)}{\gamma-\alpha}} L_t^{\frac{(1-\alpha)\gamma}{\gamma-\alpha}}. \tag{61}$$

There are a couple of things to note here: first, as $\gamma/(\gamma-\alpha) > 1$, the direct effect of A on Y increased compared to the benchmark model; secondly, the capital share is now lower, it used to be α, and now it's $\frac{\alpha(\gamma-1)}{\gamma-\alpha} < \alpha$ while the labour share increased. This also amplifies shocks because it gives more weight to the flexible input. The intuition for the higher elasticity is that individual agents now do not have to move up their labour supply, and therefore face increasing marginal disutility of working, for the aggregate amount of hours worked to increase.

This extension goes in the "more bang for the buck" direction. We now need smaller shocks to get the output movements right. It also helps the model square with the empirical observation that most variation in hours in the data is driven by the extensive, rather than the intensive margin.

Technological Returns We made many strong assumptions to get this far (and that's exactly the point of making assumptions). A particularly heroic set of restrictions we imposed is on the behaviour of the firm. A lot more on this will come later in these notes, but for now, let's just relax the assumption that technology exhibits constant returns to scale. First, let's distinguish between internal and external returns to scale. Internal returns pertain to the technology or, more in general, convexities within the boundaries of the firm (more on this in a second). External returns refer to effects outside the boundary of the firm, for example, the fact that a firm's behaviour may facilitate another firm's entry (more on this later). For now, let's assume the simplest form of internal increasing returns: increasing returns to scale in the production function. Formally, $Y = F(X)$ where F is an aggregator which is homogeneous of degree larger than 1 in the input set X. An alternative and even easier way is to say that $Y = A(Y)G(X)$ where G is homogeneous of degree 1 and $A' > 0$.[11] Now, suppose a firm is maximising profits. We get the usual conditions of $MPL = w$ and $MPL = r$. We know that by the Euler theorem for homogeneous functions $MPL \cdot L + MPK \cdot K = Y$ if the production function is CRS. Let's simplify and parametrize $A(Y) = AY^{1-1/\theta}$ with $\theta > 1$. The production function has the form $Y = (AF(L,K))^{\theta}$. Again now

[11] Reasons behind this will come back at the very end of these lecture notes.

the elasticity of Y to A is larger than 1 both mechanically and through endogenous forces similar to the ones described above.

For much much more on the topic of returns to scale, I recommend this beautiful (and hard) paper: Baqaee and Farhi (2020).

1.4.2 Frictions and Business Cycle Accounting

In this section, the goal is to do two things to our baseline model: first, introduce some simple frictions that are both reasonable and helpful in getting some facts right; second, provide a unifying framework to think about any friction in the RBC models family.

Irreversible Investment Throughout our discussion, we always allowed investment to be reversible. This implies that when a particularly bad shock hits the economy, the agent can, potentially, turn back the capital stock into consumption goods and eat it. Importantly the constraint that $I_t \geq 0, \forall t$ will have asymmetric effects. You can show that the Euler equation, which we use to study the optimal level of capital, will now include two Lagrange multipliers for the irreversibility constraint today and tomorrow. Importantly, shocks to A will have different effects on the likelihood that the constraint binds and, therefore, that the multipliers are positive.

In particular, it is immediate to note that a negative TFP shock will have larger effects on consumption than in the standard model. Intuitively the agent could disinvest capital to smooth consumption in our benchmark economy. In this setting, instead, more of the burden will fall on lower consumption today. This seems to go in the right direction as consumption was too little volatile.

Suppose instead that TFP goes up. In the standard model, we get an increase in consumption and investment. In this setting, that is still true, but the investment response will be muted. This is because, with some probability, the agent will face a binding constraint in the future and, therefore, will optimally invest less than in the benchmark model. By the budget constraint, this means that consumption will increase by more than in our standard model as most of the increase in output will be just consumed today. So we got ourselves a more volatile consumption over the cycle.

Time to Sell and Production Lags Another element that our benchmark model did not quite get right was the correlation structure between our outcomes of interest. We observed that variables had a very strong comovement. An important assumption in the model is that things are instantaneous (note that we did not specify what a period is). One way to partially break the comovement is to introduce production lags. The simplest way is to assume that firms can only sell today what they made yesterday. Denote Q_t sales, which have to clear with C_t, we can impose that $Q_t \leq S_t$ where S_t is the stock of inventories that the firms carried between period $t-1$ and t. Further, add the law of motion for inventories $S_{t+1} = S_t + Y_t - Q_t$.

If the economy is hit by a TFP shock, we will now observe that the marginal products of inputs increase today, but this can only show up in higher consumption tomorrow. This goes partially in the right direction, but we still have a very strong correlation with hours worked.

Interestingly, the introduction of inventories in RBC models was motivated by amplification/ dissipation considerations, rather than by getting correlations right across periods. Specifically, suppose competitive firms have convex production costs. This makes their problem concave in output, which, if their productivity fluctuates, induces a production smoothing motive. This is akin to the risk-averse household consumption smoothing motive. In this case, firms will use inventories to smooth out production, so they will output will typically comove less than sales with productivity. Importantly, this also implies that output is less volatile than sales, and inventories are countercyclical. This is counterfactual. We typically estimate that output is more volatile than sales and that inventories are adjusted procyclically. So inventories are a force of amplification rather than smoothing over business cycles. The way we rationalize this behaviour in our models is via a stock-out avoidance motive and fluctuations in demand. For more details on this, you can check Holt et al. (1960); Kahn (1987); Blinder and Maccini (1991); Ramey and West (1999); Bils and Kahn (2000); Khan and Thomas (2007); Alessandria et al. (2011); Ferrari (2024).

Adjustment Cost Throughout, we have assumed that adjustment of capital stocks is costless. We often think of this as a benchmark, but clearly not something that is very realistic. An easy way to get adjustment costs in the model is to assume that they are quadratic in the current stock of capital. So, for example, we can think of a firm incurring a cost of the form $\frac{\gamma}{2}I^2$. This will complicate a little the firm problem we studied before, as now the firm incurs larger costs the larger the adjustment (investment) it has to make. Here we assume firms own the capital and maximize the value of the firm, given by

$$V(A,K) = \max_{K',N} AK^{\alpha}N^{1-\alpha} - WN - I(1+\frac{\gamma}{2}I) + R^{-1}\mathbb{E}V(A',K'). \quad \textbf{st} \tag{62}$$

$$K' = (1-\delta)K + I. \tag{63}$$

Where R^{-1} is the inverse of the interest rate with which the firm discounts the future. In this setting, the labour choice is static, while the capital one is dynamic and will include the cost of adjusting the current stock. The FOC for future capital yields

$$1 + \gamma I = R^{-1}\mathbb{E}[\alpha A' K'^{\alpha-1} N'^{1-\alpha} + (1-\delta)(1+\gamma I')]. \tag{64}$$

Namely, the marginal cost of investment (including adjustment cost) must be equal to the expected MPK and forgone cost of future expected investments. The latter effect comes from the fact that an extra unit of savings today allows us to save $(1-\delta)$ fewer units tomorrow to achieve the same capital stock. Using the optimal labour choice and denoting $q \equiv 1+\gamma I$ and $\theta \equiv \alpha \left(\frac{1-\alpha}{W}\right)^{\frac{1-\alpha}{\alpha}} A^{1/\alpha}$

we get

$$q = \frac{1}{R}\mathbb{E}[\theta' + (1-\delta)q']. \tag{65}$$

Solving this recursion, while noting that θ is an index of how profitable the firm is, we obtain that investment depends on the expected future profitability of the firm

$$q_t = \frac{1}{R}\mathbb{E}_t \sum_{s=0}^{\infty} \left(\frac{1-\delta}{R}\right)^s \theta_{t+s+1}. \tag{66}$$

This is known as the Q-theory of investment. You can think of this as the firm equivalent of the permanent income hypothesis for consumption-saving problems.

Another successful stream of models noted that this approach implies that firms are constantly adjusting capital. We know from the data that things are much less continuous than that. We often observe that no firm invests small quantities. They either invest zero or at least a sizable amount. This form of inaction is typically well replicated by models in which a fixed adjustment cost is associated with investment. This means that the cost does not vary with the investment amount, and therefore, it will not be profitable to invest small sums, but it will be very profitable to invest large quantities.

To see this more formally, you can think of the problem of a firm that has a stock of capital K and has to pay a fixed cost f whenever they have a non-zero investment I. In other words, letting capital depreciate is for free, but any future capital change that is different from depreciation requires the payment of a fixed cost f. We can write the value of adjusting capital as

$$V^{Ad}(A,K) = \max_{K'} \pi - f + \beta V(A',K'), \tag{67}$$

while the value of keeping capital is given by

$$V^{Nad}(A,K) = \pi + \beta V(A',(1-\delta)K), \tag{68}$$

note that there is no optimization because capital is not changing, there is no fixed cost, and tomorrow's capital is just today's undepreciated capital. Then the full problem of the firm is given by

$$V(A,K) = \pi + \max\left\{\max_{K'} \beta V(A',K') - f, \beta V(A',(1-\delta)K)\right\}. \tag{69}$$

Under this formulation, we will have that the firm does not invest in most periods until the value of capital hits a specific lower bound, at which point it will invest a lot, up to some specific upperbound and let it depreciate to the lower bound again. This type of optimal policy is known as *s,S bounds* policy to denote the lower s and upper bound S.

Figure 3 shows the optimal policy of the investment rate for firms upon receiving a productivity shock A' relative to their previous productivity A.

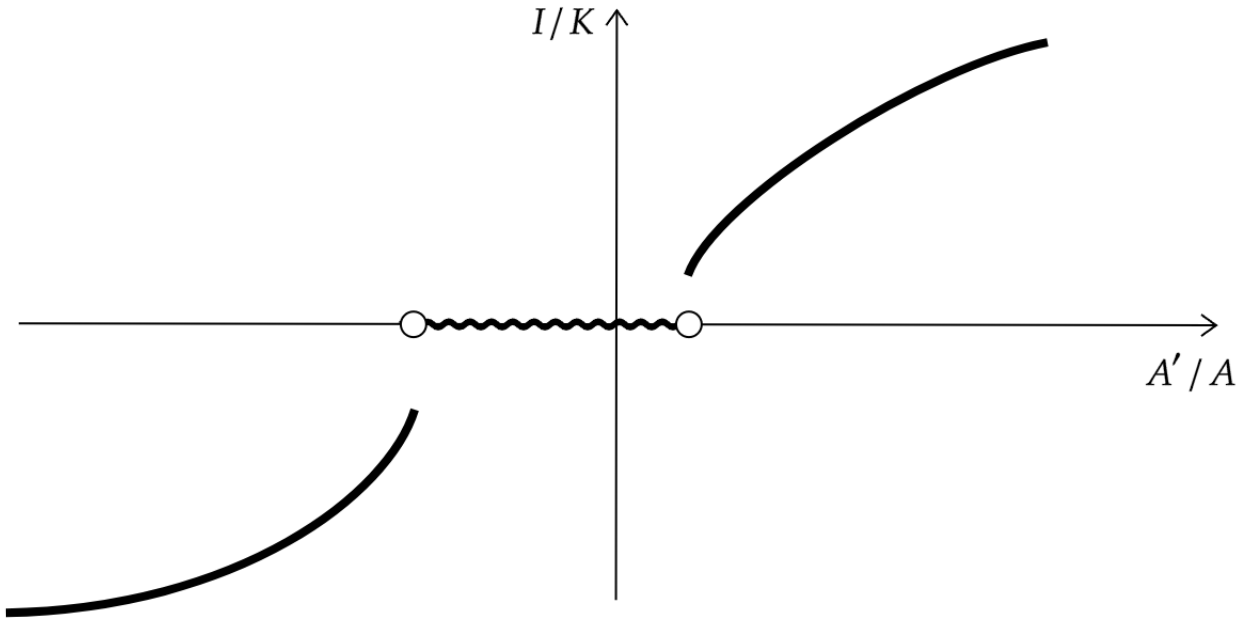
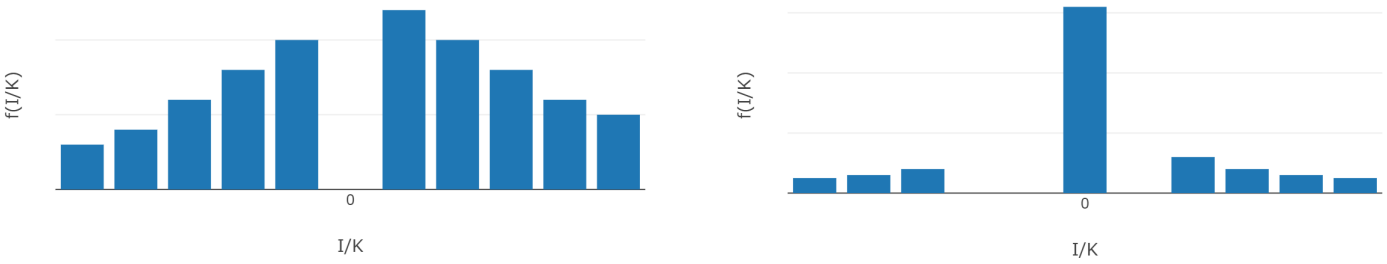


Figure 3: Investment Policy

Note a few important things. First, it is asymmetric. Capital suffers a natural decay because of depreciation, so more firms invest positive than negative amounts, even if the shock distribution is symmetric. Second, nobody invests or divests little as it is not worth paying the fixed cost for small changes in capital. As a consequence, there is a lot of mass around zero investment.

Suppose now that there is a distribution of firms that receive productivity shocks. How would the empirical investment distribution look like in the cases of smooth vs fixed adjustment costs? Figure 4 plots the two distributions.



(a) quadratic adjustment cost

(b) fixed adjustment cost

Figure 4: Investment Rate Distributions

First, note that we should expect no mass at zero investment when adjustment costs are

quadratic, while there should be a mass point at zero when they are fixed. Second, they should still be asymmetric since depreciation operates as a disincentive to dissave.

Finally, what does this look like in the data? Figure 5 plots the empirical investment rate distribution from Cooper and Haltiwanger (2006).

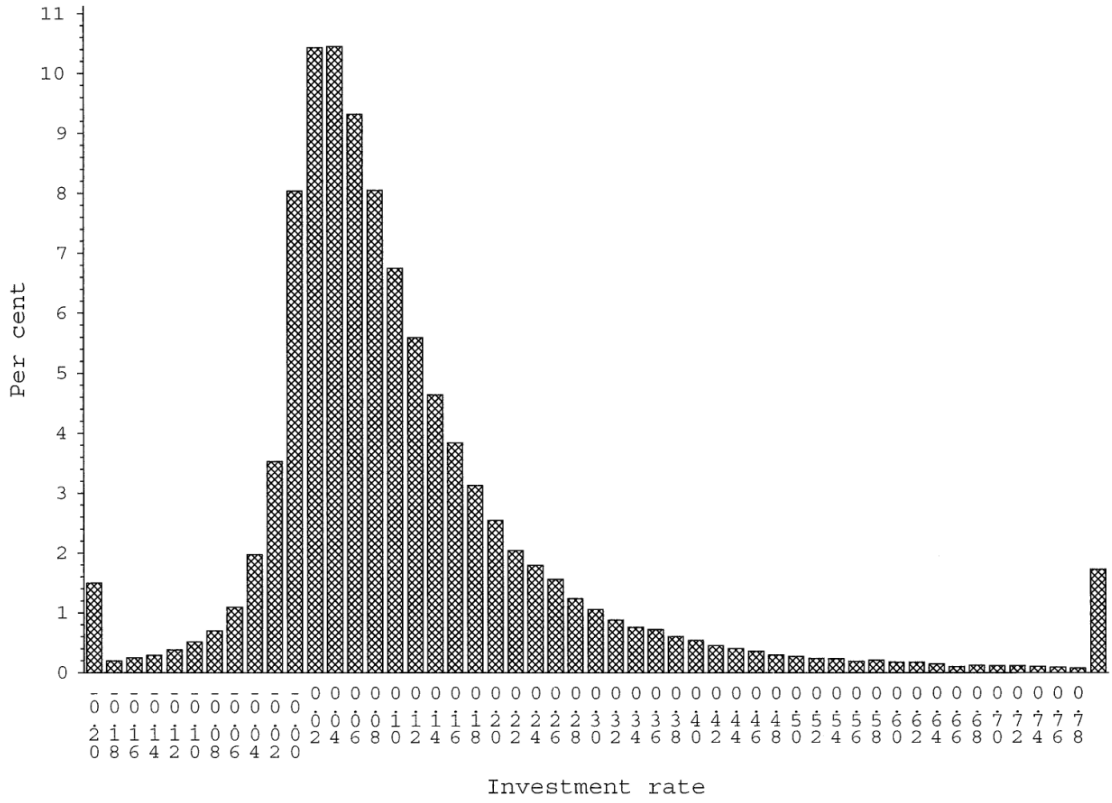


Figure 5: Empirical Investment Rate Distribution

The distribution has a significant mass point at 0, with approximately 8% of firms not investing in a given year.[12] It is right-skewed, consistently with the natural drift from depreciation. We conclude that we need non-convex adjustment costs (fixed) to replicate these empirical patterns.

Digression on Price Adjustment Costs and Sticky Prices As a last point on adjustment costs, note that this is one possible way of having nominal rigidities in a model. It is enough to move the adjustment costs to prices. More precisely, if we assume that firms have to pay some cost to change the price from p to p', we will get a natural stickiness in the economy. Exactly following the logic for investment, if we introduce quadratic costs of adjusting prices, we will get that prices move less than in the frictionless version of the economy, meaning that $|p' - p|$ is smaller than the no adjustment cost counterpart. If we instead introduce a pure menu cost, meaning that upon changing the price by any amount, we have to pay some fixed cost f, we will get an inaction

[12]The distribution is both top and bottom-coded. So you should not over-interpret the spikes at the ends of the distribution.

region. Namely, firms will only change prices of large amounts and find it optimal to keep prices constant if the optimal change is tiny. These are models of "optimal" nominal rigidities in that firms actively decide to have sticky prices. Consider the following three different model setups. In all, there is a continuum of firms. Each firm receives an idiosyncratic productivity in any given period. The state variables are their price p_{-1} from the previous period and the productivity shock A with linear production functions. Make the input x price the numeraire. Suppose, further, that firms face an isoelastic residual demand $q(p)$. To keep things as simple as possible, suppose that these price adjustment frictions are present today and tomorrow and gone thereafter.

Rotemberg (1982). A firm faces quadratic adjustment costs of the form $(p - p_{-1})^2$. Profits given in case of adjustment are given by $p^\star q(p^\star) - q(p^\star)/A - \gamma/2(p - p_{-1})^2$, where $\gamma > 0$ is a cost of adjusting the price. The profits, in case of no adjustment, are $p_{-1}q(p_{-1}) - q(p_{-1})/A$. Firms maximize the present discounted value of profits. A firm compares the two options and chooses whether to adjust and if so how much. The problem of the firm is

$$\max_{p_t, p_{t+1}, \ldots} \pi_t(p_{t-1}, p_t, A_t) + \mathbb{E}\left[R^{-1}\pi_t(p_t, p_{t+1}, A_{t+1}) + \sum_{s=2}^{\infty} R^{-t-s}\pi_{t+s}(p_{t+s-1}, p_{t+s}, A_{t+s})\right] \tag{70}$$

Because from period $t + 2$ onwards prices can be adjusted freely, we only have to care about the effect of p on profits in t and $t + 1$ and set

$$\frac{\partial \pi_t}{\partial p_t} + R^{-1}\frac{\partial \mathbb{E}\pi_{t+1}}{\partial p_t} = 0 \tag{71}$$

Based on the discussion on quadratic adjustment costs, we know all firms will adjust, and no firm will adjust a lot.

Calvo (1983) In any given period, a firm has the possibility to adjust prices with probability ϕ. This adjustment is free. With complement probability $1 - \phi$, it cannot adjust prices, as if the cost of adjustment is infinite. Firms will maximize the present discounted flow of profits taking into account the fact that there may be unable to adjust prices for many periods. You can show that in this model, the optimal pricing rule is forward-looking and depends on the expectation on productivity shocks in the future. In this model, individual firms' prices are either flexible or stuck, depending on the realization of the resetting shock. However, the aggregate price is *sticky* as it can only adjust slowly to changes in the aggregates. Importantly, the Calvo model has no selection effects, since who gets to adjust is random.

Barro (1972) A firm faces a fixed menu cost to adjust its price across periods. If the firm does not adjust, it sells at the previous period price and obtains profits $q(p_{-1})p_{-1} - p_{-1}/A$. If it adjusts it obtains $q(p^\star)p^\star - q(p^\star)/A - f$, where f is a fixed menu cost. In this model, there will be strong selection effects: the firms whose price is *off* by more are those who will reset their price. As a consequence, the cost of nominal rigidities is typically smaller, since the firms that are

furthest from the optimal price level (and therefore output) have the most incentive to pay the fixed cost and reset. Much like in the fixed investment cost model, most firms will not reset their price. Particularly, the ones who receive the smallest deviations relative to the previous period shock will elect not to pay f and keep the old price. See Golosov and Lucas Jr (2007) for more details on this topic.

End of Digression

Business Cycle Accounting So far, we wrote down a model in which we govern the sources of fluctuations and put them up against data to figure out how well we do. Chari et al. (2007) propose an alternative route. We can start from the data and ask which part of the model better explains what we observe. You can think of this exercise as a sort of variance decomposition. They write a model in which there are 4 wedges. These are distortions in optimality conditions. Foreshadowing the next section, we can write

$$MRS = (1 + \tau_n)MPN \tag{72}$$

Meaning that the MRS of the household does not equal the MPN of the firm, so the labour market is at some suboptimal clearing level. We can do the same for something that looks like the TFP shock we had, something that distorts the investment margin, or even add a government and add distortionary taxes. These wedges are something we tinker with to match the data, and then we can shut them down one at the time and figure out how much worse we do. This is like asking which independent variable has the highest explanatory power in a regression. Once we know which of the wedges seems to be the most important one, we can then dig deeper into why that is the case.

In the paper, they apply this to the Great Depression and show that if we write a model with the labour or the efficiency wedge, we get pretty close to the actual path of output in the data. On the other hand, if we allow only the investment wedge to operate, we get that output actually increased. They actually show that the efficiency wedge explains almost all of the immediate fall in output while the labour wedge explains a lot of the slow recovery.

Digression on Policy in RBC Models We conclude this section with a short discussion on the scope for policy in RBC models. We showed before that the decentralized equilibrium in the basic RBC framework is Pareto efficient. It follows immediately that any policy that distorts any agent's choice is going to reduce welfare. This is a powerful result and a useful benchmark, but also one that we should not be surprised by. We set up a model in which there is no distortion, every agent is rational and price taker. It's an immediate implication of the welfare theorems that there is nothing to be done here.

End of Digression

2 Search Theory[13]

This section introduces the basic frameworks to think about labour market frictions. First, we develop the McCall random search model, where workers and firms meet at random. In this model firms are just a posted wage, they do not really play any active role. Then we study the Diamond, Mortensen and Pissarides model, in which firms post vacancies and workers decide whether to take the job or not, after which a bargaining phase follows. Lastly, we introduce shocks to the DMP model to study how we can think of labour market frictions and their role over the business cycle.

For a broader non-technical overview of why search theory is so interesting and important, please read the Nobel lectures by Diamond, Mortensen and Pissarides in 2010 and the review article by Jim Albrecht in the course folder.[14]

2.1 Random Search

This section presents a random search model, first in discrete, then in continuous time. The model is based on McCall (1970).

2.1.1 Discrete Time

Consider an individual maximizing her expected utility

$$\mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t c_t \quad \text{st } c_t = y_t \tag{73}$$

This agent is risk neutral and is thereby only interested in maximizing the level of consumption[15]. The discount factor is such that $\beta \in (0, 1)$.

The income of the agent can take two possible values, depending on whether they are employed or unemployed

$$y_t = \begin{cases} w & \text{if employed} \\ b & \text{if unemployed} \end{cases} \tag{74}$$

Where w denotes the wage obtained from working a fixed and normalized to unity amount of hours and b is the unemployment subsidy. The state of employment is absorbing, meaning that once the worker is employed there is no chance that they will go back to unemployment. An unemployed agent instead receives work offers at given wages and can either accept or reject these. Intuitively

[13] This section is heavily inspired by Philipp Kircher's lecture notes on search theory. Any mistake, however, is most certainly mine.

[14] For a textbook analysis of most of the material covered please see Cahuc et al. (2014) or Pissarides (2000).

[15] It is possible to rationalize this assumption by having a complete set of Arrow securities, such that, conditional on the lifetime income, the agent will smooth consumption fully across periods.

the worker will have some sort of threshold rule such that if the offered wage is larger than a given threshold they will accept the offer, otherwise they will keep on searching.

To describe in more detail the problem, define the expected payoff of accepting an offer with wage w as $V(w)$ [16]. Since the state is absorbing this amounts to getting w for the rest of the worker's (infinite) life, so $V(w) = \sum_{t=0}^{\infty} \beta^t w = \frac{w}{1-\beta}$.

Define the value of being unemployed U as the sum of the unemployment benefit and the value of searching, more formally

$$U = b + \beta \mathbb{E}[\max\{V(w), U\}] \tag{75}$$

In this case, the value of search is the value of the workers' option to accept an offer w or keep on searching. Denote $\hat{V}(w) \equiv \max\{V(w), U\} = \max\left\{\frac{w}{1-\beta}, b + \beta \mathbb{E}\hat{V}(w)\right\}$ as the value of receiving an offer w that has not yet been accepted.

Given this definition, define the aforementioned threshold R as the wage level such that $V(R) = U$. Since this wage level is the one that makes the agent indifferent between accepting and rejecting the offer this wage level is referred to as the *reservation wage*. Note that since $V(w)$ is linearly increasing in w and U is independent of w, R exists and is unique. Also, R is such that if $w < R$ then $V(w) < U$ and vice-versa. Using again the absorbing state property $V(R) = \frac{R}{1-\beta}$ and the fact that $V(R) = U$, we obtain

$$R = (1-\beta)b + \beta(1-\beta)\mathbb{E}\hat{V} \tag{76}$$

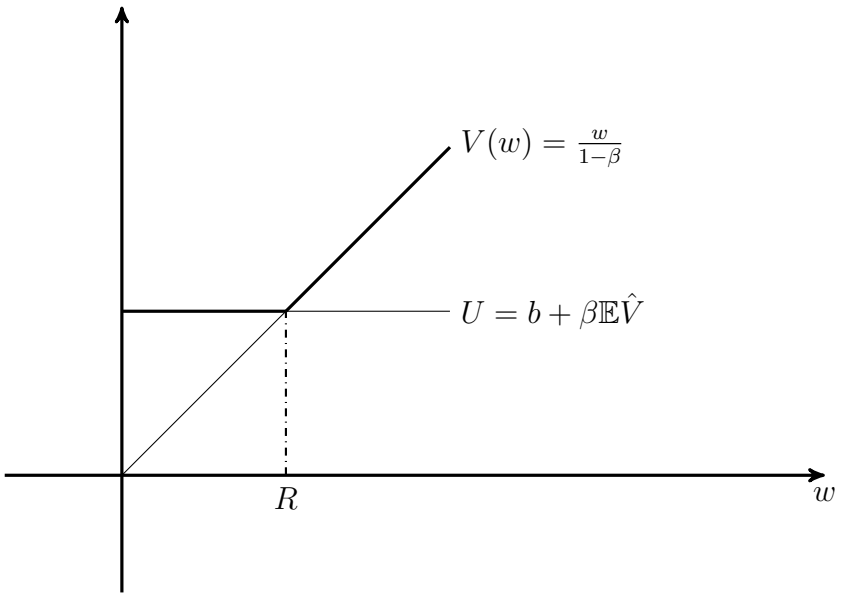


Figure 6: Value Function and Reservation Wage

[16] Notice that since this is an infinite horizon model and the wage is constant over time this expected payoff does not depend on t

Note that in Figure 6 the thick line represents the value of $\hat{V}(w)$.

Note that it is possible to rewrite $\hat{V}(w)$ as

$$\hat{V}(w) = \begin{cases} \frac{w}{1-\beta} & \text{for } w \geq R \\ \frac{R}{1-\beta} & \text{for } w < R \end{cases} \tag{77}$$

From the definition of the value of unemployment, multiplying by $1-\beta$ on both sides, we obtain:

$$(1-\beta)U = (1-\beta)b + (1-\beta)\beta\mathbb{E}[\max\{V(w), U\}] \tag{78}$$

Then by noting that $R = (1-\beta)U$ and under the assumption that $w \sim F(w), w \in [0, \infty)$

$$\begin{aligned} R &= (1-\beta)b + \beta\mathbb{E}[\max\{(1-\beta)V(w), (1-\beta)U\}] \\ R &= (1-\beta)b + \beta\mathbb{E}[\max\{w, R\}] \\ R &= (1-\beta)b + \beta\int_0^\infty \max\{w, R\}dF(w) \end{aligned} \tag{79}$$

Define a mapping $T : \mathcal{R} \to \mathcal{R}$ by

$$T(R) = (1-\beta)b + \beta\mathbb{E}\max\{w, R\} \tag{80}$$

So that there is a solution $R = T(R)$. This is assured by the Banach fixed-point theorem since $T(R)$ is a contraction. To see this, note that the definition of a contraction is as follows.

$$|T(R) - T(R')| \leq k|R - R'| \quad \text{for some } k \in [0, 1) \tag{81}$$

Let $R' < R$, then using the definition of $T(R)$ we want to show that

$$\beta\left|\int_0^\infty \max\{w, R\}dF(w) - \int_0^\infty \max\{w, R'\}dF(w)\right| < k|R - R'| \tag{82}$$

Start by noting that being $R' < R$ we have $\int_R^\infty \max\{w, R\}dF(w) = \int_R^\infty \max\{w, R'\}dF(w) = \int_R^\infty wdF(w)$, hence the LHS becomes

$$\beta\left|\int_0^R RdF(w) - \int_0^R \max\{w, R'\}dF(w)\right| \tag{83}$$

By the assumption that $R' < R$ we can get rid of the absolute value and we know that the following

is true because of the max operator

$$\beta\left[\int_0^R RdF(w) - \int_0^R \max\{w, R'\}dF(w)\right] < \beta\left[\int_0^R RdF(w) - \int_0^R R'dF(w)\right] \tag{84}$$

Using the RHS we have

$$\beta\left[\int_0^R RdF(w) - \int_0^R R'dF(w)\right] = \beta(R - R')F(R) < \beta(R - R') \tag{85}$$

Where the inequality comes from $F(R) < 1$. The proof is completed by having $k = \beta$.

Since $T(R)$ is a contraction we have that the sequence defined by $R_{n+1} = T(R_n)$ converges to the solution as $n \to \infty$.

Going back to the definition of R, it is possible to rewrite equation (76) making use of (77) under the assumption that $w \sim F(w), w \in [0, \infty)$

$$\begin{aligned}
R &= (1-\beta)b + \beta\left[R\int_0^R dF(w) + \int_R^\infty wdF(w)\right] \\
R &= (1-\beta)b + \beta\left[R\int_0^R dF(w) + \int_R^\infty wdF(w)\right] \pm \beta R \\
(1-\beta)R &= (1-\beta)b + \beta\left[R\int_0^R dF(w) + \int_R^\infty wdF(w) - R\int_0^\infty dF(w)\right] \\
(1-\beta)R &= (1-\beta)b + \beta\int_R^\infty (w - R)dF(w) \\
R &= b + \frac{\beta}{1-\beta}\int_R^\infty (w - R)dF(w)
\end{aligned} \tag{86}$$

Where the 3rd equation is derived from the fact that $1 - \int_0^R dF(w) = \int_R^\infty dF(w)$.

Equation (86) states that the utility from accepting an offer of R must be the same as the utility from rejecting the offer, which is given by the instantaneous unemployment benefit and the expected surplus from accepting a better offer tomorrow. The last element of eq. (86) is referred to as the worker surplus and is often defined as $\varphi(R) = \int_R^\infty (w - R)dF(w)$, this is analogous to the consumer surplus in a consumption model.

To conclude, in this model the optimal policy of the agent is a stationary threshold rule R. This strategy yields an expected utility, for an unemployed worker, equal to $\frac{R}{1-\beta}$.

Finally, note that the probability of accepting a job offer is independent of time and is the hazard rate $H = 1 - F(R)$ where $F(\cdot)$ denotes the cumulative density function. From H the duration of unemployment is $\frac{1}{H}$. See appendix A.2 for the derivation.

2.1.2 Continuous Time

In this section, we turn to the continuous time discussion of the basic search model.

Start by assuming that there is a random number of offers in a given period. Denote the number of offers by n and let $Prob(\#\text{ of offers } = n) = a(n)$.
Given $n > 0$ denote by $W = \max\{w_1, ..., w_n\}$ the best offer received. Also for a given $n > 0$ let $G(W, n)$ be the CDF of the best offer from the n offers.
As before one can start by looking at the value of unemployment, but for the time being assume that we do not know whether we are in the steady state or not, i.e. the value of unemployment might depend on time, so

$$U_t = b_t + \beta \sum_{n=1}^{\infty} a_t(n) \int_0^{\infty} \max\{V_{t+1}(W), U_{t+1}\} dG(W, n) + \beta a_t(0) U_{t+1} \tag{87}$$

Where again the term in brackets represents the gain above the unemployment income, or, in other words, the value of search, whereas the last term represents the value in case no offer is received.

Assume that n is generated by a stationary Poisson process, such that in any time interval Δ, the process $a(n, \Delta)$ satisfies

1. $a(1, \Delta) = \alpha\Delta + o(\Delta)$ for some $\alpha > 0$
2. $\sum_{n=2}^{\infty} a(n, \Delta) = o(\Delta)$

Where $o(\Delta)$ is any function such that $\frac{o(\Delta)}{\Delta} \to 0$ as $\Delta \to 0$.

Digression on Poisson Processes An important property of Poisson processes is that they are *memoryless*, meaning that the probability distribution of the time until the next arrival is constant, i.e. independent of history.

It is possible to show that the properties are satisfied by the Poisson process

$$a(n, \Delta) = \frac{(\alpha\Delta)^n e^{-\alpha\Delta}}{n!} \tag{88}$$

We can show that the first property is satisfied as follows: start by writing

$$\begin{aligned} a(1, \Delta) = \alpha\Delta + o(\Delta) = \alpha\Delta e^{-\alpha\Delta} \Rightarrow \\ o(\Delta) = \alpha\Delta(e^{-\alpha\Delta} - 1) \Rightarrow \\ \frac{o(\Delta)}{\Delta} = \alpha(e^{-\alpha\Delta} - 1) \xrightarrow{\Delta \to 0} \alpha(1 - 1) = 0 \end{aligned} \tag{89}$$

The second property can be shown as follows:

$$\begin{aligned} \sum_{n=2}^{\infty} a(n, \Delta) &= \sum_{n=0}^{\infty} a(n, \Delta) - a(1, \Delta) - a(0, \Delta) \\ &= 1 - \alpha\Delta e^{-\alpha\Delta} - e^{-\alpha\Delta} \end{aligned} \tag{90}$$

Dividing by Δ

$$\sum_{n=2}^{\infty} \frac{a(n,\Delta)}{\Delta} = \frac{1 - e^{-\alpha\Delta} - \alpha\Delta e^{-\alpha\Delta}}{\Delta} \tag{91}$$

When taking the limit $\Delta \to 0$ this is an indeterminate form, therefore one needs to apply L'Hopital's rule.

$$\frac{\alpha e^{-\alpha\Delta} - \alpha e^{-\alpha\Delta} + \alpha^2\Delta e^{-\alpha\Delta}}{1} \xrightarrow{\Delta\to 0} 0 \tag{92}$$

Denote the time until the next arrival by $\tilde{c}$, then $\tilde{c}$ has a CDF

$$\Phi(t) = Prob(\tilde{c} \leq t) = 1 - Prob(\tilde{c} > t) = 1 - a(0,t) = 1 - e^{-\alpha t} \tag{93}$$

The third equality comes from the fact that the "probability that the next arrival period will be larger than t" is equivalent to the "probability that until at least t there will be zero arrivals" which given the process definition is $a(0,t)$. Given the CDF described above the PDF of $\Phi'(t) = \alpha e^{-\alpha t}$. Note that both are independent of the history and of the exact choice of Δ. Given these definitions, it is possible to compute the mean time until the next arrival as

$$\begin{aligned}
\mathbb{E}\tilde{c} &= \int_0^\infty t d\Phi(t) = \int_0^\infty t\Phi'(t)dt \\
&= \int_0^\infty t\alpha e^{-\alpha t}dt = -t\alpha e^{-\alpha t}\big|_0^\infty + \int_0^\infty e^{-\alpha t}dt \\
&= 0 + -\frac{1}{\alpha}e^{-\alpha t}\Big|_0^\infty = \frac{1}{\alpha}
\end{aligned} \tag{94}$$

The parameter α is referred to as the arrival rate.

End of Digression

Suppose that the length of the period is Δ. Denote the payoff of receiving y for time of length Δ as $y\Delta$, and, similarly, $b\Delta$ and $w\Delta$ are the unemployment benefit and the wage. Also, given the length, the discounting is $\beta = e^{-r\Delta}$ (more on this in the appendix A.3). Using the previous definition in eq. (87), substituting the payoffs, subtracting $e^{-r\Delta}U_t$ on both sides and dividing by Δ yields:

$$\begin{aligned}
\frac{1 - e^{-r\Delta}}{\Delta}U_t = \frac{\Delta b}{\Delta} + e^{-r\Delta}\sum_{n=1}^{\infty}\frac{a_t(n,\Delta)}{\Delta}\int_0^\infty [max\{V_{t+\Delta}(W), U_{t+\Delta}\} - U_t]dG(W,n) \\
+ e^{-r\Delta}a(0,\Delta)\frac{(U_{t+\Delta} - U_t)}{\Delta}
\end{aligned} \tag{95}$$

Taking the limit as $\Delta \to 0$

$$rU_t = b + \alpha \int_R^\infty (V_t(w) - U_t) dF(w) + \dot{U}_t \tag{96}$$

Where $\dot{U}$ represents the time derivative.
Note that in the last equation, we can substitute $G(W, n)$ with $F(w)$ because the probability of receiving more than one offer vanishes as Δ goes to 0.

Now for the remainder of the analysis assume that we are in steady state, so we can drop the time indices and the time derivative since it will be zero, by definition of the steady state. This then lets us rewrite equation (96) as follows:

$$rU = b + \alpha \int_R^\infty (V(w) - U) dF(w) \tag{97}$$

Again the value of accepting an offer of w is

$$V(w) = \int_0^\infty e^{-rt} w dt = \frac{w}{r} \tag{98}$$

By this, we have again that $w = rV(w)$. From $R = rV(R) = rU$ we obtain the following:

$$R = b + \frac{\alpha}{r} \int_R^\infty (w - R) dF(w) = b + \frac{\alpha}{r} \varphi(R) \tag{99}$$

This is the continuous time reservation wage (in the appendix A.4 an alternative derivation is provided). Finally, as in the discrete time case, it is possible to define the hazard rate H as

$$H = \alpha(1 - F(R)) \tag{100}$$

With all this in the bag, we get to discuss one beautiful result of this theory. Assume firms all post one wage w^*. Then all the distribution would be

$$F(w) = \begin{cases} 0 & w < w^* \\ 1 & w \geq w^* \end{cases} \tag{101}$$

Using the definition of the reservation wage

$$R = b + \frac{\alpha}{r} [w^* - R] \tag{102}$$

hence

$$R = \frac{1}{1+\frac{\alpha}{r}}b + \frac{\frac{\alpha}{r}}{1+\frac{\alpha}{r}}w^* \tag{103}$$

This states that the reservation wage is a weighted average of the unemployment benefit and the posted wage. Note that $R < w^*$ if $w^* > b$. Then note that if firms were allowed to post take it or leave it offers they would all post R since at anything above the reservation wage they would be foregoing profits. This implies $w^* = b$ independently of the arrival rate of offers or of time discounting. You can reach this conclusion by noting that any firm posting $w^* > R$ would face entrants undercutting the offer till the market converges to R. Also note that $w^* = b$ means that the worker is not extracting any value, which means that if we included an $\epsilon > 0$ search cost nobody would search and the market would collapse to a monopoly outcome.

The result that firms all offer the reservation wage is known as the Diamond Paradox. It shows that, in this context, an arbitrarily small search friction is enough to give all market power to the firms and have them extract the entirety of the surplus. In this setting, workers demand a reservation wage which is given by the unemployment benefit and compensation for foregoing potentially higher wages. As firms coordinate to only offer the reservation wage there is no foregone value from taking a job at R. As a consequence, the reservation wage collapses to the unemployment benefit b. We will return to this result and how to break it after we discuss how to construct flow equations.

2.2 Flow Equations

Before introducing the DMP model, it is useful to characterise how the stock of employed and unemployed workers move in the simpler economy of the McCall model.

These quantities dynamics can be described by the evolution of *flow equations*. For simplicity, we will look at the continuous time version but the discrete counterparts follow the same intuition.

Define the stock of agents employed at wages below w as $E(w)$. Further, assume that at rate δ the stock of current jobs is destroyed, such that every second $\delta(1-u)$ workers become unemployed. Then the flow equation of this mass is

$$\dot{E}(w) = u\alpha F(w) - E(w)(\delta + \alpha_1(1 - F(w))) \tag{104}$$

This equation describes the net change of the mass of workers $E(w)$. The first term states that out of the unemployed agents some, with probability α, will receive a wage offer; out of the unemployed agents who received a wage offer, a mass $F(w)$ will receive (and accept since they are all above the reservation wage) an offer lower or equal to w; the second term states that out of the agents employed at w or less, some will have their match destroyed (with probability δ) and some other will receive, with probability α_1, a wage offer that will be above w with probability $1 - F(w)$ and

will therefore accept it.

Note that here $E(w)$ is almost like a CDF of workers employed at wages below w, the difference being that it does not integrate to 1, i.e. $E(\bar{w}) = 1 - u$.

We can then define the flow equation of unemployment as

$$\dot{u} = -u\alpha + (1-u)\delta \tag{105}$$

This is saying that the flow of agents into unemployment decreases when unemployed agents receive a job offer (and accept it) and goes up when employed agents have their match disrupted. By noting that the definition of steady state in this context is $\dot{u} = 0$

$$u = \frac{\delta}{\alpha + \delta}. \tag{106}$$

Intuitively, the steady state unemployment decreases in the offer arrival rate and increases in the job destruction rate.

2.3 On-the-Job Search, Layoffs and the Diamond Paradox

We can now go back to the Diamond Paradox. Clearly, the degenerate wage distribution is an undesirable property of the model discussed so far. To solve it we can dig deeper into the reasons behind it. The result is effectively driven by firms having market power over workers. This market power arises because today workers only get to see this one job (temporary monopoly) and because if they accept an offer they are stuck in that job forever. There is no reason for firms to try to compete to attract workers. A simple yet powerful extension to break this result is called on-the-job-search. This is the idea that at some rate new job offers arrive also for employed workers. Intuitively, this implies that workers can take an offer today and then move to a better job tomorrow. It turns out that this extension is enough to break the Diamond Paradox. Similarly one can model quits or job destruction to make the model a little closer to the real world.

In the previous models we assumed that employment was a permanent condition (absorbing state), in this section we relax this assumption by introducing the possibility for the worker to get laid off or to voluntarily quit.

One can start by assuming that there is an exogenous probability δ that the employment relationship is broken (layoffs). This clearly changes the values of the problem for the worker, since now wage offers are not permanent anymore. This gives rise to the following equation for the value of being employed [17].

$$V(w) = w + \beta[\delta U + (1-\delta)V(w)] \tag{107}$$

[17] In the following we already assume that a worker never receives more than one offer

Now in addition to there being layoffs, one can also introduce the fact that already employed workers are offered new jobs. When doing so, assume that agents have different probabilities of receiving an offer depending on their current employment status. In particular, let α denote the probability of receiving an offer if the worker is currently unemployed and let α_1 denote the probability for employed agents. Note that the latter element is what will give rise to agents switching jobs (quits). Given these definitions

$$U = b + \beta[\alpha \mathbb{E} \max\{V(w), U\} + (1 - \alpha)U] \tag{108}$$

Applying the same trick as before (subtracting βU on both sides gives)

$$(1 - \beta)U = b + \beta\alpha \mathbb{E} \max\{V(w) - U, 0\} \tag{109}$$

Where again the last element is what an agent gets on top of the unemployment benefit, in expectation, by accepting an offer.
Using this in the value of a job offer and including the possibility of getting offers while on the job

$$V(w) = w + \beta\delta U + \beta(1 - \delta)[\alpha_1 \mathbb{E}_w \max\{V(w), V(w')\} + (1 - \alpha_1)V(w)] \tag{110}$$

Where the equation states that on top of the wage itself, the worker maybe be unemployed with probability δ, if instead the match is not broken then with probability α_1 another offer will be made and the worker will be able to pick the highest of the current and the newly offered wage and, finally, with probability $1 - \alpha_1$ the worker does not receive a new offer.

It is possible to define a mapping $T(V)$ such that the problem will have a solution $V = T(V)$. T will again be a contraction, guaranteeing uniqueness of the solution. Furthermore, since T maps increasing functions into increasing functions (you can see this by noting that V is always increasing in w), the fixed point is increasing. In other words, we would like to be able to state some properties of V^* such that $T(V^*) = V^*$. By knowing that V is increasing in w and that $T(V)$ maps increasing functions into increasing functions, we know that V^* will be increasing in w.
Similarly we know that for any V that is weakly increasing, $T(V)$ is strictly increasing, so the fixed point must be strictly increasing too.

Taking eq. (110) and subtracting βV on both sides yields:

$$(1 - \beta)V(w) = w + \beta(1 - \delta)[\alpha_1 \mathbb{E}_w \max\{V(w') - V(w), 0\}] + \beta\delta(U - V(w)) \tag{111}$$

or in the formulation with periods of length Δ and with $\beta = \frac{1}{1+r\Delta}$

$$\frac{r\Delta}{1+r\Delta}V(w) = w\Delta + \beta(1-\delta\Delta)[\alpha_1\Delta\mathbb{E}_w \max\{V(w') - V(w), 0\}] + \beta\delta\Delta(U - V(w)) + o(\Delta) \tag{112}$$

Note that you can interpret the last equation as the value of an offer being the offered wage plus the discounted value of on the job search if the match is not disrupted next period plus the loss occurring in case of disruption plus a term that stands in for the value of receiving more than one offer ($o(\Delta)$). This last term is basically the value of receiving two offers multiplied by $a(2, \Delta)$ plus the value of receiving 3 multiplied by its probability and so on. We can write it as $o(\Delta)$ because as Δ vanishes the probability of receiving more than one offer vanishes too and the value of receiving more than one offer is bounded as long as the wage distribution is bounded.

$$\begin{aligned} rV(w) &= (1+r\Delta)w + (1-\delta\Delta)\alpha_1\mathbb{E}_w \max\{V(w') - V(w), 0\} + \delta(U - V(w)) + \frac{o(\Delta)}{\Delta} \\ rV(w) &= w + \alpha_1\mathbb{E}_w \max\{V(w') - V(w), 0\} + \delta(U - V(w)) \end{aligned} \tag{113}$$

For U a similar argument holds given $U = V(R)$, however, the last term will vanish, since one cannot be laid off if already unemployed, and the probability of receiving an offer is α, so

$$rU = b + \alpha\mathbb{E} \max\{V(w') - U, 0\} \tag{114}$$

Assuming, for simplicity, that the wage distribution is bounded above from 1, we have

$$rU = b + \alpha \int_R^1 (V(w') - U)dF(w') \tag{115}$$

and

$$rV(w) = w + \alpha_1 \int_w^1 (V(w') - V(w))dF(w') + \delta(U - V(w)) \tag{116}$$

Differentiating the value of an offer with respect to w [18] yields:

$$V'(w) = \frac{1}{r + \delta + \alpha_1(1 - F(w))} > 0 \tag{117}$$

Which shows that the value of working is indeed strictly increasing.

[18] One has to use the Leibniz Rule to do so.

Evaluating eq. (116) at $w = R$, and recalling that $V(R) = U$ thereby equating it to eq. (115)

$$R + \alpha_1 \int_R^1 (V(w') - U) dF(w') = b + \alpha \int_R^1 (V(w') - U) dF(w') \tag{118}$$

we have

$$R = b + (\alpha - \alpha_1) \int_R^1 (V(w') - U)\, dF(w') \tag{119}$$

Note that if the arrival rate of offers is the same independently of the current employment state ($\alpha = \alpha_1$), then the job offers are only accepted if they exceed the value of unemployment ($w > R = b$).

This also implies that a worker may accept an offer for a wage lower than the unemployment benefit if on the job search is more successful than search when unemployed ($\alpha_1 > \alpha$), whereas they may reject offers whose instantaneous payoff is larger than the unemployment benefit if it is more likely to receive an offer when unemployed than when employed.

Integrating eq. (119) by parts we have

$$\begin{aligned}
R &= b + (\alpha - \alpha_1) \int_R^1 (V(w') - U) dF(w') \\
&= b + (\alpha - \alpha_1) \left[F(w')(V(w') - V(R))|_R^1 - \int_R^1 V'(w') F(w') dw' \right] \\
&= b + (\alpha - \alpha_1) \left[(F(1)V(1) - F(R)V(R) - F(1)V(R) + F(R)V(R)) - \int_R^1 V'(w') F(w') dw' \right] \\
&= b + (\alpha - \alpha_1) \left[(F(1)V(1) - F(1)V(R)) - \int_R^1 V'(w') F(w') dw' \right] \\
&= b + (\alpha - \alpha_1) \left[V(1) - V(R) - \int_R^1 V'(w') F(w') dw' \right] \\
&= b + (\alpha - \alpha_1) \left[\int_R^1 V'(w') dw' - \int_R^1 V'(w') F(w') dw' \right] \\
&= b + (\alpha - \alpha_1) \int_R^1 V'(w')(1 - F(w')) dw' \\
&= b + (\alpha - \alpha_1) \int_R^1 \frac{1 - F(w')}{r + \delta + \alpha_1 (1 - F(w'))} dw'
\end{aligned} \tag{120}$$

Which is the reservation wage solution to the problem. Note that when we suppose that on the job search is not allowed ($\alpha_1 = 0$), the problem collapses to the previous formulation of the reservation wage.

Note that, in this setting, employed workers would accept jobs at wages higher than some reservation wage. If moving between jobs is costless, then denoting the on-the-job reservation wage as $R(w)$, such wage is determined by $V(R(w)) = V(w)$. Namely, the wage that makes the worker

indifferent between staying at the current job at wage w or moving to a new one at $R(w)$. In this setting, it is immediate that $R(w) = w$.

Suppose otherwise that there is a fixed cost k associated with changing jobs; think of re-training yourself for new tasks. Then, the reservation wage would be $R(w) : V(R(w)) - k = V(w)$. Intuitively the worker would command a reservation wage $R(w) > w$ to be compensated for the fixed cost. We can also show this mathematically. Assume that there are no layoffs to simplify the math, then the problem of the worker would be

$$rV(w) = w + \alpha_1 \int_{R(w)}^{\infty} (V(w') - k - V(w))dF(w'). \tag{121}$$

We can rewrite this as

$$V(w) = \frac{w + \alpha_1 \int_{R(w)}^{\infty}(V(w') - k)dF(w')}{r + \alpha_1(1 - F(R(w)))} \tag{122}$$

The worker will choose such reservation wage optimally by setting $\frac{\partial V(w)}{\partial R(w)} = 0$:

$$\frac{\partial V(w)}{\partial R(w)} = \frac{r + \alpha_1 \left(1 - F(R(w))\right) \left(k - V(R(w))\right) f(R(w)) + \alpha_1 f(R(w))[w + \alpha_1 \int_{R(w)}^{\infty}(V(w') - k)dF(w')]}{[r + \alpha_1(1 - F(R(w)))]^2} = 0, \tag{123}$$

which implies

$$V(R(w)) = k + \frac{w + \alpha_1 \int_{R(w)}^{\infty}(V(w') - k)dF(w')}{r + \alpha_1(1 - F(R(w)))} \tag{124}$$

So note that $V(R^\star(w)) = k + V(w)$. We can go further and note that totally differentiating equation 124, we obtain

$$V'(R^\star(w))R^{\star\prime}(w) = V'(w), \tag{125}$$

hence

$$R^{\star\prime}(w) = \frac{V'(w)}{V'(R^\star(w))} \tag{126}$$

If $V(w)$ is convex, then since $R^\star(w) > w$, $R^\star(w) - w$ is decreasing in w. To show that V is convex, note that

$$V'(w) = (r + \alpha_1(1 - F(R(w))))^{-1} > 0, \tag{127}$$

and

$$V'' = \frac{\alpha_1 f(R(w)) \left(r + \alpha_1 (1 - F(R(w)))\right)}{\left(r + \alpha_1 (1 - F(R(w)))\right)^2} > 0 \tag{128}$$

The result that $R^\star(w) - w$ is decreasing in w says that a worker with a low wage requires a higher relative premium for moving since it is more likely that he will move again in the future. Instead of paying k to accept an offer that will likely be dominated soon by yet another offer, a worker is more likely to reject and wait. Alternatively, you can think of the fact that someone with a very high wage has a very low probability of getting a better offer and to have to pay again the cost of moving. Hence, a relatively smaller difference between the current wage and the new (higher) offer is sufficient to make moving optimal.

Before moving to the next problem, it is possible to discuss, in a setting relatively similar to the ones just described, the results in Burdett and Judd (1983) and Burdett and Mortensen (1998). Recalling the afore-discussed problem of the Diamond Paradox, where, without on-the-job search, all firms would post a take-it-or-leave-it offer for $w = R$, we can now show that there will be scope for meaningful wage setting, i.e. the worker will be able to extract some surplus, due to on-the-job search.

Intuitively a firm posting wages equal to R will attract unemployed agents. However, posting a wage $w > R$ will now also attract employed workers whose current wage is below w. By thinking about the game that firms play on the wages, it is possible to restrict the class of distributions we will have in equilibrium. In particular, the distribution of wages will have the following two features: i) there are no mass points in the wage distribution (if there were at any $w < y$, then a firm posting ϵ higher would steal all the workers, making a profitable deviation, there cannot be at $w = y$ because that would imply zero profits while $w = R$ would produce positive profits due to unemployed agents receiving only one offer with positive probability); ii) there are no holes in the wage distribution (if there were, a firm would profitably undercut to fill the hole without losing any worker).

Before we move to the firms' side of the problem, we can do a direct comparison of the reservation wages in the two models with and without on-the-job search. The two reservation wage rules are

$$R = b + \frac{\alpha}{r} \int_R^\infty (w - R) dF(w), \tag{129}$$

$$R^{OTJ} = b + (\alpha - \alpha_1) \int_R^1 \frac{1 - F(w')}{r + \delta + \alpha_1 (1 - F(w'))} dw'. \tag{130}$$

First, note that in the model without OTJ, the worker would ask for a premium as it was forgoing forever the opportunity of a higher wage. Looking at the firm side, we conclude that firms would only offer the reservation wage and that it would also coincide with b. As a result, the option value

of searching is driven to 0 in equilibrium.

We can get the same result, i.e, $R = b$, by imposing $\alpha_1 = 0$ so that we remove OTJ. In the on-the-job search model, workers will demand a premium over the reservation wage as long as $\alpha > \alpha_1$, this is because off-the-job offers arrive more frequently, so taking a job comes at the additional cost of a smaller likelihood of getting a better wage. The same intuition holds in reverse by noting that if $\alpha < \alpha_1$, the worker is willing to take jobs below b as it allows them to get offers more frequently. Now consider the special case in which $\alpha = \alpha_1$. In this case, there is no difference between the continuation value, whether on or off the job. Suppose a worker gets a job offer at wage b, would they turn it down? Intuitively, they would accept it as they can get a better one with the same probability they had from unemployment. In summary, while it looks like we got the same result in the two settings, i.e., the Diamond Paradox, we will see that it does not hold true in equilibrium, i.e., once firms' optimal wage setting is taken into account.

Notice that, in the baseline model, the reservation wage was, in principle, higher than the unemployment benefit due to the forgone value of searching more. However, firms' Nash equilibrium on the wage offering side of the problem implied that the value of search went to zero, and therefore workers were only offered their reservation wage, which was also equal to the unemployment benefit. When we introduce on the job search, the foregone value of searching depends on the relative rate at which job offers arrive to employed and unemployed. If those rates are the same, then the worker is happy to take a job at wage b. This implies that the reservation wage is equal to the unemployment benefit, but we are yet to figure out whether the wage offer distribution will be degenerate at the reservation wage (Diamond Paradox).

We can now study the firm side of the problem. Assume firms want to maximize their steady-state revenue. Define $e(w)$ as the mass of workers employed exactly at wage w. Then the revenues are

$$\frac{e(w)}{Nf(w)}[y - w] \tag{131}$$

Where N is the mass of active firms and $f(w)$ is the density of offers at w. Hence the meaning of the fraction is the number of workers at w per firm. From eq. (104) we know $E(w)$. Note that this formulation of the problem implicitly assumes that firms have no capacity constraints and have a CRS production function.

$$E(w) = \frac{\frac{\alpha\delta}{\alpha+\delta}F(w)}{\delta + \alpha_1(1 - F(w))} \tag{132}$$

Where the steady-state unemployment has been substituted in. It is possible now to compute $e(w)$,

by simply taking the derivative of eq. (132)

$$e(w) = \frac{f(w)(\delta + \alpha_1)}{[\delta + \alpha_1(1 - F(w))]^2} \frac{\alpha\delta}{\delta + \alpha} \tag{133}$$

We will use this in a second, but first note that firms must make equal profits in equilibrium[19], so we can just state that

$$\frac{e(w)}{Nf(w)}[y - w] = \kappa, \tag{134}$$

for some constant κ. By substituting $e(w)$

$$\frac{1}{N} \frac{\alpha\delta}{\delta + \alpha} (\delta + \alpha_1) \frac{y - w}{[\delta + \alpha_1(1 - F(w))]^2} = \kappa \tag{135}$$

Define a new constant $\hat{K}$, as a function of the old constant, such that the following is true

$$\sqrt{y - w} = \hat{K}[\delta + \alpha_1(1 - F(w))] \tag{136}$$

Then we can solve for $F(w)$, which is now the distribution of wage offers such that the profits are equal to the equilibrium constant.

$$F(w) = -\frac{\sqrt{y - w}}{\alpha_1 \hat{K}} + \frac{\delta}{\alpha_1} + 1 \tag{137}$$

Note that now we know that $F(R) = 0$, which is already a relevant result because it states that the Diamond Paradox no longer holds. Furthermore we can use it to solve for $\hat{K}$, since by replacing w with R and $F(R) = 0$ in (137) we have

$$0 = -\frac{\sqrt{y - R}}{\alpha_1 \hat{K}} + \frac{\delta}{\alpha_1} + 1 \tag{138}$$

From here, we could solve for $\hat{K}$ and then for the constant we started with. Note that it is possible to compute $\hat{K}$ as a function of R, which is an endogenous object, so it can either be solved jointly with the reservation wage equation or, by assuming $\alpha = \alpha_1$ and knowing that this implies $R = b$, it can be solved as a function of parameters only.

To conclude, this model, due to the presence of on-the-job search, does not exhibit the Diamond Paradox. However, it still delivers an unrealistic prediction: by taking the derivative of $F(w)$ to

[19] Since firms are identical, the equilibrium wage offer distribution must be the solution to a Mixed Strategy Nash Equilibrium.

compute the density, we have

$$f(w) = \frac{1}{2\alpha_1 \hat{K}} \frac{1}{\sqrt{y-w}} \tag{139}$$

This function is such that it becomes very steep as w approaches y. Intuitively this means that we should observe larger masses of firms offering very high wages than lower ones.

This can be solved by having some heterogeneity in productivity for firms such that higher productivity firms post higher wages. A more reasonable distribution could be produced when the revenues are supermodular in y and w so firms with high y have disproportionately larger benefits from hiring workers. Hence few firms will post high wages.

2.4 Diamond, Mortensen & Pissarides

In this section, we study the Diamond, Mortensen and Pissarides model of frictional labour markets. The model is built on three main blocks. First, frictions in the market are summarized by a matching function which will take unemployed workers and firms willing to hire as inputs and produce matches as output. The second building block is a bargaining solution. A firm and worker matching produce together more than they would alone. This additional value is the surplus of the job. A bargaining solution tells us how this surplus is split between the firm and the worker, effectively pinning down profits and wages. Lastly, the model is closed by a free entry condition which pins down how many vacancies are posted at a given moment. The objects of interest in this setting are given by the level of unemployment, the amount of vacancies posted and a wage that clears the market.

2.4.1 Matching Function

Denote the stock of unemployed agents by U and the stock of posted vacancies by V. A matching function is a mapping from U, V to m, the number of matches.

$$m = m(U, V) \tag{140}$$

The function $m(U, V)$ is assumed to be increasing in both arguments, concave and homogeneous of degree 1. For the remainder of this chapter, practical examples will be carried out assuming that the matching function is a CRS Cobb-Douglas of the form

$$m(U, V) = AU^{\alpha}V^{1-\alpha} \tag{141}$$

Regarding the assumption of constant returns to scale of the matching function, a number of empirical estimates have been carried out, reviewed in Petrongolo and Pissarides (2001). These

estimates are performed on a log-linear version, consistent with the Cobb-Douglas specification: out of 13 different estimates, 8 reject CRS at 10%.

Matches are carried out in a random fashion from the sets of available agents and vacancies U, V. Therefore, vacancy filling is a Poisson process with Poisson rate $m(U,V)/V$. A convenient synthetic object to discuss is $\theta \equiv \frac{V}{U}$. We refer to this market outcome as the "market tightness". This object intuitively represents the relative size of the two sides of the market and can be thought of as a probability to trade. Also, note that this object describes a trading externality. In this model, prices are not the only allocative mechanism because there is stochastic rationing, in particular, at any point in time, there is a positive probability that a posted vacancy is not filled and a positive probability that an unemployed worker does not find a job. This externality is also referred to as congestion or search externality. One can define the probability of filling a vacancy as follows:

$$q(\theta) \equiv \frac{m(U,V)}{V} = m\left(\frac{U}{V}, 1\right) = m\left(\frac{1}{\theta}, 1\right) \tag{142}$$

Workers face the following probability of receiving an offer:

$$\alpha(\theta) = \frac{m(U,V)}{U} = m\left(1, \frac{V}{U}\right) = m(1,\theta) \tag{143}$$

Where the second equality follows from homogeneity of degree one (constant returns to scale).

Assume for now that the wages are such that all job offers are accepted, i.e. assume that $w^* \geq b$, where b denotes the unemployment benefit.

By homogeneity of degree 1, write the matching function as

$$m(U,V) = Um_U(U,V) + Vm_V(U,V) \tag{144}$$

Hence

$$1 = \eta_u + \eta_v \tag{145}$$

Where η are elasticities. Note for example that in the Cobb-Douglas case with $m(U,V) = U^{\alpha}V^{1-\alpha}$, we have that $\eta_u = \alpha$ and $\eta_v = 1 - \alpha$.

From this, it is possible to compute the elasticity of the probability of finding a job

$$\eta_{\alpha}(\theta) = \frac{\theta\alpha'(\theta)}{\alpha(\theta)} = \frac{\theta m_V(1,\theta)}{m(1,\theta)} = \eta_v > 0 \tag{146}$$

and of filling a vacancy

$$\eta_q(\theta) = \frac{\theta q'(\theta)}{q(\theta)} = -\frac{\frac{1}{\theta}m_U(\frac{1}{\theta}, 1)}{m_U(\frac{1}{\theta}, 1)} = -\eta_U < 0 \tag{147}$$

We can now go back to describing the flow equations of the model. Normalize the population to 1 and start with the dynamics of unemployment

$$\begin{aligned} u_{t+1} &= (1-\alpha(\theta)\Delta)u_t + \delta\Delta(1-u_t) + o(\Delta) \Rightarrow \\ \frac{u_{t+1}-u_t}{\Delta} &= -\alpha(\theta)u_t + \delta(1-u_t) + \frac{o(\Delta)}{\Delta} \Rightarrow \\ \dot{u} &= -\alpha(\theta)u + \delta(1-u) \end{aligned} \tag{148}$$

Which implies a steady-state unemployment

$$u = \frac{\delta}{\delta + \alpha(\theta)} \tag{149}$$

Rewrite the steady-state unemployment as

$$\left(\delta + \alpha\left(\frac{V}{U}\right)\right)u - \delta = 0 \tag{150}$$

From this

$$\begin{aligned} \frac{\partial V}{\partial U} &= -\frac{\delta + \alpha(\theta) - \theta\alpha'(\theta)}{\alpha'(\theta)} \\ &= -\frac{\delta + \alpha(\theta)[1-\eta_V]}{\alpha'(\theta)} < 0 \end{aligned} \tag{151}$$

This tells us that, in steady state, if we want to draw the equilibrium curve in the U, V plane it needs to be downward sloping. This curve is called the Beveridge Curve and is the first step toward pinning down the equilibrium in the model. We have an equilibrium relationship between market conditions (the tightness measuring the relative scarcity of unemployed and open jobs) and the amount of unemployed agents.

2.4.2 Description of the Problem

From here, we build up the rest of the economy. In particular, in this model, there is a fixed mass (normalised to 1) of risk-neutral homogeneous agents. Firms can enter the market freely but are limited to one vacancy. Unemployed agents become employed and unfilled vacancies are filled at rates coming from $m(u, v)$. Firms face exogenous costs of posting a vacancy c, and successful matches yield y. Finally, jobs are destroyed at rate δ.

Start by describing the values of unemployment and employment

$$rU = b + \alpha(\theta)(V(w) - U) \tag{152}$$

$$rV = w - \delta(V - U) \tag{153}$$

On the firm side, the values of unfilled and filled vacancies are

$$rJ = -c + q(\theta)(\Pi - J) \tag{154}$$

$$r\Pi = y - w - \delta(\Pi - J) \tag{155}$$

From these values, it is possible to define the surplus that a successful match generates as the difference between the values of a filled vacancy for both workers and firms, and the values of unfilled ones

$$S = (V + \Pi) - (U + J). \tag{156}$$

Recall that what we are after is the equilibrium wage as a function of the market tightness. Given the surplus, we need to figure out how it gets split between workers and firms. To this end, we introduce a bargaining solution.

2.4.3 Bargaining

So far we have established that in this economy when a worker and a firm meet they can create additional value (surplus). We will always assume that such surplus is positive ($S > 0$) because otherwise there are no gains from trade. We now face the problem of figuring out how this surplus is split. In particular, if you think that a worker and a firm will be able to create a total of y when cooperating, while if they do not work together the worker gets b. Suppose for a moment that we are in a static model, then the surplus $y - b$ will need to be split between the two. This fundamentally boils down to asking what is the wage, since we can think of $y - w$ going to the firm and $w - b$ going to the worker. We could assume many possible ways of splitting the surplus, in this model it is assumed that the wage is determined by Nash bargaining.

Digression on Nash Bargaining We use Nash Bargaining because it has a number of desirable properties that will become clear in a couple of lines. In general, think of the problem of two agents, denoted by 1 and 2, who can create some positive surplus $S > 0$ upon cooperating. If they cooperate they can get payoffs v_1, v_2 which we will solve for. If they do not cooperate they get their outside option value d_1, d_2.

The Nash bargaining solution can be written as solving the following problem:

$$\max_{v_1,v_2}(v_1 - d_1)(v_2 - d_2) \quad st \tag{157}$$
$$\text{the solution is feasible,}$$
$$v_i > 0, \forall i.$$

This is equivalent to asking what are the values of payoffs that put us on the Pareto frontier. This is the reason why this is a cooperative bargaining solution, we find private payoffs that maximize the joint surplus (note that this fundamentally depends on having some version of transferable utility).

The Nash Bargaining solution satisfies 4 desirable properties: i) it is Pareto optimal; ii) it is symmetric; iii) it is independent of irrelevant alternatives; iv) it is invariant to affine transformations.

Note that the symmetry result stems from valuing both agents the same. The problem in equation 157 refers to equal weights and therefore picks a specific, symmetric point, on the Pareto frontier. In general, we could maximize $(v_1 - d_1)^\alpha (v_2 - d_2)^{1-\alpha}$ and pick any point on the frontier, depending on our choice of α.

End of Digression

In our search model, we use Nash Bargaining but we dispense with the symmetry assumption. In particular, assume that workers and firms have bargaining weights of ϕ and $1 - \phi$ respectively. Then the splitting of the surplus is

$$\max_{S^w,S^f}(S^w)^\phi (S^f)^{1-\phi} \qquad st \quad S = S^f + S^w \tag{158}$$

Note that the constraint is equivalent to stating that the solution is feasible. The FOCs for this problem imply

$$\begin{aligned} &\phi(S^f) = (1-\phi)S^w \\ &\phi S = S^W = V - U, \quad (1-\phi)S = S^f = \Pi - J \end{aligned} \tag{159}$$

This tells us, given some total surplus, how we split it between agents. We now introduce the last building block to close the model by pinning down how much surplus is generated in equilibrium.

2.4.4 Free Entry Condition

So far we have a relationship between tightness and unemployment and a relationship between surplus and wages. We now have to introduce a condition that will pin down how many vacancies are created, which in turn will tell us how much surplus we generate and allow us to solve the

model. The condition we need is a staple of firm problems and has to do with the idea that if there are profits to be made in an economy, some firm currently not producing will enter the market and arbitrage them away. Similarly, if the current active firms are making losses, some of them will leave the market until the marginal firm makes non-negative profits. In the context of the DMP model the value, we need that the value of posting a vacancy is zero. The reason being that if it was positive more vacancies would be posted until firms are indifferent between posting and not posting. Hence $J = 0$ and from eq. (155)

$$\Pi = \frac{y - w}{r + \delta} \tag{160}$$

Whereas eq. (154) becomes

$$c = q(\theta)\Pi = q(\theta)\frac{y - w}{r + \delta} \tag{161}$$

This is called the Job Creation Curve (JCC) and will be one of the two curves we use to pin down the equilibrium wage and market tightness. You can think of this condition as the firm's first order condition, in that it states that the cost of opening a vacancy c must be equal to the expected value of opening a vacancy. This expected value is the firm surplus $y - w$, discounted appropriately at rate $r + \delta$ and weighted by the probability of filling the vacancy. This condition links the wage and the tightness.

2.4.5 Equilibrium

We next find another condition, from the side of the workers, linking again wages and tightness so that we can use it together with the JCC to solve for them.

Using eqs (153) and (155), solving for the wage

$$r(V + \Pi) = y - \delta[V + \Pi - U - J] \tag{162}$$

Upon noticing that the term in the square brackets is S and that $V + \Pi = S + U$

$$(r + \delta)S + rU = y \tag{163}$$

Which implies

$$S = \frac{y - rU}{r + \delta} \tag{164}$$

You can think of this as the per-period value of a job y minus the per-period outside option of the worker rU.

We can then use this result to work out the value of unemployment. Start by recalling the value

of employment and use the result of the bargaining problem

$$\begin{aligned}
rV &= w - \delta(V - U) \\
r(U + \phi S) &= w - \delta(U + \phi S - U) \\
rU + (r + \delta)\phi S &= w \\
rU + (r + \delta)\phi \frac{y - rU}{r + \delta} &= w \\
rU + \phi(y - rU) &= w
\end{aligned} \tag{165}$$

Using together eqs (152) and (153)

$$\begin{aligned}
r(V - U) &= w - b - (\delta + \alpha(\theta))(V - U) \\
V - U &= \frac{w - b}{r + \delta + \alpha(\theta)}
\end{aligned} \tag{166}$$

And using this into eq. (152)

$$rU = b + \frac{\alpha(\theta)(w - b)}{r + \delta + \alpha(\theta)} \tag{167}$$

And finally, using this result and eq. (165)

$$\begin{aligned}
rU &= w - \phi(y - rU) \\
rU(1 - \phi) + \phi y &= w \\
(1 - \phi)b + (1 - \phi)\frac{\alpha(\theta)(w - b)}{r + \delta + \alpha(\theta)} + \phi y &= w \\
-\phi b + (1 - \phi)\frac{\alpha(\theta)(w - b)}{r + \delta + \alpha(\theta)} + \phi y &= w - b \\
\phi(y - b) + (1 - \phi)\frac{\alpha(\theta)(w - b)}{r + \delta + \alpha(\theta)} &= w - b \\
w = b + (y - b)\Gamma(\theta), \qquad \Gamma(\theta) &\equiv \phi\frac{r + \delta + \alpha(\theta)}{r + \delta + \phi\alpha(\theta)}
\end{aligned} \tag{168}$$

This is the Wage Curve, and it relates the equilibrium wage to the market tightness. We can do a couple more lines of math to get a more interpretable version of the Wage Curve. Start by noting that, by the properties of the matching function, $\alpha(\theta) = \theta q(\theta)$. Secondly, by the result of the free entry condition we can write

$$\alpha(\theta) = \frac{c\theta(r + \delta)}{y - w} \tag{169}$$

Now take the Wage Curve and subtract y from both sides

$$\begin{aligned} y - w &= (y - b)(1 - \Gamma(\theta)) \\ &= (y - b)(1 - \phi)\frac{r + \delta}{r + \delta + \phi\alpha(\theta)} \\ &= (y - b)(1 - \phi)\frac{y - w}{y - w + \phi c\theta} \\ &= (y - b)(1 - \phi) - \phi c\theta \end{aligned} \tag{170}$$

Which, solving for the wage, implies

$$w = \phi y + (1 - \phi)b + \phi c\theta. \tag{171}$$

It is now clear that the wage depends on the tightness positively. This is because as the tightness increases workers have a better outside option and can thereby bargain a higher wage. You can also see how the wage is partially composed of a weighted average of the total output y and the worker outside option flow value b. If we tilt the bargaining power towards the worker, the wage increases to get closer to output.

The wage curve, together with the Job Creation curve in 161, define the equilibrium in this market. In particular, it is possible to find it exactly by solving for $w^\star$ and $\theta^\star$.

With $\theta^\star$ we can go back to the Beveridge curve in 149 and find the steady state unemployment level. We have now solved for everything we wanted in the model: θ, w, u.

2.4.6 Analysis and Comparative Statics

The model is fully pinned down by the system of Wage equation, Job Creation Curve and Beveridge Curve:

$$w = \phi y + (1 - \phi)b + \phi c\theta, \tag{WC}$$

$$w = y - \frac{c(r + \delta)}{q(\theta)}, \tag{JC}$$

$$u = \frac{\delta}{\delta + \alpha(\theta)}. \tag{BC}$$

The equilibrium, defined by the triple $\{\theta^\star, w^\star, u^\star\}$ is shown graphically in Figure 7

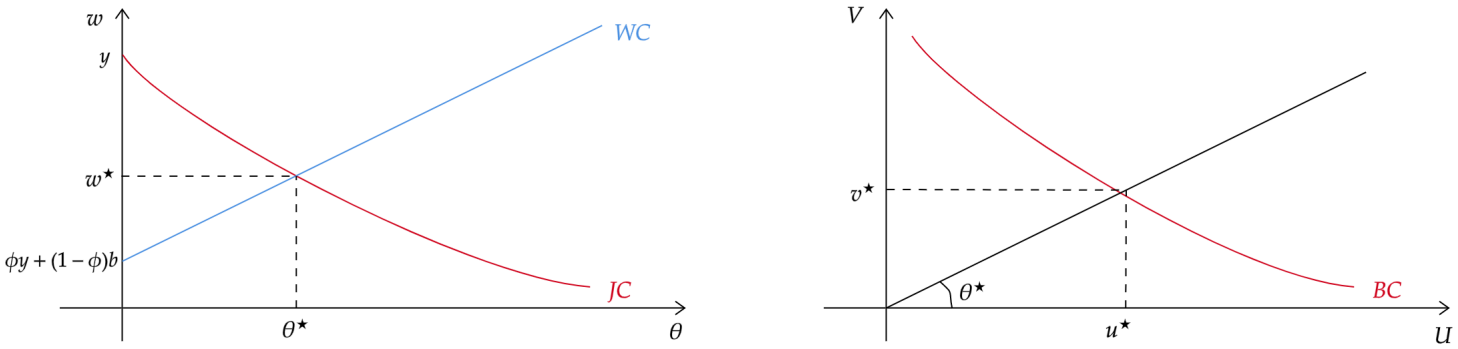


Figure 7: Equilibrium in the DMP model

With this, we can study some interesting comparative statics.

Start by considering an increase in productivity, measured by output per worker y. From both the wage curve and the job creation curve increases in y generate increases in w. If productivity increases more vacancies are created, and the surplus increases. This results in increases in θ and w. As the market tightness increases, unemployment u declines. Graphically both the WC and JC shift upwards. To show formally that the tightness increases, use the WC and JC to implicitly pin down the equilibrium market tightness:

$$F \coloneqq \phi y + (1-\phi)b + \phi c\theta^\star - y + \frac{c(r+\delta)}{q(\theta^\star)} = 0, \tag{172}$$

and use the Implicit Function Theorem to characterize $\frac{\partial \theta^\star}{\partial y}$:

$$\frac{\partial \theta^\star}{\partial y} = -\frac{\frac{\partial F}{\partial y}}{\frac{\partial F}{\partial \theta^\star}} = \frac{1-\phi}{\phi c - \frac{c(r+\delta)q'}{q(\theta^\star)^2}}. \tag{173}$$

Since $q' < 0$, every term is positive, so the tightness increases in productivity y.

Consider an increase in the unemployment benefits b. From the wage equation, this results in an increase in the wage as the workers' outside option went up. Firms have now less surplus and therefore post fewer vacancies, which pushes the market tightness down. It follows that unemployment rises. Graphically, the wage curve shifts upwards.

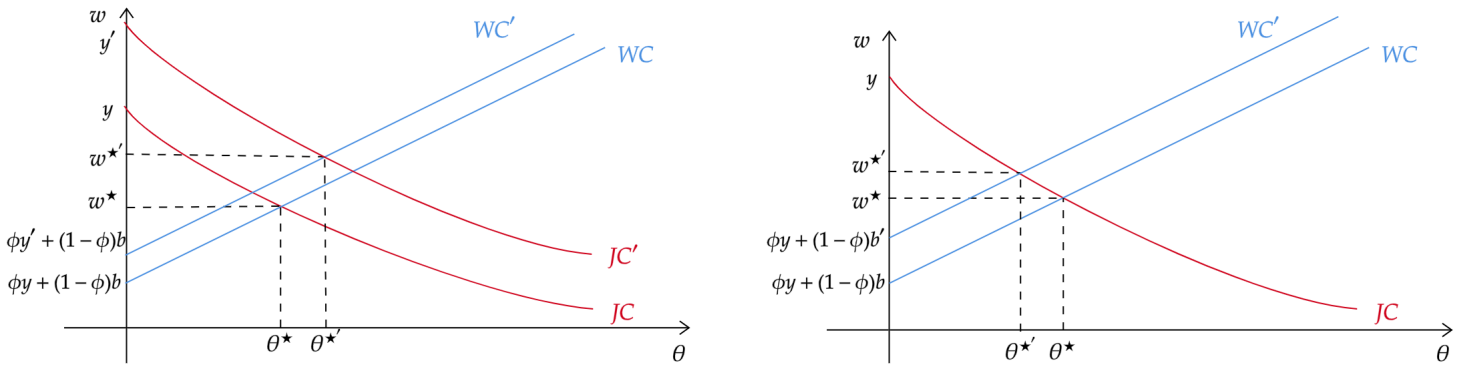


Figure 8: Comparative statics with respect to y and b

Increasing the bargaining weight of workers implies that, all else equal, posting vacancies is less profitable for firms. As a consequence, fewer vacancies are posted, the tightness increases, and workers can command a smaller premium in their wage from market conditions. However, workers obtain a larger share of the surplus, which results in a net increase in the equilibrium wage. Since the equilibrium tightness decreases, the job-finding rate declines and, therefore, steady-state unemployment increases. Graphically, the wage curve shifts up and becomes steeper.

Lastly, consider an increase in the job destruction rate δ. This will decrease the market tightness and therefore decrease the wage (to see this, you can use the first two equations to get rid of the wage and use the Implicit Function Theorem, like for y). Moreover, the Beveridge curve shifts outward as the flow into unemployment is now higher for given θ. As a result, unemployment increases. Graphically, the JC pivots around its intercept, while the WC is unchanged.

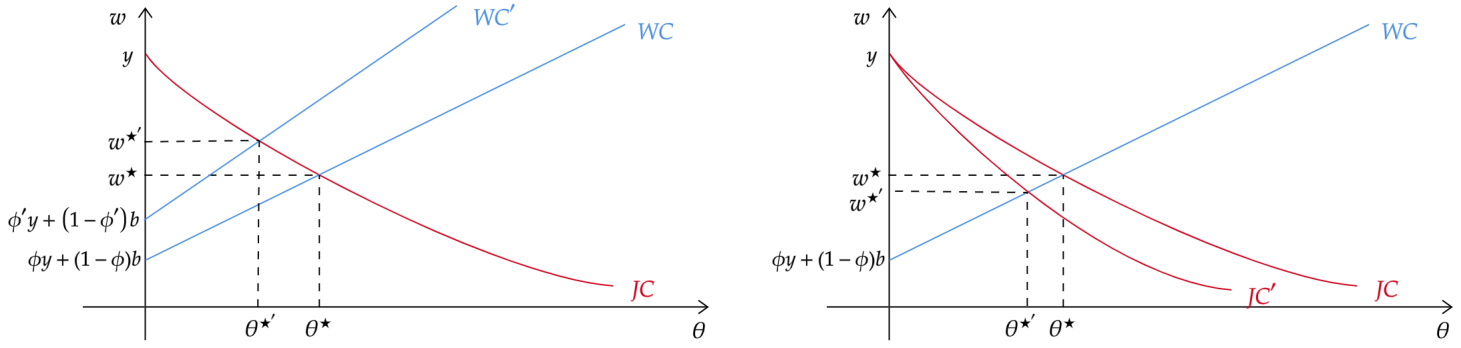


Figure 9: Comparative statics with respect to ϕ and δ

2.4.7 The DMP Model with Aggregate Shocks

In the DMP model described before firms' productivity was exogenous and fixed, now by assuming that productivity follows a stochastic process, where $y_{t+1} = f(y_t)$, we can have the model speak to the behaviour of labour markets along the business cycle.

Firms and workers now make optimal decisions in which the relevant state variables are y_t and u_t. However, since θ_t is a jump variable then u_t ceases to be a state variable.

In discrete time the model has the following values when y denotes today's productivity and y' next period's productivity:

$$U_y = b + \beta[\alpha(\theta_y)E_y V_{y'} + (1 - \alpha(\theta_y))E_y U_{y'}] \tag{174}$$

$$V_y = w_y + \beta[(1 - \delta)E_y V_{y'} + \delta E_y U_{y'}] \tag{175}$$

And for the firm

$$J_y = -c + \beta[q(\theta_y)E_y \Pi_{y'} + (1 - q(\theta_y))E_y J_{y'}] \tag{176}$$

$$\Pi_y = y - w_y + \beta[(1 - \delta)E_y \Pi_{y'} + \delta E_y J_{y'}] \tag{177}$$

Wage determination occurs again through Nash bargaining over

$$S_y = \Pi_y - J_y + V_y - U_y \tag{178}$$

Recalling that free entry implies $J_y = 0$, from the bargaining problem

$$\Pi_y = (1 - \phi)S_y \tag{179}$$

$$V_y - U_y = \phi S_y \tag{180}$$

From the value of an open vacancy and free entry

$$c = \beta q(\theta_y)E_y \Pi_{y'} = \beta q(\theta_y)(1 - \phi)E_y S_{y'} \tag{181}$$

From the value of a filled vacancy and free entry

$$(1 - \phi)S_y = y - w_y + \beta(1 - \delta)E_y \Pi_{y'} = y - w_y + \frac{1 - \delta}{q(\theta_y)}c \tag{182}$$

It is possible then to substitute into the definition of the surplus

$$S_y = y - b + \beta(1 - \delta)E_y(V_{y'} + \Pi_{y'}) + \beta E_y[\delta U_{y'} - \alpha(\theta_y)E_y V_{y'} - (1 - \alpha(\theta_y))U_{y'}] \tag{183}$$

Which can be rewritten as

$$S_y = y - b + \beta(1 - \delta)E_y[V_{y'} + \Pi_{y'} - U_{y'}] - \beta\alpha(\theta_y)E_y[V_{y'} - U_{y'}] \tag{184}$$

Note that the first square bracket is $S_{y'}$ and the second one is $\phi S_{y'}$. By free entry, we know that

$$E_y S_{y'} = \frac{c}{\beta q(\theta_y)(1-\phi)} \tag{185}$$

Hence

$$S_y = y - b + [1 - \delta - \alpha(\theta_y)\phi]\frac{c}{q(\theta_y)(1-\phi)} \tag{186}$$

Finally from eq. (182)

$$w_y = \phi y + (1-\phi)b + c\phi\theta_y \tag{187}$$

Which is the wage equation. The first two elements show that the bargaining weights move the wage closer to the output or to the unemployment benefit. Also, note that a low bargaining power for the worker keeps wages rigid since they move less with y and they move less with θ which is the variable that adjusts the fastest.

In this context, it is now possible to discuss the so-called Shimer Puzzle. The Shimer Puzzle consists of the observation that in this model, wages respond way too much to output fluctuations, thereby failing to mimic the observed volatility of vacancy posting through the business cycle. In particular, taking the Shimer (2005) parametrization (consistent with the Hosios efficiency condition, see digression on efficiency below), say that output y is normalized to 1 in normal times and assume that it goes to .98 in downturns and to 1.02 in upturns. Fix the unemployment benefit to .4, then wages turn out to be .96 in downturns and 1 in upturns. Since the profit margin for firms is approximately the same along the cycle the model fails to replicate the volatility of vacancy posting and the implied tightness of the market. The reason for this behaviour of the model is that the Hosios condition applied to the estimated matching functions requires a $\phi \approx .7$ which implies that workers are able to keep the part of the surplus that goes to firm relatively low, having wages follow output closely.

Some possible solutions to this problem have been proposed by Hagedorn and Manovskii (2008), Hall (2005) and Pissarides (2009). The former comes from the analysis of equation (187), in particular, note that in order to have wages respond less, one would need to reduce the workers' bargaining power. However, these models are always calibrated to match long-run unemployment, with a lower ϕ the only way to keep matching it is by having a higher c which, by eq. (187) would work against the goal of having more vacancies along the cycle. Therefore HM propose to increase the unemployment benefit to account for the utility of leisure, in order to get the correct cycle of vacancies and tightness, they have $b \approx .94$. Hall's solution instead consists of sticky wages, which intuitively will make wages respond less to output fluctuations, generating the observed volatility of vacancies. Finally, Pissarides (2009) argues that the failure of the model in matching the right

cyclicality of wages only applies insofar as we consider all wages. When we restrict the analysis to wages that have just been bargained, the model is able to replicate the appropriate empirical patterns.

Digression on Efficiency in the DMP Model[20] So far, we have not discussed the efficiency properties of the DMP model. Firstly, it is important to note that given any level of aggregate conditions (the market tightness), matches are pairwise efficient. This is given by the properties of Nash bargaining. The inefficiency question, therefore, needs to be about the level of market tightness and whether the θ arising from the decentralized equilibrium is the same that a social planner would choose. Embedded in this framework there are two types of externalities. First, firms posting vacancies lower the vacancy-filling rate for others, generating some version of a congestion externality, and increasing the job-finding rate of unemployed workers. Firms do not internalize these effects. These two externalities go in opposite directions in that one suggests there are too many vacancies and the other that there are too few. It is possible to show that there is a knife-edge case where the two exactly cancel out and the decentralized equilibrium is efficient. We derive this in a simple one-shot version of the economy.

Suppose the planner wants to maximize the total value of the economy by picking the vacancies per unemployed. We can write the matching function per unemployed as $m(1, v)$ and optimize

$$\max_v m(1,v)y + (1 - m(1,v))b - cv. \tag{188}$$

Where 1 is the total population and therefore $1 - m(1, v)$ is the number of unemployed workers. In this case, it is better to think of b as home production since we account for its value in the total output of the economy. The first order condition of the planner states

$$m_v(1,v)(y-b) = c. \tag{189}$$

In this one-shot model, the total surplus is simply $y - b$ and using the Nash bargaining weights, we get that

$$w = b + (y-b)\phi. \tag{190}$$

The free entry condition of firms in the decentralized equilibrium states that

$$c = \frac{m(1,v)}{v}(y-b)(1-\phi), \tag{191}$$

where the right-hand side represents the expected payoff of the firm. We now have the conditions that determine the number of vacancies under a social planner and in a decentralized equilibrium.

[20] For a deeper discussion on this see Albrecht (2011).

We can now equate them and solve for the bargaining weight to find the case in which the two exactly coincide. Formally,

$$\phi^\star = 1 - \frac{m_v(1,v)v}{m(1,v)} = 1 - \eta_v = \eta_u \tag{192}$$

This is known as the Hosios condition, from Hosios (1990). It states that if the worker bargaining weight is exactly equal to the elasticity of the matching function with respect to unemployment, then the decentralized equilibrium coincides with the planner solution. Alternatively, we can think that if the firm's bargaining weight is equal to the elasticity of the matching function with respect to vacancies, efficiency obtains. Intuitively this is the level of bargaining weight at which the firm exactly internalizes the externalities. If the bargaining weight is higher or lower, we will have that one of the two externalities will dominate the other and end up with an inefficient level of vacancy posting.

Digression on Directed Search[21] An alternative way of getting efficient outcomes in a search model is to introduce directed search. In this class of models, search is not random, rather agents queue optimally. This is also called a model of competitive search where the market makes firms internalize their vacancy posting effects.

Consider a static economy where firms looking for workers announce a wage that they commit to pay. Workers then "queue" by applying to a level of wage. Effectively, firms posting the same wage are in direct competition with one another, while they indirectly compete with firms offering different wage levels. In this sense, each wage is a somewhat separate market. Each firm, therefore, offers a contract composed by a pair $(w, \theta(w))$, where the tightness θ is a measure of queue length at wage w. Within each wage/market, there is a matching function taking as inputs the number of unemployed workers who applied for that wage and the number of vacancies at that wage posted by firms.

Denote $u(w)$, $v(w)$, $m(w)$, $\theta(w)$ the number of unemployed, vacancies, matches, and the tightness at wage w, respectively. We can define a competitive search equilibrium as a tuple of reservation utility, a set of offered wages, and market tightness $\{\mathcal{U}, \mathcal{W}, \theta\}$ such that:

1. free entry holds for each wage level:

$$q(\theta(w))[y - w] \leq c, \qquad \forall w, \text{ with equality if } w \in \mathcal{W}. \tag{193}$$

2. To sustain different wage levels (and associated tightness levels) in equilibrium, workers must

[21] This discussion is inspired by Pablo Kurlat's lecture notes on the topic.

be indifferent between queues:

$$\alpha(w)w + (1-\alpha(w))b = \alpha(w')w' + (1-\alpha(w'))b \qquad \forall w,\, w' \text{ st } \theta(w) < \infty. \tag{194}$$

3. Since all queues have to deliver the same utility, which we call $\mathcal{U}$, we must have

$$\mathcal{U} = \alpha(w)w + (1-\alpha(w))b \qquad \forall w \text{ st } \theta(w) < \infty. \tag{195}$$

4. Firms, taking as given this reservation utility, choose the profit-maximising wage and tightness to post

$$\max_{w,\theta} q(\theta(w))[y-w] \quad \text{st } \mathcal{U} = \alpha(w)w + (1-\alpha(w))b. \tag{196}$$

Note that the same equilibrium would hold if we let the workers optimize.

We want to argue that the competitive search equilibrium is efficient. Towards this, note that using the FOCs of the firm and free entry, we obtain

$$c = q(\theta)(y-w) \tag{197}$$

Using the definition of $q(\theta)$ and the properties of m recall that $q(\theta) = \frac{m(1,v)}{v}$ so we can rewrite

$$c = \frac{m(1,v)}{v}(y-w) \tag{198}$$

This is exactly the planner FOC in 189, which proves that this outcome is constrained efficient.

To understand where this result comes from, note that when firms post prices or wages we are effectively studying a special case of bargaining where the firm itself has all the bargaining power. However, in this setting, the firm is taking into account its choices on the tightness, while in the previous model, it was taking it as given. In this sense, the problem of each individual firm is much closer to that of the planner since they account for the effect of their wage choice on the tightness or, in this case, the number of applications they will receive.

For more on this topic, see the review article by Wright, Kircher, Julien and Guerrieri (2021).

End of Digression

Digression on Two-Sided Heterogeneity and Assortative Matching The model we studied so far has identical workers and identical firms matching one-to-one. One interesting extension that has had a lot of success in the literature is introducing heterogeneity. In particular, if we allow for workers and firms to be different we have to take a stance on how the quality of a match varies with the quality of workers and firms. Suppose that the quality of the matching is given

by $f(x, y)$ where x is the quality of the worker and y of the firm. It is sensible to assume that f_x, $f_y > 0$ so that the function is increasing in the firm and worker quality. But what about f_{xy}? This property turns out to be extremely important. $f_{xy} > 0$ is called *supermodularity*, while the opposite is called *submodularity*. If a function is supermodular then the two inputs are complements. Supermodularity of matching functions will generate, in equilibrium, Positive Assortative Matching (PAM). We call PAM cases in which high-quality firms match with high-quality workers. If the matching function is submodular we will have Negative Assortative Matching (NAM), namely high quality firms will match with low-quality workers and vice-versa. For a formal discussion on this see Shimer and Smith (2000).

The notion of assortative matching is a very important one in economics and can be used to explain patterns like the persistent difference in the level of development of countries or how marriages affect social mobility and inequality.

End of Digression

3 Firms in Macro

In all our discussions so far, firms were often somewhat uninteresting. They were competitive with CRS technologies, and at most, they got to decide whether to open a vacancy or not. From now on, we take the role of firms much more seriously. First, we spend some time thinking about returns and profits. We then move on to thinking about entry and exit in the workhorse model of firm dynamics. Finally, we conclude the section by studying how the extensive margin can affect macroeconomic outcomes.

3.1 Irrelevance Results under CRS

In Section 1, we worked out a number of models in which firms were competitive and had constant returns to scale. The CRS case is a very useful benchmark for more complex models. The usefulness, however, coincides with some key limitations. In the CRS model, we can derive a number of equivalence (or irrelevance) results. For example, recall that it does not matter who owns the capital, whether the firm or households, and that it does not matter who owns the firm, as profits are zero. Furthermore, in CRS economies, the size of the firms is purely demand-driven as the firm's optimal choice only pins down the relative input usage.

It is easy to show that if firms are homogeneous with CRS production functions, then the distribution of factors across firms is irrelevant. Note, first, that the first order conditions of capital and labor imply

$$\frac{\alpha}{1-\alpha}\frac{w}{r} = \frac{K^\star}{L^\star}, \tag{199}$$

with α being the output elasticity with respect to capital. Denote $\bar{k} = K^\star/L^\star$ at the optimum. If all firms share the same input market, and, therefore, face the same w and r, they will also choose the same capital-labor-ratio $\bar{k}$.

Suppose that aggregate output is just the sum over all individual firms' output. Under the continuum assumption, this is given by

$$Y_t = \int_i y_{it} di = \int_i K_{it}^\alpha L_{it}^{1-\alpha} di = \int_i \bar{k}^\alpha L_{it} di = \bar{k}^\alpha L_t. \tag{200}$$

Noting that $K_t = \int_i K_{it} di = \int_i \bar{k} L_{it} di = \bar{k} \int_i L_{it} di = \bar{k} L_t$, we can use this to derive

$$Y_t = \bar{k}^\alpha L_t = K_t^\alpha L_t^{-\alpha} L_t = K_t^\alpha L_t^{1-\alpha}, \tag{201}$$

which is the aggregate production function. We conclude that the distribution of resources across firms is irrelevant since they all share the same optimal capital-labor ratio.

A second irrelevance result concerns the distribution of claims to firms' profits. By the Euler

theorem, we know that the profits of atomistic firms operating in competitive markets with CRS production functions are zero. It is, therefore, irrelevant who owns the firms since we can distribute claims to zero profits any way we want without changing the allocation.

Lastly, we have already shown the Modigliani-Miller theorem, which says that ownership of capital is irrelevant if capital markets are efficient.

To break these irrelevance results, we study the case of decreasing returns to scale.

3.2 Decreasing Returns to Scale

In this section, we depart from this benchmark case and discuss how firms behave under decreasing returns to scale (DRS). To make matters simple, we assume DRS directly in the production function

$$y = zk^{\alpha}n^{\beta}, \quad \alpha + \beta < 1. \tag{202}$$

Then the firm maximizes profits $\pi = y - rk - wn$. The standard FOCs for Cobb-Douglas give the usual constant expenditure shares

$$\alpha y = rk \quad \text{and} \quad \beta y = wn. \tag{203}$$

We immediately note that

$$\pi = y - \alpha y - \beta y > 0. \tag{204}$$

So profits are positive in equilibrium and, therefore, it matters who owns the firm. Furthermore, solving the firm's FOCs, we get

$$y = z^{\frac{1}{1-\alpha-\beta}} \left(\frac{\beta}{w}\right)^{\frac{\beta}{1-\alpha-\beta}} \left(\frac{\alpha}{r}\right)^{\frac{\alpha}{1-\alpha-\beta}}. \tag{205}$$

As $1-\alpha-\beta > 0$, this expression pins down the level of output uniquely. You can also immediately see how the CRS case would generate indeterminacy of the size of the firm.

An alternative way to see that we will have positive equilibrium profits is to study average and marginal costs. First, write the problem of a firm choosing how much to produce

$$\max_{y} py - C(y), \tag{206}$$

where $C(Y)$ is the cost function. Trivially, it will choose $y : p = C'$ since this is equivalent to $\pi' = 0$. Recall that profits can be written as $\pi = py - C(y) = y(p - C(y)/y) = y^{\star}(C' - C(y^{\star})/y^{\star})$, which says that the profit rate is given by the difference between the marginal and average cost at the optimal level of output. It is easy to show that if returns to scale are decreasing, then it has

to be that $C'(y^\star) > C(y^\star)/y^\star$, and therefore, the profit rate is positive. To bring the point home, we can derive it easily in the Cobb-Douglas case. Start by deriving the optimal amount of capital and labor used by the firm. We know that $k^\star = n^\star \frac{w\alpha}{r\beta}$. Using this in the production function

$$y = z\left(\frac{w}{r}\frac{\alpha}{\beta}\right)^{\alpha} n^{\star\alpha+\beta}, \tag{207}$$

so that

$$n^\star = y^{\star\frac{1}{\alpha+\beta}} z^{-\frac{1}{\alpha+\beta}} \left(\frac{w}{r}\frac{\alpha}{\beta}\right)^{-\frac{\alpha}{\alpha+\beta}}. \tag{208}$$

Similarly for capital

$$k^\star = y^{\star\frac{1}{\alpha+\beta}} z^{-\frac{1}{\alpha+\beta}} \left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{-\frac{\beta}{\alpha+\beta}}. \tag{209}$$

Plugging these into the cost function

$$C(y^\star) = rk^\star + wn^\star = y^{\star\frac{1}{\alpha+\beta}} z^{-\frac{1}{\alpha+\beta}} \left[w\left(\frac{w}{r}\frac{\alpha}{\beta}\right)^{-\frac{\alpha}{\alpha+\beta}} + r\left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{-\frac{\beta}{\alpha+\beta}}\right]. \tag{210}$$

It is immediate that the average cost of this firm is

$$\frac{C(y^\star)}{y^\star} = y^{\star\frac{1}{\alpha+\beta}-1} z^{-\frac{1}{\alpha+\beta}} \left[w\left(\frac{w}{r}\frac{\alpha}{\beta}\right)^{-\frac{\alpha}{\alpha+\beta}} + r\left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{-\frac{\beta}{\alpha+\beta}}\right], \tag{211}$$

while the marginal cost is

$$C'(y^\star) = \frac{1}{\alpha+\beta} y^{\star\frac{1}{\alpha+\beta}-1} z^{-\frac{1}{\alpha+\beta}} \left[w\left(\frac{w}{r}\frac{\alpha}{\beta}\right)^{-\frac{\alpha}{\alpha+\beta}} + r\left(\frac{\beta}{\alpha}\frac{r}{w}\right)^{-\frac{\beta}{\alpha+\beta}}\right], \tag{212}$$

and, therefore $\frac{C(y^\star)}{y^\star} < C'(y^\star) \Leftrightarrow \alpha + \beta < 1$. This result is an important foundation for what we will study next. When firm size depends on firms' characteristics and not just on demand, then firm heterogeneity shapes aggregate outcomes. In the next section, we build exactly such a model.

3.3 Firms Dynamics, Heterogeneity and the Extensive Margin - Hopenhayn (1992)

We lay out the model developed in Hopenhayn (1992). This is the workhorse setting for firm dynamics, and even the seminal trade model in Melitz (2003) is based on it, The great feature of this model is that it is a relatively simple setting to think about entry and exit of firms.

The economy is populated by a continuum of competitive firms producing with decreasing returns. Their productivity is idiosyncratic and follows a Markov process. We abstract from aggregate risk.

Firms use labour as input and take prices as given. Firms profits are

$$\pi(z) = \max_n \{pzF(n) - wn - c_f\}, \tag{213}$$

where c_f is a per period cost for producing. Note that without this cost we would not have exit. A firm is indexed by its productivity z. We now separate between firms that are in the market (incumbents) and firms that are out (entrants). Incumbents' productivity evolves according to some Markov process with a transition density $f(z'|z)$. Entrants draw an initial productivity from $g(z)$ upon paying a sunk cost c_e.

The timing is that inside a period incumbents decide whether to stay or leave and entrants whether to enter or not. Firms pay their sunk costs c_f, c_e. After paying the cost firms learn their new productivities and produce. We pick $w = 1$ as the numeraire. Denote the distribution of firms in the market by $\mu(z)$. The supply curve is then given by

$$Y(p) = \int y(z,p)\mu(z)dz. \tag{214}$$

And, given some exogenous demand function $D(p)$ the market clears at $Y(p) = D(p)$. In this setting we can study the value of the firm $v(z,p)$, given by the Bellman equation

$$v(z,p) = \pi(z,p) + \max\left\{0, \beta \int v'(z',p)f(z'|z)dz'\right\}. \tag{215}$$

We can conjecture the existence of an exit threshold $z^\star(p)$ such that a firm with $z < z^\star(p)$ exits. This threshold is the solution $z^\star$ to

$$0 = \int v'(z',p)f(z'|z^*)dz'. \tag{216}$$

As $\pi(z)$ is strictly increasing so is $v(z,p)$ for $p \geq 0$. It follows trivially that such threshold exists for interior cases (it can also be that nobody ever exits).

Entrants are ex ante identical, then they pay their sunk cost c_e and draw their current productivity. The free entry condition is given by:

$$\int v'(z,p)g(z)dz \leq c_e. \tag{217}$$

If the mass of entrants $m > 0$, then the condition holds with equality. We can now characterise

the law of motion of the distribution of firms productivities:

$$\mu(z') = \int \phi(z'|z)\mu(z)dz + mg(z'), \tag{218}$$

where $\phi(z'|z) = f(z'|z)\mathbb{1}[z \geq z^\star]$ is the productivity distribution conditional on not exiting the market.

We can now study the properties of this economy by looking at the comparative statics on sunk costs. First, note that as we increase the entry cost c_e, trivially, the mass of entrants goes down. Less entry implies more profits for the incumbents, which in turn imply lower exit rates. Immediately, a higher entry cost generates lower turnover, importantly, also on the exit margin. We can see that the higher c_e generates less selection as more unproductive incumbents are not replaced by more productive entrants. The price also increases.

Finally, note that the effect on the firm size distribution is ambiguous. The increased price implies incumbents expand output and employment. On the other hand, there is a negative selection effect with more unproductive firms staying in. The net effect is not clear.

3.3.1 A "quasi-static" version of the Hopenhayn Model[22]

Take the following simplified version of the Hopenhayn model. Firms are not subject to productivity draws after they have entered. This implies that we can just think about the entrant distribution $g(z)$ and not be worried about the $f(z'|z)$ component. To simplify matters even further, assume that the production function is given by $F(n) = n^\alpha$ with $\alpha \in (0, 1)$. The firm profits can be written as

$$\pi(z, p) = pzn^\alpha - wn - c_f. \tag{219}$$

We can use the wage as the numeraire (which means we will have to solve for the output price as the key aggregate object) and get the first order condition

$$p\alpha z n^{\alpha-1} = 1, \tag{220}$$

which implies that the optimal amount of labour hired is

$$n^\star = (z\alpha p)^{\frac{1}{1-\alpha}} \tag{221}$$

and the optimal size of the firm

$$y^\star = z(z\alpha p)^{\frac{\alpha}{1-\alpha}}. \tag{222}$$

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

With this solution, we can compute the firm profits at the optimum so that we can then impose the zero profit condition and pin down the exit cutoff. First, profits are given by

$$\begin{aligned}\pi(z,p) &= pz(z\alpha p)^{\frac{\alpha}{1-\alpha}} - (z\alpha p)^{\frac{1}{1-\alpha}} - c_f \\ &= (1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}(pz)^{\frac{1}{1-\alpha}} - c_f. \end{aligned} \tag{223}$$

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. We can now 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.

The value for a firm with productivity z is given by

$$v(z,p) = \max\left\{0, \sum_{t=0}^{\infty} \beta^t \pi(z,p)\right\} = \max\left\{0, \frac{\pi(z,p)}{1-\beta}\right\}. \tag{224}$$

Where we can replace the sum of future profits with its annuity value since there is no residual uncertainty.[23] It follows that a potential entrant will face a value which is the expectation over all possible z realizations of $v(z,p)$, hence

$$v^E(p) = \int v(z,p) g(z) dz. \tag{225}$$

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 obtain

$$(1-\beta)c_e = \int_{z^\star(p)}^{\infty} ((1-\alpha)\alpha^{\frac{\alpha}{1-\alpha}}(pz)^{\frac{1}{1-\alpha}} - c_f) g(z) dz. \tag{226}$$

The zero profit condition $\pi(z^\star, p^\star) = 0$ implies

$$p^\star = c_f^{1-\alpha}(1-\alpha)^{\alpha-1}\alpha^{-\alpha} z^{\star -1}. \tag{227}$$

Plugging this solution into (226), we get

$$(1-\beta)\frac{c_e}{c_f} = \int_{z^\star(p)}^{\infty} \left[\left(\frac{z}{z^\star}\right)^{\frac{1}{1-\alpha}} - 1\right] g(z) dz. \tag{228}$$

Equation (228) provides the value of the exit cutoff. Plugging this into (227), we obtain the equilibrium price, which completes the characterization of the equilibrium. To study the comparative statics of the solution, note that the RHS of eq. (228) is decreasing in $z^\star(p)$. As a consequence,

[23] Note that in this setting, we can equivalently write the exit decision at the beginning or at the end of the period since there is no change in productivity across periods.

we immediately note that increasing the cost of entry will generate less entry, and therefore, less productive incumbents will be able to stick around ($z^\star \downarrow$). Through the price equation in eq. (227), we see that the price will increase as the average productivity in the economy declines ($p^\star \uparrow$). By a similar logic, an increase in the per-period production cost will increase the exit cutoff; this, however, has ambiguous effects on the equilibrium price. We can differentiate the price equation with respect to the per-period cost c_f

$$\frac{\partial p^\star}{\partial c_f} = (1-\alpha)c_f^{1-\alpha-1}(1-\alpha)^{\alpha-1}\alpha^{-\alpha}z^{\star-1} - c_f^{1-\alpha}(1-\alpha)^{\alpha-1}\alpha^{-\alpha}\frac{\partial z^\star}{\partial c_f}z^{\star-2} \tag{229}$$

$$= p^\star\left(\frac{1-\alpha}{c_f} - \frac{\partial z^\star}{\partial c_f}z^{\star-1}\right). \tag{230}$$

We can rewrite this in terms of elasticities to obtain

$$\frac{\partial p^\star}{\partial c_f}\frac{c_f}{p^\star} = 1-\alpha-\frac{\partial z^\star}{\partial c_f}\frac{c_f}{z^\star} \lessgtr 0. \tag{231}$$

The sign is ambiguous, but we can say more about what drives the effect. The first term comes from the direct effect of an increase in c_f on the price: firms need to charge higher prices to break even the higher per-period fixed cost. The partial elasticity of the price to c_f is given by the extent of decreasing returns $1-\alpha$. The second term, which is negative, represents the selection effect of a higher fixed cost. Only higher productivity firms will be able to break even. As a consequence, the average productivity in the economy increases and, therefore, the price declines. Consider an economy in which the cutoff $z^\star$ is insensitive to the change in the fixed cost. Then, the selection force is small, and the direct effect dominates to generate a price increase. Suppose instead that the cutoff is very elastic, then a small increase in the fixed cost induces strong selection in the economy and this can induce a price drop as firms are much more productive.

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 the per-period production cost, we would have that firms stay in forever as entry costs are sunk. 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.

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 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 10, from Traina (2018) shows exactly this point.

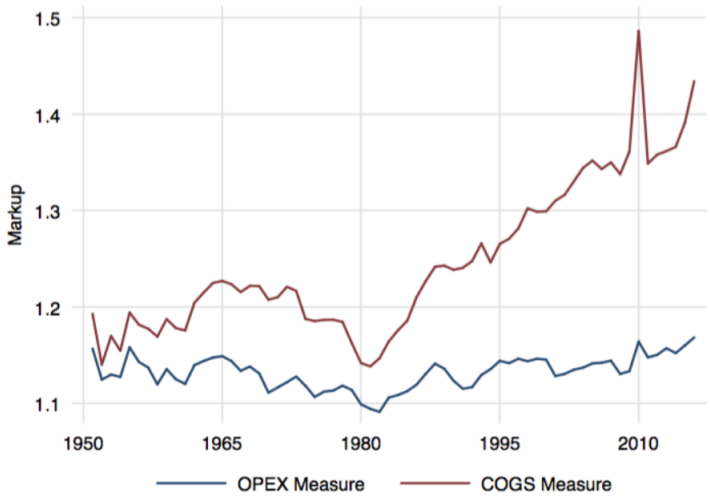


Figure 10: Markups Trends from Traina (2018).

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 high theory level, we do not need to make such distinction, particularly if it turns out to be problematic when we try to look at the real world.

End of Digression

Digression on Heterogeneity and Mean-Preserving Spreads An important question that we ask in macro these days is how important heterogeneity is. We have made a lot of progress in recent years in modelling more complex and realistic settings in which firms or agents are heterogeneous. Suppose we take our heterogeneous agent model. Can we say something about how our economy would look like with more heterogeneity? Answering this question at a general level is not trivial, but there are a couple of tricks we can use. First, we can use Mean-Preserving Spreads (MPS). MPS are special cases of "spreads" or of second order stochastic dominance. In particular, take two distributions μ and $\tilde{\mu}$ on the same support $[a, b]$, fix the means to $\bar{\mu}$ for both distributions. We say $\tilde{\mu}$ is a MPS of μ if

$$\int_a^y [\tilde{\mu}(x) - \mu(x)]dx \geq 0, \quad \forall y \leq b. \tag{232}$$

An alternative and much more interesting for us the definition of MPS: $\tilde{\mu}$ is an MPS of μ if

$$\mathbb{E}[f(\tilde{\mu})] \geq \mathbb{E}[f(\mu)], \tag{233}$$

for any convex function f. This is a very powerful and useful way to think about it. Suppose we take our heterogeneous firms model. We know we can write, say, output as

$$\int f(\theta)\mu(\theta)d\theta, \tag{234}$$

where f is the firms' policy function from their productivities to output decisions. If we want to know what happens to output when we make our firms more heterogeneous, i.e. we apply an MPS to the firm productivity distribution, all we have to figure out is whether the firms' policy function is concave or convex. This is a pretty powerful tool and one that holds in more general and complicated problems, see Jensen (2018).

End of Digression

3.4 Market Power - Monopolistic Competition

So far, we have always worked with firms operating in a competitive environment. 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 do so gradually: first, we look at so-called "monopolistic competition". In this model, there is a continuum of firms, but they are all monopolists on their own variety. This is very convenient because, again, there is no strategic interaction, but firms now face a downward-sloping demand curve. Note that the continuum assumption implies that no firm is "large" in the granular sense.

Before diving into complicated models of imperfect competition is worth asking why this is a promising and important avenue for research. We start by this, by now, very famous graph from De Loecker, Eeckhout and Unger (2020). The authors tell us that over the last 40 years, markups have been increasing substantially. Understanding why this happened and what this means for the economy is a central problem in modern macro.

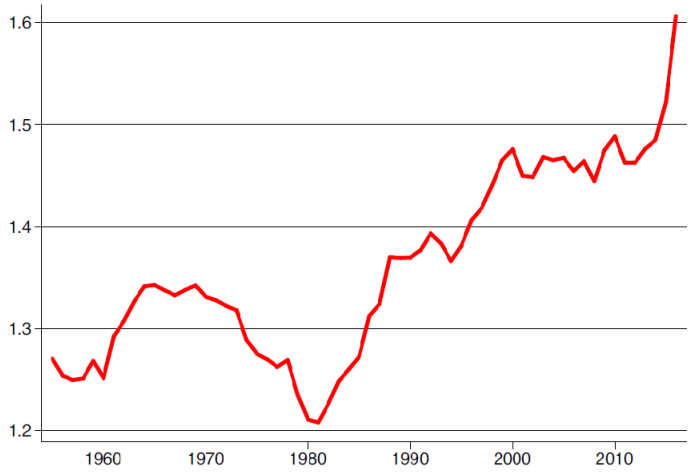


Figure 11: Markups Trends from De Loecker, Eeckhout and Unger (2020).

Further, we have some evidence that markups are quite costly for the economy. Intuitively market power distorts many margins of an economy (as we will discuss at length later), from compressed output to lower new varieties and misallocation of factors. In a recent paper Edmond, Midrigan and Xu (2018) put a number to this cost:

	efficient	uniform output subsidy	remove misallocation	entry subsidy
log deviation from benchmark, $\times 100$				
output, Y	35.9	33.3	1.0	3.8
consumption, C	28.8	28.7	1.2	5.3
employment, L	16.5	15.6	−0.3	2.9
mass of firms, N	13.1	6.3	−2.9	17.2
capital, K	49.8	47.3	1.0	3.9
aggregate efficiency, E	2.9	1.0	0.3	2.8
welfare gains, CEV, %	6.6	4.9	1.3	0.5

Figure 12: Cost of Markups from Edmond, Midrigan and Xu (2018)

Digression on Markups on Marginal vs Average Cost In a previous digression, we discussed the subtleties of thinking about what is a marginal vs a fixed cost in a world of indivisibilities. We concluded that this is a consequential problem when thinking about what is a markup and how it has evolved over time. Since marginal and fixed costs are nowhere to be found in a balance sheet, a natural question is, why don't we define markups over average cost, rather than marginal cost. If we were to look at average cost, we could basically (for single product firms) just take the bottom line profits, divide them by revenues and we would immediately back out markups since

$$\frac{pq - \bar{c}q}{pq} = 1 - \frac{1}{\tilde{\mu}} \tag{235}$$

If fixed costs are negligible, this approach works perfectly. The question is what this measure of markups tells us about the degree of competitiveness in the economy in settings where fixed costs are not negligible. Take any economy in which the number of active firms is pinned down by some free entry condition. This, by definition, implies that $\pi - f = pq - cq - f = 0$, in terms of per unit of output, we get $p - c - f/q = 0$. If we define the average cost as $c + f/q$, as we would in the approach described above, we get that the markup $\tilde{\mu} \equiv \frac{p}{c+f/q} = 1$ independently of the competitive structure of the economy. It follows that this markup is not at all informative about whether we live in a monopoly or in perfect competition. For a deeper discussion on markups, how to compute them, and their evolution, see Hall (2018), Basu (2019) and Syverson (2019).

End of Digression

3.4.1 Monopolistic Competition

We start from a pretty general formalization of market power and then pick the special case of monopolistic competition. To start with note that we define a firm i as having market power if for some good's price p, $\partial p/\partial y_i \neq 0$.

When this is the case, we will typically find the optimal quantity by equating marginal revenues and marginal cost c and get

$$p + p' y_i = c. \tag{236}$$

Defining the price elasticity of demand as $\epsilon_D = -y'p/y$ and the markup as $\mu = p/c$ we get

$$\mu = \frac{\epsilon_D}{\epsilon_D - 1} > 1. \tag{237}$$

Recall that elasticities are point-specific concepts, so in principle, they depend on where we are in the residual demand. Indeed we typically find that larger firms tend to charge higher markups (something known as "Marshall Second Law of Demand").

Further note that by the Euler Theorem for homogeneous of degree δ functions, we can write

$$\pi_i = p(y_i)y_i - c\delta y_i = p(y_i)y_i(1 - \delta/\mu).^{24} \tag{238}$$

Which also implies that the average profits are given by $(1 - \delta/\mu)$. Note a couple of things here: a firm with decreasing returns makes positive profits even at a markup of 1; a firm with increasing returns makes positive profits only if $\mu > \delta$, which imposes a condition between the technology returns to scale and the elasticity of demand.

These results are true independently of what happens on the demand side. To keep things simple, we adopt the most common set of preferences in models of market power known as the Dixit-Siglitz preferences. The convenience of this utility, also known as Constant Elasticity of Substitution (CES), is that ϵ_D is a constant and is identical for all firms. This is the most common set of preferences in trade or New Keynesian models. The utility is given by

$$\sum_t \beta^t u(C_t) = \sum_t \beta^t \left[\int C_{it}^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}. \tag{239}$$

Here varieties (products) are differentiated. These preferences have the property that $\epsilon_D = \sigma$. Furthermore, they exhibit what is called "love for variety," which states that as the single good marginal utility is decreasing, consumers find it optimal to divide their consumption into as many varieties as possible (for a fixed expenditure level).

Note that assuming CES is not enough to get the constant markup result, the next section actually provides a counterexample. The key is the combination of CES and monopolistic competition. Without the latter, we would have that the pricing behaviour of a single firm can affect the price index, and the markup would not be constant anymore.

We can start by solving the consumer problem to derive a demand schedule and then move on to the firm problem. We start from a household with a CES aggregator, which combines many different consumption goods

$$C = \left(\int_i c_i^{\frac{\sigma-1}{\sigma}} di \right)^{\frac{\sigma}{\sigma-1}}, \tag{240}$$

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

$$\int_i p_i c_i di = I, \tag{241}$$

[24] To see this, take, for example, a production function $y = k^\delta$. If the rental rate is r, the FOC on the profit maximization will be $r = (p(y) + p'(y)y)\delta k^{\delta-1}$. The profits will be given by $\pi = p(y)y - k(p(y) + p'(y)y)\delta k^{\delta-1} = p(y)y - (p(y) + p'(y)y)\delta k^\delta = p(y)y - (p(y) + p'(y)y)\delta y = p(y)y(1 - \delta - \delta/\epsilon_D) = p(y)y(1 - \delta/\mu)$.

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

$$\frac{\sigma-1}{\sigma}c_i^{\frac{\sigma-1}{\sigma}-1}\frac{\sigma}{\sigma-1}\left(\int_i c_i^{\frac{\sigma-1}{\sigma}}di\right)^{\frac{\sigma}{\sigma-1}-1}-\lambda p_i=0, \tag{242}$$

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

$$c_i=\left(\frac{p_i}{p_j}\right)^{-\sigma}c_j. \tag{243}$$

We can now define the ideal price index P, defined as the price of a unit of C when consumers are optimally choosing across varieties. This price index is designed so that $\int_i p_i c_i di = PC$ at the optimal choice of the consumer. We can obtain the price index by noting 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. (240). 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

$$P=\left(\int_i p_i^{1-\sigma}di\right)^{\frac{1}{1-\sigma}}.$$

We can now go back to the optimal splitting rule in eq. (243), multiply both sides by p_i and integrate over i to obtain

$$PC = P^{1-\sigma}p_j^{\sigma}c_j, \tag{244}$$

which, finally, implies

$$c_j = \left(\frac{p_j}{P}\right)^{-\sigma} C. \tag{245}$$

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

$$p = \frac{\sigma}{\sigma - 1} c. \tag{246}$$

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.

Digression on the Socially Optimal Number of Firms An important result proved by Dixit and Stiglitz (1977) and extended by Dhingra and Morrow (2019) is that the CES + monopolistic competition economy is constrained efficient in the presence of entry costs and a free entry condition. Before studying the result and discussing its intuition, we can look at a more general framework to think about efficiency. This is the problem studied by Mankiw and Whinston (1986).

Consider an economy populated by identical firms and, for now, assume that they produce a homogeneous good. When there are N firms in the economy, each individual firm produces q_N so that the total quantity is $Q = Nq_N$ and the price of the good is given by the demand schedule $P(Q) = P(Nq_N)$. Each firm has to pay an entry cost K. Production comes with a cost function $c(q_N)$. Profits are defined by $\pi_N = P(Nq_N)q_N - c(q_N) - K$. A free entry equilibrium is given by $\pi_{N^e} = 0$, where N^e is the number of firms in the free entry equilibrium.

We can define social welfare as

$$\mathcal{W}(N) = \int_0^{Nq_N} P(s)ds - Nc(q_N) - NK,$$

where the first term is the consumer surplus from consuming a quantity Nq_N of good, the second term is the cost of producing q_N for all N firms and the third term is the entry cost for each firm.[25]

We make the following three assumptions for the problem to be well-defined:

A.1 $Q(N) > Q(N'), \forall N > N'$.

A.2 $\frac{\partial q_N}{\partial N} \leq 0, \forall N$.

A.3 $P(Nq_N) - c'(q_N) \geq 0, \forall N$.

The first condition states that the aggregate quantity is increasing in the number of firms. The second one, which is the most important, implies that entry is *business-stealing*. Adding a firm to the economy reduces the quantity produced by each individual firm (the new firm is "stealing business" from existing firms). The last condition simply states that prices have to be weakly larger than marginal costs. The free entry equilibrium is defined by $\pi_{N^e} = 0$ while the constrained efficient allocation is defined by $\frac{\partial \mathcal{W}(N)}{\partial N} = 0$. We can start with the latter and consider a constrained planner that cannot choose the allocation but can choose the number of firms in the economy. Once the firms enter, they are free to choose their optimal quantity/price. Optimizing over the number of varieties N we obtain

$$\frac{\partial \mathcal{W}(N)}{\partial N} = P(Nq_N)\left[q_N + N\frac{\partial q_N}{\partial N}\right] - c(q_N) - Nc'\frac{\partial q_N}{\partial N} - K = 0. \tag{247}$$

The first term is the change in the consumer surplus induced by adding one firm. There is the direct effect $P(Nq_N)q_N$ from the extra firm and the change induced on the surplus generated by all other N firms: $P(Nq_N)N\frac{\partial q_N}{\partial N}$. The third and fourth terms represent the change, both direct and indirect, on the cost of producing the total quantity, and the last term is the additional fixed cost of K. Note that we can rewrite this change in social welfare, using the definition of profits, as

$$\frac{\partial \mathcal{W}(N)}{\partial N} = \underbrace{P(Nq_N)\left[q_N - c(q_N)\right] - K}_{\pi_N} + N\frac{\partial q_N}{\partial N}\left[P(Nq_N) - c'\right] \tag{248}$$

$$\pi_N + N\frac{\partial q_N}{\partial N}\left[P(Nq_N) - c'\right] = 0. \tag{249}$$

This condition for the socially optimal number of firms is very useful. First, note that the additional firm appropriates profits π_N. Second, note that the *business-stealing* effect of entry affects the other firms. The additional firm reduces the output of all other N firms.

[25] Note that it would not be correct for the second term to be $c(Nq_N)$ since we have not made homogeneity assumptions. We would need to impose that $c(q_N)$ is homogeneous of degree 1 for the step to go through.

Turning to whether the free entry equilibrium generates the efficient number of varieties, note that free entry is defined as $\pi_{N^e} = 0$. The condition in eq. (249) instead defines the socially optimal number of varieties, call it $N^\star$. It is immediate to note that $N^e = N^\star$ if and only if the second term in eq. (249) is 0. This is true if either i) there is no business-stealing effect of entry: $\frac{\partial q_N}{\partial N} = 0$ or ii) firms are pricing at marginal cost $P(Nq_N) = c'$. In perfectly competitive environments with a fixed cost, the first condition is true, which is why the competitive equilibrium coincides with the planner allocation. In general, however, whenever entry is business-stealing, and we are away from marginal cost pricing, the equilibrium will feature too many firms $N^e > N^\star$. You can conclude this by noting that when we evaluate eq. (249) at the free entry number of firms

$$\frac{\partial \mathcal{W}(N)}{\partial N} = \underbrace{\pi_{N^e}}_{0} + N^e \frac{\partial q_N^e}{\partial N} \left[P(N^e q_{N^e}) - c'\right] = N^e \frac{\partial q_{N^e}}{\partial N} \left[P(N^e q_{N^e}) - c'\right] < 0, \tag{250}$$

so the planner would like to take out some firms since the social marginal value of entry is negative. Intuitively, why are there too many firms? Because when firms enter, they do not internalize the *business-stealing externality* they have on other firms. This is a negative externality which the planner accounts for and therefore chooses to have fewer firms than in the free entry equilibrium.

Consider now the same economy but with differentiated products. Without specifying preferences, we can keep it more general and state that consumers value the consumption of each variety q_i through

$$U = G\left[\sum_i f(q_i)\right], \tag{251}$$

where both the subutility $f(\cdot)$ and the aggregator $G(\cdot)$ are increasing and concave. With these assumptions, the constrained planner solves

$$\max_N \mathcal{W}(N) = G[Nf(q_N)] - Nc(q_N) - NK. \tag{252}$$

The consumer maximization problem implies that their willingness to pay is $G'[Nf(q_N)]f'(q_N)$. Firms with market power will price the consumers at their willingness to pay and, therefore, obtain profits

$$\pi_N = G'[Nf(q_N)]f'(q_N)q_N - c(q_N) - K. \tag{253}$$

A free entry equilibrium is such that $\pi_{N^e} = 0$. The constrained social planner set

$$\frac{\partial \mathcal{W}(N)}{\partial N} = G'\left(Nf'\frac{\partial q_N}{\partial N} + f\right) - c(q_N) - Nc'(q_N)\frac{\partial q_N}{\partial N} - K = 0. \tag{254}$$

Rearranging and using the definition of profits

$$\frac{\partial \mathcal{W}(N)}{\partial N} = \pi_N + N(G'f' - c')\frac{\partial q_N}{\partial N} + G'(f - f'q_N) = 0. \tag{255}$$

Note that now, additional entry carries an extra effect. The first term represents the profits that the additional entrant can appropriate. The second term is the *business-stealing* effect (recall that $G'f'$ is the price) and is negative. The third term represents the fact that entry increases the willingness to pay for all other varieties. This effect, which is positive, is induced by the *love-for-variety* property of the aggregator $G(\cdot)$. This effect is itself an externality, in that it is not appropriated by the marginal entrant through their profits. In this case, it is not possible to unambiguously state whether the equilibrium features too many or too few firms. If the *love-for-variety* effect dominates, the equilibrium will have too few firms, relative to the constrained optimal solution. Finally, if the two externalities exactly cancel out, then the free entry equilibrium is constrained efficient and $N^e = N^\star$.

We can now go back to the monopolistic competition + CES case. CES demand functions have a knife-edge property: the elasticity of substitution and the preference for variety are governed by the same parameter σ. In this special case, the two forces cancel out exactly, and the free entry equilibrium is constrained efficient. This is the result from Dixit and Stiglitz (1977) and extended to the heterogeneous firms case in Dhingra and Morrow (2019).

The intuition is as follows: a constrained social planner can choose the number of firms by paying the entry cost. In the free entry equilibrium, firms charge markups to cover the entry cost. Are these markups efficient from the perspective of the planner? They are because without them, nobody would enter, and because of the CES properties, these are identical across firms, so there is no cross-sectional misallocation. The free entry condition eliminates pure rents and leaves only quasi-rents (profits to cover the entry cost). Since there are no pure rents and there is no misallocation, the two allocations coincide.

Put differently, markups are constant and identical, which implies that the economy is too small relative to the second-best scenario, but the allocation is efficient in relative terms. You would like to increase the size of all firms by the same amount, not change the relative size. Why is it efficient for the economy to be small? Because part of the output is used to pay the entry costs so, when the planner is constrained by having to pay these costs too, it picks a uniform markup for firms just so that the costs are covered. This is exactly what a free entry condition does in the CES+monopolistic competition model, hence the two coincide.

For further extensions of this result to economies with incumbents see Bajo et al. (2024) and to variable elasticity of substitution see Parenti et al. (2017).

End of Digression

You can easily (but tediously) solve the rest of the model and see that output is lower than its efficient level (which can actually be restored by a subsidy which exactly undoes the markup) if you do not have free entry. This model, however, has the somewhat limiting property of having all firms with the same constant markup. First, we know that markups are very different between firms, as shown in Figure 13. Secondly, it's hard to believe that markups do not respond to the business cycle. This result comes from firms being "small" relative to the economy. To address these shortcomings, we need to move away from the monopolistic competition setting to one in which firms are still small in the economy but large in their sector. This allows us to have markups move over the cycle and change the economy's response to shocks. We study this model in Section 4.4.

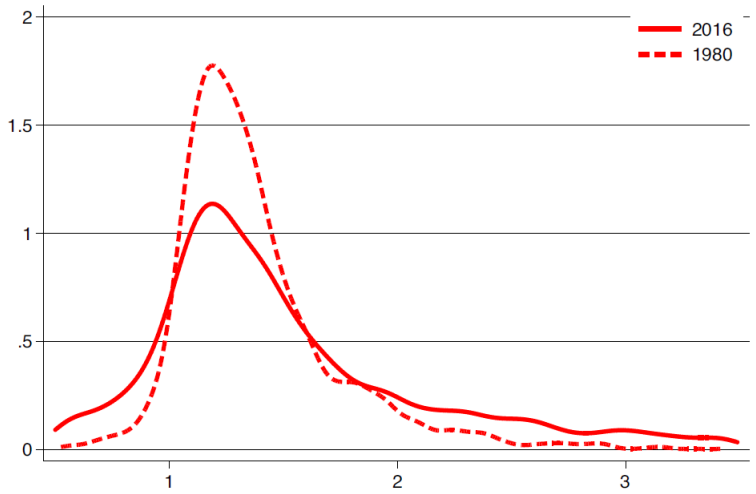


Figure 13: Markups Distribution from De Loecker, Eeckhout and Unger (2020).

Digression on Factor Market Power So far, we have discussed the problem of market power in product (output) markets. There is a large and growing literature concerned instead with market power on factor markets. To keep the discussion simple, consider the problem of a local labour market: for some specialized professions, there might be relatively few potential employers within a city. Workers considering larger markets would have to potentially incur moving costs. The presence of such costs gives local employers some market power over workers. The key ingredient we need is an upward-sloping residual labour supply (the equivalent of the downward-sloping residual demand on product markets). We can generate such an upward-sloping supply via preferences for employers/locations. Denote the inverse residual labour supply of firm i as $w(L_i)$. Denote the labour supply elasticity $\epsilon > 0$, meaning that $L = w^{\epsilon}$. Consider the profit maximization problem of firm i

$$\max_{L_i} L_i^{\beta} - w(L_i)L_i. \tag{256}$$

The optimizing firm chooses labor as

$$\beta \frac{Y_i}{L_i} = w(L_i) + w' L_i = w \frac{\epsilon + 1}{\epsilon}. \tag{257}$$

To benchmark our model, recall that the competitive version of the optimal labour choice is

$$\beta \frac{Y_i}{L_i} = w. \tag{258}$$

Solving for $L_i^\star$ and using the labour market clearing condition $w = L^{1/\epsilon}$ we obtain

$$L_i^\star = \beta^{\frac{1}{\frac{\epsilon+1}{\epsilon} - \beta}}, \tag{259}$$

and, therefore, output is given by

$$Y_i^\star = \beta^{\frac{\beta}{\frac{\epsilon+1}{\epsilon} - \beta}}. \tag{260}$$

In the case of factor market power, combining (257) with the labour supply equation we obtain

$$\tilde{L}_i^\star = \left(\frac{\epsilon \beta}{\epsilon + 1} \right)^{\frac{1}{\frac{\epsilon+1}{\epsilon} - \beta}}, \tag{261}$$

which implies

$$\tilde{Y}_i^\star = \left(\frac{\epsilon \beta}{\epsilon + 1} \right)^{\frac{\beta}{\frac{\epsilon+1}{\epsilon} - \beta}}. \tag{262}$$

Note two properties of the solution. First, we can define the labour share Θ_L in the two economies

$$\Theta_L = \beta \quad \text{and} \quad \tilde{\Theta}_L = \frac{\epsilon \beta}{\epsilon + 1} < \beta. \tag{263}$$

So the presence of market power on the factor market shrinks the labour share and, therefore, increases its complement to 1, which in this case is the profit share. Next, note that we can compare the optimal firm size in the two economies

$$\frac{Y_i^\star}{\tilde{Y}_i^\star} = \left(\frac{\epsilon}{\epsilon + 1} \right)^{\frac{\beta}{\frac{\epsilon+1}{\epsilon} - \beta}} > 1. \tag{264}$$

The presence of factor market power reduces the optimal size of the firm. Intuitively, firms facing an upward-sloping labour supply internalize the fact that, by hiring an extra worker, they increase the wage of all other employees. Therefore, the marginal cost of labour is higher than in the competitive case. As the marginal cost is higher at the given market price for the output good

(normalized to 1), the firm will optimally choose to operate at a smaller scale. Lastly, note that as ϵ increases, $Y_i^\star/\tilde{Y}_i^\star$ converges to 1 from above. Namely, as the labour supply becomes ever more elastic, the gap between the competitive and market power optimal size shrinks to zero.

For a recent macro paper studying the role of factor (labour) market power, see Berger et al. (2022). For a paper combining both product and factor market power, see Lo Bello and Pesaresi (2024). Finally, for a paper studying the quantitative importance of product vs factor market power, see Deb et al. (2022).

End of Digression

Digression on Market Power from Consumer Search We tend to think of market power as coming from differentiation. This is the backbone of the monopolistic competition model. It is not obvious, however, that we should observe no market power when goods are homogeneous. To see this, note that we can apply the Diamond Paradox to the context of goods rather than labour. In a world in which consumers spend time shopping (a search cost) in all but the first firm, firms with homogeneous goods will be able to extract all the surplus and behave like monopolists. This immediately implies that the market unravels. No consumer visits any firm but the first, and since there is no further search, there is no competition. Recall that in the Diamond Paradox setting, all jobs were homogeneous, yet firms extracted all the surplus. The same argument works for consumer search and homogeneous goods.

End of Digression

Digression Hopenhayn (1992) and Melitz (2003) At this point, we have solved the Hopenhayn (1992) model and the Melitz (2003) (in your problem set). For comparison, consider the quasi-static Hopenhayn (1992) we studied above. We found that the solution to the case where there is positive entry is governed by eq. 228. In the process of solving Melitz (2003) you have certainly found a similar-looking condition like

$$\int_{\varphi^*} \left[\left(\frac{\varphi}{\varphi^*} \right)^{\sigma-1} - 1 \right] g(\varphi) d\varphi = \frac{\delta f_e}{f}. \tag{265}$$

These two look exactly identical, provided that we set $\frac{1}{1-\alpha} = \sigma - 1$. The characterization of the aggregate economy is given simply by this condition (determining the cutoff) and a labor market clearing condition, which is identical in the two economies. Should we then conclude that these economies are isomorphic?

Counterintuitively, we should not. First note that by the Euler Theorem, the profit rate in Hopenhayn (1992) is $\frac{\pi}{y} = 1 - \alpha$. Conversely, in the Melitz (2003) model we have that $\frac{\pi}{py} = \frac{1}{\sigma}$. Hence, if we impose the parametric restriction $\frac{1}{1-\alpha} = \sigma - 1$, the profit rates of firms are different in the two economies.

Next, consider the aggregate production function in the two economies. In Hopenhayn (1992), we can simply sum individual firms' output since they make perfectly subsitutable goods. Hence,

$$Y_H = L^{\alpha}\left(\int z^{\frac{1}{1-\alpha}} dF(z)\right)^{1-\alpha}. \tag{266}$$

Note that $F(z)$ does not integrate to 1, but rather to the number of firms, which we can denote M. We can take this measure outside the integral to turn F into the equivalent probability measure $\mathcal{F}$ and write

$$Y_H = L^{\alpha}M^{1-\alpha}\left(\int z^{\frac{1}{1-\alpha}} d\mathcal{F}(z)\right)^{1-\alpha}. \tag{267}$$

The corresponding aggregate production function in Melitz (2003) is given by plugging individual output y into the CES aggregator to obtain

$$Y_M = LM^{\frac{1}{\sigma-1}}\left(\int z^{\sigma-1} dF(z)\right)^{\frac{1}{\sigma-1}}. \tag{268}$$

Again setting $1-\alpha = \frac{1}{\sigma-1}$, it is immediate that they are not equivalent in the aggregate since the returns to aggregate labour L are different.

To recover the aggregate equivalence, we need to change the aggregate returns to scale in the Hopenhayn (1992) production function. We can do this by introducing an aggregate externality and replacing $y = zn^{\alpha}$ with $y = Y^{1-\alpha}(zn)^{\alpha}$.[26] Following the same steps as before to find individual firms' output and aggregating across them, we obtain

$$Y_H = LM^{\frac{1-\alpha}{\alpha}}\left(\int z^{\frac{1}{1-\alpha}} d\mathcal{F}(z)\right)^{\frac{1-\alpha}{\alpha}}. \tag{269}$$

We can make the aggregate production functions coincide by setting $\sigma - 1 = \frac{\alpha}{1-\alpha}$. Importantly, we still do not recover equivalent firm-level outcomes. We conclude that, despite the similar aggregate behaviour, the two economies are not, in fact, isomorphic.

End of Digression

3.5 Misallocation - Hsieh and Klenow (2009)[27]

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

[26] Note that replacing $y = zn^{\alpha}$ with $y = (zn)^{\alpha}$ does not change anything since it is just a rescaling of what productivity z is.

[27] This subsection is based on Chris Edmond's lecture slides.

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

$$\pi_i = p_i y_i - w L_i - r K_i. \tag{270}$$

Suppose further that the firm faces a downward sloping demand derived from CES preferences. We know (see Appendix A.5) that, denoting c the marginal cost for $A_i = 1$, the price is given by

$$p_i = \frac{\sigma}{\sigma - 1} \frac{c}{A_i}. \tag{271}$$

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

$$p_i = \frac{\sigma}{\sigma - 1} \left(\frac{r}{\alpha}\right)^{\alpha} \left(\frac{w}{1-\alpha}\right)^{1-\alpha} A_i^{-1}. \tag{272}$$

We can then derive the optimal capital-labour ratio

$$\frac{K_i}{L_i} = \frac{\alpha}{1-\alpha} \frac{w}{r}. \tag{273}$$

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

$$y_i = p_i^{-\sigma} P^{\sigma} Y, \tag{274}$$

using the price equation, we can write that

$$y_i = \left(\frac{\sigma}{\sigma - 1} \left(\frac{r}{\alpha}\right)^{\alpha} \left(\frac{w}{1-\alpha}\right)^{1-\alpha} \right)^{-\sigma} A_i^{\sigma} P^{\sigma} Y, \tag{275}$$

which immediately implies

$$y_i \propto A_i^{\sigma}. \tag{276}$$

Meaning that output is proportional to firm productivity with curvature given by σ. In this model, firms will be heterogeneous in size, as a direct consequence of productivity heterogeneity.

Note also that the markup does not distort the cross-sectional distribution of firm size. This is a property of the special case of CES and monopolistic competition. Because, as we have shown before, the markup is identical across firms, the presence of the markup can distort the size but not the relative size of firms. Whether the actual size is inefficient or not depends on whether we impose a free entry condition as part of the equilibrium. If we do, we are back to the Dixit and Stiglitz (1977) results, if we do not, then firms are too small by a factor $\left(\frac{\sigma}{\sigma-1}\right)^{-\sigma}$. You can prove this result by solving the planner problem and noting that it can be decentralized via a production subsidy $(\sigma-1)/\sigma$ that exactly undoes the markup so that firms are back to marginal cost pricing.

In summary, while the economy is too small, there is no cross-sectional misallocation in this model. Another way to see this is to note that, by the firm's optimization problem, we have that

$$MRPK_i = r, \tag{277}$$

$$MRPL_i = w, \quad \forall i. \tag{278}$$

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.

Our 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

$$Y = AK^{\alpha}L^{1-\alpha}, \tag{279}$$

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

$$rK_i = \frac{\alpha c}{A_i} y_i, \tag{280}$$

$$wL_i = \frac{(1-\alpha)c}{A_i} y_i. \tag{281}$$

where $\frac{c}{A_i} y_i$ is the total cost with $c = \left(\frac{r}{\alpha}\right)^{\alpha}\left(\frac{w}{1-\alpha}\right)^{1-\alpha}$. It follows that aggregate capital and labour,

as given by

$$K = \sum_i K_i = \frac{\alpha c}{r} \sum_i \frac{y_i}{A_i}, \tag{282}$$

$$L = \sum_i L_i = \frac{(1-\alpha)c}{w} \sum_i \frac{y_i}{A_i} \tag{283}$$

We can plug these into the aggregate production function

$$A = Y K^{-\alpha} L^{\alpha-1} \tag{284}$$

$$= Y \left(\frac{\alpha c}{r} \sum_i \frac{y_i}{A_i} \right)^{-\alpha} \left(\frac{(1-\alpha)c}{w} \sum_i \frac{y_i}{A_i} \right)^{\alpha-1} \tag{285}$$

$$= Y \left(\frac{\alpha}{r} \right)^{-\alpha} \left(\frac{1-\alpha}{w} \right)^{\alpha-1} c^{-1} \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{286}$$

$$= Y \left(\frac{\alpha}{r} \right)^{-\alpha} \left(\frac{1-\alpha}{w} \right)^{\alpha-1} \left(\frac{r}{\alpha} \right)^{-\alpha} \left(\frac{w}{1-\alpha} \right)^{\alpha-1} \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{287}$$

$$= Y \left(\sum_i \frac{y_i}{A_i} \right)^{-1} \tag{288}$$

Inverting this condition, we obtain

$$A^{-1} = Y^{-1} \sum_i \frac{y_i}{A_i}, \tag{289}$$

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

$$A^{-1} = \sum_i \left(\frac{p_i}{P} \right)^{-\sigma} A_i^{-1}. \tag{290}$$

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

$$\frac{p_i}{P} = \frac{\frac{\sigma}{\sigma-1}\frac{c}{A_i}}{\left(\sum_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{291}$$

$$= \frac{\frac{\sigma}{\sigma-1}\frac{c}{A_i}}{\left(\sum_j \left(\frac{\sigma}{\sigma-1}\frac{c}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{292}$$

$$= \frac{\sigma}{\sigma-1}\frac{c}{A_i}\frac{\sigma-1}{\sigma c}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{293}$$

$$= \frac{1}{A_i}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}}. \tag{294}$$

We can plug this into (290) and obtain

$$A^{-1} = \sum_i A_i^{\sigma-1}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{-\sigma}{\sigma-1}}, \tag{295}$$

which implies

$$A = \left(\sum_i A_i^{\sigma-1}\right)^{-1}\left(\sum_j A_j^{\sigma-1}\right)^{\frac{\sigma}{\sigma-1}} \tag{296}$$

$$= \left(\sum_j A_j^{\sigma-1}\right)^{\frac{1}{\sigma-1}}. \tag{297}$$

That is, the aggregate productivity of the economy is a geometric 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

$$\pi_i = (1-\tau_i^Y)p_i(y_i)y_i - wL_i - (1+\tau_i^K)rK_i, \tag{298}$$

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

$$\pi_i = \tilde{p}_i(y_i)y_i - wL_i - \tilde{r}_i K_i. \tag{299}$$

The FOCs yield

$$\tilde{r}_i = \lambda_i \alpha A_i K_i^{\alpha-1} L_i^{1-\alpha} \tag{300}$$

$$w = \lambda_i (1-\alpha) A_i K_i^{\alpha} L_i^{-\alpha}. \tag{301}$$

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

$$\frac{K_i}{L_i} = \frac{\alpha}{1-\alpha}\frac{w}{r}\frac{1}{1+\tau_i^K}, \tag{302}$$

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

$$MRPL_i = w\frac{1}{1-\tau_i^Y}, \tag{303}$$

$$MRPK_i = r\frac{1+\tau_i^K}{1-\tau_i^Y}. \tag{304}$$

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

$$p_i = \frac{\sigma}{\sigma - 1} \frac{c}{A_i} \frac{(1 + \tau_i^K)^\alpha}{1 - \tau_i^Y}. \tag{305}$$

To obtain this result start from the cost minimization problem

$$\min_{K_i, L_i} wL_i + (1 + \tau_i^K) r K_i + \lambda(\bar{y} - A_i K_i^\alpha L_i^{1-\alpha}), \tag{306}$$

recalling that λ is the marginal cost of the firm. Next, you can maximize the profits in (298) using the marginal cost you just derived to obtain the optimal price in (306). 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

$$(1 - \tau_i^Y) p_i = \frac{\sigma}{\sigma - 1} \lambda. \tag{307}$$

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

$$K_i = \frac{\lambda_i \alpha y_i}{\tilde{r}_i}, \tag{308}$$

$$L_i = \frac{\lambda_i (1 - \alpha) y_i}{w}. \tag{309}$$

Plugging into the production function

$$y_i = A_i \left(\frac{\lambda_i \alpha y_i}{\tilde{r}_i} \right)^\alpha \left(\frac{\lambda_i (1 - \alpha) y_i}{w} \right)^{1-\alpha}, \tag{310}$$

which implies

$$\lambda_i^{-1} = A_i \left(\frac{\alpha}{\tilde{r}_i} \right)^\alpha \left(\frac{(1 - \alpha)}{w} \right)^{1-\alpha}. \tag{311}$$

Therefore the marginal cost is given by

$$\lambda_i = \frac{\tilde{c}_i}{A_i} = A_i^{-1} \left(\frac{\alpha}{\tilde{r}_i} \right)^{-\alpha} \left(\frac{(1 - \alpha)}{w} \right)^{\alpha - 1}. \tag{312}$$

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 (307), we obtain (305).

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

$$y_i = p_i^{-\sigma} P^{\sigma} Y, \tag{313}$$

which, using the price, implies

$$y_i = \left(\frac{\sigma}{\sigma - 1} c\right)^{-\sigma} A_i^{\sigma} \left(\frac{1 - \tau_i^Y}{1 + \tau_i^K}\right)^{\sigma}. \tag{314}$$

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

$$y_i \propto A_i^{\sigma} \left(\frac{1 - \tau_i^Y}{1 + \tau_i^K}\right)^{\sigma}. \tag{315}$$

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

$$(1 + \tau_i^K) r K_i = \alpha \frac{c}{A_i} (1 + \tau_i^K)^{\alpha} y_i, \tag{316}$$

$$w L_i = (1 - \alpha) \frac{c}{A_i} (1 + \tau_i^K)^{\alpha} y_i. \tag{317}$$

Therefore the effective capital and labour shares are

$$\Theta_K = \frac{rK}{PY} = \frac{\alpha}{\frac{\sigma}{\sigma - 1}} \sum_i \frac{1 - \tau_i^Y}{1 + \tau_i^K} \frac{p_i y_i}{PY} \tag{318}$$

$$\Theta_L = \frac{wL}{PY} = \frac{1 - \alpha}{\frac{\sigma}{\sigma - 1}} \sum_i (1 - \tau_i^Y) \frac{p_i y_i}{PY}. \tag{319}$$

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

$$PY = \left(\frac{r}{\Theta_K}\right)^{\alpha} \left(\frac{w}{\Theta_L}\right)^{1-\alpha} K^{\alpha} L^{1-\alpha}. \tag{320}$$

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

$$A = Y K^{-\alpha} L^{\alpha - 1}, \tag{321}$$

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

$$p_i A_i = \frac{\sigma}{\sigma - 1} c \frac{(1 + \tau_i^K)^{\alpha}}{1 - \tau_i^Y}. \tag{322}$$

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

$$PA = \left(\frac{r}{\Theta_K}\right)^{\alpha}\left(\frac{w}{\Theta_L}\right)^{1-\alpha} \equiv TFPR. \tag{323}$$

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

$$P = \left(\sum_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}} \tag{324}$$

$$= \left(\sum_j \left(\frac{TFPR_j}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}, \tag{325}$$

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

$$A = TFPR\left(\sum_j \left(\frac{TFPR_j}{A_j}\right)^{1-\sigma}\right)^{\frac{1}{\sigma-1}} \tag{326}$$

$$= \left(\sum_j \left(A_j \frac{TFPR}{TFPR_j}\right)^{\sigma-1}\right)^{\frac{1}{\sigma-1}} \tag{327}$$

So the aggregate productivity A in this economy is not the same as in the efficient economy. If $TFPR = TFPR_j$, $\forall j$ then (327) collapses to (297). 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, defined as $s_i = \frac{p_i y_i}{PY} = (y_i/Y)^{\frac{\sigma-1}{\sigma}}$, see Appendix A.5 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

$$A_i = y_i K_i^{-\alpha} L_i^{\alpha-1}, \tag{328}$$

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

$$A_i = K_i^{-\alpha} L_i^{\alpha-1} \left(\frac{p_i y_i}{PY}\right)^{\frac{\sigma}{\sigma-1}} Y \tag{329}$$

$$= K_i^{-\alpha} L_i^{\alpha-1} (p_i y_i)^{\frac{\sigma}{\sigma-1}} \chi, \tag{330}$$

where $\chi \coloneqq (PY)^{\frac{\sigma}{1-\sigma}} Y$ is just a function of aggregates. We cannot recover the level of productivities A_i since these aggregates are a free parameter. However, we can normalize it to 1 and study relative productivities. From (330), 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

$$TFPR_i = (p_i y_i) K_i^{-\alpha} L_i^{\alpha-1}. \tag{331}$$

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

3.6 Endogenous Amplification

So far, these models of firms have nothing to say about the business cycle. The reason is two-fold. First, we assumed no aggregate uncertainty; firms were identical and small.

In the next section, we will remove the "small", but first, we can already say something interesting while firms are still price-takers by introducing aggregate fluctuations.

The general goal will always be to characterize the elasticity $\epsilon_y \equiv \frac{\partial \log y}{\partial log z}$ where z is the aggregate shock. We start by noting that, with DRS δ, it is immediate, in partial equilibrium, that firms will respond

$$\epsilon_y = \frac{\partial \log y}{\partial \log z} = \frac{1}{1-\delta} > 1. \tag{332}$$

In general equilibrium, we know that the wage will respond. To characterize that, pick $y = zn^{\delta}$.[28]

[28] These results generalize to any production function homogeneous of degree δ.

We know that

$$y = z^{\frac{1}{1-\delta}} \left(\frac{\delta}{w}\right)^{\frac{\delta}{1-\delta}}. \tag{333}$$

Using the implicit function theorem, we get that

$$\epsilon_y = \frac{1}{1-\delta+\delta\frac{w'y}{w}}, \tag{334}$$

where $w'y/w$ is the elasticity of the wage with respect to output.[29] This is likely to be positive as higher output implies a higher labour demand and, therefore, higher wages. Hence the GE elasticity is smaller than the PE elasticity because the wage responds endogenously. From this result is also clear that when we take the limit to CRS, the output elasticity uniquely depends on the wage elasticity. The magnitude of the economy's response, therefore, only depends on the elasticity of factor supply, in this case, labour. In this model, the clearing of the labour market determines the marginal cost of firms and, therefore, their optimal size. For example, if we assume that the labour supply is inelastic at $\bar{n}$, we get that the aggregate production function is given by

$$y = z\bar{n}^{\delta} \tag{335}$$

and therefore, the elasticity is 1, meaning that wages absorb all the extra gains and real output changes one-to-one with the increase in productivity.[30] This is intuitive because output only depends on the total supplied labour, which is a constant. Such labour is transformed into output at rate z, hence increasing z by 1%, as labour does not change, increases output by 1%. This is not the case if you have elastic labour supply, for example, $n = w^{\psi}$ for some $\psi > 0$. In this case, the labour market clearing, obtained by equating firms' labour demand and workers' labour supply, is

$$w^{\psi+\frac{1}{1-\delta}} = (z\delta)^{\frac{1}{1-\delta}}. \tag{336}$$

Plugging the equilibrium wage back into the optimal level of output and taking the usual elasticity with respect to z we have

$$\epsilon_y = \frac{1+\psi}{1+\psi(1-\delta)}. \tag{337}$$

Checking the $\psi = 0$ case yields back the inelastic supply case of unitary elasticity. Intuitively, for large values of ψ, where the labour supply responds significantly to changes in wages, output changes the most (you can verify that the elasticity is increasing in ψ). When $\psi > 0$, the elasticity is strictly larger than 1. This can be decomposed by a direct effect of z on output, such that, with

[29] To get this result, it might be convenient to do a change of variable into $x \equiv \log y$. and then study $\partial x/\partial \log z$.
[30] To see this, you can use the labour market clearing to solve for the wage and then check its elasticity.

constant labour, output increases one-to-one. On top of this effect, the equilibrium level of labour also increased. This, in turn, adds to the unitary elasticity, endogenously amplifying the effect of the exogenous shock.

In a similar fashion, we can ask what happens with the extensive margin. Suppose that we live in the Hopenhayn (1992) world with idiosyncratic productivities and aggregate shocks. Let's, however, assume that idiosyncratic productivities are fixed so that all uncertainty is coming from the aggregate state. Then GDP is by aggregating policy functions

$$Y(p,z) = \int_{\phi > \phi^\star(p,z)} y(\phi, p, z)\mu(\phi)d\phi = D(p). \tag{338}$$

On top of all previous considerations, we now have to ask ourselves what a change in z does to $\phi^\star$. Intuitively, a positive aggregate shock will make entry more appealing (thereby lowering the minimum productivity threshold). On the other hand, as wages respond, they might fully undo such an effect. With labour supplies that are not too elastic, the net contribution of entry will be positive. Nonetheless, we still have that the new entrants are not as productive as incumbents, which lowers the efficiency of the economy. In a model like the one just described, augmented with physical capital accumulation, Clementi and Palazzo (2016) show that the extensive margin is responsible for about 20% of the economy's response to aggregate shocks.

So far, we have studied amplification in the case of competitive and small firms. We now proceed to relax these two assumptions by looking at how firms can be "large" in an economy and how market power can affect business cycle fluctuations.

Digression on Convex Supply Curves If we are interested in understanding how an economy responds to shocks, an important issue is whether adjustment occurs mostly through prices or quantities. This matters, particularly, in terms of state-dependence: does the economy respond the same way to large and small shocks? Suppose that the supply curve is very concave. If the economy is currently small (therefore in the steep part of the supply curve), an upward shift in demand has small quantity effects and large price effects; if the economy is already large (in the flat part), then changes in demand will mostly be absorbed by quantities. Suppose otherwise that the supply curve is very convex. Then exactly the opposite holds: small economies absorb shocks via larger quantity responses than prices responses. Then we can ask when are supply curves concave/convex and how do they look in the data?

Consider a competitive model with constant returns to scale. The supply curve of each firm is given by $p = c$, which is therefore flat in the (y, p) space. Firms' supply is infinitely price elastic. Any shift in demand is absorbed fully by quantities.

Consider instead a competitive model where firms have decreasing returns to scale $\delta < 1$. Firms'

supply is given by $p = c(y)$. Formally, suppose the firm minimises

$$\mathcal{L} = \sum_x p_x x + \lambda[Y - f(x)], \tag{339}$$

Then the FOCs imply $p_x = \lambda f_x(x)$, $\forall x$ and therefore $x^\star = f_x^{-1}(\frac{p_x}{\lambda})$. Homogeneous functions have two convenient properties (see Appendix A.1): i) if $f(x)$ is homogeneous of degree δ, then f_x is homogeneous of degree $\delta - 1$ and ii) f^{-1} is homogeneous of degree $1/\delta$. From this, $x^\star = \lambda^{\frac{1}{1-\delta}} f_x^{-1}(p_x)$. Plugging into the production function $f(x^\star) = f(\lambda^{\frac{1}{1-\delta}} f_x^{-1}(p_x)) = \lambda^{\frac{\delta}{1-\delta}} f(f_x^{-1}(p_x))$. Hence $\lambda = y^{\frac{1-\delta}{\delta}} [f(f_x^{-1}(p_x))]^{\frac{\delta-1}{\delta}}$. Clearly, now the marginal cost $c(y) = \lambda(y^\star)$ is increasing in output, and therefore the supply curve is upward sloping. The convexity/concavity of the supply curve is determined by whether $\delta \lessgtr 1/2$. To see that note that $p'' = \frac{(1-\delta)(1-2\delta)}{\delta^2} y^{\frac{1-3\delta}{\delta}} [f(f_x^{-1}(p_x))]^{\frac{\delta-1}{\delta}}$, which is positive for $\delta < 1/2$.

In a recent paper, Boehm and Pandalai-Nayar (2022) show that, at the industry level, supply curves are convex. They rationalise this through a model in which firms have constant returns to scale but have capacity constraints $\bar{y}$. This implies that the supply curve of each individual firm is given by $p = mc(y)$, as long as $y < \bar{y}$, but they become vertical for all $y > \bar{y}$. If different firms in the same industry have different constraints $\bar{y}_i$, then the industry-level supply curve will be convex when the constraint starts binding for some firms. Intuitively, as the change in demand increases in size, more and more firms will be in the vertical portion of their supply curves. As a consequence, the economy will be in the steeper part of the industry-level supply curve. Notice the difference between these two approaches: above, we characterized when individual firms have convex supply curves. Then, we got the same result through aggregation: we could have a collection of firms whose supply curves are first infinitely elastic and then infinitely inelastic, and, as long as the capacity constraint is heterogeneous, aggregation will smooth the kink out into a a convex aggregate supply.

End of Digression

Digression on Cleansing Recessions Since Schumpeter (1939) theory of business cycles, we hold the idea that, despite the short-run cost, recessions can have positive long-run effects. The idea behind the *liquidationist* view of business cycles is that recessions are a time in which unproductive firms are forced to exit and that, as the economy recovers, they are replaced by better firms. These forces are often referred to as the *cleansing effects* of recessions. This idea is formalized in Caballero and Hammour (1994) in the context of a *putty-clay* model of capital. In this class of models, capital has *vintages*: capital created in 2024 is more productive than capital produced in 1980. Therefore, in recessions, old capital is the first to turn unprofitable and leave the economy. This is then replaced by new capital. As a consequence, recessions improve the average productivity of capital in the economy.

Consider instead a different, more familiar model from Bajo et al. (2024). Firms are heterogeneous in their permanent productivity as in the quasi-static Hopenhayn (1992) model. They operate a

constant return technology using labor, produce a differentiated variety, and compete in monopolistic competition as in Melitz (2003). These varieties are assembled by a final good producing competitive firm that buys the intermediate differentiated goods and combines them into a single final good consumed by the household. This firm has the following production function

$$Y = M^{q-\frac{1}{\sigma-1}} \left[\int y(z)^{\frac{\sigma-1}{\sigma}} \mu(z) dz \right]^{\frac{\sigma}{\sigma-1}}, \tag{340}$$

where M is the number of varieties in the economy and σ is the elasticity of substitution. This aggregator is a generalization of the CES aggregator in Dixit and Stiglitz (1977) first used in Dixit and Stiglitz (1975); Ethier (1982); Benassy (1996). To understand why this aggregator is appealing, consider the notion of *love-for-variety*. This is often defined as the elasticity of welfare to the number of varieties in the economy. In a CES economy, this elasticity is $\frac{1}{\sigma-1}$. Note that the same parameter that governs love-for-variety also governs the elasticity of substation across varieties and the elasticity of demand. The aggregator in 340 conversely has two distinct parameters: q governs love-for-variety while σ governs the elasticities of substitution and demand. When $q = \frac{1}{\sigma-1}$, the aggregator collapses to the standard CES aggregator in 240.

Suppose that, as in Hopenhayn (1992), firms pay an entry cost to draw their permanent productivity and a fixed cost to operate in each period. Both are paid in terms of labour. Note that the presence of M in the production function of the final good producer represents an externality from the perspective of differentiated good firms, since they do not take it into account when they make entry decisions. From the perspective of each individual firm, the solution of Melitz (2003) applies since the aggregate externality does not affect individual choices. The equilibrium in this economy is defined by a productivity cutoff $\underline{z}$ and a number of firms M.

Consider now fluctuations in the fixed cost f^c that induce business cycles. In particular, suppose that suddenly the fixed cost jumps up. On impact, the cutoff rises so that the zero profit condition holds at the higher level of fixed cost. This results in the exit of the least productive firms in the economy. The economy reaches a new steady state with fewer firms that are, on average, more productive.

Suppose now that the fixed cost reverts to its initial level. It is possible to show that the cutoff reverts to its initial equilibrium level. However, the post-crisis economy features fewer firms that are, on average, more productive. This dynamics is shown in Figure 14.

First, note that Schumpeterian forces are at play: unproductive firms are replaced after the recession by entrants that are, on average, more productive. Furthermore, since there are fewer firms in the economy, fewer fixed costs need to be paid, which frees up some labour for production. However, M decreased after the recession. To study the effect on GDP and welfare, totally differentiate the aggregator, plugging in the individual firm optimal production choice $y_\tau(z) =$

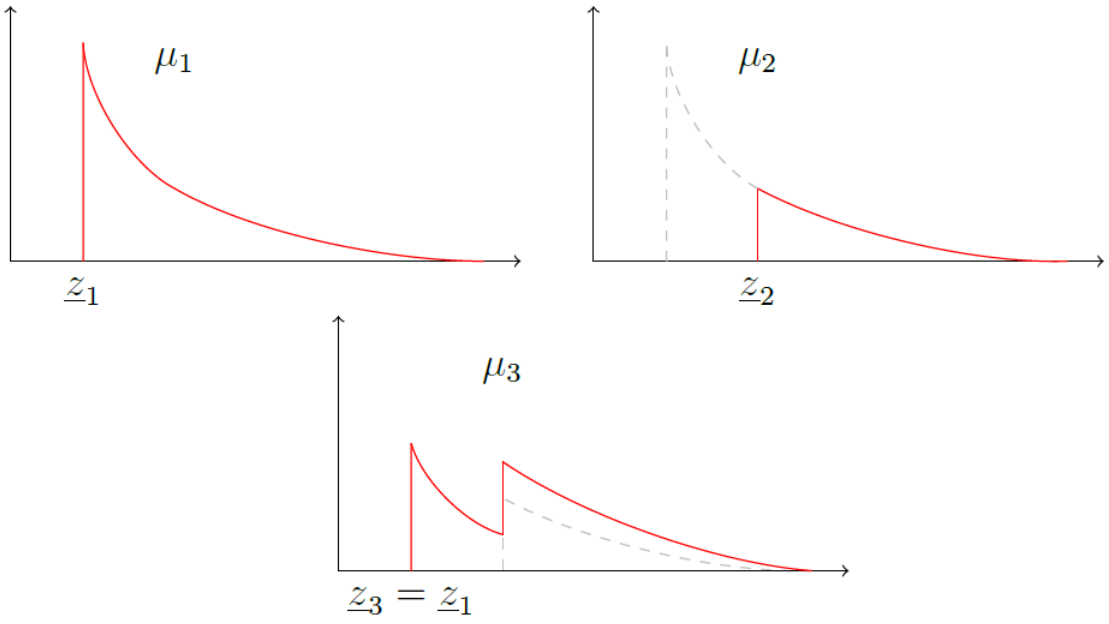


Figure 14: The figure shows the entry and exit dynamics over the business cycle. Panel (A) shows the distribution μ_1 before the shock hits. Upon impact, the left tail of firms with productivity less than $\underline{z}^2$ leave, creating distribution μ_2 (B). Finally, after fixed costs return to pre-shock levels and new firms drawn from the baseline distribution (μ^0) enter, μ_3 becomes the distribution of productivities in the market (C).

$z\mathcal{I}\frac{\sigma-1}{\sigma}\frac{z^{\sigma-1}}{\int z^{\sigma-1}\mu_\tau(z)dz}$ in 340, where $\mathcal{I}$ is the total income of the household:

$$\partial \log Y = \left(q - \frac{1}{\sigma - 1}\right)\delta \log M + \delta \log L + \delta \log Z, \tag{341}$$

where $Z \equiv \left(\int z^{\sigma-1}\mu(z)dz\right)^{\frac{1}{\sigma-1}}$ is aggregate productivity. The last two terms are positive after a recession since some labour was freed up for production, and there was some positive selection, which increased the average productivity of firms operating in the economy. However, there are fewer firms: $\delta \log M < 0$. Suppose that $q \gg \frac{1}{\sigma-1}$, namely there is much more love-for-variety than implied by CES. Then, it is possible that trading off fewer, on average more productive firms induces a net output and welfare loss. For more details on this, see Bajo et al. (2024).

End of Digression

4 Large Firms, Networks and Oligopoly 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. From this, we can discuss the Lucas (1977) argument for the irrelevance of idiosyncratic fluctuations. We then break this result by letting firms have granular effects as in Gabaix (2011). In the second part of the section we study economies that preserve the macro behaviour of granular models but have much richer micro-level behaviour through production networks. Finally, we discuss the role of imperfect competition and market power, looking at the recent literature on oligopoly in macro.

In this section, we discuss three distinct elements of recent advances in the field of firms in macro. 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

We are interested in understanding how idiosyncratic shocks to firms turn into aggregate fluctuations. We start by deriving this in an efficient benchmark up to a first order approximation. Consider an economy in which output Y is given by an aggregator $Y(c_1, \dots, c_n)$ whose price index is normalized to 1. Suppose, further, that households inelastically supply capital and labour. Further, suppose that firms produce with a production function f using as input labour, capital, and the output of other firms, denoting x_{ij} the amount of output if firm i used in production by firm j. Lastly, assume that firms have a Hicks-neutral productivity A_i. The planner problem is given by

$$\max_{\{c_i\}_i, K, L} Y(c_1, \dots, c_n), \quad \text{st} \tag{342}$$

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

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

$$\sum_i k_i \leq K. \qquad (\lambda_K)$$

The Lagrangian of the planner problem is

$$\max_{\{c_i\}_i,K,L} Y(c_1,\ldots,c_n) + \sum_i \mu_i \left(A_i F_i(\cdot) - \sum_j x_{ij} - c_i \right) + \lambda_L \left(L - \sum_i l_i \right) + \lambda_K \left(K - \sum_i k_i \right) \tag{343}$$

Then, by the Envelope Theorem, we have that a change in the productivity of individual firms A_i induces

$$\frac{dY}{dA_i} = \sum_i \mu_i F_i(\cdot). \tag{344}$$

Note that, by the First Welfare Theorem, $\mu_i = p_i$, $\forall i$. Then, using the fact that $s_i = \frac{y_i p_i}{Y}$ is the sales share of GDP, or Domar weight, and dividing both sides by Y, we obtain

$$d\log Y = \sum_i s_i d\log A_i. \tag{345}$$

This is known as the **Hulten theorem**. This very general result which states that in efficient economies, locally, the aggregate effect of an idiosyncratic shock to the productivity of firm i is governed by their sales share of GDP. Note that, in general, the sales share of GDP need not sum to 1 in the presence of intermediate inputs. If the economy features an Input-Output structure, the sum of sales shares of GDP is larger than 1.

Note where we used the efficiency and where the first order effects. We used the First Welfare Theorem to go from multipliers of the planner problem to prices in the competitive equilibrium. By invoking the envelope theorem, this result only holds locally, or up to first order. We know that, in general, this result will not hold at higher orders when the economy is allowed to reallocate resources.

Irrelevance of Micro Shocks - Lucas (1977) 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

$$\sigma_{GDP} = (Var(d\log Y))^{\frac{1}{2}} = \left(\sum_i s_i^2 \right)^{\frac{1}{2}} \sigma. \tag{346}$$

If there is a large number of firms N of similar sizes, the term $(\sum_i s_i^2)^{\frac{1}{2}}$ 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 - Gabaix (2011)[31]

Throughout the course, we looked at economies where firms were small. The definition of small was that no individual firm's choice or shock was 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.[32] 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

$$y_t = \sum_{i=1}^{N} s_{it}, \tag{347}$$

which immediately implies that GDP growth is

$$\frac{\Delta y_{t+1}}{y_t} = \frac{1}{y_t} \sum_{i=1}^{N} \Delta s_{it+1} = \sum_{i=1}^{N} \sigma_i \frac{s_{it}}{y_t} \epsilon_{it+1}. \tag{348}$$

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

$$\sigma_y = \left(\sum_{i=1}^{N} \sigma_i^2 \left(\frac{s_{it}}{y_t} \right)^2 \right)^{1/2}. \tag{349}$$

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

[32] 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".

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

$$\sigma_y = \sigma h, \tag{350}$$

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

$$\sigma_y = \frac{\sigma}{\sqrt{N}}. \tag{351}$$

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

$$\hat{\sigma}_y \approx 0.012\% \text{ per year.} \tag{352}$$

The aggregate measured volatility of US GDP is approximately 1%, so we are off by orders of magnitude. More in general, 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 $\zeta > 1$, then, as $N \to \infty$,

$$\begin{aligned} \sigma_y &\sim \frac{\nu_\zeta}{\ln N} && \zeta = 1 \\ \sigma_y &\sim \frac{\nu_\zeta}{N^{1-1/\zeta}} && 1 < \zeta < 2 \\ \sigma_y &\sim \frac{\nu_\zeta}{N^{1/2}} && \zeta \geq 2, \end{aligned} \tag{353}$$

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 it 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.

Digression on 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*

$$\Gamma_t \equiv \sum_{i=1}^{K} \frac{s_{it-1}}{y_{t-1}} \hat{\epsilon}_{it}, \tag{354}$$

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.[33] In this case, we would then have

$$\Gamma_t \equiv \sum_{i=1}^{K} \frac{s_{it-1}}{y_{t-1}} (g_{it} - \bar{g}_t). \tag{355}$$

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 into 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).

Digression on Power Laws, Properties and Genesis[34] 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

$$\log(Rank_y) = \log a + \beta \log X \tag{356}$$

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

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

Where $Rank_y$ is the rank of observation Y in terms of order statistics. 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 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

Digression on Granular Firm Dynamics So far we have studied firm dynamics and granular effects in isolation. A growing literature is merging these two to think about how an economy with granular firms evolves over time and what are its aggregate properties. Carvalho and Grassi (2019) study an economy populated by a finite number of firms whose productivity is heterogeneous and evolves according to a discrete Markov process. Firms operate with decreasing returns to scale and in perfect competition. Labour is elastically supplied. Effectively this is Hopenhayn (1992) without the continuum of firms assumption. If a firm has productivity φ, for a given aggregate state of the economy μ (in the models of section 3.3 the aggregate state was either the price or the wage), it maximises $\pi(\mu, \varphi) = \max_n \varphi n^\delta - w(\mu)n - c_f$. This implies an optimal size $n(\mu, \varphi) = \left(\frac{\delta\varphi}{w(\mu)}\right)^{\frac{1}{1-\delta}}$. In this model, the aggregate state is determined by the distribution of firms μ. If μ_s is the mass of firms

with productivity φ^s, then the aggregate productivity of the economy is $A = \left(\sum_s^S (\varphi^s)^{\frac{1}{1-\delta}} \mu_s\right)^{1-\delta}$.[35] To close the model we add an elastic labour supply $L = Mw^\gamma$, where M is the number of potential entrants. This last assumption implies that the equilibrium wage is independent of the number of potential entrants, which is a rather intuitive property. The labour market clearing implies the equilibrium wage $w = \left(\frac{\alpha A}{M^{1-\alpha}}\right)^{\frac{1}{1+\gamma(1-\alpha)}}$.

From here, note that a change in a productivity bin φ_s induces a change in output of all firms since firms can have aggregate effects. Formally: $\frac{\partial \log y_i}{\partial \log \varphi_s} = \frac{\partial \log y_i}{\partial \log w}\frac{\partial \log w}{\partial \log \varphi_s} = \frac{\partial \log y_i}{\partial \log w}\frac{\partial \log w}{\partial \log A}\frac{\partial \log A}{\partial \log \varphi_s} = -\frac{\delta}{1-\delta}\frac{1}{1+\gamma(1-\delta)}(1-\delta)\frac{\mu_s + \sum_k \varphi_k \frac{\partial \mu_k}{\partial \varphi_s}}{A} = \frac{\delta}{\gamma(\delta-1)-1}\frac{\mu_s + \sum_k \varphi_k \frac{\partial \mu_k}{\partial \varphi_s}}{A}$. We would like to characterize the law of motion of the distribution μ_t. In the continuum case of Hopenhayn (1992), this has a deterministic law of motion given by (218). In the discrete case, the law of motion has a stochastic component, depending on the specific realization of the Markov chain. Carvalho and Grassi (2019) show that in this economy, without aggregate shocks, aggregate productivity has a stochastic process with persistence. They also show that the volatility of the economy is time-varying and depends on the dispersion of firm size. Quantitatively, granular shocks explain approximately a quarter of the observed volatility of aggregate output.

End of Digression

4.3 Network Economies - Acemoglu, Carvalho, Ozdaglar and Tahbaz-Salehi (2012)

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 the equilibrium and recover the Hulten Theorem to 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

[35] To obtain this, define the aggregate production function $Y = AL^\delta$, with $Y = \sum_j y_j$ and $L = \sum_j n_j$. Then note that $L = (\alpha/w)^{\frac{1}{1-\delta}} \sum_j \varphi_j^{\frac{1}{1-\delta}}$ and therefore $Y = A\,(\alpha/w)^{\frac{\alpha}{1-\delta}} \left(\sum_j \varphi_j^{\frac{1}{1-\delta}}\right)^\alpha$. Finally, note that $Y = \sum_j y_j = (\alpha/w)^{\frac{\alpha}{1-\alpha}} \sum_j \varphi_j \frac{1}{1-\alpha}$. You can combine these two to back out A.

(2019).[36] There is a representative consumer with 1 unit of labour, supplied inelastically. The consumer aggregates goods by

$$u(\underline{c}) = \sum_{i=1}^{n} \beta_i \ln(c_i/\beta_i). \tag{357}$$

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

$$x_i = z_i \Gamma_i l_i^{\alpha_i} \prod_{j=1}^{n} x_{ji}^{a_{ji}}, \tag{358}$$

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 collection $\mathcal{A}$ of a_{ij} 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 $in_i = \sum_j a_{ji} = 1 - \alpha_i$. 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

$$y_i = c_i + \sum_j x_{ij}. \tag{359}$$

We start by noting that the minimization of expenditure for firms implies that $p_j x_{ji} = a_{ji} p_i y_i$ and $w l_i = \alpha_i p_i y_i$. 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 359 (after we multiply both sides by p_i) we get

$$p_i y_i = \beta_i E + p_i \sum_j \alpha_{ij} \frac{p_j y_j}{p_i} = \beta_i E + \sum_j \alpha_{ij} p_j y_j. \tag{360}$$

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

$$\lambda_i = \beta_i + \sum_{j=1}^{n} a_{ij} \lambda_j, \tag{361}$$

This first important result states that a firm's sector share over GDP is given by its importance

[36] 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).

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

$$\Lambda = [I - \mathcal{A}]^{-1} B. \tag{362}$$

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),[37] therefore we can express it as

$$L = [I - \mathcal{A}]^{-1} = \sum_{j=0}^{\infty} \mathcal{A}^j. \tag{363}$$

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

$$\lambda_i = \sum_{j=1}^{n} \ell_{ij} \beta_j. \tag{364}$$

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

$$y_i = z_i \Gamma_i \left(\frac{\alpha p_i y_i}{w} \right)^{\alpha_i} \prod_j \left(\frac{\alpha_{ji} p_i y_i}{p_j} \right)^{\alpha_{ji}}. \tag{365}$$

From here, we note that since $\alpha_i + \sum_{ji} a_{ji} = 1$, 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.

[37] 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.

Solving for the price we get

$$p_i^{-1} = z_i \Gamma_i \left(\frac{\alpha_i}{w}\right)^{\alpha_i} \prod_j \left(\frac{\alpha_{ji}}{p_j}\right)^{\alpha_{ji}}. \tag{366}$$

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. From here, you can also see how a change in the productivity of firm i will affect other firms j. When i becomes more productive, its price will decrease. This price is part of the marginal cost of other firms j, which implies that their marginal cost will decrease. Since all these firms price at marginal cost, the price of firms j will decrease, thereby amplifying the effect of the change in the productivity of i. You can also immediately see that productivity shocks travel from an industry to their customer, therefore going downstream.

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 problem.

We can go further to solve this problem. First, we take logs of 366. 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

$$\ln p_i = -\epsilon_i + \alpha_i \ln w + \sum_j a_{ji} \ln p_j - \ln \Gamma_i - \Delta_i, \tag{367}$$

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

$$\ln p_i = -\epsilon_i + \alpha_i \ln w + \sum_j a_{ji} \ln p_j \tag{368}$$

We can use again the fact that $\alpha_i + \sum_j a_{ji} = 1$ to get

$$\ln(p_i/w) = -\epsilon_i + \sum_j \alpha_{ji} \ln(p_j/w). \tag{369}$$

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

$$\hat{P} = -\epsilon + \mathcal{A}' \hat{P} \tag{370}$$

to get

$$\hat{P} = -[I - \mathcal{A}']^{-1}\epsilon. \tag{371}$$

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 $\ln(p_i/w) = -\sum_j \epsilon_j \ell_{ji}$, 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 $\ln(p_i/GDP) = -\sum_j \epsilon_j \ell_{ji}$. Recall that from before we had $p_i y_i/GDP = \sum_{j=1}^n \ell_{ij}\beta_j$, taking logs

$$\ln y_i + \ln(p_i/GDP) = \ln(\sum_{j=1}^{n} \ell_{ij}\beta_j). \tag{372}$$

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

$$\ln y_i = \sum_j \epsilon_j \ell_{ji} + C \tag{373}$$

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 the productivity of all firms, denoted $d\epsilon$. The effect of this productivity shock on the output of firm i is

$$d\ln y_i = d\epsilon_i + \sum_j (\ell_{ji} - \mathbb{1}_{i=j}) d\epsilon_j. \tag{374}$$

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. In other words

$$\frac{d\ln y_i}{d\ln z_j} = \ell_{ji}. \tag{375}$$

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

$$\ln(p_i/w) = -\epsilon_i + \sum_j \alpha_{ji} \ln(p_j/w) = -\sum_j \epsilon_j \ell_{ji}, \tag{376}$$

we can multiply by β_i and sum over i.

$$\sum_i \beta_i \ln(p_i/w) = -\sum_{j=1} \epsilon_j \sum_{i=1}^{n} \ell_{ji}\beta_i, \tag{377}$$

$$-\ln w = -\sum_{j=1} \epsilon_j \sum_{i=1}^{n} \ell_{ji}\beta_i - \sum_i \beta_i \ln p_i. \tag{378}$$

Note that the last term 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 last term becomes zero as it is the log of 1. Changing the sign we get

$$\ln GDP = \sum_{j=1}^{n} \epsilon_j \sum_{i=1}^{n} \ell_{ji}\beta_i = \sum_{j=1}^{m} \epsilon_j \lambda_j. \tag{379}$$

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.

Upon noting that the weight of the firm is still given by its sales share, we know that this is the statement of the Hulten Theorem. Formally the theorem states that

$$\frac{d \ln GDP}{d \ln z_j} = \lambda_j. \tag{380}$$

The intuition is that in a world with input-output linkages, the effect of a productivity shock to a firm depends not only on how it changes value added directly but also on how it increases it via changes in other connected firms' responses. These two effects are jointly summarized by a firm's sales rather than by its value added. An important corollary of the Hulten theorem is that the micro details of the economy do not matter for aggregate movements, provided that the sales shares distribution is the same. For example, we could take economies with different production networks which generate the same set of λ_js and the two would experience the same aggregate fluctuations. Importantly, they would differ in their micro-level behaviour. We come back to this shortly.

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 share 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 the aggregate volatility of this economy based on how 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 with the key difference that the relevant metric will not be the value added of a firm anymore (recall that in Gabaix (2011) sales and value added were the same thing) but rather by its sales to account for the indirect effects through the network.

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

$$\sigma_{\ln GDP} = \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{1/2} \sigma, \tag{381}$$

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

$$\sigma_{\ln GDP} = \sigma \left(\sum_{i=1}^{n} \lambda_i^2 \pm \left(\sum_{i=1}^{n} \lambda_i \right)^2 \right)^{1/2} = \tag{382}$$

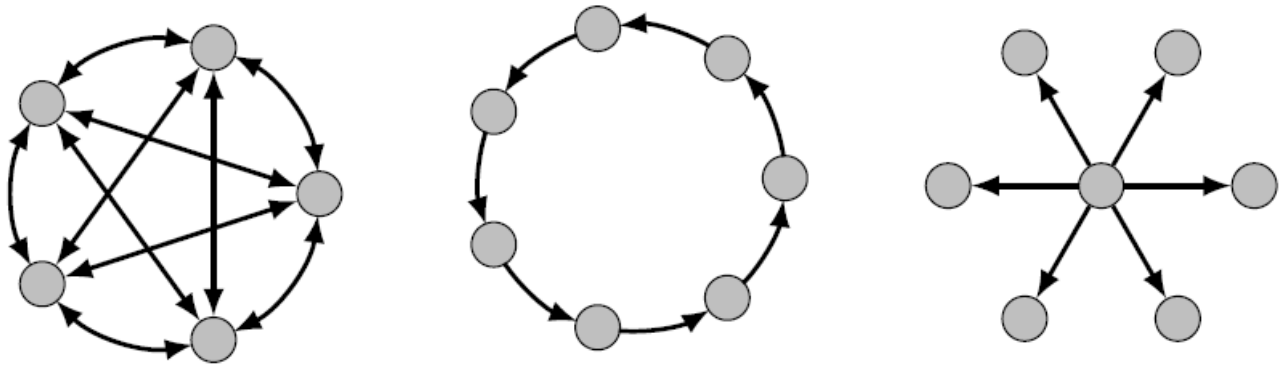
$$= \sigma \left(\frac{1}{n\alpha^2} + n \text{var}(\underline{\lambda}) \right)^{1/2} = \tag{383}$$

$$= \frac{\sigma}{\alpha\sqrt{n}} \left(1 + n^2 \alpha^2 \text{var}(\underline{\lambda}) \right)^{1/2}. \tag{384}$$

From this formulation, we note a number of important insights. First, if the Domar weights are symmetric, 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 15. 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 15: 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.

We conclude this chapter by discussing fragility and tail risk. We will mostly go through the simple model in Levine (2012). For a much richer model, check the beautiful (and hard) paper by Elliott, Golub and Leduc (2020). Levine (2012) adopts a straightforward and elegant approach to figuring out how reliable chains are. Note that this question is very salient these days. The model takes a strong stance on complementarities to characterize a clear upper bound of fragility. We assume that production is Leontief. We assume individuals are endowed with a unit of intermediate input. Intermediate goods are transformed at rate β into other intermediate inputs or a final good. We assume that, alternatively, agents can specialize in making intermediate input $j-1$ from intermediate input j at rate λ_j. We call a *k-production chain* for $k \geq 1$ a generalist turning inputs into final good and $j = 1...k-1$ specialists. The output of the chain is then given by

$$f(k) = \beta \prod_{j=1}^{k-1} \lambda_j. \tag{385}$$

Now we introduce the probability p that any individual fails. The goal here is to study the fragility of the chain, defined by its failure. It is immediate to see that the production structure implies that the failure of any individual in the chain will generate a failure of the whole chain. In this sense, the Leontief assumption implies an upper bound of fragility. The opposite bound would be given by an economy in which we frictionlessly shift between perfectly substitutable inputs. This is not the interesting bit of the paper, the real question is what the optimal chain length is. To figure this out, define $R(k)$ how reliable a k-chain is. Then we can define aggregate output as $y = f(k)R(k)$, and we pick the optimal k. Intuitively this will depend on how fast reliability decays with specialization and the returns to specialization.

Kremer (1993) uses a similar intuition to build a model of o-ring production, see also Demir, Fieler, Xu and Yang (2020) for a more recent application to production networks. The fundamental assumption made by Kremer (1993) is positive assortative matching between workers. A key implication of these strong complementarities implies large multipliers effects. Whereas if we have an unproductive worker in an otherwise productive workers chain, we lose a lot of output. The o-ring structure is an extreme case of this.

Another important element we have been silent on so far is the role of adjustment costs in networks. We basically assume that networks are exogenous (or as little endogenous as they can be since they are fully given by technology). More recent models of this problem look at network formation. An important feature of these networks is that they respond frictionlessly to changes. For example, suppose that some element of the input-output matrix a_{ij} changes, then the optimal input mix immediately adjusts. What happens instead if there is a fixed cost of forming or breaking a

link? The model will naturally have more inertia but also larger endogenous amplification. Suppose that a firm gets a very bad productivity draw, its price will spike up and, in a frictionless world, all other firms will basically stop buying the product. Suppose instead that there is a large fixed cost of stopping the purchase from such a firm (maybe because we tailored the production to be efficient with that input) then, firms will not drop the unproductive supplier. This will immediately generate a propagation of the shock in the network, increasing its multiplier on GDP. You can show that when goods are very strongly complementary or there are large fixed costs in adjusting links, the economy has a larger volatility. An example of such a model is Huneeus (2018).

The model in Elliott, Golub and Leduc (2020) looks at this problem but in the case in which firms can invest to make links more robust to shocks. They show that firms are concerned with disruptions and, therefore, source the same input from many firms. They also show that supply networks can be "fragile", meaning that small aggregate shocks can generate large drops in output.

In general, the problem of network formation and how it interplays with shocks has been the focus of a recent literature. A non-exhaustive list of papers is Lim (2018); Acemoglu and Azar (2020); Kopytov et al. (2022); Acemoglu and Tahbaz-Salehi (2023); Ferrari and Pesaresi (2024). Similarly, these models have also been applied to settings with price rigidities in Ghassibe (2021); Ferrari and Ghassibe (2024).

4.4 Oligopolistic Competition

In this section, we work out a simple version of oligopolistic competition with the goal to then embed it in our standard RBC model. The formulation used here is a version of the model in Atkeson and Burstein (2008), specifically used in Ferrari and Queirós (2024). We start from the production side as the demand side is somewhat standard. Assume there is a final good, the numeraire, which is a CES aggregate of I different industries $Y_t = \left(\sum_{i=1}^{I} y_{it}^{\rho}\right)^{1/\rho}$, where y_{it} is the quantity of industry $i \in [0,1]$, and $\sigma_I = 1/(1-\rho) > 1$ is the elasticity of substitution across industries. I is assumed to be large so that each individual industry has a negligible size in the economy. The output of each industry i is itself a CES composite of differentiated goods or varieties $y_{it} = \left(\sum_{j=1}^{n_{it}} y_{jit}^{\eta}\right)^{1/\eta}$, where n_{it} is the number of active firms in industry i at time t (to be determined endogenously) and $\sigma_G = 1/(1-\eta) > 1$ is the within-industry elasticity of substitution. Following Atkeson and Burstein (2008), we assume that goods are more easily substitutable within industries than across industries: $0 < \rho < \eta \leq 1$.

Given these assumptions, the inverse demand for each variety j in industry i is given by

$$p_{ijt} = \left(\frac{Y_t}{y_{it}}\right)^{1-\rho} \left(\frac{y_{it}}{y_{ijt}}\right)^{1-\eta}. \tag{386}$$

We assume that in every industry $i \in \{0, \ldots, I\}$ there is a maximum number of entrepreneurs $N \in \mathbb{N}$, so that $n_{it} \leq N$. Entrepreneurs j can produce their variety by combining capital k_{ijt} and

labour l_{ijt} through a Cobb-Douglas technology

$$y_{ijt} = \underbrace{e^{a_t}\,\gamma_{ij}}_{\tau_{ijt}} (k_{ijt})^{\alpha} (l_{ijt})^{1-\alpha}. \tag{387}$$

γ is permanently drawn from a distribution while a follows an AR(1).

Firms pay some sunk cost c_f to enter the market. Conditional on entry, firms play a static Cournot game, meaning that they simultaneously choose quantities taking competitors' output as given. The key difference with the previous model is that now we cannot neglect the effect of a firm's choices on the price index within the industry. Hence, after normalizing the price index of the final good to 1, a firm j solves

$$\begin{aligned} \max_{y_{ijt}} \quad & \left(p_{ijt} - \frac{\Theta_t}{\tau_{ijt}}\right) y_{ijt} \qquad \text{s.t.} \quad p_{ijt} = \left(\frac{Y_t}{y_{it}}\right)^{1-\rho} \left(\frac{y_{it}}{y_{ijt}}\right)^{1-\eta} \\ & \qquad\qquad\qquad\qquad\quad y_{it} = \left(\sum_{k=1}^{n_{it}} y_{ikt}^{\eta}\right)^{\frac{1}{\eta}}. \end{aligned} \tag{388}$$

Where Θ_t is the factor cost index once the firm has optimized the input sourcing problem by minimizing expenditure. We can write the first order condition of the problem before substituting in the constraints as

$$\frac{\partial \pi_{ijt}}{\partial y_{ijt}} = p_{ijt} - \frac{\Theta_t}{\tau_{ijt}} + y_{ijt} p'_{ijt} = 0 \tag{389}$$

From the constraints, we have that

$$p_{ijt} = Y_t^{1-\rho} \left(\sum_{k=1}^{n_{it}} y_{ikt}^{\eta}\right)^{\frac{\rho-\eta}{\eta}} y_{ijt}^{\eta-1}, \tag{390}$$

hence we can write p'_{ijt} as

$$p'_{ijt} = Y^{1-\rho} \left[(\eta-1) y_{ijt}^{\eta-2} \left(\sum_{k=1}^{n_{it}} y_{ikt}^{\eta}\right)^{\frac{\rho-\eta}{\eta}} + \frac{\rho-\eta}{\eta} \left(\sum_{k=1}^{n_{it}} y_{ikt}^{\eta}\right)^{\frac{\rho-\eta}{\eta}-1} \eta y_{ijt}^{\eta-1} y_{ijt}^{\eta-1}\right] \tag{391}$$

$$= \underbrace{Y_t^{1-\rho} \left(\sum_{k=1}^{n_{it}} y_{ikt}^{\eta}\right)^{\frac{\rho-\eta}{\eta}} y_{ijt}^{\eta-2}}_{p_{ijt} y_{ijt}^{-1}} \left[\eta - 1 + (\rho-\eta) \frac{y_{ijt}^{\eta}}{\sum_k y_{ikt}^{\eta}}\right] \tag{392}$$

$$= p_{ijt} y_{ijt}^{-1} \left[\eta - 1 + (\rho-\eta) \frac{y_{ijt}^{\eta}}{\sum_k y_{ikt}^{\eta}}\right]. \tag{393}$$

Next, note that we can write the market share of firm j in industry i, s_{ijt} as [38]

$$s_{ijt} = \frac{p_{ijt} y_{ijt}}{\sum_j p_{ijt} y_{ijt}} \tag{394}$$

$$= \frac{y_{ijt} \left(\frac{Y_t}{y_{it}}\right)^{1-\rho} \left(\frac{y_{it}}{y_{ijt}}\right)^{1-\eta}}{\sum_j \left(\frac{Y_t}{y_{it}}\right)^{1-\rho} \left(\frac{y_{it}}{y_{ijt}}\right)^{1-\eta} y_{ijt}} \tag{395}$$

$$= \frac{y_{ijt}^{\eta}}{\sum_j y_{ijt}^{\eta}}. \tag{396}$$

Hence

$$p'_{ijt} = p_{ijt} y_{ijt}^{-1} \left[\eta - 1 + (\rho - \eta) s_{ijt}\right]. \tag{397}$$

Substituting back into the first order condition, we obtain

$$p_{ijt} - \frac{\Theta_t}{\tau_{ijt}} + p_{ijt} \left[\eta - 1 + (\rho - \eta) s_{ijt}\right] = 0. \tag{398}$$

Hence, the solution to (388) yields a system of n_{it} non-linear equations in $\{p_{ijt}\}_{j=1}^{n_{it}}$ (one for each firm)

$$p_{ijt} = \underbrace{\frac{1}{\eta - (\eta - \rho)\, s_{ijt}}}_{\mu_{ijt}} \frac{\Theta_t}{\tau_{ijt}}, \tag{399}$$

where s_{ijt} is the market share of firm j and μ_{ijt} is the markup over the marginal cost Θ_t / τ_{ijt}.

Equation (399) establishes a positive relationship between market shares and markups. This happens because firms internalize the impact of size on the price they charge p_{ijt}; large firms end up restricting output disproportionately more (relative to productivity), thereby charging a high markup. Moreover, as equation (399) also highlights, market shares are themselves a positive function of RTFP $p_{ijt}\, \tau_{ijt}$. Our model thus features a positive association between revenue productivity, size, and markups. Therefore, a shock that generates larger productivity differences across firms will also lead to larger markup dispersion.

To conclude the description of the industry equilibrium, we need to determine the number of active firms n_{it}. To this end, let $\Pi(j, n_{it}, \mathcal{F}_{it}, X_t) \coloneqq (p_{it} - \Theta_t/\tau_{ijt})\, y_{ijt}$ denote the equilibrium profits of firm $j \leq n_{it}$ in industry i (gross of the fixed production cost) when there are n_{it} active firms, given a productivity distribution $\mathcal{F}_{it} \coloneqq \{\gamma_{i1}, \gamma_{i2}, \ldots\}$ and a vector of aggregate variables $X_t \coloneqq [a_t, Y_t, \Theta_t]$. The equilibrium number of firms must be such that (i) the profits of each active firm are not lower than the fixed cost c_i and (ii) if an additional firm were to enter, its

[38] See Appendix A.5 for additional derivations on the market shares in a CES economy.

profits would be lower than the fixed cost. Mathematically, an interior solution $n_{it}^* < N$ to the equilibrium number of firms must satisfy

$$\left[\Pi\left(n_{it}^*, n_{it}^*, \mathcal{F}_i, X_t\right) - c_f\right]\left[\Pi\left(n_{it}^* + 1, n_{it}^* + 1, \mathcal{F}_i, X_t\right) - c_f\right] \leq 0. \tag{400}$$

This is a free entry condition stating that the last firm in makes positive profits and the first firm out would make losses up entry. Note that this is a slackness condition (meaning that it is derived by combining two inequalities) because we assumed that $n_{it} \in \mathbb{N}$. If we had allowed $n_{it} \in \mathbb{R}$, the condition would just be that there exists a firm that is indifferent between being in or out of the market.

With this in the bag, we can take home a couple of points. First, the number of firms in a market is endogenous, hence we will have fluctuations at the extensive margin over the business cycle. Second, markups are not just a constant, they depend on i) the number of firms, ii) a firm's productivity (which moves over the cycle), and iii) the GE effects through the cost of inputs.

We can go further in solving the model. For simplicity, assume $\eta = 1$ so that we only have to care about one layer of aggregation. As this is a Cournot game, we can solve for the unique industry price. We can use equation (399) for generic firms k and m to get

$$s_k = (1-\rho)^{-1}\left[1 - \frac{\gamma_m}{\gamma_k}[1 - (1-\rho)s_m]\right] \tag{401}$$

By definition $\sum_k s_k = 1$, hence if we sum over the last equation

$$\frac{n - (1-\rho)}{\sum_k^n \gamma_k^{-1}} = \gamma_m[1 - (1-\rho)s_m] \tag{402}$$

Using this in the FOC of firm m we immediately get

$$p = \frac{\sum_k^n \gamma_k^{-1}}{n - (1-\rho)}\Theta \tag{403}$$

From here, we can already see what changes within this industry will do to the economy (if the industry is small so that we can take Θ as given). Increases in n will decrease the price. Increases in the dispersion of the productivity distribution (MPS) will, in general, increase the price. Decreasing ρ, which is tantamount to making the goods less substitutable, will increase the price. Note that all these comparative statics go through the markup, not the marginal cost, which we held fixed.

We can now start building the macro version of this to think about what the aggregates are going to look like as a function of the details at the micro level. Note that the following holds for any price-making behaviour (we go back to the Atkeson and Burstein (2008) model in a bit).

Firms with productivity γ_{ij} solve

$$\max \quad p_{ij}\,\gamma_{ij}F\left(k_{ij},l_{ij}\right)-R\,k_{ij}-W\,l_{ij}. \tag{404}$$

The first order condition with respect to capital is

$$\frac{dp_{ij}}{dy_{ij}}F_k\left(k_{ij},l_{ij}\right)\gamma_{ij}F\left(k_{ij},l_{ij}\right)+p_{ij}\gamma_{ij}F_k\left(k_{ij},l_{ij}\right)=R \tag{405}$$

$$p_{ij}F_k\left(k_{ij},l_{ij}\right)=\underbrace{\left[\frac{dp_{ij}}{dy_{ij}}\frac{y_{ij}}{p_{ij}}+1\right]^{-1}}_{\mu_{ij}}\frac{R}{\gamma_{ij}}, \tag{406}$$

where

$$\mu_{ij}=\left[\frac{dp_{ij}}{dy_{ij}}\frac{y_{ij}}{p_{ij}}+1\right]^{-1} \tag{407}$$

is the markup, which is equal to the inverse of the firm's factor share

$$\omega_{ij}=\frac{1}{\mu_{ij}}=\left[\frac{dp_{ij}}{dy_{ij}}\frac{y_{ij}}{p_{ij}}+1\right]. \tag{408}$$

Let Θ be the unit variable cost for a firm with unit productivity, or, equivalently, $\lambda_{ij}\gamma_{ij}=\Theta$, where λ_{ij} is the Lagrange multiplier of the cost minimization problem for a firm with productivity γ_{ij}. Then we have

$$p_{ij}F_k\left(k_{ij},l_{ij}\right)=\mu_{ij}\frac{R}{\gamma_{ij}} \tag{409}$$

$$\underbrace{\mu_{ij}\frac{\Theta}{\gamma_{ij}}}_{p_{ij}}F_k\left(k_{ij},l_{ij}\right)=\mu_{ij}\frac{R}{\gamma_{ij}} \tag{410}$$

$$\Theta F_k\left(k_{ij},l_{ij}\right)=R. \tag{411}$$

Because of CRS, $F_k\left(k_{ij},l_{ij}\right)$ only depends on the capital-labour ratio. Since all firms face the same factor prices, we have

$$\frac{k_{ij}}{l_{ij}}=\frac{K}{L} \quad \Rightarrow \quad F_k\left(k_{ij},l_{ij}\right)=F_k\left(K,L\right), \tag{412}$$

which implies

$$\Theta F_k\left(K,L\right)=R. \tag{413}$$

This result states that the rental rate of the economy can be written as the unit variable cost index Θ times the marginal product of capital.

Recall that at the firm level, the payments towards factors are given by the cost index of a single unit divided by productivity. The same holds at the aggregate level in this economy. Θ is the unit variable cost index, denoting Φ aggregate productivity and Ω the factor share (namely the complement of the profit share). We can show that $\Theta = \Omega\Phi$. Note that the following identity should hold in equilibrium

$$\int_{ij} \omega_{ij} p_{ij} \gamma_{ij} F\left(k_{ij}, l_{ij}\right) = \Omega\, \Phi F\left(K, L\right). \tag{414}$$

In words, the sum of factor payments from all firms should be equal to aggregate factor payments in the economy. Note that

$$\omega_{ij} p_{ij} \gamma_{ij} = \underbrace{\frac{1}{\mu_{ij}}}_{\omega_{ij}} \underbrace{\mu_{ij} \frac{\Theta}{\gamma_{ij}}}_{p_{ij}} \gamma_{ij} = \Theta. \tag{415}$$

Substituting we obtain

$$\int_{ij} \Theta F\left(k_{ij}, l_{ij}\right) = \Omega\, \Phi F\left(K, L\right). \tag{416}$$

Let η_{ij} denote the input share of each firm, namely the fraction of a given input used by firm ij out of the total amount of that input in the economy. Formally,

$$\eta_{ij} = \frac{k_{ij}}{K} = \frac{l_{ij}}{L} \tag{417}$$

Because of CRS, we have that

$$F\left(k_{ij}, l_{ij}\right) = \eta_{ij} F\left(K, L\right) \tag{418}$$

and hence

$$\int_{ij} \Theta\, \eta_{ij} F\left(K, L\right) = \Omega\, \Phi F\left(K, L\right) \tag{419}$$

$$\Theta\, F\left(K, L\right) \underbrace{\int_{ij} \eta_{ij}}_{=1} = \Omega\, \Phi F\left(K, L\right) \tag{420}$$

$$\Theta = \Omega\, \Phi \tag{421}$$

Combining the last equation with equation (413) we obtain

$$\Omega\, \Phi F_k\left(K, L\right) = R. \tag{422}$$

This tells us that the rental rate of the economy depends on the feature of the production function, a measure of aggregate productivity and a measure of the factor share (which is the inverse of the aggregate markup). We could do the same for labour and effectively build an aggregate production function in which aggregate productivity is a complicated function of individual productivities, market power and within-industry equilibria. To give you an example, in the Atkeson and Burstein (2008) economy, aggregate productivity is given by

$$\Phi\left(\mathbf{\Gamma}, \mathbf{N}_t\right) = \left[\sum_{i=1}^{I}\left(\sum_{j=1}^{n_{it}} \omega_{ijt}^{\eta}\right)^{\frac{\rho}{\eta}}\right]^{\frac{1}{\rho}} \left(\sum_{i=1}^{I}\sum_{j=1}^{n_{it}} \frac{\omega_{ijt}}{\tau_{ijt}}\right)^{-1}, \tag{423}$$

where

$$\omega_{ijt} \coloneqq \left[\sum_{k=1}^{n_{it}}\left(\frac{\mu_{ikt}}{\tau_{ikt}}\right)^{\frac{\eta}{1-\eta}}\right]^{\frac{\eta-\rho}{\eta}\frac{1}{1-\rho}} \left(\frac{\tau_{ijt}}{\mu_{ijt}}\right)^{\frac{1}{1-\eta}}. \tag{424}$$

Note two things here: i) if I, the number of industries, is not very large, changes in productivity of a single firm can change aggregate TFP; ii) however, we assumed that firms do not internalize such effect. A similar assumption is made in the recent paper by Burstein et al. (2020). Lastly, note that we have an endogenous number of firms n_{it}, which can respond to fluctuations over the business cycle.

4.4.1 Aggregation

Let $\mathbb{C}_t \coloneqq W_t\, L_t + R_t\, K_t$ represent aggregate variable costs. We can write the aggregate factor share $\Omega\left(\cdot\right) \coloneqq \mathbb{C}_t / Y_t$ as a function of individual markups and market shares[39]

$$\Omega\left(\mathbf{\Gamma}, \mathbf{N}_t\right) = \sum_{i=1}^{I}\sum_{j=1}^{n_{it}} s_{it}\; s_{ijt}\; \mu_{ijt}^{-1}. \tag{425}$$

Combining the previous equation with the first order condition in (399), we can also write the aggregate factor share as a negative function of all industry-level HHI of concentration

$$\Omega\left(\mathbf{\Gamma}, \mathbf{N}_t\right) = \sum_{i=1}^{I} s_{it}\; \left[\eta - \left(\eta - \rho\right)\; HHI_{it}\right]. \tag{426}$$

[39]Note that the aggregate factor share is equal to the inverse of the aggregate markup $\mu\left(\cdot\right) \coloneqq Y_t / \mathbb{C}_t$.

where $HHI_{it} \coloneqq \sum_{j=1}^{n_{it}} s_{ijt}^2$. The converse of this result is about the firm, sectoral and aggregate markups which we derive in detail. Start by noting that for a firm, the markup is given by

$$\mu_{ij} = \frac{p_{ij}}{\lambda_{ij}} \tag{427}$$

by definition, where λ_{ij} is the marginal cost. Next, we can define an industry-level markup as

$$\mu_i = \frac{p_i}{\lambda_i}, \tag{428}$$

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

$$\mu_i = \frac{p_i}{\sum_j \lambda_{ij} \frac{y_{ij}}{y_i}} = \left(\sum_j \lambda_{ij} \frac{y_{ij}}{p_i y_i} \right)^{-1} \tag{429}$$

$$= \left(\sum_j \frac{\lambda_{ij}}{p_{ij}} \frac{p_{ij} y_{ij}}{p_i y_i} \right)^{-1} = \left(\sum_j \frac{\lambda_{ij}}{\mu_{ij} \lambda_{ij}} s_{ij} \right)^{-1} = \left(\sum_j \mu_{ij}^{-1} s_{ij} \right)^{-1}. \tag{430}$$

Using the firm's markup, we obtain

$$\mu_i = \left(\sum_j (\eta + (\rho - \eta) s_{ij}) s_{ij} \right)^{-1} = \left(\eta \sum_j s_{ij} + (\rho - \eta) \sum_j s_{ij}^2 \right)^{-1} \tag{431}$$

$$= \left(\eta + (\rho - \eta) HHI_i \right)^{-1}, \tag{432}$$

where I have used the definition $HHI_i = \sum_j s_{ij}^2$. 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

$$\mu = \frac{P}{\lambda} = \left(\sum_i \lambda_i \frac{y_i}{PY} \right)^{-1} = \left(\sum_i \frac{\lambda_i}{p_i} \frac{p_i y_i}{PY} \right)^{-1} = \left(\sum_i \mu_i^{-1} s_i \right)^{-1}. \tag{433}$$

Using the industry-level markup derived above

$$\mu = \left(\eta \sum_i s_i + (\rho - \eta) \sum_i s_i HHI_i\right)^{-1} = \left(\eta + (\rho - \eta) \sum_i s_i HHI_i\right)^{-1}. \tag{434}$$

This result, which is identical to the findings of Grassi (2017) and Burstein et al. (2020), highlights two important relationships. First, industries with higher concentration have larger markups. When highly concentrated industries have large shares in the economy (large s_{it}), the economy's average markup is also high. This is important to understand how markups at the firm, industry and economy level move along the business cycle, which turns out to be a key statistic for amplification or dampening of shocks.

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.[40] 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, they 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

Digression on the Inefficiency of Markup Heterogeneity The Hsieh and Klenow (2009) result, as well as the outcomes of oligopolistic competition models, suggest that markup heterogeneity á la Edmond et al. (2018) is inefficient.

We can trace our steps towards this conclusion. First, we considered economies where firms face households with the same constant elasticity demand for each variety. We noted that markups in these settings arise because of the differentiation between goods. In general, we think of these markups as potentially inefficient as they restrict the quantity consumed to $q(p) < q(c)$ and, therefore, a lost opportunity to exploit gains from trade. We saw that if there are fixed costs to entry, these may not be inefficient as they represent quasi- rather than pure rents.

However, when markups are heterogeneous despite homogeneous elasticities in the preferences, we conclude that these are a symptom of inefficiency. These were driven by exogenous wedges

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

in Hsieh and Klenow (2009) and by firms exploiting their non-infinitesimal size in Atkeson and Burstein (2008).

There is, however, a completely different approach that would read the same data in a very different light. This is the search model in Menzio (2024).

Suppose there is a measure 1 of identical sellers, producing a homogeneous good at marginal cost c. The payoff from selling q units at price p is $q(p-c)$. On the other side of the market, there is a measure b of buyers for each seller. Each buyer demands a single unit of the good and enjoys a payoff of $u-p$, where u is the willingness to pay. The key element is that there are frictions in the market as each buyer only meets a subset of the sellers. Each buyer meets n sellers at random, where $n \sim Poisson(\lambda)$. Buyers meeting $n > 1$ sellers only buy from the cheapest option.

Clearly, no seller will post a price above u. The payoff at price p for sellers is

$$V(p) = \left[\sum_k b_k \pi_k(p)\right](p-c), \tag{435}$$

where $\pi_k(p)$ is the probability that a buyer who has met k other firms buys from this seller. From the Poisson assumption, b_k, the measure of buyers meeting the seller *and* k other competitors is given by b times the probability of a buyer meeting $k+1$ sellers:

$$b_k = b\frac{e^{-\lambda}\lambda^{k+1}}{(k+1)!}(k+1). \tag{436}$$

Similarly, we can write $\pi_k(p)$, the probability that between the seller and the k competitors, the seller is chosen. This is equivalent to the probability that the seller has the lowest price among the set of sellers the buyer has met. Denote F the equilibrium offered price distribution, then:

$$\pi_k(p) = (1-F(p))^k + \sum_{j=1}^{k} \frac{\chi(p)(1-F(p))^{k-j}}{j}, \tag{437}$$

where $\chi(p)$ is the fraction of buyers posting price p. The first term captures the probability that every other seller posted a price larger than p. The summation instead deals with ties. Suppose that the seller and one competitor both posted price p, then the buyer randomizes and buys from the seller with probability $1/2$. The summation then accounts for the probability that j competitors posted price p while $k-j$ posted prices above p. In this instance, the seller is chosen with probability $1/j$. We then sum over all possible values of j up to k.

In this setting, we can immediately establish a few things. First, much like in the Burdett and Judd (1983), we cannot have holes in the offered price distribution. Next, it has to be that the support of the offered price distribution $F(p)$ is an interval of the type $[p_l, p_h]$ with $p_h = u$. Third, since the sellers are identical, the only way in which we can sustain equilibrium price dispersion is

if, at all posted prices, the payoff is the same. First we can substitute b_k and π_k to get

$$V(p) = b\lambda e^{-\lambda F(p)}(p - c). \tag{438}$$

By the last observation, we must have that $V(p) = V^\star$, $\forall p \in [p_l, u]$. Evaluating the payoff at $p = u$, and noting that F(u)=1 since it is the end point of the support

$$V^\star = b\lambda e^{-\lambda}(u - c). \tag{439}$$

Since we must have that

$$V(p) = b\lambda e^{-\lambda F(p)}(p - c) = V^\star, \tag{440}$$

we can combine them and solve for $F(p)$:

$$F(p) = 1 - \frac{1}{\lambda} \log \left(\frac{u - c}{p - c} \right). \tag{441}$$

Noting that at $p = p_l$ we must have $F(p_l) = 0$ we obtain

$$p_l = c + e^{-\lambda(u-c)}, \tag{442}$$

which completes the characterization of the candidate equilibrium. For this equilibrium to be unique, we note that i) by construction $V(p) = V^\star$, $\forall p \in [p_l, p_h]$, and ii) $V(p') < V^\star$, $\forall p' \notin [p_l, p_h]$.

Now note that the equilibrium described above is efficient. The social surplus generated by each transaction is $u - c$ independently of who the buyer buys from. As long as trade happens, the market and the planner deliver the same outcome. A planner constrained by the same search frictions also delivers the same probability of trade (they cannot undo the frictions).

Finally, we can discuss markups. Rank the sellers according to their price. A seller in the x^{th} quantile of the price distribution has $F(p(x)) = x$ by definition. Hence it charges a price

$$p(x) = c + (u - c)e^{-\lambda(1-x)} \tag{443}$$

and, therefore, charges a markup

$$\mu(x) = 1 + \left(\frac{u}{c} - 1 \right) e^{-\lambda(1-x)}. \tag{444}$$

This is easily interpretable: $u/c - 1$ is the markup that a monopolist would charge. The term $e^{-\lambda(1-x)}$ encapsulates the competitive threat of competitors ranked below quantile x. This term is 1 for the firm charging the largest price while strictly smaller than 1 for everyone else. The rate at which markups decline along the price distribution is governed by the extent of search frictions λ.

We have established that, despite the equilibrium being efficient, we have positive and heterogeneous markups. This should be surprising in a context in which firms are homogeneous in productivity and producing a perfect substitute good. Why are markups positive and heterogeneous? First, markups are positive because firms meet some *captive* buyers, meaning buyers that did not meet any other firm. The seller can then extract surplus from them via positive markups. Next, markups have to be heterogeneous because prices have to be heterogeneous. The reason being that the distribution cannot have mass points: the whole support needs to be filled with mass. Otherwise, there would be profitable deviations. The general formula based on demand elasticities still holds, so we can recast our problem in those terms by noting that at price $p \in [p_l, p_h]$, firms face a demand equal to

$$q(p) = b\lambda e^{-\lambda F(p)}, \tag{445}$$

hence the elasticity is $\epsilon(p) = \lambda f(p)p$. Note that this elasticity is an equilibrium object since it depends on the equilibrium offered price distribution. Next, note that the price elasticity decreases in the price. Something we discussed earlier on under the name of "Marshall Second Law of Demand". Finally, markups decline in the intensity of competition λ. When λ is large, fewer customers are *captive*. Hence firms charge smaller markups.

Suppose you observed data on heterogeneous markups. Depending on whether you look at this data through the lenses of a CES oligopolistic framework or through this search-theoretic one, you draw very different conclusions. In one case, markup heterogeneity is inefficient; in the other, it is efficient.

End of Digression

4.4.2 Imperfect Competition and the Business Cycle

We start the discussion on the implication of imperfect competition on the business cycle by thinking about a hypothetical economy in which entry and exit are endogenous, and markups are not constant. Two ways to get this economy are using CES with a finite number of firms (see Jaimovich, 2007) or by oligopoly models with endogenous N, like Atkeson and Burstein (2008).

The intuition is common to both models. Suppose an economy is populated by monopolies and hit by a positive aggregate shock. Firms enter the market, and, if we were in a stochastic version of Hopenhayn (1992), we would conclude that entry has some positive effect on output on top of the direct technology effect. In this context, we get an extra kick from entrants, reducing incumbents' market power. Suppose after the shock, the economy is populated by duopolies, and the old monopolists' markups shrink as their market shares go down. This generates a further increase in output that we would not have with constant markups.

[40] Note that with finite N, even with CES preferences, markups are not constant, they depend negatively on N.

This also implies that such an amplification mechanism goes through markups being countercyclical, as net entry is procyclical.

Jaimovich and Floetotto (2008) study this in the context of the Jaimovich (2007) economy. They argue that this interaction between entry and competition actually gives us a lot of amplification. They show that depending on how large the steady state markup is, we get between 60 and 160% more amplification, as measured by the ratio between the standard deviation of output and the shock in the basic RBC model vs the endogenous entry/exit model.

This channel, empirically, shows up in endogenous movements of aggregate TFP. We will discuss this in detail further down, but when economies have aggregate production functions which are not additive in the firms' production function, then aggregate TFP can respond endogenously to movements in the exogenous aggregate shock. Markups are an obvious example, another one is reallocation. If, after a shock, activity is reallocated from low to high-productivity firms, we will get further amplification.

Another example of this channel, taken from Ferrari and Queirós (2024), is the following graph of the economy's response to a negative aggregate TFP shock.

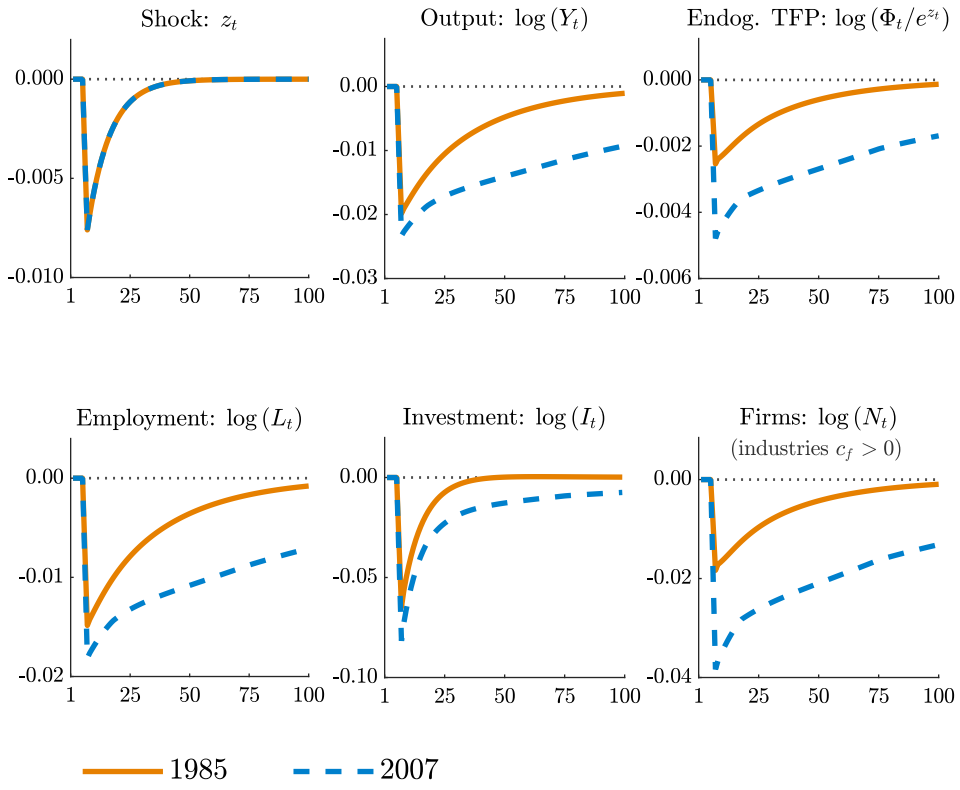


Figure 16: Impulse Response to Negative TFP shock.

The two IRFs shown in the figure correspond to an economy calibrated to match the US 1985 markup distribution and one for the 2007 distribution. The graphs show that there is significantly more endogenous amplification in 2007 when the underlying firm distribution is much more heterogeneous. The bottom right panel shows that this is driven by a more pronounced

response of the extensive margin. In turn, this depresses output further as incumbents' market power increases. As there is more market power, the return to investment and, therefore, investment itself is lower. This implies a slower recovery as the capital stock takes more time to go back to steady-state levels.

Digression on Markup Cyclicality with Granular Firms This discussion holds true in models with countercyclical markups (typically because of procyclical net entry) when the business cycle is driven by aggregate shocks. Suppose instead that we write down a granular version of the Atkeson and Burstein (2008) economy without entry and exit. This is the model in Burstein et al. (2020). Here the business cycle is driven productivity shocks to large firms. Suppose that a large firm in an industry gets a positive shock. It will certainly increase its market share and its markup. Given our previous aggregation results, the industry markup increases because i) the high-markup firm is getting bigger as a share of the industry; ii) the firm is increasing its markup. What happens to industry output? If the firm increases its markup but still reduces its price (because its marginal cost decreased), then the industry price will decline, and industry output will expand. Therefore the industry markup is procyclical relative to industry output. Aggregating further, as the industry is expanding as a share of the economy (recall that these are now granular industries, so they move aggregates), then we have that the economy markup is increasing because i) the industry markup is increasing; ii) the industry getting a bigger share of the economy than before.

In summary, the granular version of these models, as in Burstein et al. (2020), can generate the opposite markup cyclicality prediction. Note that this implies that when the economy is booming, output increases by less than it would in a constant markup economy. This is because as market power is procyclical, part of the productivity increase turns into higher rents (through higher prices) rather than higher quantities.

End of Digression

In summary, you can think of these channels as operating as endogenous TFP in an economy that admits an aggregate production function representation of the type

$$Y = A(Y)F(K, L), \tag{446}$$

where $F(K, L)$ is the CRS production function of firms, and we are loading all the micro interactions in $A(Y)$. If markups are countercyclical, they effectively imply that $A' > 0$ and that the economy will have endogenous amplification (much like in the DRS case we looked at before). If instead, market power operates as a shock dampener, we will have that $A' < 0$.

In the next section, we dig a bit deeper into the potential of this representation to think about coordination and strategic complementarity in macro models.

5 Coordination and Multiplicity in Macro Models

So far we have worked through a set of models in which, by appropriate assumption, we made everything well-behaved and made solutions typically unique.

In this section, we consider a departure from this case. We briefly discuss simple settings in which multiplicity may arise, both in terms of equilibria and, more interestingly, of steady states.

Multiplicity is often an undesirable feature of models. Mostly because it implies that things are indeterminate or strongly depend on starting conditions. Nonetheless, sometimes it's a useful thing to have to match data. For example, Figure 17 shows the deviations from trend of some key macro moments. It is virtually impossible to generate this kind of ergodic distribution without some sort of multiplicity.

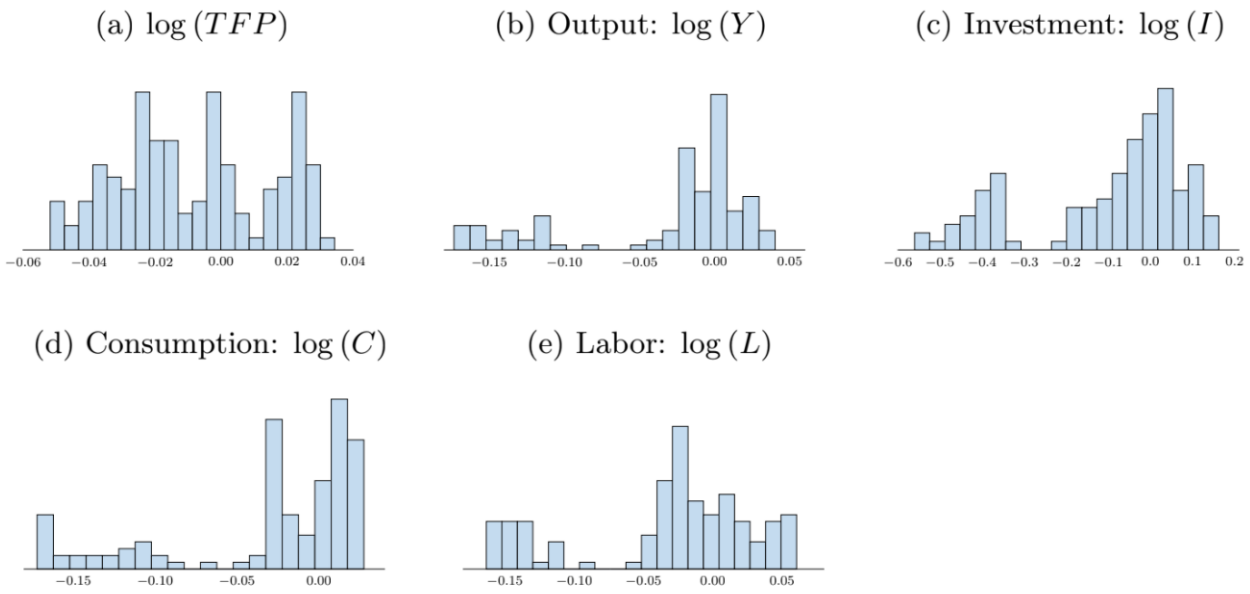


Figure 17: Deviations from trend in the data form Schaal and Taschereau-Dumouchel (2018)

5.1 Equilibria Multiplicity in the Entry Game[41]

We start by discussing equilibria multiplicity. The first rule of multiplicity is that there must be some complementarity somewhere in the model. In the entry, game suppose we need to figure out how many firms there are in the economy. We are used to thinking of free entry conditions as stating that the expected value of entry equals the cost of entry. Fixing notation, suppose firms can either produce one unit of the consumption good or zero (not entering). We assume that to enter, firms pay a fixed cost f.

Denote $p(N)$ the price at which the good is sold and some constant marginal cost c. Importantly, assume that the markup $\mu(N)$ is not strictly decreasing in N like we typically do. Assume there

[41] This section is inspired by insights from ongoing work with Matteo Escudé.

is some $N^\star$ such that for $N > N^\star$, $\mu'(N) > 0$ until $N^{\star\star}$, after which it's decreasing again.

This implies that there is a part of the problem in which further entry makes every incumbent more profitable. In such a case, it can be that there is multiplicity of equilibria, as in Figure 18.

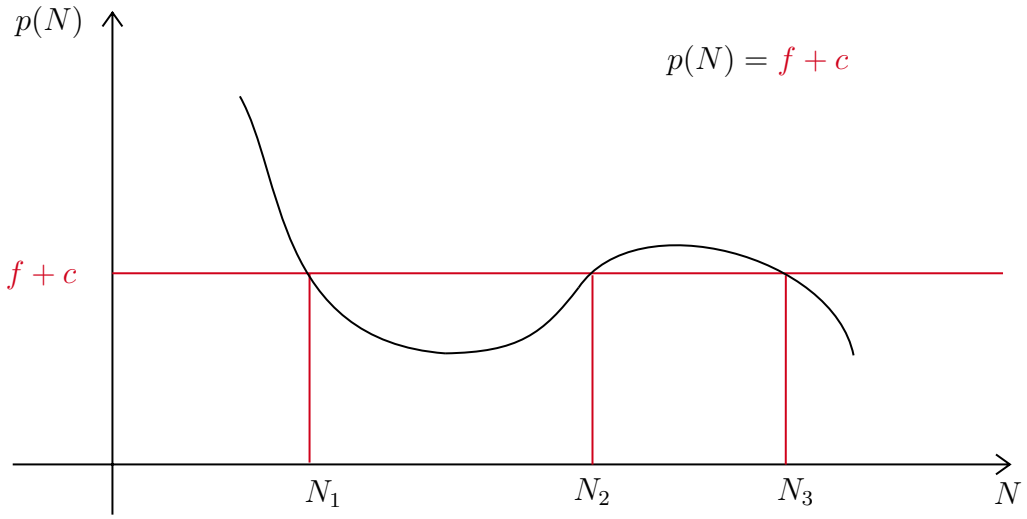


Figure 18: Entry Multiplicity loosely based on Ferrari and Escudé (2020).

In this economy, an equilibrium is a solution to $p(N) = f + c$. Since $p(N)$ is not monotone, there can be multiple solutions and equilibria. Importantly, not all equilibria are the same.

A way to discuss this is to think deeper about the fundamental equilibrium condition of free entry.

Digression on Free Entry Conditions A free entry condition is typically written in the form $f = \mathbb{E}(V^E)$ where f is the cost of entry and V^E is the value of entry. To simplify things, suppose the economy is static and deterministic, then $V^E = \pi$ and typically π is a function of N. So the free entry condition says $f = \pi(N)$. If we look at the real world, it turns out that N is constrained to be a natural number (hard to have 1.5 firms). If that is true, we cannot be quite sure that the free entry condition holds exactly. We get around that by using the slackness condition we used when looking at discrete oligopolies: $(\pi(N) - f)(\pi(N+1) - f) < 0$, with the idea that, if $\pi' < 0$, this is equivalent to saying $\pi(N) > f$ and $\pi(N+1) < f$. You can already see that this becomes a bit odd if there are points in which $\pi' > 0$. If that's the case, we get that actually with $\pi(N) - f < 0$ and $\pi(N+1) - f > 0$, the slackness condition is verified, and we got ourselves an equilibrium. This is, however, an equilibrium in which adding more firms makes all incumbents better off and removing firms makes all incumbents worse off. By definition, this is not stable. Very loosely, suppose we add or remove a firm, do we go back to this equilibrium? No, we actually end up in one of the neighbouring ones.

This is to point out that when we impose free entry conditions, we need to be careful about implicit assumptions on the derivative of profits. If there is no entry spillover or complementarity,

this is typically given by the pro-competitive effects of entry and decreasing returns, such that $\pi' < 0$ anyway.

End of Digression

In this sense, looking at the fact that free entry holds can be not enough to study equilibria. The "way" in which it holds is much more informative. In particular, when we have equilibria in which the free entry "holds from below", these will typically be unstable in the trembling hand sense.

Now the natural question is why can it be useful to think of multiple equilibria. It is useful to think about the role of policy in determining which world we live in. Importantly, if we can Pareto-rank equilibria, we can determine which world the planner would make us live in. These models, particularly in macro, are put under the umbrella of models with *sunspot equilibria*. Loosely speaking, these are models in which an exogenous change in beliefs can make the economy move from one equilibrium to another. Examples of these include models of bank runs, bubbles, coordination between households and firms, and debt and currency crises. In all these settings, active policy can be viewed as an equilibrium selection device.

5.2 Steady State Multiplicity[42]

In this section, we study what we can do if we get rid of all the assumptions that deliver us uniqueness. As in the case of multiple equilibria, we need to be quite careful in doing comparative statics. The general idea here is that we need some complementarity to generate multiplicity. Importantly, the complementarity needs to involve capital so that the law of motion is non-monotone. More specifically, recall that a steady state can be thought of as a solution to the capital fixed point problem

$$K' = G(K) = K \tag{447}$$

Where we are putting the whole model into the function $G(K)$. Our typical assumptions on how $G(\cdot)$ looks like, in particular strict concavity, ensure uniqueness of the fixed point. In this section, we write a very general economy in which such strict concavity is not guaranteed, and therefore we can have multiple fixed points.

To continue the theme of the role of firms in macro, we write an economy in which the demand side is very simple and standard, while the production side features possible complicated interactions between firms.

The economy consists of a representative household, a final good, a set of industries, and a set of potential producers. The household supplies labour and capital elastically, earning rental and

[42]This section is heavily inspired by work I have been doing with Francisco Queirós. For an extended version of this discussion, see Ferrari and Queirós (2024).

wage rates r_t and w_t. All active firms need to pay a fixed cost in units of the final good to produce. They purchase factors of production in competitive factor markets but imperfectly compete in final product markets. We assume that firms produce with a CRS production function with decreasing returns on each input

$$F(L, K) = AL \cdot f(k).$$

Where k is capital per worker. The technology set is further described by a $(I \times N)$ matrix of Hicks-neutral idiosyncratic productivities Π and a $(I \times N)$ matrix of fixed costs $\mathcal{C}$. We use $I \in \mathbb{N}^+$ to denote the number of industry types and $N \in \mathbb{N}^+$ the maximum number of firms per industry. The elements of each row of Π appear in decreasing order so that $\pi_{ij} \geq \pi_{ij+1}$ (i.e. more productive firms are ranked first).[43] To simplify the exposition, we assume that firms have identical fixed costs $c_{ij} = c$ and that, within an industry, firms enter sequentially in decreasing order of productivity.[44]

Call n the $(I \times 1)$ vector containing the number of firms in each industry. Finally, note that the technology set consists of

$$\Lambda = [\,\Pi\,,\; c\,].$$

We now define the equilibrium in this economy.

Definition 3 (Equilibrium)**.** *An equilibrium in this economy is a set of policies such that i) all agents optimize; ii) all active firms make no loss; iii) inactive firms would make a loss upon entry and iv) all markets clear.*

We restrict our attention to cases in which there exists a subset of the primitives' space featuring a unique equilibrium and an arbitrary number of steady states. We further assume that the economy admits an aggregate production function.

Definition 4 (Aggregate Production Function)**.** *An aggregate production function of the economy is given by*

$$F(\Lambda, n, k) = \Phi(\Lambda, n)\, F(K, L), \tag{448}$$

where Λ is the technology set (as defined above), n is the mass of active firms, and $\Phi(\Lambda, n)$ is aggregate TFP. Finally, $F(K, L)$ is the firm-specific CRS production function.

Note that $\Phi(\Lambda, n)$ will reflect two terms: the (weighted) average productivity of each technology π_{ij} and love for variety. These are discussed in greater detail below. From the aggregate production

[43] These assumptions are with no loss of generality. If an industry can only have a maximum of $M < N$ firms, this can be modelled as $\pi_{ij} = 0 \; \forall j \geq M + 1$.

[44] Therefore, if firm π_{ij} is active, all firms k with $\pi_{ik} > \pi_{ij}$ are also active.

function, it is possible to characterize the economy's inverse demand for capital. Let Ω denote the aggregate factor share (i.e. the ratio of total labour and capital payments over gross output Y). Then the equilibrium rental rate is given by

$$r(\Lambda, n, k) = A\Omega(\Lambda, n)\Phi(\Lambda, n)\, F_K(K, L(\Lambda, n, K)). \tag{449}$$

It will be convenient to define $g(\Lambda, n) \equiv \Omega(\Lambda, n)\Phi(\Lambda, n)$. Furthermore, the mass of active firms shall also be treated as an equilibrium outcome, in particular $n(\Lambda, K)$.

Let r^{ss} be the steady-state rental rate, from the utility maximization problem of the representative household.[45] Therefore the existence of multiple steady states in this economy boils down to the following equation having multiple solutions

$$r^{ss} = Ag(\Lambda, n(\Lambda, K))\, F_K(K, L(\Lambda, n, K)). \tag{450}$$

We assume that the following Inada conditions hold

$$\lim_{K\to 0} r\,(\Lambda, K) = \infty \quad \text{and} \quad \lim_{K\to\infty} r\,(\Lambda, K) = 0.$$

This trivially implies that the economy features an odd number of steady states.[46]

With all this machinery, we are now ready to think about what all the possible reasons such that 450 can admit multiple solutions. All the objects in 449 have clear economic interpretations which allows us to think through the forces behind any multiplicity.

5.2.1 Taxonomy of Endogenous TFP

We start by noting that a necessary condition for multiplicity of steady states is that $\exists K^\star$: $r_K(K^\star) > 0$. It follows that the necessary condition can be rewritten as
$\exists K^\star : (\partial/\partial K)\, \Omega(\Lambda, n(\Lambda, K^\star))\Phi(\Lambda, n(\Lambda, K^\star))F_K(K^\star, L(\Lambda, n, K^\star)) > 0$. There are four main mechanisms underlying the possible locally increasing returns to capital.

[45] Assuming that the representative households have time-separable preferences, a discount factor β and a constant depreciation rate δ, the steady-state rental rate is equal to $r^{ss} = \beta^{-1} - (1 - \delta)$.

[46] We assume away irregular steady states.

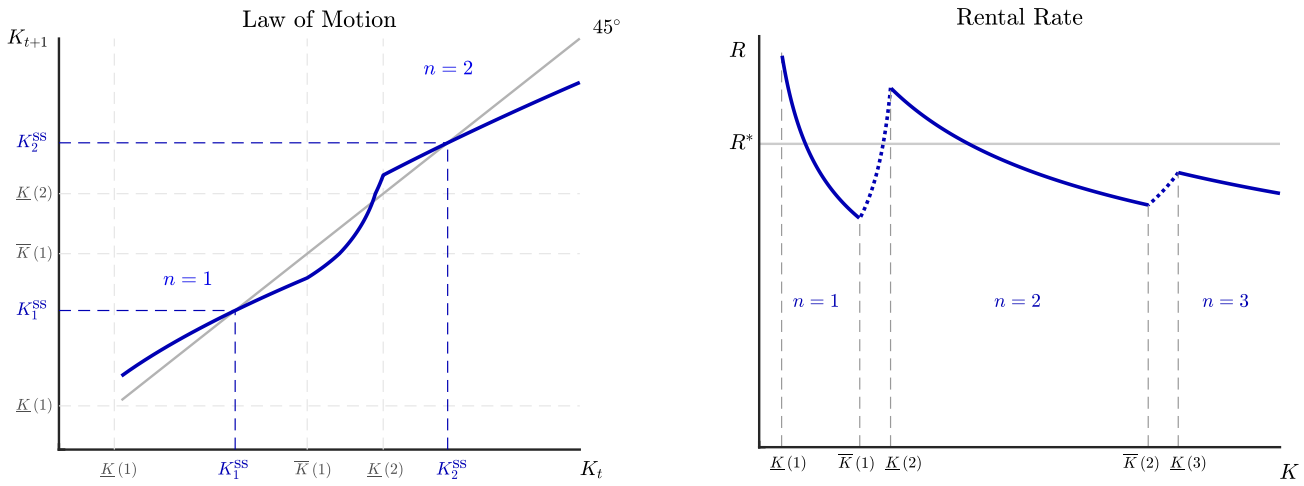


Figure 19: Steady State Multiplicity from Ferrari and Queirós (2024).

Average Firm TFP In an economy with heterogeneous technologies and no love for variety, aggregate TFP can be written as a weighted average of firm-level productivities. The weights will depend on market shares. Average firm-level TFP can be increasing in k if a larger capital favours a reallocation towards more productive types.

Love for Variety In models with product differentiation, aggregate TFP typically increases in the number of available varieties. This reflects the fact that utility/welfare are themselves increasing in the number of available goods (i.e. there is *love for variety*). Take for simplicity an economy where firms operate with the same level of productivity $\pi_{ij} = \pi$, but with possibly different fixed costs c_{ij}. Each firm produces a differentiated good. A larger capital stock k can increase the incentives for the entry of new firms/goods, thereby making Φ (weakly) increasing in k. Examples of papers highlighting this channel as a source of multiple equilibria/steady-states include Schaal and Taschereau-Dumouchel (2019).[47]

Market Power In models featuring imperfect competition and variable markups, changes in the number of active players can have an impact on the distribution of income across factors of production and oligopoly rents. Take, for example, an economy where firms have identical fixed costs $c_{ij} = c$ but possibly different productivities π_{ij} (this is the model we studied in the previous section). Assume further that firms enter sequentially in reverse order of productivity. If profit levels are increasing in the aggregate capital stock k (for a given set of players), a larger capital stock will result in a larger number of firms and lower markups. Lower markups, in turn, translate in a higher factor share Ω. This can establish a positive relationship between Ω and k. Examples

[47] Without relying on multiple equilibria or multiple steady-states, Cooper and John (2000) and Bilbiie, Ghironi and Melitz (2012) show that a combination of imperfect competition with endogenous entry can generate endogenous amplification and persistence of aggregate fluctuations.

of papers highlighting this channel as a source of multiple equilibria/steady-states include Pagano (1990), Chatterjee, Cooper and Ravikumar (1993), Galí and Zilibotti (1995) and Jaimovich (2007).

Endogenous labour Supply An elastic labour supply is another force that can counterbalance decreasing returns to capital. Note, however, that this channel alone is not enough to make $F_K(\cdot)$ increasing in K: if labour supply is infinitely elastic, $F_L(\cdot)$ and $F_K(\cdot)$ will be constant; but for finite elasticities, $F_K(\cdot)$ needs to be decreasing in K.[48] However, if some of the other channels described above (e.g. higher TFP or lower markups) are active and have a positive effect on the equilibrium wage, endogenous labour supply will provide an amplification mechanism to such a channel. To sum up, an elastic labour supply provides an additional mechanism that can counterbalance decreasing returns but is not enough to make $F_K(K, L)$ increasing in K.

Bonus: Increasing Returns In this economy I have assumed that the production function and, therefore, the aggregate production function have technological decreasing returns. Clearly, I can generate multiplicity if I assume that technological returns are increasing. This is, however, not an *endogenous* force like the other ones described above.

For an extended discussion of steady-state multiplicity and its practical uses, you can check Ferrari and Queirós (2024). In that paper, we take a specific model of oligopolistic competition and embed it in an RBC setting. We apply the insights in this section to the US economy and the aftermath of the Great Recession.

A graphical example of how thinking about multiplicity can be useful is in Figure 20. In these graphs, two economies are hit with the same aggregate shock. One economy behaves like a standard macro model with endogenous amplification through the extensive margin and market power. The second one, on the other hand, changes steady state. You can tell because the response features a permanent deviation from the shock levels. In this specific case, the difference between the two economies is given by a higher level of firm heterogeneity in the 2007 model. After the negative shock, so many firms exit that the economy does not revert to the old steady state. It converges to a new one with lower capital, competition and output.

[48] Since $F(K, L)$ is homogeneous of degree one, its partial derivatives are homogeneous of degree zero. This implies that they are solely functions of the capital-labour ratio K/L: $F_L(K, L)$ increases in K/L, whereas $F_K(K, L)$ decreases in K/L. Therefore, $F_K(K, L)$ cannot increase when $F_L(K, L)$ increases (and vice-versa). It is, however, possible that both $F_K(K, L)$ and $F_L(K, L)$ are constant. This happens if labour supply is infinitely elastic and the economy exhibits constant K/L.

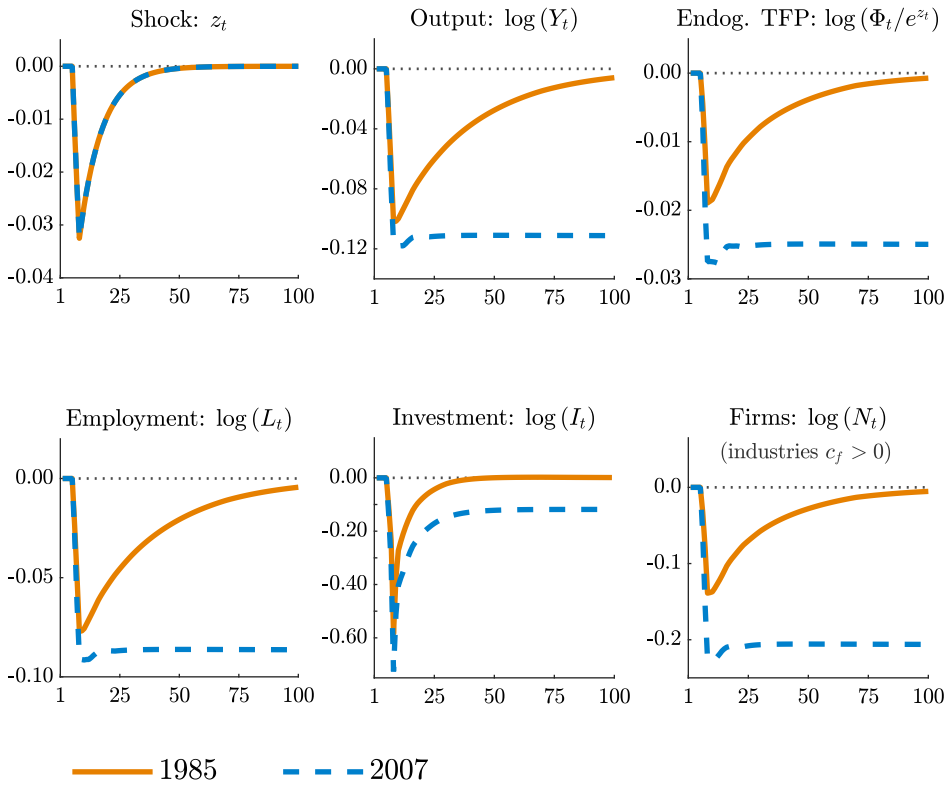


Figure 20: Impulse Response to Negative TFP shock

References

Acemoglu, Daron and Alireza Tahbaz-Salehi, "The Macroeconomics of Supply Chain Disruptions," *NBER Working Paper*, 2023, *27565.*

_ **and Pablo D Azar**, "Endogenous production networks," *Econometrica*, 2020, *88* (1), 33–82.

_ **, 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.

Adda, Jerome and Russell Cooper, *Dynamic economics: quantitative methods and applications*, MIT press, 2003.

Albrecht, James, "Search theory: The 2010 Nobel memorial prize in economic sciences," *Scandinavian Journal of Economics*, 2011, *113* (2), 237–259.

Alessandria, George, Joseph P Kaboski, and Virgiliu Midrigan, "US trade and inventory dynamics," *American Economic Review*, 2011, *101* (3), 303–307.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-market, trade costs, and international relative prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

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

Bajo, Igli, Frederik H. Bennhoff, and Alessandro Ferrari, "Not-So-Cleansing Recessions," *Working paper*, 2024.

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

_ **and** _ , "Entry vs. Rents," Technical Report, National Bureau of Economic Research 2020.

Barro, Robert J, "A theory of monopolistic price adjustment," *The Review of Economic Studies*, 1972, *39* (1), 17–26.

Basu, Susanto, "Are price-cost markups rising in the United States? A discussion of the evidence," *Journal of Economic Perspectives*, 2019, *33* (3), 3–22.

Bello, Salvatore Lo and Lorenzo Pesaresi, "Equilibrium Effects of the Minimum Wage: The Role of Product Market Power," *Working Paper*, 2024.

Benassy, Jean-Pascal, "Taste for variety and optimum production patterns in monopolistic competition," *Economics Letters*, 1996, *52* (1), 41–47.

Berger, David, Kyle Herkenhoff, and Simon Mongey, "Labor market power," *American Economic Review*, 2022, *112* (4), 1147–1193.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz, "Endogenous Entry, Product Variety, and Business Cycles," *Journal of Political Economy*, 2012, *120* (2), 304–345.

Bils, Mark and James A Kahn, "What inventory behavior tells us about business cycles," *American Economic Review*, 2000, *90* (3), 458–481.

Blinder, Alan S and Louis J Maccini, "The resurgence of inventory research: what have we learned?," *Journal of Economic Surveys*, 1991, *5* (4), 291–328.

Boehm, Christoph E and Nitya Pandalai-Nayar, "Convex supply curves," *American Economic Review*, 2022, *112* (12), 3941–69.

Burdett, Kenneth and Dale T Mortensen, "Wage differentials, employer size, and unemployment," *International Economic Review*, 1998, pp. 257–273.

_ **and Kenneth L Judd**, "Equilibrium price dispersion," *Econometrica: Journal of the Econometric Society*, 1983, pp. 955–969.

Burnside, Craig and Martin Eichenbaum, "Factor-hoarding and the propagation of business-cycle shocks," *The American Economic Review*, 1996, *86* (5), 1154.

Burstein, Ariel, Vasco M Carvalho, and Basile Grassi, "Bottom-up markup fluctuations," Technical Report, National Bureau of Economic Research 2020.

Caballero, Ricardo J and Mohamad L Hammour, "The Cleansing Effect of Recessions," *American Economic Review*, 1994, *84* (5), 1350–1368.

Cahuc, Pierre, Stéphane Carcillo, and André Zylberberg, *Labor economics*, MIT press, 2014.

Calvo, Guillermo A, "Staggered prices in a utility-maximizing framework," *Journal of monetary Economics*, 1983, *12* (3), 383–398.

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

_ **and Basile Grassi**, "Large firm dynamics and the business cycle," *American Economic Review*, 2019, *109* (4), 1375–1425.

Chari, Varadarajan V, Patrick J Kehoe, and Ellen R McGrattan, "Business cycle accounting," *Econometrica*, 2007, *75* (3), 781–836.

Chatterjee, Satyajit, Russell Cooper, and B Ravikumar, "Strategic Complementarity in Business Formation: Aggregate Fluctuations and Sunspot Equilibria," *The Review of Economic Studies*, 1993, *60* (4), 795–811.

Clementi, Gian Luca and Berardino Palazzo, "Entry, exit, firm dynamics, and aggregate fluctuations," *American Economic Journal: Macroeconomics*, 2016, *8* (3), 1–41.

Cooper, Russell and Andrew John, "Imperfect Competition and Macroeconomics : Theory and Quantitative Implications," *Cahiers d'Économie Politique*, 2000, *37* (1), 289–328.

Cooper, Russell W and John C Haltiwanger, "On the nature of capital adjustment costs," *The Review of Economic Studies*, 2006, *73* (3), 611–633.

Deb, Shubhdeep, Jan Eeckhout, Aseem Patel, and Lawrence Warren, "What Drives Wage Stagnation: Monopsony or Monopoly?," *Journal of the European Economic Association*, 2022, *20* (6), 2181–2225.

Demir, Banu, Ana Cecilia Fieler, Daniel Yi Xu, and Kelly Kaili Yang, "O-Ring Production Networks," Technical Report, Technical Report 2018. Unpublished manuscript 2020.

Dhingra, Swati and John Morrow, "Monopolistic competition and optimum product diversity under firm heterogeneity," *Journal of Political Economy*, 2019, *127* (1), 196–232.

Dixit, AK and J Stiglitz, "Monopolistic competition and optimum product diversity, University of Warwick," *Economic Research Paper*, 1975, *64*.

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

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.

Elliott, Matthew, Benjamin Golub, and Matthew V Leduc, "Supply network formation and fragility," *Available at SSRN 3525459*, 2020.

Ethier, Wilfred J, "National and international returns to scale in the modern theory of international trade," *The American Economic Review*, 1982, *72* (3), 389–405.

Ferrari, Alessandro, "Inventories, demand shocks propagation and amplification in supply chains," *arXiv preprint arXiv:2205.03862*, 2024.

_ **and Francisco Queirós**, "Firm Heterogeneity, Market Power and Macroeconomic Fragility," *Working Paper*, 2024.

_ **and Lorenzo Pesaresi**, "Specialization, Complexity & Resilience in Supply Chains," *Working Paper*, 2024.

_ **and Matteo Escudé**, "Trade Policy Uncertainty with Irreversible Investment," *mimeo*, 2020.

_ **and Mishel Ghassibe**, "A Disaggregated Economy with Optimal Pricing Decisions," *Working paper*, 2024.

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.

Galí, Jordi and Fabrizio Zilibotti, "Endogenous Growth and Poverty Traps in a Cournotian Model," *Annals of Economics and Statistics*, 1995, *37-38*, 197–213.

Ghassibe, Mishel, "Endogenous production networks and non-linear monetary transmission," *Working paper*, 2021.

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.

Golosov, Mikhail and Robert E Lucas Jr, "Menu costs and Phillips curves," *Journal of Political Economy*, 2007, *115* (2), 171–199.

Grassi, Basile, "Io in io: Size, industrial organization, and the input-output network make a firm structurally important," *Work. Pap., Bocconi Univ., Milan, Italy*, 2017.

Hagedorn, Marcus and Iourii Manovskii, "The cyclical behavior of equilibrium unemployment and vacancies revisited," *American Economic Review*, 2008, *98* (4), 1692–1706.

Hall, Robert E, "Employment fluctuations with equilibrium wage stickiness," *American economic review*, 2005, *95* (1), 50–65.

_ , "New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy," Technical Report, National Bureau of Economic Research 2018.

Hansen, Gary D, "Indivisible labor and the business cycle," *Journal of monetary Economics*, 1985, *16* (3), 309–327.

Holt, Charles C, Franco Modigliani, John Muth, and Herbert Simon, *Planning Production, Inventories, and Work Force.* 1960.

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

Hosios, Arthur J, "On the efficiency of matching and related models of search and unemployment," *The Review of Economic Studies*, 1990, *57* (2), 279–298.

Houthakker, Hendrik, "The Pareto Distribution and the Cobb-Douglas Production Function in Activity Analysis," *Review of Economic Studies*, 1955, *23* (1), 27–31.

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.

Huneeus, Federico, "Production network dynamics and the propagation of shocks," *Princeton University*, 2018.

Jaimovich, Nir, "Firm dynamics and markup variations: Implications for sunspot equilibria and endogenous economic fluctuations," *Journal of Economic Theory*, November 2007, *137* (1), 300–325.

_ **and Max Floetotto**, "Firm dynamics, markup variations, and the business cycle," *Journal of Monetary Economics*, 2008, *55* (7), 1238–1252.

Jensen, Martin Kaae, "Distributional comparative statics," *The Review of Economic Studies*, 2018, *85* (1), 581–610.

Kahn, James A, "Inventories and the volatility of production," *The American Economic Review*, 1987, pp. 667–679.

Khan, Aubhik and Julia K. Thomas, "Inventories and the Business Cycle: An Equilibrium Analysis of (S, s) Policies," *American Economic Review*, September 2007, *97* (4), 1165–1188.

King, Robert G and Sergio T Rebelo, "Resuscitating real business cycles," *Handbook of macroeconomics*, 1999, *1*, 927–1007.

Kopytov, Alexandr, Bineet Mishra, Kristoffer Nimark, and Mathieu Taschereau-Dumouchel, "Endogenous production networks under supply chain uncertainty," *Available at SSRN*, 2022.

Kremer, Michael, "The O-ring theory of economic development," *The Quarterly Journal of Economics*, 1993, *108* (3), 551–575.

Levine, David K, "Production chains," *Review of Economic Dynamics*, 2012, *15* (3), 271–282.

Lim, Kevin, "Endogenous Production Networks and the Business Cycle," *Working Paper*, 2018.

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

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

Mankiw, N Gregory and Michael D Whinston, "Free entry and social inefficiency," *The RAND Journal of Economics*, 1986, pp. 48–58.

Mas-Colell, Andreu, Michael Dennis Whinston, Jerry R Green et al., *Microeconomic theory*, Vol. 1, Oxford university press New York, 1995.

McCall, J. J., "Economics of Information and Job Search," *The Quarterly Journal of Economics*, 1970, *84* (1), 113–126.

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

Menzio, Guido, "Markups: A Search-Theoretic Perspective," Technical Report, National Bureau of Economic Research 2024.

Pagano, Marco, "Imperfect Competition, Underemployment Equilibria and Fiscal Policy," *Economic Journal*, June 1990, *100* (401), 440–463.

Parenti, Mathieu, Philip Ushchev, and Jacques-François Thisse, "Toward a theory of monopolistic competition," *Journal of Economic theory*, 2017, *167*, 86–115.

Petrongolo, Barbara and Christopher A Pissarides, "Looking into the black box: A survey of the matching function," *Journal of Economic literature*, 2001, *39* (2), 390–431.

Pissarides, Christopher A, *Equilibrium unemployment theory*, MIT press, 2000.

_ , "The unemployment volatility puzzle: Is wage stickiness the answer?," *Econometrica*, 2009, *77* (5), 1339–1369.

Ramey, Valerie A and Kenneth D West, "Inventories," *Handbook of macroeconomics*, 1999, *1*, 863–923.

Rogerson, Richard, "Indivisible labor, lotteries and equilibrium," *Journal of monetary Economics*, 1988, *21* (1), 3–16.

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

Rotemberg, Julio J., "Sticky Prices in the United States," *Journal of Political Economy*, 1982, *90* (6), 1187–1211.

Sargent, Thomas J and Lars Ljungqvist, *Recursive macroeconomic theory*, MIT press, 2012.

Schaal, Edouard and Mathieu Taschereau-Dumouchel, "Coordinating business cycles," *Working Paper*, 2018.

_ **and** _ , "Aggregate Demand and the Dynamics of Unemployment," *Working Paper*, 2019.

Schumpeter, Joseph Alois, *Business cycles: A theoretical, historical and statistical analysis of the capitalist process*, McGraw Hill Book Company, 1939.

Shimer, Robert, "The cyclical behavior of equilibrium unemployment and vacancies," *American Economic Review*, 2005, *95* (1), 25–49.

_ **and Lones Smith**, "Assortative matching and search," *Econometrica*, 2000, *68* (2), 343–369.

Syverson, Chad, "Macroeconomics and market power: Context, implications, and open questions," *Journal of Economic Perspectives*, 2019, *33* (3), 23–43.

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

Wright, Randall, Philipp Kircher, Benoît Julien, and Veronica Guerrieri, "Directed search and competitive search equilibrium: A guided tour," *Journal of Economic Literature*, 2021, *59* (1), 90–148.

A Appendix

A.1 Mathematical Results

Here I prove some results used in the text whose proofs I could not easily find.

Lemma 1. *Consider a function $f(\underline{x})$, homogeneous of degree δ. Then all $\frac{\partial f(\underline{x})}{\partial x_i}$ are homogeneous of degree $\delta - 1$.*

Proof. If f is homogeneous of degree δ we can write

$$\delta f(\underline{x}) = \sum_i x_i \frac{\partial f(\underline{x})}{\partial x_i}. \tag{A.1}$$

Taking a derivative of this with respect to some x_j

$$\delta \frac{\partial f(\underline{x})}{\partial x_j} = \sum_i x_i \frac{\partial^2 f(\underline{x})}{\partial x_j \partial x_i} + \frac{\partial f(\underline{x})}{\partial x_j}. \tag{A.2}$$

Where the second term on the RHS comes from the fact that when the summation is evaluated at x_j the derivative also includes $\frac{\partial x_j}{\partial x_j}\frac{\partial f(\underline{x})}{\partial x_j} = \frac{\partial f(\underline{x})}{\partial x_j}$.

Denote $g(\underline{x}) = \frac{f(\underline{x})}{\partial x_j}$ we have

$$(\delta - 1) g(\underline{x}) = \sum_i x_i \frac{\partial g(\underline{x})}{\partial x_i}. \tag{A.3}$$

Which implies that $g(\cdot)$ is homogeneous of degree $\delta - 1$. □

Lemma 2. *Consider a function $f(\underline{x})$, homogeneous of degree δ. Then all $f^{-1}(\underline{x})$ is homogeneous of degree $1/\delta$.*

Proof. By the definition of homogeneous function, for some $s > 0$

$$f(s\underline{x}) = s^\delta f(\underline{x}). \tag{A.4}$$

Applying the inverse to both sides

$$f^{-1}(f(s\underline{x})) = f^{-1}(s^\delta f(\underline{x})) \tag{A.5}$$

$$s\underline{x} = f^{-1}(s^\delta f(\underline{x})) \tag{A.6}$$

Denote $y = f(\underline{x})$, then

$$s\underline{x} = f^{-1}(s^\delta y) \tag{A.7}$$

$$s f^{-1}(y) = f{-}1(s^\delta y) \tag{A.8}$$

Denote $\kappa = s^{\delta}$, we have

$$\kappa^{1/\delta} f^{-1}(y) = f^{-1}(\kappa y), \tag{A.9}$$

which proves the statement. □

A.2 Duration

This section shows that the duration in the baseline McCall model is $1/H$.

Start by noting that the probability an agent will be unemployed for arbitrary d periods from now is

$$Prob(D = d) = (1 - H)^{d-1} H \tag{A.10}$$

This is saying that the probability of being unemployed d periods is equal to the probability of rejecting $d - 1$ offers and accepting 1.
From this, the cumulative distribution is

$$F = \sum_{d=1}^{\infty} (1 - H)^{d-1} H = 1 \tag{A.11}$$

Redefine the cumulative, simply by having $t = d - 1$ and changing the starting point of the sum accordingly, as

$$\sum_{t=0}^{\infty} (1 - H)^{t} H = 1 \tag{A.12}$$

Differentiating with respect to H

$$\begin{aligned} f &= \sum_{t=0}^{\infty} [-t(1 - H)^{t-1} H + (1 - H)^{t}] = 0 \\ \sum_{t=0}^{\infty} -t(1 - H)^{t-1} H &= -\sum_{t=0}^{\infty} (1 - H)^{t} \\ \sum_{t=1}^{\infty} t(1 - H)^{t-1} H &= \frac{1}{H} \end{aligned} \tag{A.13}$$

Where the change of index in the last line is allowed cause the element of the summation for $t = 0$ is equal to 0.

Note that this is the expected duration, you can see this by noting that this takes the form

$\mathbb{E}(x) = \sum xp(x)$. Hence $\mathbb{E}(d) = \frac{1}{H}$.

A.3 Discounting in Continuous Time

This brief section describes why, in continuous time, discounting is carried out in the form of $\beta = e^{-r\Delta}$.

Start by recalling that given an interest rate r, $\beta = \frac{1}{1+r}$

Think of an asset with value v_t in discrete time, then

$$\begin{aligned} v_t(1+r) &= v_{t+1} \\ r &= \frac{\Delta v_t}{v_t} \end{aligned} \tag{A.14}$$

In continuous time the instantaneous rate of change of the value is $\frac{d}{dt}v(t)$, so that

$$r = \frac{\frac{d}{dt}v(t)}{v(t)} \tag{A.15}$$

or, equivalently

$$r = \frac{d}{dt} ln(v(t)) \tag{A.16}$$

Use this definition to compare values that are Δ time apart, integrate both sides

$$\begin{aligned} \int_t^{t+\Delta} r ds &= \int_t^{t+\Delta} \frac{d}{dt} ln(v(s)) ds \\ r\Delta &= ln(v(t+\Delta)) - ln(v(t)) \\ r\Delta &= ln\left(\frac{v(t+\Delta)}{v(t)}\right) \end{aligned} \tag{A.17}$$

This implies that

$$e^{r\Delta} = \frac{v(t+\Delta)}{v(t)} \tag{A.18}$$

rearranging

$$v(t)e^{r\Delta} = v(t+\Delta) \tag{A.19}$$

This looks like the equation with started from in discrete time so that $e^{r\Delta}$ plays the role of $(1+r)$. This implies that

$$\beta = e^{-r\Delta} \tag{A.20}$$

A.4 Alternative Derivation of Continuous Time Reservation Wage

Let $\beta = \frac{1}{1+r\Delta}$, then

$$U = b\Delta + \frac{1}{1+r\Delta}[\alpha\Delta\mathbb{E}\max\{V(w), U\} + (1-\alpha\Delta)U + o(\Delta)] \tag{A.21}$$

This equation reads the value of unemployment is the payoff from unemployment benefits plus the value of receiving one offer (first element in bracket) weighted by the probability of receiving that offer plus the value of receiving zero offers, i.e. unemployment again, weighted by the corresponding probability plus an error, standing in for more offers.
Again we can define the value of multiple offers as $o(\Delta)$ because the likelihood of it will vanish as $\Delta \to 0$ and the value of multiple offers, which will take the form $V(W)$ where $W = \max\{w_1, ..., w_n\}$ is bounded if the distribution of wages is bounded.
Multiplying by $1 + r\Delta$ and subtracting U

$$r\Delta U = (1+r\Delta)b\Delta + \alpha\Delta\mathbb{E}\max\{V(w) - U, 0\} + \frac{o(\Delta)}{\Delta} \tag{A.22}$$

Dividing by Δ and letting $\Delta \to 0$

$$rU = b + \alpha\max\{V(w) - U, 0\} \tag{A.23}$$

Which is again the definition of the reservation wage in continuous time.

A.5 CES + Monopolistic Competition

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.[49] Second, note that while there is dispersion in market

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

shares, this is no indication of market power. To see that, start with the definition of market share:

$$s_i = \frac{p_i y_i}{PY} = \frac{y_i y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y\left(\int_j p_j^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{A.24}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y\left(\int_j (y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} P)^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{A.25}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}} Y^{\frac{1}{\sigma}} P}{Y Y^{\frac{1}{\sigma}} P\left(\int_j (y_i^{-\frac{1}{\sigma}})^{1-\sigma}\right)^{\frac{1}{1-\sigma}}} \tag{A.26}$$

$$= \frac{y_i^{\frac{\sigma-1}{\sigma}}}{Y\left(\int_j y_i^{\frac{\sigma-1}{\sigma}}\right)^{\frac{1}{1-\sigma}}} \tag{A.27}$$

$$= y_i^{\frac{\sigma-1}{\sigma}} Y^{-1}\left(\int_j y_i^{\frac{\sigma-1}{\sigma}}\right)^{\frac{1}{\sigma-1}} \tag{A.28}$$

$$= y_i^{\frac{\sigma-1}{\sigma}} Y^{-1} Y^{\frac{1}{\sigma}} = \left(\frac{y_i}{Y}\right)^{\frac{\sigma-1}{\sigma}}. \tag{A.29}$$

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

$$s_i = \frac{p_i y_i}{PY} = \frac{p_i p_i^{-\sigma} P^{\sigma} Y}{PY} \tag{A.30}$$

$$= p_i^{1-\sigma} P^{\sigma-1} = p_i^{1-\sigma}\left(\int_j p_j^{1-\sigma}\right)^{-1} \tag{A.31}$$

$$= \left(\frac{\sigma}{\sigma-1}\frac{c}{A_i}\right)^{1-\sigma}\left(\int_j \left(\frac{\sigma}{\sigma-1}\frac{c}{A_j}\right)^{\sigma-1}\right)^{-1} \tag{A.32}$$

$$= \frac{A_i^{\sigma-1}}{\int_j A_j^{\sigma-1}}. \tag{A.33}$$

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