Chapter 22

Heterogeneous firms

Toshihiko Mukoyama

22.1 Introduction

In most macroeconomic models (and earlier in this textbook), it is assumed that there exists an aggregate production function

$$Y = F(K, L),$$

where Y is aggregate output, K is aggregate capital, and L is aggregate labor. This assumption is justified if the production functions for all firms are homogeneous. In reality, firms are heterogeneous in many dimensions. There are large and small firms, young and old firms, growing and contracting firms, productive and unproductive firms, and of course firms produce goods that are typically not perfect substitutes.

Answering many macroeconomic questions requires explicitly taking firm heterogeneity into account. For example, how should we encourage (or discourage) the entry of new firms? Should we support growing firms, and if so how? Should we be concerned about the growing prominence of "mega-firms"? What are the causes and consequences of the recent decline in firm entry and reallocation of resources across firms? Analysis based on the aggregate production function is not helpful in answering these questions. More generally, in light of firm heterogeneity, does the representative-firm assumption give misleading answers to standard policy questions?

In this chapter, we consider such questions. Looking at the data, we will see some indications that the prominence of large firms in the U.S. economy has been rising in recent years. The reallocation of resources through the entry and exit of firms, as well as the expansion or contraction of firms, seem to be slowing down in recent years (often referred to as the "decline in business dynamism"). These phenomena have potentially important consequences in the macroeconomic context. The rise of big firms may be associated with an increase in their market power. The market power in the product and the labor market could be linked to market distortions and changes in the labor share. The lack of reallocation of resources from unproductive firms to productive firms may lead to lower aggregate productivity due to "misallocation" (resource allocation that is suboptimal). Misallocation may also lead to a slower rate of innovation and aggregate productivity growth. In addition, the dominance of large firms may have implications for other macroeconomic phenomena, such as business cycle fluctuations.

In an effort to break out of the aggregate production function approach, this chapter covers basic facts, models, and methods for analyzing an economy with heterogeneous firms.

22.2 A simple model

We begin by considering a simple example where firm heterogeneity matters for macroeconomic analysis.[1] Suppose that there is a unit mass of firms. The firms produce a homogeneous good under perfect competition. The production function of firm i (where i is the index of firms: $i \in [0,1]$) is

$$y_i = a_i F(\mathbf{x}_i)^\gamma,$$

where y_i is the output of firm i and $\gamma \in (0,1)$. Firms are heterogeneous in their productivity: a_i is different across firms. $\mathbf{x}_i$ is the input vector for firm i. Assume that $F(\mathbf{x}_i)$ exhibits constant returns to scale. Then, because $\gamma < 1$, the overall production of y_i exhibits decreasing returns to scale in inputs $\mathbf{x}_i$. The decreasing returns property is important. With constant returns, the most productive firm(s) (e.g., with the largest a_i) takes over the entire economy's production, and the outcome is either (i) a monopoly or oligopoly of one or a few firms, which would contradict the perfect-competition assumption; or (ii) only the most efficient firms with common a_i operate as price takers, which would replicate the homogeneous-firms scenario. Let $\mathbf{X}$ be the endowment vector of inputs in the economy.

Due to the constant-returns property of $F(\mathbf{x}_i)$, we can solve the firm's problem in two steps: first, solve the cost-minimizing combination of inputs for one unit of $F(\mathbf{x})$ and, second, decide on the optimal scale of production. The first stage is common across firms:

$$\min_{\mathbf{x}} \ \mathbf{p}\mathbf{x}$$

subject to

$$F(\mathbf{x}) = 1,$$

where $\mathbf{p}$ is the vector of input prices. Let the solution of this problem be $\mathbf{x}^*$ and the minimized unit cost be $c \equiv \mathbf{p}\mathbf{x}^*$.

Let $m_i = F(\mathbf{x}_i)$ be the choice of the firm i's combined inputs. The constant-returns property implies that the optimal input choice is $\mathbf{x}_i = m_i \mathbf{x}^*$ and the cost of production is cm_i. The second stage optimization problem is

$$\max_{m_i} \ a_i m_i^\gamma - c m_i. \tag{22.1}$$

The first-order condition for this problem is

$$a_i m_i^{\gamma-1} = \frac{c}{\gamma}. \tag{22.2}$$

Therefore, $y_i = (c/\gamma) m_i$ for all i. Adding up for all i,

$$Y = \frac{c}{\gamma} M \tag{22.3}$$

[1] A similar framework is used by Hopenhayn (2014a). Some of the results below overlap with his.

holds, where

$$Y = \int y_i di \tag{22.4}$$

is the total output and

$$M = \int m_i di. \tag{22.5}$$

Note that, in equilibrium, $M = \int F(\mathbf{x}_i)di = F(\mathbf{X})$ has to hold. Let us define

$$A \equiv \left(\int a_i^{\frac{1}{1-\gamma}} di \right)^{1-\gamma}. \tag{22.6}$$

From (22.2),

$$A = \frac{c}{\gamma} M^{1-\gamma}$$

holds. Combining with (22.3) and $M = F(\mathbf{X})$,

$$Y = AF(\mathbf{X})^{\gamma}. \tag{22.7}$$

In this environment, this relationship can be viewed as the aggregate production function.[2] The heterogeneity of firms matters through the aggregation (22.6): the aggregate outcome is influenced by the distribution of a_i to the extent that it yields different values of A in (22.6).

To illustrate, suppose that a_i follows a lognormal distribution, where $\log(a_i) \sim N(\nu - \sigma^2/2, \sigma^2)$. From the property of the lognormal distribution, the average of a_i, $\int a_i di$, is $\exp(\nu)$. However, the following can also be established.[3]

$$A = \exp\left(\nu + \frac{\gamma}{1-\gamma} \frac{1}{2} \sigma^2 \right). \tag{22.8}$$

Therefore, the dispersion parameter σ influences the level of A even when the average productivity $\exp(\nu)$ is constant. This result holds because the productive resources are endogenously allocated: a productive firm uses more input than an unproductive firm and therefore has a greater presence in aggregate production than merely having higher productivity. When the dispersion parameter σ is larger, the economy has more room to allocate resources to the highly productive firms in the right tail. Allocation of inputs is the key to analyzing heterogeneous firms: when the inputs are not allocated properly, aggregate productivity, and hence output, can be raised by the mere reallocation of resources (inputs). In this chapter, we always keep two questions in mind: (i) how the distribution of a_i is determined, and (ii) how the economy allocates resources to different firms.

[2] An example of this aggregation is when the production function is $y_i = a_i(k_i^{\alpha} \ell_i^{1-\alpha})^{\gamma}$, where k_i is firm i's capital input, ℓ_i is the firm i's labor input, and $\alpha \in (0,1)$. In this case, the aggregate production function is

$$Y = A(K^{\alpha} L^{1-\alpha})^{\gamma},$$

where A is given by (22.6). The rental rate of capital in equilibrium is $r = \gamma\alpha A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha-1}L^{1-\alpha}$, the wage rate is $w = \gamma(1-\alpha)A(K^{\alpha}L^{1-\alpha})^{\gamma-1}K^{\alpha}L^{-\alpha}$, and the unit cost of production is $c = (r/\alpha)^{\alpha}(w/(1-\alpha))^{1-\alpha}$.

[3] See Appendix 22.A.1 for derivation.

22.3 Firm heterogeneity in the data

This section describes some facts related to firm heterogeneity. We will focus on U.S. data. The statistics presented here are based on publicly available data.[4] The natural first question is: how heterogeneous are U.S. firms? Figure 22.1 shows the firm size distribution as the number of firms in each size category, as a fraction of the total number of firms. Here, firm size is measured by the number of employees.

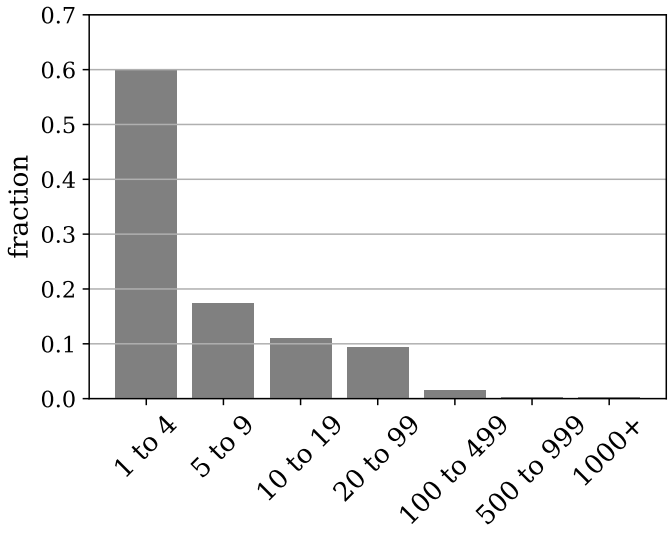

Figure 22.1: Distribution of firm size in 2022.

Source: Business Dynamics Statistics.

Figure 22.1 shows that the firm size distribution is quite dispersed. There are over 5 million firms in the U.S., and the majority are very small firms with 1 to 4 employees. At the same time, there are over 10,000 large firms with more than 1,000 employees, as well as over 1,000 firms with 10,000 employees or more.

The fact that very small firms account for the majority of firms does not imply that large firms are unimportant. Figure 22.2 plots the employment share of each size category. Approximately half of all employees work at firms with 1,000 or more employees. In fact, approximately 30% of workers are employed by very large firms with 10,000 or more employees.

The firm dynamics literature often uses data at the establishment level. An establishment is a fixed physical location where economic activity occurs; it is more straightforward to identify an establishment than a firm. A firm is a collection of establishments under common ownership, and it is often difficult to identify a firm in an administrative dataset. Establishments are also heterogeneous. Figure 22.3 is the establishment size distribution. There are over 7 million establishments in the U.S. economy, and approximately half are very small establishments with 1 to 4 employees.

Figure 22.4 plots the number of establishments that are owned by each firm size cate-

[4] All figures in this section are drawn from the U.S. Census Bureau's Business Dynamics Statistics. See `https://bds.explorer.ces.census.gov/`.

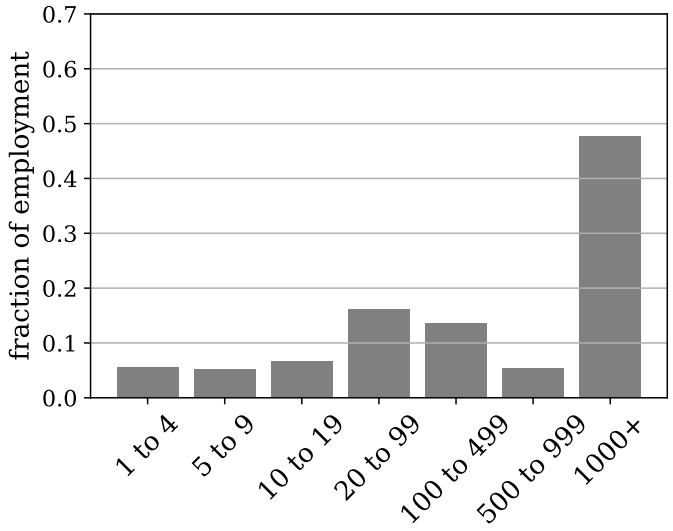

Figure 22.2: Employment share of each size category in 2022.

Source: Business Dynamics Statistics.

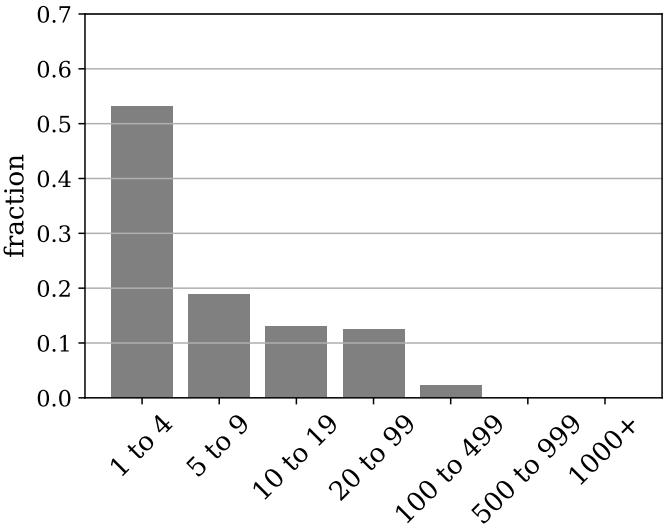

Figure 22.3: Distribution of establishment size in 2022.

Source: Business Dynamics Statistics.

gory. It shows that many establishments are owned by large firms (approximately 16% of all establishments are owned by firms in the 1,000+ category). Whereas almost all "1 to 4" category firms own only one establishment, the firms in the 1,000+ category own 100 establishments on average. Very large firms with 10,000 or more employees own approximately 600 establishments on average.[5]

Aside from the cross-sectional heterogeneity, U.S. firms conduct significant adjustments

[5]See Cao, Hyatt, Mukoyama, and Sager (2022) for a detailed analysis of the number of establishments per firm and its time-series properties.

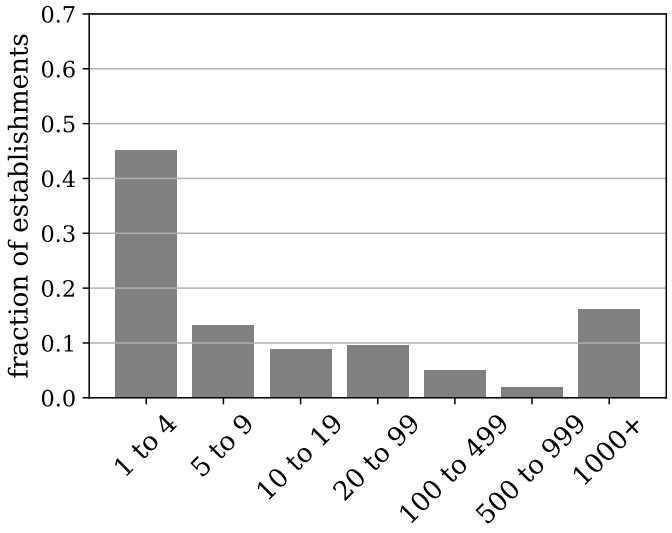

Figure 22.4: Fraction of establishments owned by each firm size category in 2022.

Source: Business Dynamics Statistics.

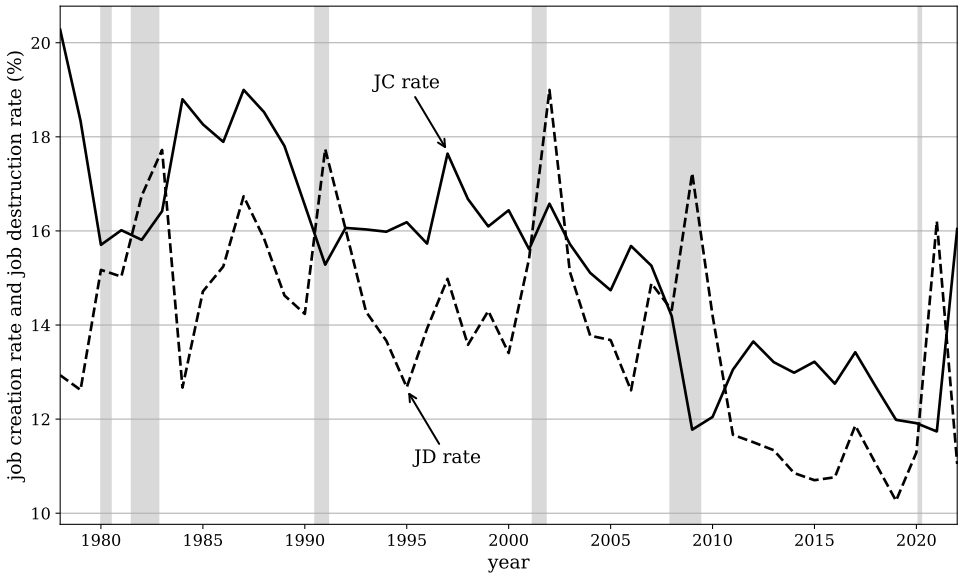

Figure 22.5: Annual job creation and destruction rates (establishment level).

Source: Business Dynamics Statistics.

over time. One measure of a firm's size adjustment is job creation and destruction. Job creation (JC) refers to the expansion of firms or establishments, whereas job destruction (JD) refers to the contraction of firms or establishments. BDS publishes the establishment-

level JC and JD rates. The JC rate is defined as

$$JC_t \equiv \frac{\sum_{i:\ell_{it}>\ell_{i,t-1}}(\ell_{it}-\ell_{i,t-1})}{\bar{L}_t}, \tag{22.9}$$

where ℓ_{it} is the employment of establishment i at year t, L_t is the total employment at year t (which is the sum of ℓ_{it}), $\bar{L}_t \equiv (L_t + L_{t-1})/2$. In words, the JC rate is the sum of employment increases in all expanding establishments, divided by total employment (the average for time t and $t-1$). The JD rate is similarly defined as

$$JD_t \equiv \frac{\sum_{i:\ell_{it}<\ell_{i,t-1}}(\ell_{i,t-1}-\ell_{it})}{\bar{L}_t}.$$

The JD rate is the sum of the employment decrease by contracting establishments, divided by total employment (the average for time t and $t-1$). JC and JD, often called gross job flows, measure the magnitudes of labor reallocation across establishments. Figure 22.5 plots the JC and JD rates from the BDS dataset. The shaded area is the recession period defined by the National Bureau of Economic Research (NBER).[6] Three properties are notable. First, the magnitudes of JC and JD are large. Both the JC and JD rates exceed 10% in any given year. Second, both rates are cyclical. When a recession arrives, the JC rate declines and the JD rate increases. Third, there is a general declining trend in both the JC and JD rates. It is, moreover, well known that a wide range of indicators of reallocation, including the JC and JD rates, have declined in recent years. Some researchers call this trend the "declining business dynamism" of the U.S. economy.

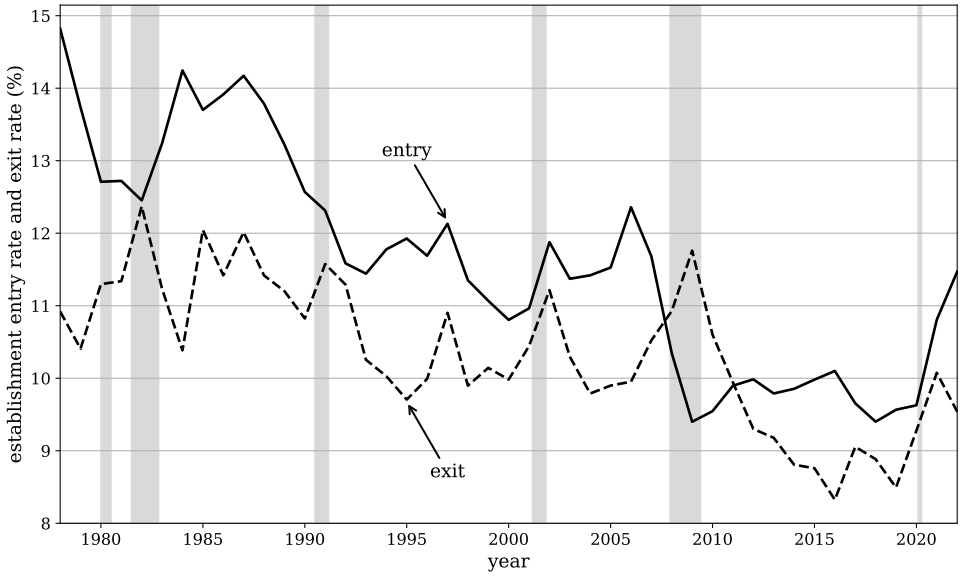

Figure 22.6: Annual establishment entry and exit rates.

Source: Business Dynamics Statistics.

Another measure of reallocation is the rate of entry and exit. Many firms and establishments enter and exit every year. Figure 22.6 plots the entry rate and exit rates of

[6] See `https://www.nber.org/research/business-cycle-dating`.

establishments. The entry rate is defined as the number of entering establishments between $t-1$ and t divided by the total number of establishments (the average of time $t-1$ and t). The exit rate is defined as the number of exiting establishments between $t-1$ and t divided by the total number of establishments (the average of time $t-1$ and t). One can observe similar properties here as in the JC and JD rates: the entry and exit rates are large, cyclical, and there are overall declining trends.

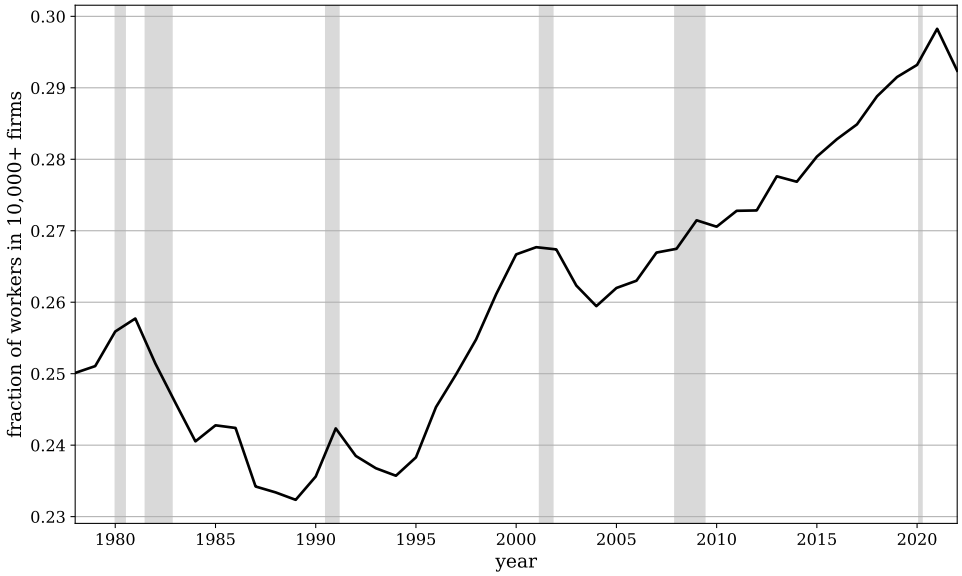

Figure 22.7: Fraction of employees working in firms with 10,000+ employee.

Source: Business Dynamics Statistics.

Over the last few decades, there have been significant changes in the heterogeneity among U.S. firms. In addition to the "declining dynamism" described above, one topic that caught researchers' attention is the dominance of large firms. Figure 22.7 plots the fraction of workers employed by firms in the 10,000+ size category.[7] This fraction has increased steadily since the early 1990s, indicating that large firms are starting to dominate the U.S. economy. This dominance raised concerns about the market power of large firms, and is consistent with another strand of research that tries to measure the trend of market power in the U.S. economy. Figure 22.8 reproduces Figure 1 of De Loecker et al. (2020). It measures the trend of the average markup (i.e., price over the marginal cost) of U.S. public firms in the Compustat dataset. The markup series exhibits an increasing trend since the 1980s.[8]

[7] In drawing this figure, the distinction between firms and establishments is very important. See Appendix 22.A.2.

[8] The evolution of market power, both in the product market and the factor market, remains an active research topic. The studies that follow De Loecker et al. (2020) highlight important methodological limitations and industry heterogeneity. Useful surveys include Miller (2025) and Syverson (2025).

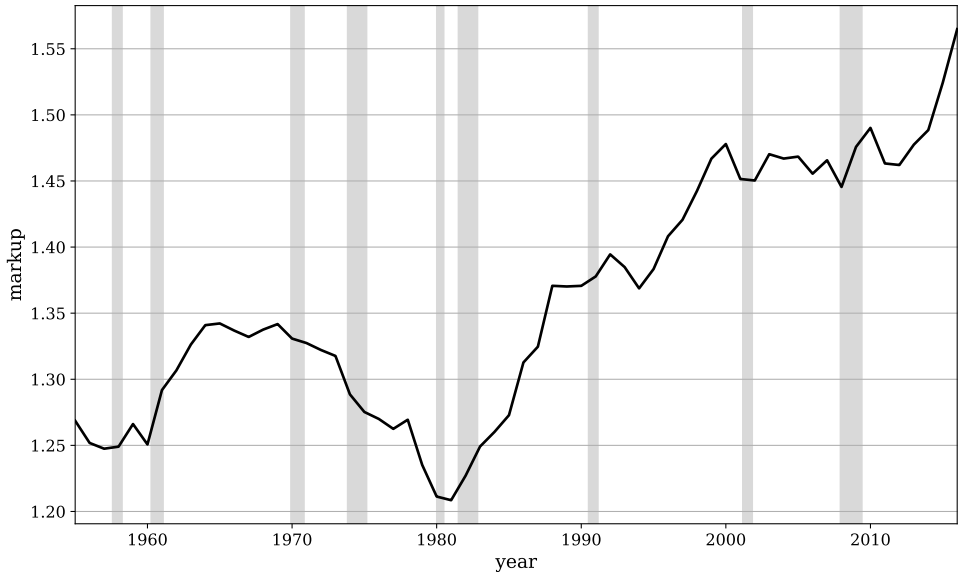

Figure 22.8: Markup of the U.S. public firms.

Source: De Loecker et al. (2020).

22.4 Reallocation and misallocation

The previous section shows that there is a large degree of reallocation among U.S. firms. How much does reallocation matter for aggregate productivity? In Section 22.1, we have seen that firm heterogeneity affects the aggregate outcome through the endogenous allocation of inputs. In a dynamic economy where firm productivity changes over time, one can imagine that the constant reallocation of inputs can have an important impact on aggregate productivity.

Foster, Haltiwanger, and Krizan (2001) illustrate the quantitative impact of reallocation through the following simple accounting framework. Let us denote the productivity of establishment i (they use establishment-level data and not firm-level data) at time t as a_{it}. Output-weighted average productivity $\bar{A}_t$ is defined as

$$\bar{A}_t \equiv \sum_i s_{it} a_{it},$$

where s_{it} is the output share of establishment i. Then, by denoting the $x_t - x_{t-1}$ by Δx_t,

$$\begin{aligned}\Delta \bar{A}_t = & \sum_{i \in C} s_{it-1} \Delta a_{it} + \sum_{i \in C} (a_{it-1} - \bar{A}_{t-1}) \Delta s_{it} + \sum_{i \in C} \Delta a_{it} \Delta s_{it} \\ & + \sum_{i \in N} s_{it} (a_{it} - \bar{A}_{t-1}) - \sum_{i \in X} s_{it-1} (a_{it-1} - \bar{A}_{t-1})\end{aligned}$$

holds, where C is the set of continuing establishments (establishments that exist both at time $t-1$ and t), N is the set of new establishments (establishments that enter between time $t-1$ and t), and X is the set of exiting establishments (establishments that exit between time $t-1$ and t). The increase in average productivity can occur for five distinct reasons. First, each of the existing establishments can increase its productivity. Second, an establishment with

higher-than-average productivity can increase its market share. Third, the first two effects can be magnified if both occur at the same time (i.e., a high-productivity establishment raises the share and its own productivity). Fourth, the entering establishment can be better than the average. Fifth, the exiting establishment can be worse than the average. All factors except for the first one can be interpreted as the contribution of reallocation. That is, if $\Delta s_{it} = 0$ and there are no entry and exit, the only way for the aggregate productivity to increase is for each establishment to increase its productivity. Using U.S. manufacturing data from 1977 to 1987, Foster et al. (2001) estimate (see their Table 8.4) that 45% of the aggregate change in multifactor productivity (the change in output that is not accounted for by the change in capital, labor, and intermediate goods) is accounted for by the first factor. The remaining 55% is the contribution of reallocation. This decomposition highlights the importance of reallocation in determining aggregate productivity growth.

Recently, a large body of literature has evaluated the role of various frictions that hinder the optimal allocation of resources. This literature emphasizes the existence of the *misallocation* of productive inputs as the source of low aggregate total factor productivity. A subset of literature, such as Restuccia and Rogerson (2008) and Hsieh and Klenow (2009), emphasizes firm-specific distortions as the source of misallocation. To see how firm-specific distortions can affect aggregate productivity, consider the model of Section 22.1 and add an assumption that the government taxes the output of firm i at the rate of τ_i. Thus, instead of the problem (22.1), the firm solves

$$\max_{m_i} (1-\tau_i) a_i m_i^{\gamma} - c m_i.$$

The rest of the model is the same, and GDP is still measured as $Y = \int y_i di$. After going through similar steps as in Section 22.1, one can show that the aggregate production function still takes the form of (22.7), but A is modified to satisfy

$$A = \frac{\int a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{\gamma}{1-\gamma}} di}{\left(\int a_i^{\frac{1}{1-\gamma}} (1-\tau_i)^{\frac{1}{1-\gamma}} di \right)^{\gamma}}. \tag{22.10}$$

One can easily see that this A is identical to (22.6) when $\tau_i = 0$ for all i. Now, suppose that a_i and $1-\tau_i$ follow a bivariate lognormal distribution. In particular, $(\log(a_i), \log(1-\tau_i)) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\nu_a - \sigma_a^2/2, \nu_\tau - \sigma_\tau^2/2)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_a^2 & \rho\sigma_a\sigma_\tau \\ \rho\sigma_a\sigma_\tau & \sigma_\tau^2 \end{bmatrix}.$$

Inserting this formulation into (22.10), we obtain the following.[9]

$$A = \exp\left(\nu_a + \frac{\gamma}{1-\gamma} \frac{1}{2} (\sigma_a^2 - \sigma_\tau^2) \right). \tag{22.11}$$

[9]See Appendix 22.A.3 for derivation. Hsieh and Klenow (2009) obtain a similar expression. This case is special in that ρ does not appear in the expression for aggregate productivity. In general, the correlation between $1-\tau_i$ and a_i matters for aggregate productivity, and Restuccia and Rogerson (2008) emphasize the importance of this correlation. See Hopenhayn (2014b) for related discussions.

One can easily see that this expression is identical to (22.8) when $\sigma_\tau = 0$. A large dispersion in $1 - \tau_i$ is detrimental to aggregate productivity. Intuitively, when $1 - \tau_i$ is dispersed, some productive firms do not expand, because $(1 - \tau_i)a_i$ is low even when a_i is large. At the same time, some unproductive firms employ a large amount of the input because $(1 - \tau_i)a_i$ is large even though a_i is small. Note that, in this setting, ν_τ does not influence aggregate productivity because it does not distort the allocation of input across firms, and the total supply of inputs is fixed.

Another subset of literature examines various specific policies and institutions that cause misallocation. Examples include size-specific taxes and regulations, entry regulations, and regulations governing hiring and firing.

22.5 Firm heterogeneity in general equilibrium

When firms are forward-looking, frictions affecting reallocation have further effects through firms' behavior. Hopenhayn and Rogerson (1993) highlight this mechanism in the context of firing taxes and quantify the outcome in a general equilibrium framework. In the following, we introduce the Hopenhayn and Rogerson (1993) framework with slightly different notation. In addition to the experiments on firing taxes (replicating the Hopenhayn and Rogerson, 1993 exercises), this section conducts experiments with entry barriers.[10]

22.5.1 Setup

There is a continuum of firms in the economy. We focus on a steady state, where prices and aggregate quantities (employment, output, and the number of firms) remain constant over time. In this section, we omit the firm's index i when there is no risk of confusion. Each firm uses only labor ℓ_t as an input. Firms behave competitively and maximize their profit facing a wage w_t. The firm's production function is $y_t = a_t \ell_t^\gamma$, where a_t is (exogenous) idiosyncratic productivity. In addition to wages, firms must pay c_f units of goods as the fixed operation cost every period. The firing taxes imposed by the government take the form of $\tau \max(0, \ell_{t-1} - \ell_t)$, where $\tau > 0$ is the firing tax for dismissing one worker. The government transfers all firing taxes back to the representative consumer. Therefore, the firm's flow profit is

$$\pi(\ell_{t-1}, \ell_t, a_t) = a_t \ell_t^\gamma - w_t \ell_t - c_f - \tau \max(0, \ell_{t-1} - \ell_t). \tag{22.12}$$

Note that the output price is normalized to 1, and the only endogenous price in each period is w_t.

The timing for the firms within a period is as follows. At the beginning of each period, the incumbent firm from the last period decides whether to exit. If the firm exits, it pays the firing cost $\tau \ell_{t-1}$. If it stays, it receives the current period value of a_t from the stochastic process

$$\log(a_t) = \alpha + \rho \log(a_{t-1}) + \varepsilon_t,$$

[10] The analysis of entry barriers is not in Hopenhayn and Rogerson (1993) but is subsequently conducted by, for example, Moscoso Boedo and Mukoyama (2012).

where α and $\rho \in [0, 1)$ are parameters and $\varepsilon_t \sim N(0, \sigma^2)$. After observing a_t, the firm decides on its employment and then produces.

Note that, unlike the model in Section 22.1, the firm's employment decision is dynamic in the presence of a positive firing tax. When the firm decides to hire a worker, it foresees that it has to pay the firing cost when it wants to shed workers in the future due to a negative productivity shock. This effect makes the firm reluctant to hire a worker when it receives a positive productivity shock. The dynamic programming problem for the firm is

$$W(a, \ell_{-1}) = \max_{\ell} \pi(\ell_{-1}, \ell, a) + \beta \max \{\mathbb{E}[W(a', \ell)|a], -\tau\ell\}, \tag{22.13}$$

where the subscript -1 represents the previous period value and prime ($'$) represents the next period value. $\beta \in (0, 1)$ is the consumer's discount factor (which is equal to the firm's discount factor, since the steady-state interest rate will take on this value) and $\mathbb{E}[\cdot|a]$ represents the expected value given a.

We assume free entry; that is, anyone can enter as long as the entry cost is paid. After the entry cost is paid, the firm draws productivity, employs workers, and produces. The entry cost is assumed to be $c_e + \kappa$, where c_e is the technological entry cost, including the investment required when entering, and κ is additional (wasteful) policy-related cost that we interpret as "entry barriers." Free entry implies

$$W^e = c_e + \kappa, \tag{22.14}$$

where W^e is the value of the entry that satisfies

$$W^e = \int (W(a, 0) + c_f) d\nu(a).$$

In the integral, $\nu(a)$ is the exogenous distribution of a for a new entrant. Note that we assume entrants do not have to pay the fixed operation cost c_f in the period they enter. Thus c_f in the first period, which is included in the profit expression (22.12), is "added back" here.

The representative consumer owns the firms, works, and consumes. The utility is

$$\sum_{t=0}^{\infty} \beta^t [u(C_t) - \chi L_t^s],$$

where $u(C_t)$ is increasing and concave utility from consumption C_t, $\chi > 0$ is a parameter, and L_t^s is the labor supply. In the steady state, the consumer's problem is static.[11] It reads

$$\max_{C, L^s} \; u(C) - \chi L^s$$

subject to

$$C \leq wL^s + \Pi + R,$$

[11] For completeness, one can allow borrowing and lending and then the interest rate will be set such that these amounts are both zero. In a steady state, consumption will be constant and the Euler equation will then deliver a gross interest rate $1/\beta$.

where Π is the total profit of the firms and R is the total transfer. The first-order condition is

$$wu'(wL^s + \Pi + R) = \chi. \tag{22.15}$$

Therefore, labor supply is a function of w and $\Pi + R$.

As Kaas (2021) points out, the competitive equilibrium of this model has a structure often referred to as "block recursive." That is, the equilibrium price (wage in this model) can be computed without the information on the distribution of state variables across incumbent firms. To see this, note that $W(a, \ell_{-1})$ can be computed from (22.13) once the value of w is known. Thus, W^e can be computed as a function of w. The free-entry condition (22.14) can be used to pin down the equilibrium value of w. For a given mass of entry M, the decision rule of (22.13) can be used to compute the stationary distribution of firms across different state variables a and ℓ_{-1}. With the stationary distribution, one can compute total labor demand L^d, total profits Π, and the total firing tax R. All L^d, Π, and R are functions of the entry mass M. Thus, the labor supply equation (22.15) can be used to determine the level of entry mass M that is consistent with the labor market equilibrium $L^d = L^s$.

Hopenhayn and Rogerson (1993) calibrate the model with $\tau = 0$ to the U.S. economy and examine the effect of τ quantitatively. The model here is identical to theirs except that (i) some notations are different and (ii) they normalize the wage as 1 and the market equilibrium determines the product price p, which corresponds to $1/w$ in our notation. We also set the baseline $\kappa = 0$.

The calibration procedure follows Hopenhayn and Rogerson (1993). One period is set at five years. First, the functional form of the utility function for consumption is assumed to be natural log: $u(c) = \log(c)$. Some parameters are set ex ante. The discount factor is set at $\beta = 0.8$, corresponding to the value of 4% per year. The production function parameter γ is 0.64, corresponding to the labor share.

To assign values to the remaining parameters, we assume that the $(\tau, \kappa) = (0, 0)$ case corresponds to the U.S. economy and find the parameter values so that various statistics from the model-generated data match the corresponding data moments. The parameters for the productivity process are set using the property of the model that the property of the productivity shock is directly reflected in the firm's employment decision. Using plant-level data from U.S. manufacturing, we set $\alpha = 0.076$ so that the average size of firms is 61.7 (the actual model moment is 62.4); $\rho = 0.93$, so that the autocorrelation of $\log(\ell)$ is 0.93; and $\sigma = 0.253$, so that the variance of the growth rate for ℓ is 0.53. The operation cost $c_f = 18.0$ is set to match an exit rate of 37% (the actual model moment is 34%). The entrant's productivity distribution ν is set so that the size distribution of young firms matches U.S. data. The entry cost c_e is set to 9.04 so that the free-entry condition holds with $w = 1$. The disutility of working χ is set so that the steady-state labor supply L is 0.6.

22.5.2 The effects of firing taxes

Table 22.1 summarizes the steady-state outcomes of the model with $\tau = 0$, $\tau = 0.1$, and $\tau = 0.2$, keeping $\kappa = 0$. Because one period is assumed to be five years, and the period wage (earnings per worker) with $\tau = 0$ is 1, $\tau = 0.1$ corresponds to six months' salary of

Table 22.1: Model results with firing taxes

	$\tau = 0$	$\tau = 0.1$	$\tau = 0.2$
Wage	1.000	0.977	0.957
Total output	100	97.7	95.7
Total employment	100	98.3	97.4
Labor productivity	100	99.4	98.3
$JC(= JD)$ rate	0.28	0.25	0.21

a worker.[12] For total output, total employment, and labor productivity, the $\tau = 0$ case is normalized to 100.[13]

There are several important points to note. First, it is not a priori obvious whether the equilibrium employment L goes up or down when τ increases. The reason is that the effect on firing (firms fire less because of the taxes) brings L up, whereas the effect on hiring (firms do not hire much even with a positive a shock, given that, in the future, they may have to fire these extra workers) brings L down. Which one dominates is a quantitative question; here, the latter effect dominates, and L decreases when τ increases to $\tau = 0.1$ and $\tau = 0.2$.

Second, labor productivity, Y/L, declines with τ. The reason is the misallocation mentioned in Section 22.4. Because a firm with a good a shock does not expand as much as in the first-best allocation, and a firm with a bad a does not fire as many workers with the firing tax, labor is not allocated properly across firms. These incentives imply that the marginal product of labor is dispersed (in the first best, the marginal product of labor is equalized). The difference from Section 22.4 is that the misallocation stems from the firm's dynamic decisions, especially for hiring. Firing and exit decisions are also affected by dynamic considerations. In general equilibrium, misallocation also affects the wage level and firm entry.

Third, the job creation (JC) rate, defined as (22.9), decreases with τ.[14] The reallocation of labor across firms is reduced because of the reluctance to hire and fire described above. The lack of reallocation is, therefore, closely linked to the productivity loss due to misallocation.

22.5.3 The effects of entry barriers

Table 22.2 describes the model outcome for $\kappa = 0$, $\kappa = 0.5$, and $\kappa = 5.0$, keeping $\tau = 0$.[15] As in Table 22.1, for total output, total employment, and labor productivity, the $\kappa = 0$ case

[12] Moscoso Boedo and Mukoyama (2012) computes the costs of business regulations corresponding to the τ that explicitly shows up in the World Bank's Doing Business dataset. The cross-country median of τ is about eight months of annual wages, and the average for low-income countries is 1.2 times the annual wages.

[13] Although the calibration is the same as in Hopenhayn and Rogerson (1993), the numbers are not exactly the same. The reason for the discrepancy is likely due to detailed differences in computation.

[14] Here, because the economy is in steady state, the job creation rate is equal to the job destruction rate.

[15] Moscoso Boedo and Mukoyama (2012) measures the costs of entry regulations corresponding to κ in the World Bank's Doing Business dataset. The cross-country median value of κ is 3.4 times the annual wages (about 0.7 times the five-year wages), and the average of the low-income countries is 29.9 times the annual wages (corresponding to 6 times the five-year wages. Note that although κ in the current model is not in terms of annual wages, the baseline annual wage is set at 1.0, and thus the units are comparable.

Table 22.2: Model results with entry barriers

	$\tau = 0$	$\kappa = 0.5$	$\kappa = 5.0$
Wage	1.000	0.986	0.879
Total output	100	98.6	87.9
Total employment	100	99.5	96.2
Labor productivity	100	99.1	91.4
$JC(= JD)$ rate	0.28	0.28	0.28

is normalized to 100.

Entry barriers have a substantial effect on outcomes. A higher cost of entry implies that the value of a firm has to be higher in equilibrium (see equation (22.14)), and a high firm value implies that the equilibrium wage has to be low. A low wage affects labor productivity through three channels. First, a low wage implies that low-productivity firms are less likely to exit. This effect pushes down aggregate productivity. Second, a low exit rate means that the entry rate is also low in steady state. Because entrants are less productive than incumbents, this effect increases aggregate productivity. Finally, the size of incumbents is larger because of lower wages, and due to decreasing returns to scale, a large scale implies lower productivity. The first and third effects push down aggregate productivity, and the second effect pushes it up; the former force dominates quantitatively.

The outcome of this exercise also highlights how policy has heterogeneous effects across firms. Whereas entry barriers harm entry, they increase the value of incumbent firms. When considering the entry policies, a significant conflict arises between incumbent firms and potential entrants.[16]

22.6 Alternative market arrangements

The above discussions have assumed that all markets are perfectly competitive. We have seen in earlier chapters that many macro models consider alternative market arrangements. This section introduces two alternative market arrangements with market power in the context of firm heterogeneity. There are two takeaways from this section. First, the insights on misallocation in Section 22.4 go through with minor modifications. Second, the inclusion of market power enables us to examine how firm heterogeneity interacts with other macroeconomic variables of interest, such as the aggregate level of markups.

22.6.1 Monopolistic competition

A popular alternative formulation in the macroeconomic context is monopolistic competition (Section 6.3.5). In this setting, firms produce differentiated goods, and only one firm produces each good.

[16]See Mukoyama and Popov (2014) for a politico-economic analysis of policies on firm entry. Mukoyama and Popov (2014) demonstrate that the political lobbying of incumbent firms and the economic benefits from limited entry may reinforce each other, potentially generating multiple steady states.

The standard setting considers two types of goods, the *final good* and the *intermediate goods*. The final good is produced in a perfectly competitive sector with constant returns to scale technology:

$$Y = \left[\int y_i^{\frac{\sigma-1}{\sigma}} di \right]^{\frac{\sigma}{\sigma-1}}. \tag{22.16}$$

where σ is the elasticity of substitution parameter. We assume $\sigma > 1$ so that the monopoly problem of each intermediate-good producer is well-defined. The aggregate Y in (22.16) can alternatively be considered as the utility by a consumer. This aggregation (22.16) is sometimes referred to as the Dixit-Stiglitz utility function.

Consider the setting in Section 22.1, except that (22.16) replaces (22.4) and that each good i is now monopolistically produced by an intermediate-good producer i. The intermediate-good producer's production structure is the same as in Section 22.1. The intermediate-good producers use the same inputs, and the input market is perfectly competitive. The aggregation of input (22.5) remains the same.

First, consider the cost-minimization problem of the final good producer:

$$\min_{\{y_i\}} \int p_i y_i di$$

subject to (22.16) for a given Y. Letting the Lagrange multiplier of the constraint be λ, the first-order condition is

$$p_i = \lambda y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}, \tag{22.17}$$

which implies

$$\int p_i y_i di = \lambda Y$$

and thus λ represents the (minimum) cost of producing one unit of the final good. Because the final-good market is perfectly competitive, λ is also the price of the final good. Let us call this price P. From (22.17),

$$P = \left[\int p_i^{1-\sigma} di \right]^{\frac{1}{1-\sigma}}$$

holds. As in Section 22.1, normalize the final-good price to be 1. Therefore, $P = \lambda = 1$ and the inverse demand function for good i is, from (22.17),

$$p_i = y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}}. \tag{22.18}$$

Keep in mind here that Y here is a shorthand that represents the production of all other goods in (22.16).

The monopolistic producer i maximizes profit given the inverse demand function (22.18) and its production function. The problem, which corresponds to (22.1) in Section 22.1, is

$$\max_{m_i} \left(a_i m_i^{\gamma}\right)^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} a_i m_i^{\gamma} - c m_i. \tag{22.19}$$

In a Nash equilibrium, each intermediate good firm i has to solve this problem given the other firms' input choices: m_j for all $j \neq i$. The important assumption in the monopolistic

competition is that each firm is small compared to the aggregate, so that it ignores the effect of m_i (therefore y_i) on Y. Therefore, each firm takes Y as given. With a similar step as in Section 22.1, we can obtain the relationship

$$Y = \frac{c}{\gamma}\frac{\sigma}{\sigma-1}M$$

instead of (22.3), and the same expression for the aggregate production function (22.7) where A is now modified to

$$A \equiv \left(\int a_i^{\frac{1}{\frac{\sigma}{\sigma-1}-\gamma}} di\right)^{\frac{\sigma}{\sigma-1}-\gamma}$$

instead of (22.6). Note that we have $\sigma/(\sigma-1)$ instead of 1, reflecting that each firm faces another factor (in addition to the decreasing returns to scale) that limits firm size. One important difference between this formulation and the basic model in Section 22.1 is that we can now accommodate constant returns to scale (or even some increasing returns to scale).

22.6.2 Oligopoly and endogenous markups

In the model of monopolistic competition with the constant elasticity of substitution aggregation (22.16), intermediate-good producers set a constant markup. To see this, first take a look at the first-order condition of the problem (22.19):

$$\left(1-\frac{1}{\sigma}\right)\gamma a_i^{1-\frac{1}{\sigma}} m_i^{\gamma-\frac{\gamma}{\sigma}-1} Y^{\frac{1}{\sigma}} = c. \tag{22.20}$$

Using (22.18), $y_i = a_i m_i^{\gamma}$, and the fact that the marginal cost $\mathcal{M} \equiv \partial(cm_i)/\partial y_i$ can be expressed as

$$\mathcal{M} = \frac{c m_i^{1-\gamma}}{\gamma a_i}, \tag{22.21}$$

(22.20) can be rewritten as

$$p_i = \frac{\sigma}{\sigma-1}\mathcal{M}. \tag{22.22}$$

Therefore, $\sigma/(\sigma-1)$ is the markup and is constant as long as σ is constant.

In many contexts, this constant markup property is a convenient model feature. However, this feature also imposes some limitations: the model cannot be used to analyze the endogenous changes in markups when the economic environment or policies change. The question of markup determination is particularly relevant in the recent U.S. economy. As mentioned in Section 22.3, De Loecker et al. (2020) observe that the level of markups has increased since the 1980s in the U.S. economy. Subsequent research has followed (at least) three different paths: (i) moving away from the monopolistic competition assumption; (ii) moving away from the CES assumption; and (iii) considering endogenous differences in productivity across firms. This section provides a brief introduction to the first approach, which is based on the formulation presented in Atkeson and Burstein (2008).

Consider the same setting as in Section 22.6.1, but where each intermediate good itself is the combination of several products. Following Atkeson and Burstein (2008), let us use

the term *sector* i to denote the collection of J firms that produce inputs for intermediate good i. Within each sector, let us index each firm by j and call a particular firm's product a *brand*. The production of intermediate good i is dictated by the function

$$y_i = \left[\sum_{j=1}^{J} q_{ij}^{\frac{\eta-1}{\eta}} \right]^{\frac{\eta}{\eta-1}}. \tag{22.23}$$

We assume that $\eta < \infty$, that is, brands are imperfect substitutes. We also assume that $\eta > \sigma > 1$, that is, brands (within a sector) are more substitutable than are goods (across sectors). The final-good production function is (22.16). We keep assuming that each intermediate good is small compared to the entire economy so that each firm does not consider the influence of its production decision on final good production Y (or the general price level). However, it is sufficiently large within each sector so that it is aware of the influence on y_i (and the price of intermediate good i).

In this setting, the final good producer has to solve two layers of the cost-minimization problem: (i) find the best combination of y_i for a given Y, and (ii) find the best combination of q_{ij} for a given y_i. The first cost-minimization problem is identical to the one in Section 22.6.1. The inverse demand function for y_i is given by (22.18). The second-stage cost-minimization problem can be solved similarly, and the result is

$$\frac{\hat{p}_{ij}}{p_i} = q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}}, \tag{22.24}$$

where $\hat{p}_{ij}$ is the price of the brand j in sector i and p_i is now the price of the sector-i good:

$$p_i = \left[\sum_{j=1}^{J} \hat{p}_{ij}^{1-\eta} \right]^{\frac{1}{1-\eta}}.$$

Note that (22.23) and (22.24) imply

$$p_i y_i = \sum_{j=1}^{J} \hat{p}_{ij} q_{ij}, \tag{22.25}$$

which is a consequence of (22.23) being a constant returns to scale function. Let the production function for firm j in sector i be

$$q_{ij} = a_{ij} m_{ij}^{\gamma}, \tag{22.26}$$

where m_{ij} is the "combined input" as before. The firm maximizes profit, $\hat{p}_{ij} q_{ij} - c m_{ij}$. Using the inverse demand function, the problem the firm solves is

$$\max_{q_{ij}, m_{ij}} \; q_{ij}^{-\frac{1}{\eta}} y_i^{\frac{1}{\eta}} y_i^{-\frac{1}{\sigma}} Y^{\frac{1}{\sigma}} q_{ij} - c m_{ij}.$$

subject to (22.23) (with q_{ij} for the other firms taken as given) and (22.26). As in the monopolistic competition case, the firm takes Y as given. From the first-order condition, noting the relationships (22.18) and (22.24), the pricing rule can be derived as

$$\hat{p}_{ij} = \frac{\varepsilon(s_{ij})}{\varepsilon(s_{ij}) - 1} \mathcal{M}, \tag{22.27}$$

where $\mathcal{M}$ is the marginal cost (analogous to (22.21))

$$\mathcal{M} = \frac{cm_{ij}^{1-\gamma}}{\gamma a_{ij}}$$

and

$$\varepsilon(s_{ij}) = \left[\frac{1}{\eta}(1 - s_{ij}) + \frac{1}{\sigma}s_{ij}\right]^{-1}. \tag{22.28}$$

Here, s_{ij} is

$$s_{ij} = \frac{\hat{p}_{ij}q_{ij}}{p_i y_i} = \frac{\hat{p}_{ij}q_{ij}}{\sum_{h=1}^{J}\hat{p}_{ih}q_{ih}}, \tag{22.29}$$

where the second equation utilizes the relationship (22.25). Thus, s_{ij} is the sales share of the firm j in sector i. Because $s_{ij} \in [0, 1]$ and $\sigma < \eta$, $\varepsilon(s_{ij})$ takes the value between σ and η. In particular, $\varepsilon(s_{ij})$ is η when $s_{ij} = 0$, monotonically decreases with s_{ij}, and reaches $\varepsilon(s_{ij}) = \sigma$ when $s_{ij} = 1$.

Comparing (22.22) and (22.27), the difference is that σ is replaced by $\varepsilon(s_{ij})$. In the current model, the markup of firm j can vary depending on the sales share. When firms are symmetric,

$$\varepsilon(s_{ij}) = \left[\frac{1}{\eta}\frac{J-1}{J} + \frac{1}{\sigma}\frac{1}{J}\right]^{-1}.$$

It is intuitive that the markup is the highest with $\varepsilon(s_{ij}) = \sigma$ when $J = 1$ (monopoly within the sector), which is exactly the monopolistic competition case (22.22). The markup decreases as J increases and $\varepsilon(s_{ij}) \to \eta$ as $J \to \infty$. This framework allows the monopoly power (represented by the number of firms J) to be linked to the markup. When firms are not symmetric (for example, because the a_{ij}s are different across firms), firm heterogeneity can feed into markup heterogeneity.

In the above derivation, we made the Cournot competition assumption: each producer chooses its quantity given the quantities of the other producers in the same sector. Alternatively, we can make the Bertrand competition assumption: each producer chooses its price given the prices of the other producers in the same sector. It can be shown that the formula (22.27) still holds, but with a different $\varepsilon(s_{ij})$:

$$\varepsilon(s_{ij}) = \eta(1 - s_{ij}) + \sigma s_{ij}$$

instead of (22.28), where s_{ij} is still defined by (22.29). The intuition is similar to that in the Cournot competition case. See Appendix 22.A.4 for the details of the derivation.

22.7 Business cycles and heterogeneous firms

As we saw in Section 22.3, many statistics on firm behavior, such as job creation and destruction rates, as well as entry and exit rates, exhibit clear cyclical patterns. It is natural, therefore, to think of the causes and consequences of such cyclicality in firm dynamics.

Note that in the model of Section 22.5, firms face idiosyncratic shocks, but the aggregate economy is stationary. The basic logic is simple: firm-level shocks are smoothed out by being

summed up across firms to create GDP. To illustrate, suppose that the GDP Y_t is the sum of firm-level output y_{it}, $i = 0, 1, ..., N$.[17] The growth rate of y_{it} is identically and independently distributed with mean 0 and variance σ^2. That is,

$$\frac{y_{i,t+1} - y_{it}}{y_{it}} = \sigma\varepsilon_{i,t+1},$$

where $\varepsilon_{i,t+1}$ is a random variable with mean zero and variance one. Then, the growth rate of Y_t is

$$\frac{Y_{t+1} - Y_t}{Y_t} = \frac{1}{Y_t}\sum_{i=1}^{N} \Delta y_{i,t+1} = \sum_{i=1}^{N} \frac{y_{it}}{Y_t}\sigma\varepsilon_{i,t+1}.$$

Thus, the standard deviation of GDP growth rate, σ_Y, is

$$\sigma_Y = \sigma\sqrt{\sum_{i=1}^{N}\left(\frac{y_{it}}{Y_t}\right)^2}. \tag{22.30}$$

When firms are of equal size, that is, $y_{it}/Y_t = 1/N$, $\sigma_Y = \sigma/\sqrt{N}$. When there are one million firms in the economy (there are over 5 million firms in U.S. data), $1/\sqrt{N} = 0.001$. A typical standard deviation of the firm-level volatility is between 10% and 20%; thus, the effect of idiosyncratic shocks on aggregate volatility is about two orders of magnitude smaller than is the volatility of GDP.[18] In other words, it is negligible in the context of business cycle fluctuations.

22.7.1 Aggregate shocks and firm dynamics

A strand of literature takes the law of large numbers as given and adds aggregate shocks in analyzing aggregate fluctuations. In that case, the model in Section 22.5 can be modified to have a production function for firm i as

$$y_{it} = z_t s_{it} \ell_{it}^{\gamma}.$$

Here, the variable z_t is added. As in the standard real business cycle model, z_t can be interpreted as the aggregate productivity shock. The model can then be calibrated and computed. As discussed earlier, the Hopenhayn and Rogerson (1993) model has a *block-recursive* structure. Therefore, computing equilibria for this type of model is often substantially easier than it is when computing equilibria in standard heterogeneous-consumer models.

It is known that the modified Hopenhayn and Rogerson (1993) model performs well in replicating the aggregate fluctuations in statistics such as job creation and destruction rates, as well as in entry and exit rates. Some other firm-level statistics are difficult to replicate by a simple modification of the Hopenhayn and Rogerson (1993) model.[19]

[17] This illustration is based on the exposition of Gabaix (2011).
[18] Gabaix (2011) estimates GDP volatility to be 12% in U.S. data.
[19] See Lee and Mukoyama (2018) for a detailed analysis.

22.7.2 Can idiosyncratic shocks generate aggregate fluctuations?

Despite the law of large numbers result, some researchers believe that micro-level shocks can have an important role in generating aggregate fluctuations. The calculation at the beginning of this section made two assumptions: (i) the distribution of idiosyncratic shocks has a finite variance, and (ii) there are no input-output networks. This subsection introduces the analysis of the economic environment where one of these two assumptions does not hold.

This research agenda is attractive because, since the outset of the real business cycle research agenda, fluctuations in aggregate shocks have often been criticized for being a "black box." If we know that the cycles stem from idiosyncratic shocks, one can imagine that an effective stabilization policy would target the firms whose idiosyncratic shocks matter for aggregate fluctuations.

Hulten's theorem

Before discussing how micro shocks can matter for macroeconomic fluctuations, it is useful to introduce a simple theorem by Hulten (1978). Let us consider a setting where there are N different sectors, indexed by i. Here, we use the terminology "sectors" because we would like to think of a competitive equilibrium. Using "firms" would yield the same result as long as firms behave as price-takers. Sector i's production is y_i. The production function is

$$y_i = a_i F(k_i, \ell_i, x_{i1}, x_{i2}, ..., x_{iN}),$$

where a_i is the TFP, x_{ij} is the sector-j product used in sector i, k_i is capital used in sector i, and ℓ_i is labor used in sector i. Note that the total sales $\sum_i p_i y_i$ is different from the total value added (i.e., the GDP), which is equal to $\sum_i p_i c_i$, because some of the output is used as intermediate goods. Let $Y = \sum_i p_i c_i$.

Hulten's (1978) theorem states that the first-order output effect of marginal changes in firm productivity levels is

$$\frac{dY}{Y} = \sum_i D_i \frac{da_i}{a_i},$$

where D_i is a sales share for sector i and labeled its Domar weight:

$$D_i = \frac{p_i y_i}{\sum_i p_i c_i}. \tag{22.31}$$

D_i's denominator is total value added, whereas the numerator is the sales of sector i. The proof of the theorem can be found in Appendix 22.A.5.

Two pieces of intuition are key to understanding Hulten's theorem. First, why does only a_i matter (and not information about inputs)? The reason is the envelope theorem. Because the economy achieves the first best, the input allocation is already optimized. Therefore, to a first-order approximation, the adjustment of inputs due to shocks to a_i does not have an impact on welfare. With homothetic utility, the welfare result can be mapped to GDP. Second, why is the numerator of the weight measured as sales? This question seems natural, especially because the denominator is in value added, which implies that the Domars weight do not necessarily sum up to 1. The intuition is that, when there are input-output networks, the improvement in TFP in a downstream firm also raises the value of intermediate inputs.

To see the second point more clearly, consider the following simple example. Suppose that there are two sectors, sectors 1 and 2. Sector 1 produces the consumption good, whose price is normalized to 1. The production function is $y_1 = a_1 x^{1-\gamma}\ell^{\gamma}$, where y_1 is the output, a_1 is the TFP, x is the intermediate input (purchased from sector 2), and ℓ is the labor input. The parameter γ is in (0,1). Sector 2 produces the intermediate good using capital: $y_2 = a_2 k$. The capital supply is fixed at K and the labor supply is fixed at L. In a competitive equilibrium, $x = y_2$ and we obtain $Y = a_1(a_2 K)^{1-\gamma}L^{\gamma}$, which gives

$$\frac{dY}{Y} = \frac{da_1}{a_1} + (1-\gamma)\frac{da_2}{a_2}. \tag{22.32}$$

Note that the weights in front of both TFP growth rates do not sum up to 1. To confirm Hulten's theorem, let us compute the Domar weight of each sector. Let the price of the intermediate good be p. In competitive equilibrium, $p = (1-\gamma)a_1 x^{-\gamma}L^{\gamma}$ holds, where $x = a_2 K$. Thus, the value added of sector 2 is

$$V_2 = (1-\gamma)a_1(a_2 k)^{1-\gamma}L^{\gamma}.$$

The value added of sector 1 is

$$V_1 = Y - px = \gamma a_1(a_2 k)^{1-\gamma}L^{\gamma}.$$

Thus, the Domar weight of sector 2 (because in this sector, sales are equal to value added) is $V_2/Y = 1-\gamma$. The Domar weight of sector 1 (because the sales are Y) is $Y/Y = 1$. Both correspond to the coefficients on the right-hand side of (22.32), confirming Hulten's theorem. Note that if we consider the value added instead of sales in sector 1, the coefficient is computed as $V_1/Y = \gamma$.

Large firms

As we discussed in Section 22.3, there are many large firms in the U.S. economy. In fact, the firm size distribution is "fat-tailed" and can be well approximated by a Pareto distribution at the right tail. Figure 22.9 reproduces the bin data plot of Figure 2A in Carvalho and Grassi (2019). It plots the size of a firm, measured by employment, against the firm size "percentile" (the fraction of firms whose size is larger than the value on the x-axis). The figure uses two datasets: BDS, in triangles (using the same data as is behind Figure 22.1), and Compustat data on publicly traded firms, in circles.[20] The Compustat data has large firms but the BDS table does not contain finer information on large firms, which is why both datasets are used in the figure. One can see that the plot with log-log axes is close to a straight line, indicating that the distribution is well approximated by a Pareto distribution.

Gabaix (2011) argues that when the distribution of firm size is fat-tailed, that is, there is a considerable presence of large firms (as in the Pareto distribution), the formula (22.30) implies a significant impact of σ on σ_Y. In particular, he considers the case where the firm size distribution is Pareto:

$$\Pr[y_i > x] = \chi x^{-\zeta}, \tag{22.33}$$

[20] The time period for the BDS data in Figure 22.9 is 1977–2012, whereas the time period for Figure 22.1 is 2019.

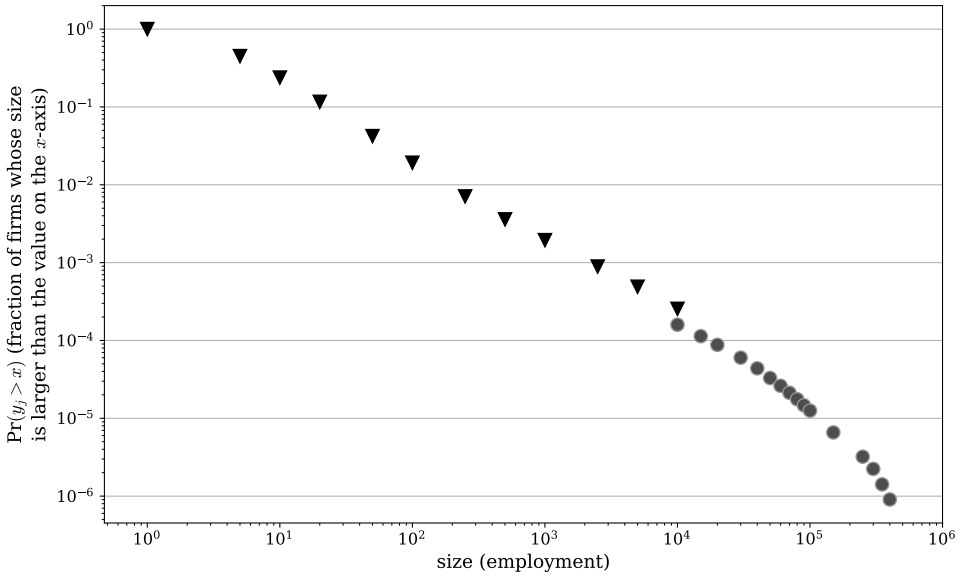

Figure 22.9: Distribution of firm size using log-log axes.

Source: Business Dynamics Statistics and Compustat. Reproduced from Carvalho and Grassi (2019).

where χ and ζ are constants. The variable y_i here is employment, and $\Pr[y_i > x]$ is the fraction of firms whose sizes are larger than x. The empirically relevant case for the U.S. is a ζ close to 1. (Because (22.33) implies $\log(\Pr[y_i > x]) = -\zeta \log(x) + \log(\chi)$, this can be seen from the slope of Figure 22.9 being close to -1.) For this case, he derives

$$\sigma_Y \sim \frac{v_\zeta}{\log(N)}\sigma$$

as $N \to \infty$, that is, $\sigma_Y \log(N)$ converges to $v_\zeta \sigma$ in distribution, where v_ζ is a random variable that does not depend on N or σ. This formula implies that aggregate volatility is proportional to $1/\log(N)$, rather than to $1/\sqrt{N}$. The function $1/\log(N)$ declines more slowly than $1/\sqrt{N}$ as N becomes large; for example, $1/\log(1,000,000) \approx 0.072$ whereas $1/\sqrt{1,000,000} = 0.001$. Gabaix (2011) calculates that the coefficient on σ in the above formula would be approximately 0.12 instead of 0.001. The effect of idiosyncratic shocks is, therefore, two magnitudes larger than for the identical-firm case; thus, the volatility induced by the idiosyncratic shock can have an effect of a similar magnitude to observed aggregate business-cycle volatility. In recent work, Carvalho and Grassi (2019) build a quantitative business cycle model like the one in Section 22.5 and analyze the business cycle dynamics driven by idiosyncratic shocks to large firms.

Effects of production networks

An additional important factor that can magnify the effect of idiosyncratic shocks on the aggregate economy is that the Domar weight (22.31) divides the firm sales by the aggregate value added. Many large firms, such as Walmart, Amazon, and GM, have significantly larger sales than their value added, magnifying their contribution to aggregate fluctuations. These

firms are at the downstream of the production network, and the impact of their productivity shocks on aggregate GDP therefore goes beyond the impact of their value added.

Some researchers examine potentially important further implications of the production network. Note that Hulten's theorem builds on two assumptions: (i) the economy is efficient, and (ii) the first-order approximation is sufficiently accurate. When these two assumptions are not appropriate, network structures can play a role in considering the aggregate effects of idiosyncratic shocks. For example, in a recent paper, Baqaee and Farhi (2019) argue that there are situations where the second-order effect is quantitatively important. In another paper, Baqaee and Farhi (2020) consider economies with distortions, and there the network structure also plays a role. In fact, Baqaee and Farhi (2020) derive a formula based on a first-order approximation that extends Hulten's theorem to a very general accounting framework, where the Solow residual can be decomposed into two terms, where one term captures how misallocation—as that discussed earlier in this chapter—changes over time and the remaining term describes "pure" technology growth. We now turn to the determinants of pure technology growth on the firm level.

22.8 Endogenous productivity

Given the importance of idiosyncratic shocks in generating heterogeneity across firms and their dynamics, it is natural to wonder what factors influence idiosyncratic productivity. One natural framework that can be used to analyze this issue is the models of endogenous productivity change, introduced in the economic growth chapter (Chapter 13). In particular, Klette and Kortum (2004) introduced a framework that generates entry, exit, expansion, and contraction of firms due to endogenous innovation. Below, we explain the Klette and Kortum (2004) model using discrete time.[21]

Consider an economy with a continuum of products on the unit interval $[0, 1]$. We will focus on the balanced-growth path of the economy, where all aggregate variables grow at the same rate. The representative consumer's utility function is

$$\sum_{t=0}^{\infty} \beta^t \log(C_t),$$

where $\beta \in (0, 1)$ is the discount factor and C_t is defined as

$$C_t = \exp\left(\int_0^1 \log\left(\sum_{k=-1}^{J_t(j)} q_t(j,k) c_t(j,k) \right) dj \right). \tag{22.34}$$

Here, $j \in [0, 1]$ is the index of the product. The aggregation with natural log implies that the elasticity of substitution across goods is 1. The index k is an integer that runs from -1 to $J_t(j)$. The value of k represents the generation of the product: a newer generation (i.e., a larger k) product is of higher quality. $J_t(j)$ is the cutting-edge (state-of-the-art) generation of good j. Generation k of product j (call it product (j, k) for short) contributes

[21] Ates and Saffie (2021) also develop a discrete-time version of the Klette and Kortum (2004) model.

to consumption in the form of $q_t(j,k)c_t(j,k)$. Here, $q_t(j,k)$ is the *quality* of product (j,k) and $c_t(j,k)$ is the *quantity* of product (j,k). The fact that the aggregation is additive across different generations implies that different generations are perfect substitutes. If generation k' has twice the quality of generation k, that is, $q_t(j,k')/q_t(j,k) = 2$, consuming one unit of product (j,k') is equivalent to consuming two units of product (j,k).

The consumer faces two layers of problems, one intratemporal and one intertemporal. The intratemporal problem is to determine how to allocate expenditure across different goods at each point in time. The intertemporal problem is to decide how much to spend across time.

Let us start with the intratemporal problem. Let E_t be the expenditure at period t and $p_t(j,k)$ be the price of product (j,k). First note that within product j, it is optimal to purchase only the generation whose quality-adjusted price, $p_t(j,k)/q_t(j,k)$, is the lowest. Therefore, let us only consider such a generation k for each j. Then, the intratemporal problem is

$$\max_{c_t(j,k)} \int_0^1 \log\left(q_t(j,k)c_t(j,k)\right) dj$$

subject to

$$\int_0^1 p_t(j,k)c_t(j,k)dj \leq E_t. \tag{22.35}$$

The solution, given the assumed preferences, is equal shares for all goods:

$$c_t(j,k) = \frac{E_t}{p_t(j,k)}. \tag{22.36}$$

Given the solution to the intratemporal problem, aggregate consumption in period t can be rewritten as

$$C_t = E_t \exp\left(\int_0^1 [\log(q_t(j,k)) - \log(p_t(j,k))]dj\right).$$

This relationship can be rewritten as

$$P_t C_t = E_t,$$

where the price index for consumption is

$$P_t \equiv \exp\left(\int_0^1 [\log(p_t(j,k)) - \log(q_t(j,k))]dj\right).$$

As in Section 22.6.1, normalize $P_t = 1$. This normalization implies

$$\int_0^1 \log(p_t(j,k))dj = \int_0^1 \log(q_t(j,k))dj \tag{22.37}$$

at any t.

The intertemporal problem is now

$$\max_{C_t} \sum_{t=0}^{\infty} \beta^t \log(C_t)$$

subject to

$$\sum_{t=0}^{\infty}\left(\frac{1}{1+r}\right)^{t} C_t \leq \mathcal{A}_0,$$

where $\mathcal{A}_0$ is the present discounted value of all future labor and asset incomes. The asset income in this economy comes from the claim to the profit of the firms. Here, we are imposing that the interest rate $r > 0$ is constant, as we will focus on a balanced growth path. The optimization results in a standard Euler equation:

$$\frac{1}{C_t} = \beta(1+r)\frac{1}{C_{t+1}}.$$

Along the balanced-growth path, C_{t+1}/C_t grows at a constant rate. Let us define $\gamma \equiv (C_{t+1} - C_t)/C_t$. Then,

$$\frac{1}{1+r} = \frac{\beta}{1+\gamma}. \tag{22.38}$$

Each product (j,k) is produced by one firm. However, one firm can own several product lines. A *firm* here is indeed defined as the collection of product lines that it produces: a small firm owns only a few product lines, and a large firm owns many product lines. Although firms are heterogeneous in this dimension, the analysis of the firm decision in the Klette and Kortum (2004) model is relatively simple because the model has a structure that allows each of the firm's product lines to make decisions independently. In the following, we exploit this property and analyze the firm's decisions at the product line level.

First, consider the production decision. Producing one unit of a product takes one unit of labor.[22] Thus, the unit production cost is w_t. Given that the production cost is the same, the cutting-edge producer (the "leader") has an advantage over other producers (with the older generation of product j). Because the demand elasticity is 1, the optimal pricing is to set the price as high as possible. Here, the cutting-edge producer cannot increase the price in an unlimited manner. Once the price is set sufficiently high, the $J_t(j)-1$ generation producer can enter profitably. The highest price the cutting-edge producer can charge is

$$p_t(j, J_t(j)) = \lambda w_t, \tag{22.39}$$

where $\lambda > 1$ represents the technology step $q_t(j,k+1)/q_t(j,k)$. Here λ coincides with the markup rate. This pricing behavior is called *limit pricing*. Thus, if this price is chosen, the closest follower would make zero profits if it produced, and we will assume it does not produce.

Given the price, the period profit from one product line for the leader is

$$\pi_t \equiv (p_t(j, J_t(j)) - w_t)\frac{C_t}{p_t(j, J_t(j))} = \left(1 - \frac{1}{\lambda}\right) C_t. \tag{22.40}$$

From (22.37) and (22.39),

$$\log(\lambda w_0) = \int_0^1 \log(q_0(j, J_0(j))dj.$$

[22] As we saw in Section 22.6.1, the production scale remains finite even with constant returns to scale because of the monopoly power.

By normalizing $q_0(j, J_0(j)) = \lambda$ for all j, this relationship implies $w_0 = 1$.

Second, consider the innovation decision. For each product line a firm has, it conducts innovation activity. For a product line, then, the firm decides on the intensity of innovation, η. Its required labor input is $R(\eta)$, where $R(\cdot)$ is an increasing and convex function. The cost of innovating at intensity η is therefore $w_t R(\eta)$. The intensity η represents the probability that the firm gains another product line. When innovation is successful, the firm can start producing using another product line. This newly-added product is assumed to be λ times better than the current cutting-edge product. The newly-added product is randomly chosen from $[0, 1]$ and each firm is infinitesimally small compared to the entire economy, the probability of innovating in its own product is zero. Hence, the new innovation always takes the market away from another firm. Because other firms also innovate, each firm can also lose product lines. Let μ be the probability that another firm innovates and takes over the current product line. Along the balanced-growth path, μ is constant over time.

Denoting the value of a leader product line by V_t, the Bellman equation for the firm is

$$V_t = \max_{\eta} \ \pi_t - w_t c(\eta) + \frac{1}{1+r}(1 + \eta - \mu)V_{t+1}. \tag{22.41}$$

The final term is the expected future value of the product line. There are four possible scenarios for the current leader of the product j: (i) it innovates and is not taken over by another firm; (ii) it fails to innovate and is taken over by another firm; (iii) both occur; and (iv) neither occurs. The probability of (i) is $\eta(1 - \mu)$ and the future value is $2V_{t+1}$. The probability of (ii) is $\mu(1 - \eta)$ and the future value is 0. The probability of (iii) is $\mu\eta$ and the future value is V_{t+1}. The probability of (iv) is $(1 - \mu)(1 - \eta)$ and the future value is V_{t+1}. Therefore, the expected future value is computed as $(1 + \eta - \mu)V_{t+1}$, which can be seen in the final term.

Along the balanced-growth path, V_t, π_t, and w_t all grow at the common gross rate $1 + \gamma$. Dividing both sides of (22.41) by $(1 + \gamma)^t$ and using (22.38), (22.40), and $w_0 = 1$,

$$v = \max_{\eta} \ \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1 + \eta - \mu)v, \tag{22.42}$$

where $v \equiv V_t/(1 + \gamma)^t$. The first-order condition is

$$R'(\eta) = \beta v. \tag{22.43}$$

There are many potential entrants in the economy. As in Section 22.5, we assume free entry. An entrant can hire c_e units of labor and enter: it then obtains a product line for sure. Because the number of product lines is fixed to $[0, 1]$, this means that it improves on an existing product line and takes it over from another firm. The free-entry condition is

$$V_t = w_t c_e.$$

Dividing by $(1 + \gamma)^t$, we obtain

$$v = c_e. \tag{22.44}$$

Let the entry rate (the amount of entry at each period) be ν. Given our assumptions, the fraction of product lines taken over, μ, will result from either entry or incumbent innovation, so that

$$\mu = \eta + \nu. \tag{22.45}$$

There are three types of labor demand: (i) production, (ii) innovation by incumbents, and (iii) entry. From (22.36) and $E_t = C_t$, production at time 0 is $c_0(j,k) = C_0/\lambda$ and thus labor demanded for production is C_0/λ. For innovation by incumbents, $R(\eta)$ units of labor are used. For entry, ν units are demanded. The labor supply is fixed at L. Thus, the labor-market equilibrium condition (using (22.45)) is

$$\frac{C_0}{\lambda} + R(\eta) + \nu = L. \tag{22.46}$$

In sum, the general equilibrium of the model involves four unknowns (v, η, C_0, and μ) with four equations:

$$v = \left(1 - \frac{1}{\lambda}\right) C_0 - R(\eta) + \beta(1 + \eta - \mu)v,$$

which is from (22.42), the first-order condition (22.43), the free-entry condition (22.44), and the labor-market equilibrium condition (22.46).

Notice that, given our functional-form assumptions, the economy's growth rate, γ, does not appear in any of the remaining equations. However, it is endogenous and can be computed. First, note that consumption C_t in (22.34) is equal to E_t, which is equal to aggregate expenditures on goods (see (22.35)). Along the balanced-growth path, the growth rate of E_t is also equal to the growth rate of w_t. From (22.37) and (22.39),

$$\begin{aligned}
(\gamma \approx) \log(w_{t+1}) - \log(w_t) &= \int_0^1 \log(p_{t+1}(j, J_{t+1}(j))) - \log(p_t(j, J_t(j)))dj \\
&= \int_0^1 \log(q_{t+1}(j, J_{t+1}(j))) - \log(q_t(j, J_t(j)))dj \\
&= \int_0^1 (\log(\lambda^{J_{t+1}(j)}) - \log(\lambda^{J_t(j)}))dj \\
&= \mathbb{E}[\log(\lambda^{J_{t+1}})] - \mathbb{E}[\log(\lambda^{J_t})] \\
&= \mathbb{E}[J_{t+1}]\log(\lambda) - \mathbb{E}[J_t]\log(\lambda) \\
&= \mu(t+1)\log(\lambda) - \mu t \log(\lambda) \\
&= \mu \log(\lambda).
\end{aligned}$$

The first equality follows from (22.39), the second is from (22.37), and the third is from the definition that $J_t(j)$ is the cutting-edge generation at industry j. In the fourth equality, we utilize the law of large numbers. Because each industry is subject to the i.i.d. shock and there is a continuum of industries, we can replace the cross-sectional average with the expected value when we interpret J_t as a random variable. The next inequality uses the fact that J_t is viewed as a sum of Bernoulli trials with winning probability μ (i.e., every period, the probability that a product is taken over by another firm is μ). The growth rate of the economy thus depends on μ, which is driven by the innovation intensity by incumbents (η) and by entrants (ν), as well as by the innovation step λ.

A major strength of the Klette and Kortum (2004) model over traditional endogenous growth models is that the definition of a firm is clear, allowing for the analysis of the dynamics of firms and the firm-size distribution. For the current discrete-time model, the analysis of firm-size distribution is somewhat more complex than for the Klette-Kortum

model in its original form (which is formulated in continuous time), because in a discrete-time formulation, many events can occur to a firm during the same period; the details are described in Appendix 22.A.6. Although the firm-size distribution is not analytically straightforward in the discrete-time version, it is straightforward to solve for it numerically on a computer. The advantage of this model over models with exogenous idiosyncratic shocks, such as the ones considered in Section 22.5, is that it allows us to analyze how policies and changes to the economic environment affect the productivity process itself.[23]

One statistic that we can compute in closed form is the average growth rate of the firm size. First, note that because each product line produces the same quantity and employs the same number of workers, the firm size distribution coincides with the distribution of product lines across firms. As can be seen in the discussion of (22.42), the average number of product lines next period per line this period is $1+\eta-\mu$. That is, the expected net growth rate of any firm's size is $\eta-\mu$. From (22.45), $\eta-\mu=-\nu<0$. Thus, the average growth rates of firms, small as well as large, is negative; the fact that average growth is negative is of course a consequence of the simplifying assumption that the total number of product lines is fixed and that there is entry (entrants of course grow, by definition). The property that large and small firms have a common average growth rate is called Gibrat's Law, although it is usually stated in the context of the positive average growth rate.

There are several counterfactual predictions from the Klette and Kortum (2004) model. Recent literature has made progress in modifying the model to replicate the salient features of the data. First, in the data, we associate firms with very advanced products making higher-than-average profits. In the model, a firm with a higher-quality product does not earn a higher profit on that product line than on its (or others') lower-quality lines. This feature stems from two assumptions: (i) the elasticity of substitution across goods is 1, and (ii) the technology is not cumulative. On the first point, the utility specification implies that the leader's revenue is the same regardless of prices and quality levels. With higher substitutability, both can matter for the size and profit of the firm. On the second point, any outside firm can innovate over the state-of-the-art product at the same cost as paid by the incumbent. Suppose the model is extended so that the incumbent firm can improve its own product quality. In such a case, there is an additional determinant of the size and profit differences (and, therefore, for idiosyncratic productivity shock distributions in Section 22.5) across firms.

Second, the firm-size distribution does not feature a Pareto tail. Intuitively, it is challenging to create many large firms in this economy because the firm size contracts on average. One alternative formulation is one where, instead of a negative growth rate, a large firm has a positive constant growth rate g. Suppose, in addition, that all firms experience an exit shock with the probability $\delta\in(0,1)$. Now consider a very large firm so that we can ignore the integer constraints on product lines. Let us start from the firms between size n and $n+\Delta$, where Δ is a small number relative to n. When the stationary density at n is $h(n)$, the mass of firms between these sizes is approximated by $h(n)\Delta$. In the next period, the surviving mass is $(1-\delta)h(n)\Delta$. After one period, size n will grow to $(1+g)n$ and size $n+\Delta$ will grow to $(1+g)(n+\Delta)$. Thus, the mass between these new sizes will be $(1+g)h((1+g)n)\Delta$.

[23] See Mukoyama and Osotimehin (2019) for an example of such policy analysis.

Therefore, in the stationary distribution,

$$(1+g)h((1+g)n)\Delta = (1-\delta)h(n)\Delta$$

has to hold. Guess that the distribution is Pareto: $h(n) = Fn^{-(\zeta+1)}$, where $F > 0$ and $\zeta > 0$ are parameters. In particular, ζ is the tail index that showed up in Section 22.7.2. The above equation can then be rewritten as

$$(1+g)F((1+g)n)^{-(\zeta+1)}\Delta = (1-\delta)Fn^{-(\zeta+1)}\Delta.$$

This equality holds for any n and Δ when

$$\zeta = -\frac{\log(1-\delta)}{\log(1+g)} > 0.$$

Thus, we verified that the firm dynamics with positive (and constant) growth, combined with a constant exit rate, can be consistent with a stationary distribution that is Pareto.[24] The tail index ζ is small (i.e., a thick tail) when δ is small or g is large. A natural question is then how we can modify the Klette and Kortum (2004) model so as to have positive firm growth rate at the right tail. One possibility is to break equation (22.45).[25] For example, if some innovation *creates* new products, there can be firm expansion without contributing to μ. Suppose that the new product creation among the total innovation is ξ (i.e., among the total $\eta+\nu$ innovation, ξ creates new products, and $\mu = \eta+\nu-\xi$ replaces existing products). Then, the average growth rate of a firm, which is still $\eta - \mu$, is now equal to $\xi - \nu$ (instead of just $-\nu$). If ξ is sufficiently large, $\xi - \nu$ can be positive.

[24] See Mukoyama and Osotimehin (2019) for a similar derivation.
[25] Luttmer (2011) discusses related insights.