NBER WORKING PAPER SERIES

HOW COSTLY ARE MARKUPS?

Chris Edmond
Virgiliu Midrigan
Daniel Yi Xu

Working Paper 24800
http://www.nber.org/papers/w24800

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

We particularly thank our discussants Salomé Baslandze and Gino Gancia for their insightful feedback. We also thank seminar participants at Duke University, the FRB New York, FRB St Louis, Keio University, the LSE, Notre Dame, NYU, the University of Adelaide, and the University of Chicago for their comments. Edmond thanks the Australian Research Council for financial support under grant DP-150101857. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

How Costly Are Markups?
Chris Edmond, Virgiliu Midrigan, and Daniel Yi Xu
NBER Working Paper No. 24800
July 2018
JEL No. D04,E02,L1,O4

ABSTRACT

We study the welfare costs of markups in a dynamic model with heterogeneous firms and endogenously variable markups. We find that the welfare costs of markups are large. We decompose the costs of markups into three channels: (i) an aggregate markup that acts like a uniform output tax, (ii) misallocation of factors of production, and (iii) an inefficiently low rate of entry. We find that the aggregate markup accounts for about two-thirds of the costs, misallocation accounts for about one-third, and the costs due to inefficient entry are negligible. We evaluate simple policies aimed at reducing the costs of markups. Subsidizing entry is not an effective tool in our model: while more competition reduces individual firms' markups it also reallocates market shares towards larger firms and the net effect is that the aggregate markup hardly changes. Size-dependent policies aimed at reducing concentration can reduce the aggregate markup but have the side effect of greatly increasing misallocation and reducing aggregate productivity.

Chris Edmond
Department of Economics
University of Melbourne
Parkville VIC 3010
AUSTRALIA
chris.edmond@gmail.com

Virgiliu Midrigan
Department of Economics
New York University
19 W. 4th St.
New York, NY 10012
and NBER
virgiliu.midrigan@nyu.edu

Daniel Yi Xu
Department of Economics
Duke University
213 Social Science Bldg
419 Chapel Drive
Box 90097
Durham, NC 27708-0097
and NBER
daniel.xu@duke.edu

1 Introduction

How large are the welfare costs of product market distortions? What kinds of policies can best overcome these distortions? We answer these questions using a dynamic model with heterogeneous firms and endogenously variable markups. In our model, the welfare costs of markups can be decomposed into three channels. First, the *aggregate markup* — the cost-weighted average of firm-level markups — acts like a uniform output tax levied on all firms. Second, there is cross-sectional markup dispersion because larger firms effectively face less competition and so charge higher markups than smaller firms. This markup dispersion gives rise to *misallocation* of factors of production across firms. Third, there is an *inefficiently low rate of entry*. Our goal is to quantify these three channels using US data and to evaluate policies aimed at reducing the costs of markups.

Our model features heterogeneous firms engaged in monopolistic competition with non-CES demand, as in Kimball (1995). Within a given industry, more productive firms are, in equilibrium, larger and face less elastic demand and so charge higher markups than less productive firms. As a consequence, changes in the environment that allow more productive firms to grow at the expense of less productive firms will be associated with an increase in the aggregate markup and a decline in the aggregate labor share. In this sense, our model is consistent with the literature's recent emphasis on the reallocation of production from firms with relatively high measured labor shares to firms with relatively low measured labor shares (Autor, Dorn, Katz, Patterson and Reenen, 2017a,b; Kehrig and Vincent, 2017) and the observation that firms with high markups have been getting larger, driving up the average markup (Baqaee and Farhi, 2018). Importantly, markups in our model are returns to past *sunk investments* in developing new products and in acquiring capital. Policies aimed at reducing markup distortions can have the unfortunate side-effect of distorting these investment decisions.

We calculate the welfare costs of markups by asking how much the representative consumer would benefit if the economy transitioned from a steady state with markup distortions to an efficient steady state. We calibrate the initial steady state to match the levels of concentration in sales in US data as well as the firm-level relationship between sales and the wage bill. We find that the total welfare costs of markups are large. For our benchmark calibration, the representative consumer would gain 7.5% in consumption-equivalent terms if they transitioned from the initial distorted steady state to the efficient steady state.

We then turn to quantifying the relative importance of the three channels by which markups reduce welfare in our model. We find that the aggregate markup distortion is the most important channel, accounting for about two-thirds of the total costs in our benchmark model. Misallocation accounts for about one-third of the total costs. The costs due to inefficient entry are negligible.

We calibrate our benchmark model to match an aggregate markup of 1.15, corresponding to the 2012 estimate of Barkai (2017) for the US private sector. Recently De Loecker and Eeckhout (2017) have estimated a 2012 economy-wide markup of about 1.6 based on Compustat data. Their economy-wide markup of 1.6 is the *sales-weighted* average of firm-level markups. But theory implies that it is the *cost-weighted* average of firm-level markups that is the relevant statistic that summarizes the distortions to employment and investment decisions. That is, the 'wedges' in the aggregate employment and investment optimality conditions are proportional to the cost-weighted average markup, not the sales-weighted average. When we calculate this cost-weighted average using the same Compustat data we obtain an aggregate markup of 1.25. In an alternative calibration of our model that matches this higher level of markups we find that the representative consumer would gain 18.6% in consumption-equivalent terms from the removal of all markup distortions, of which about four-fifths are due to the aggregate markup and one-fifth is due to misallocation. Since the Compustat data samples only the very largest firms in the US, we think of these larger losses as an upper bound on the total costs of markups.

Although the losses from misallocation in our model are sizeable, accounting for an aggregate TFP loss of about 1.2% in our benchmark model, they are nonetheless small relative to standard estimates in the literature (Restuccia and Rogerson, 2008; Hsieh and Klenow, 2009). Misallocation losses are relatively small because high-productivity firms who charge high markups do so precisely because they face low demand elasticities. With these low demand elasticities, the aggregate technology features a kind of '*near-satiation*' where there are strongly diminishing returns to increasing the output of an individual firm. This feature of the technology implies that a benevolent planner cannot achieve large gains by reallocating factors of production towards high-productivity firms. By contrast, if we incorrectly impose constant elasticity demand, as standard in the literature, then the dispersion of markups in our model would imply much larger losses from misallocation. With constant demand elasticities a planner *can* achieve large gains by reallocating factors of production towards high-productivity firms, so a given amount of markup dispersion is much more costly.

Our decomposition of the relative importance of the three channels by which markups reduce welfare also helps us evaluate policies aimed at reducing markups. A sufficiently sophisticated scheme of interventions can of course achieve the efficient allocation, but here we are interested in *simple* policies that may be more practical. One such policy is subsidizing entry (or reducing barriers to entry) so as to increase competition. We find that subsidizing entry is *not* an effective policy tool in our model. In particular, we find that even a large increase in the number of firms has a negligible effect on both the aggregate markup and the amount of misallocation.[1] To understand this, recall that the aggregate markup is a cost-weighted average of firm-level markups. An increase in the number of firms has two

[1]There are however standard love-of-variety gains from increasing the number of firms.

effects on this weighted average. The direct effect is a reduction in the markups of each firm. But there is also an important compositional effect. In our model, small firms face more elastic demand and are vulnerable to competition from entrants. Large firms face less elastic demand and are less vulnerable to competition. So when there is an increase in the number of firms, small, low markup firms contract by more than large, high markup firms and the resulting reallocation means high markup firms get relatively more weight in the aggregate markup calculation. In our model, this offsetting compositional effect is almost exactly as large as the direct effect so that overall the aggregate markup falls by a negligible amount.[2]

Another example of a simple policy is the use of size-dependent taxes to reduce within-industry concentration and thereby reduce the markups of large producers. We find that taxes that fall disproportionately on large firms, which is a simple way to model antitrust policy, can substantially reduce the aggregate markup in our model, but they come at considerable cost. This is because in our model the distorted allocation actually features *too little concentration* relative to the efficient allocation and a further reduction in concentration increases misallocation thereby reducing aggregate productivity.

This result has implications for the design of policy responses to the simultaneous rise in concentration and markups. Regardless of whether the rise in concentration and markups is due to changes in regulation, as in Peltzman (2014) and Grullon, Larkin and Michaely (2017), or changes in the scalability of technology, as in Haskel and Westlake (2017), or some mix of the two, size-dependent policies aimed at reducing concentration in order to bring down the overall level of markups may backfire because of the resulting increase in misallocation. Empirically, this also suggests that if the rise in concentration and markups observed in recent US data is due to a reduction in, say, antitrust enforcement, then it may be the case that the overall level of markups rose yet at the same time misallocation fell. This is speculative, but is consistent with Baqaee and Farhi (2018) who document that the increase in concentration and markups in the US has been accompanied by an improvement in allocative efficiency.

Finally we show that our key results are not driven by our assumptions about market structure. Our benchmark model uses *monopolistic competition* with non-CES demand. But in our robustness section we study an alternative model in which variable markups arise due to *oligopolistic competition* among a finite number of heterogeneous firms, as in Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015). When this model with oligopolistic competition is calibrated to match the same US concentration facts as our benchmark model, we again find that the losses from misallocation are relatively small and that even large increases in the number of firms have small effects on the aggregate markup and misallocation.

[2]These offsetting direct and compositional effects are reminiscent of results in the trade literature, e.g., Bernard, Eaton, Jensen and Kortum (2003) and Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2017).

Existing results on costs of markups. The starting point for discussion of the welfare costs of markups is Dixit and Stiglitz (1977), though the literature goes back to Lerner (1934). Recent work such as Zhelobodko, Kokovin, Parenti and Thisse (2012), Dhingra and Morrow (2016) and Behrens, Mion, Murata and Suedekum (2018) studies variable markups in static models with heterogeneous firms. In contrast, ours is a dynamic model where markups are returns to past investments. Though policies that reduce markups may be beneficial in the short run, they are costly overall because they depress the returns to such investments. Like us, Bilbiie, Ghironi and Melitz (2008) study a dynamic model and quantify the costs of markups but they assume a representative firm. We find, however, that accounting for firm heterogeneity plays a crucial role in evaluating policies aimed at reducing markup distortions.[3]

Markups and misallocation. In our model markups increase with firm size. This is one form of misallocation in the sense of Restuccia and Rogerson (2008), and Hsieh and Klenow (2009). We find that the losses from this form of misallocation are on the order of 1 to 3%. This suggests that size-dependent subsidies can increase aggregate productivity by at most 1 to 3%. We view these numbers as an upper bound on the gains from size-dependent subsidies since we attribute all of the systematic relationship between firm revenue productivity and firm size to market power, and not to, say, overhead costs as in Autor, Dorn, Katz, Patterson and Reenen (2017b) and Bartelsman, Haltiwanger and Scarpetta (2013). Because of this we are likely somewhat overstating the true relationship between markups and firm size and overstating the losses from this form of misallocation.

It is important to recognize that we abstract from all other sources of markup variation that may cause misallocation. Firms may operate in different locations or sell different products in different sectors and charge different markups depending on the amount of competition they face in those different markets. In principle policies that condition on location or other relevant market details may be able to address these forms of misallocation too. But finely-tuned policies that condition on details of market conditions location-by-location would require large amounts of information and may need to be adjusted frequently based on changing conditions. All this seems challenging in practice. For this reason we have limited our analysis to size-dependent markup variation and we find that the gains from eliminating misallocation due to size-dependent markup variation are likely no more than 2 to 3%.

The paper proceeds as follows. Section 2 presents the model. Section 3 characterizes the efficient allocation against which we assess the costs of markups. Section 4 explains how we calibrate the model to match US concentration facts. Section 5 presents our results on the costs of markups. Section 6 discusses the robustness of our results. Section 7 concludes.

[3]Other related work includes Atkeson and Burstein (2010, 2018) who provide a welfare analysis of innovation policies in firm dynamics models but who abstract from variable markups and Peters (2016) who studies innovation, firm dynamics, and variable markups but who does not evaluate the welfare costs of markups.

2 Model

The economy consists of a representative consumer with preferences over final consumption and labor supply and who owns all the firms. The final good is produced by perfectly competitive firms using a bundle of differentiated intermediate inputs. The differentiated inputs are produced by monopolistically competitive firms using capital, labor and materials. To enter the differentiated input market a firm must expend a fixed quantity of labor to develop a blueprint. Upon entry and after it learns its productivity, the firm makes a once-and-for-all decision about how much to invest in its capital. There is no aggregate uncertainty. We focus on characterizing the steady state and transitional dynamics after a policy change.

Representative Consumer. The representative consumer seeks to maximize

$$\sum_{t=0}^{\infty} \beta^t \Big(\log C_t - \psi \frac{L_t^{1+\nu}}{1+\nu} \Big), \tag{1}$$

subject to the budget constraint

$$C_t = W_t L_t + \Pi_t,$$

where C_t denotes consumption of the numeraire final good, L_t denotes labor supply, W_t denotes the real wage, and Π_t denotes aggregate firm profits, net of intangible investment and the cost of creating new firms. The representative consumer's labor supply satisfies

$$\psi C_t L_t^{\nu} = W_t.$$

Since firms are owned by the representative consumer they use the one-period discount factor $\beta C_t / C_{t+1}$ to discount future profit flows.

Final good producers. Let Y_t denote aggregate production of the final good. This can be used for consumption C_t, investment in intangible capital X_t, or as materials B_t, so that

$$C_t + X_t + B_t = Y_t.$$

The use of the final good as materials gives the model a simple *roundabout* production structure, as in Jones (2011) and Baqaee and Farhi (2018).

The final good Y_t is produced by perfectly competitive firms using a bundle of differentiated intermediate inputs $y_t(\omega)$ for $\omega \in [0, N_t]$ where N_t denotes the mass of available varieties. This bundle of inputs is assembled into final goods using the *Kimball aggregator*

$$\int_0^{N_t} \Upsilon\Big(\frac{y_t(\omega)}{Y_t}\Big)\, d\omega = 1, \tag{2}$$

where the function $\Upsilon(q)$ is strictly increasing, strictly concave, and satisfies $\Upsilon(1) = 1$. The CES aggregator is the special case $\Upsilon(q) = q^{\frac{\sigma-1}{\sigma}}$ for $\sigma > 1$.

Taking the prices $p_t(\omega)$ of the inputs as given and normalizing the price of the final good to 1, final good producers choose $y_t(\omega)$ to maximize profits

$$Y_t - \int_0^{N_t} p_t(\omega) y_t(\omega)\, d\omega,$$

subject to the technology (2). The optimality condition for this problem gives rise to the demand curve facing each intermediate producer

$$p_t(\omega) = \Upsilon'\Big(\frac{y_t(\omega)}{Y_t}\Big)\, D_t, \tag{3}$$

where

$$D_t := \left(\int_0^{N_t} \Upsilon'\Big(\frac{y_t(\omega)}{Y_t}\Big)\frac{y_t(\omega)}{Y_t}\, d\omega\right)^{-1} \tag{4}$$

is a *demand index*. In the CES case $\Upsilon(q) = q^{\frac{\sigma-1}{\sigma}}$ this index is a constant $D_t = \sigma/(\sigma-1)$ so that (3) reduces to the familiar constant elasticity demand curve $p_t(\omega) = (y_t(\omega)/Y_t)^{-\frac{1}{\sigma}}$.

Klenow-Willis specification. For our benchmark model we use the Klenow and Willis (2016) specification

$$\Upsilon(q) = 1 + (\sigma-1)\exp\left(\frac{1}{\varepsilon}\right)\varepsilon^{\frac{\sigma}{\varepsilon}-1}\left[\Gamma\left(\frac{\sigma}{\varepsilon}, \frac{1}{\varepsilon}\right) - \Gamma\left(\frac{\sigma}{\varepsilon}, \frac{q^{\varepsilon/\sigma}}{\varepsilon}\right)\right], \tag{5}$$

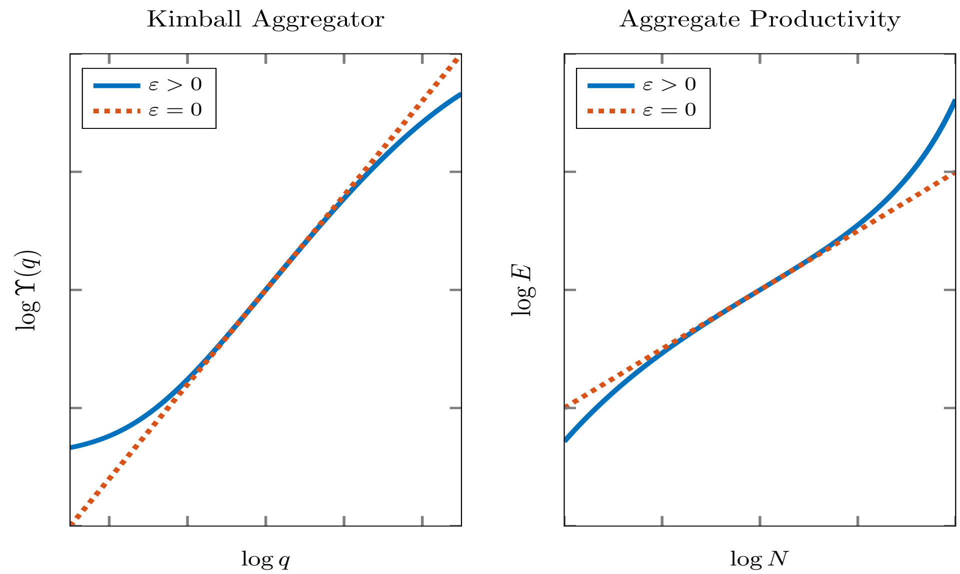
with $\sigma > 1$ and $\varepsilon \geq 0$ and where $\Gamma(s, x)$ denotes the upper incomplete Gamma function

$$\Gamma(s, x) := \int_x^\infty t^{s-1} e^{-t} dt.$$

The left panel of Figure 1 shows the shape of $\Upsilon(q)$. Setting $\varepsilon = 0$ gives the CES case $\Upsilon(q) = q^{\frac{\sigma-1}{\sigma}}$. When $\varepsilon > 0$, the elasticity of substitution is lower for firms with higher relative quantity $q = y/Y$, implying that larger firms choose higher markups. We view this as a parsimonious and tractable way of modeling the forces that arise in models of oligopolistic competition of the type studied by Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015). In those models larger firms face less competition in their own industries, have lower demand elasticities and choose higher markups. Indeed, as we show in our robustness section, many of the results in our setting with monopolisic competition extend to an environment with oligopolistic competition.

Love of variety. This specification of the production function implies *love of variety* in the sense that aggregate productivity increases with the number of firms. To see this, suppose that there are N firms in the economy with a constant returns technology in labor, $y = l$. Assuming a total labor L available for production, in a symmetric equilibrium $y = L/N$ so

Figure 1: Love of variety with Kimball aggregator



that the total amount of final output is given by $N\Upsilon(y/Y) = N\Upsilon(L/(NY)) = 1$. Aggregate productivity $E = Y/L$ is then implicitly determined by $N\Upsilon(1/(NE)) = 1$. In the CES special case $\varepsilon = 0$ we get the familiar solution $E = N^{\frac{1}{\sigma-1}}$. When $\varepsilon > 0$, aggregate productivity E is more sensitive to the number of varieties N, as shown in the right panel of Figure 1.

Intermediate input producers. Each variety ω is produced by a single firm. Firms are created by paying a sunk cost κ in units of labor. On entry, a new firm obtains a one-time productivity draw $e \sim G(e)$. Firms exit with exogenous probability δ per period. We focus on a symmetric equilibrium where producers with the same e will make the same decisions so henceforth we will simply index firms by e. On entry and after drawing e, a new firm makes a one-time irreversible investment in capital, $x_t(e)$. This capital does not depreciate, so the amount of capital available to a producer of age $i = 1, 2, ...$ is

$$k_{it}(e) = x_{t-i}(e).$$

The assumption that capital is chosen once-and-for-all is a simple way of introducing adjustment costs that prevent capital from reallocating across firms after policy reforms. This assumption also lets us capture investments in intangible capital, whose resale value is much lower than that of tangible capital (as in Haskel and Westlake, 2017).

A firm of age i and productivity e uses its capital $k_{it}(e) = x_{t-i}(e)$, hires labor l and purchases materials b to produce output according to

$$y_{it}(e) = e\, k_{it}(e)^{1-\eta}\, v_{it}(e)^{\eta}, \tag{6}$$

where v is a constant-returns-to-scale composite of the variable inputs

$$v = \left(\phi l^{\frac{\theta-1}{\theta}} + (1-\phi) b^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}}, \tag{7}$$

where ϕ determines the share of the two factors in production and θ is the elasticity of substitution between labor and materials.

We break the firm's problem into two steps, first solving a static profit maximization problem taking as given the initial investment, and then solving the firm's dynamic choice of whether to enter and how much capital to acquire at entry.

Static problem. First observe that a firm that chooses $v(e)$ units of the composite variable input will allocate that amongst labor and materials according to

$$l(e) = \phi^{\theta} \left(\frac{W}{P_v}\right)^{-\theta} v(e),$$

and

$$b(e) = (1-\phi)^{\theta} \left(\frac{1}{P_v}\right)^{-\theta} v(e)$$

where P_v is the unit price of the composite variable input

$$P_v = \left(\phi W^{1-\theta} + (1-\phi)\right)^{\frac{1}{1-\theta}}.$$

Each firm maximizes profits taking as given the production function (6) and the demand curve (3). Letting $z := e\, k_{it}(e)^{1-\eta}$ denote the firm's effective productivity, we can write the static profits of a firm of type z as

$$\pi(z) = \max_{y \geq 0} \left[\Upsilon'\left(\frac{y}{Y}\right) y - P_v \left(\frac{y}{z}\right)^{\frac{1}{\eta}} \right]. \tag{8}$$

Let $y(z)$ denote the solution to the firm's static problem and let $q(z) = y(z)/Y$ denote their relative output. The firm's price $p(z)$ can be written as a markup $\mu(q(z))$ over marginal cost

$$p(z) = \mu(q(z)) \times \frac{P_v}{\eta} \left(\frac{y(z)}{z}\right)^{\frac{1}{\eta}} \frac{1}{y(z)}. \tag{9}$$

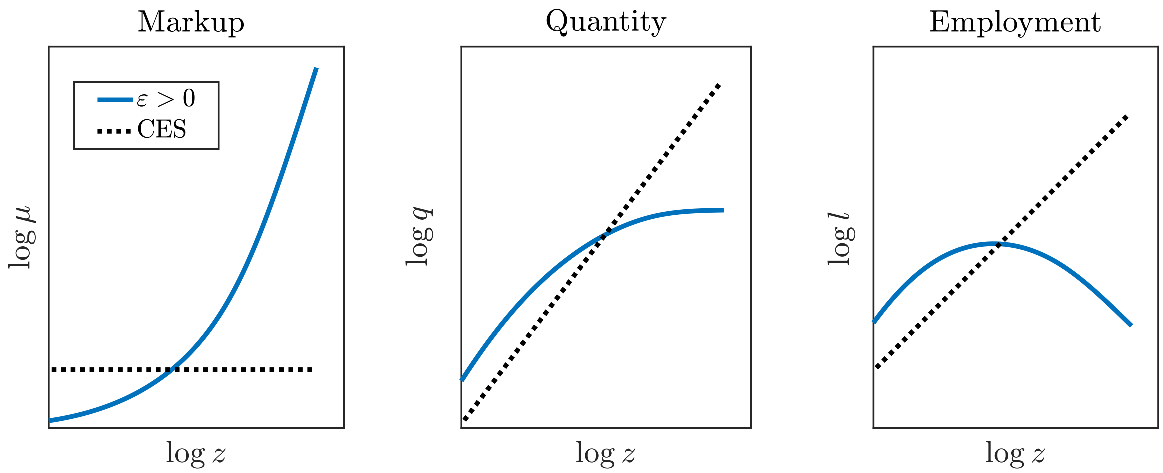
The Klenow-Willis specification in (5) gives

$$\Upsilon'(q) = \frac{\sigma-1}{\sigma} \exp\left(\frac{1-q^{\frac{\varepsilon}{\sigma}}}{\varepsilon}\right), \tag{10}$$

which implies the demand elasticity

$$-\frac{\Upsilon'(q)}{\Upsilon''(q)q} = \sigma q^{-\frac{\varepsilon}{\sigma}} \tag{11}$$

Figure 2: Static choices



which in turn implies the markup function

$$\mu(q) = \frac{\sigma}{\sigma - q^{\frac{\varepsilon}{\sigma}}}. \tag{12}$$

When $\varepsilon = 0$, this reduces to the familiar CES markup $\mu = \sigma/(\sigma - 1)$. When $\varepsilon > 0$, larger firms find it optimal to choose higher markups. The extent to which a firm's markup increases with its relative size is determined by ε/σ. The ratio of these two parameters is therefore critical in shaping how markups and quantities change with productivity and competition.

Figure 2 illustrates these static choices, plotting the markup $\mu(z)$, relative quantity $q(z)$ and employment $l(z)$, as a function of effective productivity z. When ε is relatively high, the markup increases more with productivity, implying that the quantity increases less with productivity. Indeed, when productivity is sufficiently high, employment may actually *decrease* with productivity because of strongly diminishing marginal revenue productivity.

Note that the firm's quantity choice is bounded. A profit-maximizing firm will not increase production to the point where the elasticity of demand from (11) is less than one. This implies a bound on the relative quantity equal to

$$q < \sigma^{\frac{\sigma}{\varepsilon}}.$$

The model therefore implies a threshold level of productivity $\bar{z}$ above which all firms produce the same amount of output and respond to an increase in productivity z by simply reducing the amount of variable inputs needed to produce a fixed amount of output.

Also note that

$$\pi(z) = p(z)y(z) - P_v v(z) \tag{13}$$

and we can rewrite the first order condition (9) as

$$\frac{P_v v(z)}{p(z)y(z)} = \frac{\eta}{\mu(q(z))}. \tag{14}$$

Since markups are increasing in relative size $q(z)$ this implies that a firm's variable input share in sales and well as the sales share of payments to each factor are decreasing in $q(z)$.

Dynamic problem. Now consider a firm at time t that has paid the sunk cost κW_t to enter and drawn $e \sim G(e)$. From (8), a firm with effective productivity z will have flow profits $\pi_{t+i}(z)$ at age $i = 1, 2, \ldots$. Choosing investment $x_t(e)$ at entry determines their effective productivity $z = e\, x_t(e)^{1-\eta}$ going forward and delivers the profit stream $\pi_{t+i}(e\, x_t(e)^{1-\eta})$ for $i = 1, 2, \ldots$. So having drawn e, a firm that enters at date t choose $x_t(e)$ to maximize

$$-\, x_t(e) + \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \pi_{t+i}\left(e\, x_t(e)^{1-\eta}\right), \tag{15}$$

where profits are discounted according to $\beta^i C_t / C_{t+i}$ and the firm exits at exogenous rate δ.

Using the definition of $\pi_t(z)$ from (8) and the envelope condition, the first order condition for $x_t(e)$ can be written

$$x_t(e) = \frac{1-\eta}{\eta} \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} P_{v,t+i}\, v_{t+i}(e\, x_t(e)^{1-\eta}), \tag{16}$$

where we make explicit the dependence of future sales (and therefore the variable input v_{t+i}) on the firm's initial investment. The solution to the fixed-point problem in (16) gives the firm's optimal investment choice $x_t(e)$. Using (14) we can also write this as

$$x_t(e) = (1-\eta)\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \frac{p_{t+i}(e) y_{t+i}(e)}{\mu_{t+i}(e)}, \tag{17}$$

where $\mu_{t+i}(e)$, say, is shorthand for $\mu_{t+i}(e x_t(e)^{1-\eta})$. This expression shows that the optimal investment is a function of the future sales scaled by the firm's markup at each future date.

Free-entry condition. Let M_t denote the mass of entrants in period t. Free entry drives the expected profits of potential entrants to zero. Since the sunk entry cost κW_t is paid prior to the realization of the productivity draw e, the free-entry condition is

$$\kappa W_t = \int \left(\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \pi_{t+i}\left(e\, x_t(e)^{1-\eta}\right) - x_t(e) \right) dG(e), \tag{18}$$

which, using (13), (14) and (16) can be written

$$\kappa W_t = \int \left(\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \left(1 - \mu_{t+i}(e)^{-1}\right) p_{t+i}(e) y_{t+i}(e) \right) dG(e). \tag{19}$$

In short, a firm's incentives to enter are determined by its operating profits, net of investment, and are therefore a function of markups and the firm's overall sales. Both markups and a firm's sales decrease with additional entry so that entry occurs to the point at which the expected profits are equal to the cost of creating a new variety.

Equilibrium. Let $H_t(z)$ denote the measure of firms with effective firm productivity z in period t. Let $N_t = \int dH_t(z)$ denote the overall mass of firms in period t. Given an initial measure $H_0(z)$, a recursive equilibrium is a sequence of firm prices $p_t(z)$ and allocations $y_t(z)$, $v_t(z)$, $l_t(z)$, $b_t(z)$, $x_t(z)$, mass of new entrants M_t, wage rate W_t, aggregate output Y_t, consumption C_t, and labor supply L_t, as well as measure of effective productivity $H_t(z)$, such that firms and consumers optimize and the labor and goods markets all clear.

The total mass of firms evolves according

$$N_{t+1} = (1-\delta)N_t + M_t,$$

while the measure of effective productivity evolves according to

$$H_{t+1}(z) = (1-\delta)H_t(z) + M_t \int \mathbb{1}\{\, e\, x_t(e)^{1-\eta} \leq z \,\}\, dG(e), \tag{20}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

Labor market clearing requires

$$L_t = \int l_t(z)\, dH_t(z) + M_t\kappa. \tag{21}$$

Similarly, goods market clearing requires

$$Y_t = C_t + M_t \int x_t(e)\, dG(e) + \int b_t(z)\, dH_t(z), \tag{22}$$

where the second-last term on the RHS reflects investment by the new entrants and the last term on the RHS reflects purchases of materials by all firms.

Aggregation. We now derive an aggregate production function for this economy and show how aggregate productivity and the aggregate input choices relate to the cross-sectional distribution of markups. These aggregation results motivate a two-step approach that we use to compute an equilibrium. First, given a distribution $H_t(z)$ of individual firms' effective productivity, we solve for the relative quantities $q_t(z) = y_t(z)/Y_t$ that maximize firm profits. Second, given these choices, we solve for all aggregate prices and quantities.

Let Z_t denote the *aggregate productivity* of this economy, implicitly defined by an aggregate production function that relates the total amount of final goods Y_t to the total amount of the composite variable input V_t used in production:

$$Y_t = Z_t V_t^{\eta}. \tag{23}$$

Here $V_t = \int v_t(z)\, dH_t(z)$ is an aggregate index of variable inputs given by

$$V_t = \left[\phi \tilde{L}_t^{\frac{\theta-1}{\theta}} + (1-\phi) B_t^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}}, \tag{24}$$

where $\tilde{L}_t = \int l_t(z)\, dH_t(z)$ denotes the quantity of labor *used in production*.

Let $\mathcal{M}_t$ denote the *aggregate markup* of this economy, implicitly defined by

$$\frac{P_{v,t} V_t}{Y_t} = \frac{\eta}{\mathcal{M}_t}. \tag{25}$$

This aggregate markup acts like a wedge in the choice of variable inputs and reduces the share of payments to variable factors below their production elasticity η. In turn, $\mathcal{M}_t$ reduces the share of each variable input. For example, the share of labor in production is

$$\frac{W_t \tilde{L}_t}{Y_t} = \frac{\eta}{\mathcal{M}_t} \times \phi^{\theta} \left(\frac{W_t}{P_{v,t}}\right)^{1-\theta}. \tag{26}$$

Some algebra shows that the aggregate productivity Z_t can be expressed in terms of firm-level productivities z according to

$$Z_t = \left(\int \left(\frac{q_t\,(z)}{z}\right)^{\frac{1}{\eta}} dH_t\,(z)\right)^{-\eta}. \tag{27}$$

The aggregate markup $\mathcal{M}_t$ is a *cost-weighted* arithmetic average[4] of firm-level markups

$$\mathcal{M}_t = \int \mu_t(z)\, \frac{v_t(z)}{V_t}\, dH_t(z). \tag{28}$$

We find it instructive to further decompose aggregate productivity Z_t into a term that captures the exogenous efficiency of individual producers and a term that summarizes their past investment choices. To this end, let $n_{it} = (1-\delta)^{i-1} M_{t-i}$ denote the measure of surviving producers of age i in period t. Aggregate capital K_t is

$$K_t = \sum_i n_{it} \int k_{it}(e)\, dG(e)$$

where $k_{it}(e) = x_{t-i}(e)$. We can then write the aggregate production function

$$Y_t = E_t K_t^{1-\eta} V_t^{\eta},$$

where aggregate productivity is $Z_t = E_t K_t^{1-\eta}$ and where E_t is a measure of *aggregate efficiency*

$$E_t = \left(\sum_i n_{it} \int \frac{q_{it}(e)}{e}\, dG(e)\right)^{-1},$$

that is, a quantity-weighted harmonic average of firm-level efficiency e.

[4]Or a sales-weighted *harmonic* average as in Edmond, Midrigan and Xu (2015). See Appendix A.

Solution algorithm. We now outline how we solve the model. We use the aggregation results above to calculate the aggregate production function and evaluate the representative consumer's optimality conditions, which are functions solely of aggregate variables, including the aggregate markup $\mathcal{M}_t$ and productivity Z_t. Given a sequence of $\mathcal{M}_t$ and Z_t we can solve for the equilibrium of this economy at each date. We also note that for a given measure of producers $H_t(z)$, computing $\mathcal{M}_t$ and Z_t is relatively straightforward. In particular, we can scale the profit function in (8) by the demand index D_t and aggregate output Y_t and write

$$\tilde{\pi}_t(z) = \max_{q \geq 0} \left[\Upsilon'(q)q - A_t \left(\frac{q}{z} \right)^{\frac{1}{\eta}} \right], \tag{29}$$

where A_t is an aggregate statistic that summarizes the conditions relevant for an individual firm, in particular

$$A_t := \frac{P_{v,t}}{D_t} Y_t^{\frac{1-\eta}{\eta}}.$$

We find the optimal relative quantity $q(z, A)$ for a firm of type z for any arbitrary value of A by solving the first order condition

$$\Upsilon'(q(z, A))q(z, A) = \mu(q(z, A)) \frac{A}{\eta} \left(\frac{q(z, A)}{z} \right)^{\frac{1}{\eta}}. \tag{30}$$

We then pin down the equilibrium value of A_t using the Kimball aggregator

$$\int \Upsilon(q(z, A_t)) \, dH_t(z) = 1,$$

which then gives us the equilibrium relative quantities, demand index, and prices

$$q_t(z) = q(z, A_t),$$

$$D_t = \left(\int \Upsilon'(q_t(z)) q_t(z) \, dH_t(z) \right)^{-1},$$

$$p_t(z) = \Upsilon'(q_t(z)) D_t \, .$$

From these we can compute the aggregate markup $\mathcal{M}_t$ and productivity Z_t.[5]

Given an initial conjecture for the sequence of $H_t(z)$ during the transition, we can compute the aggregate prices and quantities at each date and then use these, together with the free entry condition (19) and an entrant's optimal investment choice (17), to obtain an updated sequence $H_t(z)$. We then iterate on the implied fixed-point problem in the sequence $H_t(z)$ until convergence.

[5] See Gopinath and Itskhoki (2010) and Amiti, Itskhoki and Konings (2017) for more details on solving for the equilibrium in this setting.

Steady state entry and capital stock. To build intuition, we briefly characterize the steady-state capital stock $K = N \int x(e)\, dG(e)$ and mass of firms N. Using (17) and aggregating across all firms gives

$$\frac{K}{Y} = \frac{1-\eta}{\frac{1}{\beta} - 1 + \delta}\,\frac{1}{\mathcal{M}}, \tag{31}$$

so that the steady state capital stock is distorted by the aggregate markup $\mathcal{M}$, just as the static choices are. Similarly, evaluating (19) at steady state and simplifying gives

$$\frac{N}{Y} = \frac{1}{\kappa W}\,\frac{1}{\frac{1}{\beta} - 1 + \delta}\left(1 - \frac{1}{\mathcal{M}}\right), \tag{32}$$

where the first term is the inverse of the cost of entering and the second and third term give the expected discounted value of entering, which increases with the aggregate markup.

3 Efficient allocation

In this section we derive the efficient allocation in our economy by considering the problem of a benevolent planner who faces the same technological and resource constraints as in the decentralized economy. Comparing the efficient allocation chosen by the planner to the decentralized allocation reveals three channels through which markups distort outcomes in the decentralized economy: (i) the aggregate markup acts like a uniform output tax, (ii) markup dispersion gives rise to misallocation of factors of production, and (iii) markups distort the entry margin.

Planner's problem. The planner chooses how many varieties to create, how to allocate factors of production, how much to invest, consume, and work so as to maximize the representative consumer's utility subject to the resource constraints for labor (21) and goods (22), the law of motion for the distribution of productivity (20), the production functions (6) and (7), and the Kimball aggregator (2). The initial condition for this problem is the initial distribution of productivities $H_0(z)$.

We use asterisks to denote the planner's allocation. It turns out to be convenient to solve the planner's problem by expressing aggregate output as a function of the history of past entry M^*_{t-i} and investment x^*_{t-i} choices. With this change of variables, the planner's problem can be written as maximizing

$$\sum_{t=0}^{\infty} \beta^t \left(\log C^*_t - \psi \frac{(\tilde{L}^*_t + \kappa M^*_t)^{1+\nu}}{1+\nu} \right) \tag{33}$$

subject to the resource constraint for goods

$$C^*_t + X^*_t + B^*_t = \left(\sum_{i=1}^{\infty} (1-\delta)^{i-1} M^*_{t-i} \int \left(\frac{q^*_{it}(e)}{e\, x^*_{t-i}(e)^{1-\eta}} \right)^{\frac{1}{\eta}} dG(e) \right)^{-\eta} V(\tilde{L}^*_t, B^*_t)^{\eta}, \tag{34}$$

and the Kimball aggregator

$$\left(\sum_{i=1}^{\infty} (1-\delta)^{i-1} M_{t-i}^* \int \Upsilon\left(q_{it}^*(e)\right) \, dG(e)\right) = 1, \tag{35}$$

where $q_{it}^*(e)$ is the relative quantity of a productive unit that began i periods earlier with draw e. In writing these two constraints we have used the constant exit rate δ and the expression for aggregate productivity Z_t in equation (27).

We again break the problem into two steps, first solving a static allocation problem and then determining the remaining variables.

Planner's static allocation. To determine $q_{it}^*(e)$ we first recognize that it is sufficient to determine $q_t^*(z)$, since age i only matters through the choice of initial investment, which in this notation is summarized by z. Then let $\lambda_{1,t}^*$ denote the multiplier on the planner's resource constraint (34) and $\lambda_{2,t}^*$ denote the multiplier on the Kimball aggregator (35). The first order condition that determines $q_t^*(z)$ can be written

$$\Upsilon'(q_t^*(z))q_t^*(z) = A_t^* \left(\frac{q_t^*(z)}{z}\right)^{\frac{1}{\eta}}, \tag{36}$$

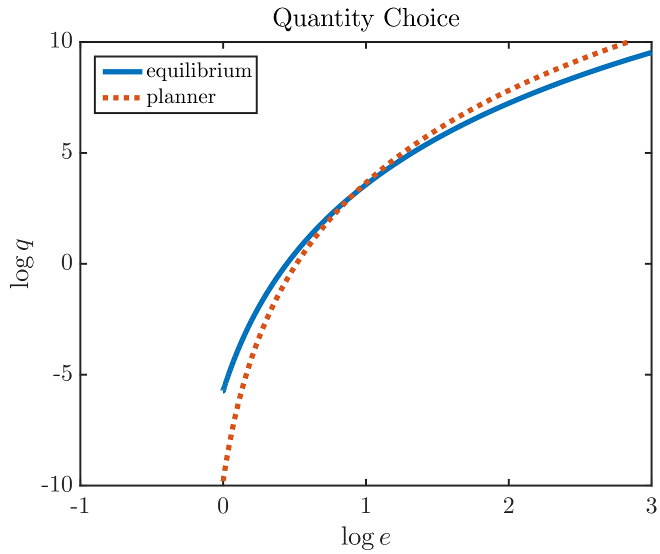
where

$$A_t^* = \frac{\lambda_{1,t}^*}{\lambda_{2,t}^*} Y_t^* Z_t^{*\frac{1}{\eta}}. \tag{37}$$

As in the decentralized equilibrium, the distribution of individual productivities only affects $q_t^*(z)$ through the aggregate A_t^*. We can therefore solve (36) for an arbitrary value of this statistic and then find the specific value of A_t^* that satisfies the Kimball aggregator (35).

Misallocation. Comparing the equilibrium allocation in (30) and the planner's allocation in (36) reveals the misallocation among existing firms in the decentralized equilibrium. Since more productive firms have higher markups, they produce too little compared to the social optimum and employ too little of the variable factors. Figure 3 illustrates the misallocation by comparing the relative sizes of firms in the decentralized equilibrium to the relative size the planner would allocate for them. The planner's allocation is not log-linear in productivity, as it would be with CES demand. The extra concavity reflects strongly diminishing marginal productivity as relative quantity increases. As we discuss below, this feature of the model implies that the gains from reallocating factors of production are not as high in this economy as would be the case in an economy with CES demand (for a given distribution of markups).

Figure 3: Equilibrium and planner's allocations compared



Planner's initial investment choice. Now consider the planner's choice of how much to invest in each new variety. Using $\lambda^*_{1,t} = 1/C^*_t$ and $X^*_t = M^*_t \int x^*_t(e)\, dG(e)$, the planner's first order condition for investment $x_t(e)$ can be written

$$x^*_t(e) = (1-\eta)\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C^*_{t+i}}{C^*_t}\right)^{-1} Y^*_{t+i} Z^{*\frac{1}{\eta}}_{t+i} \left(\frac{q^*_{t+i}(e)}{z^*_{t+i}(e)}\right)^{\frac{1}{\eta}}, \tag{38}$$

where $z^*_{t+i}(e) = e\, x^*_t(e)^{1-\eta}$ and $q^*_{t+i}(e)$ is shorthand for $q^*_{i,t+i}(z^*_{t+i}(e))$, etc.

This implies that the planner's steady state capital/output ratio is

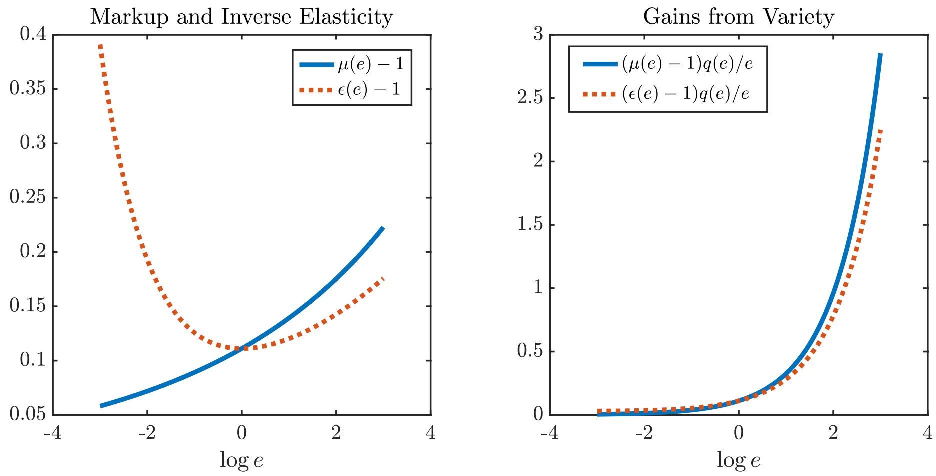
$$\frac{K^*}{Y^*} = \frac{1-\eta}{\frac{1}{\beta} - 1 + \delta}. \tag{39}$$

Comparing this with the steady state capital/output ratio in the decentralized equilibrium, given in (31) above, we see that the planner's capital/output ratio is higher than in the decentralized equilibrium. In short, the decentralized equilibrium features too little investment because of the aggregate markup distortion.

Planner's choice of new varieties. Now consider the planner's choice of new varieties M^*_t. As shown in Appendix B, the optimality condition that determines M^*_t can be written

$$\kappa\psi C^*_t L^{*\nu}_t = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C^*_{t+i}}{C^*_t}\right)^{-1} \int \left[\epsilon^*_{t+i}(e) - 1\right] p^*_{t+i}(e) y^*_{t+i}(e)\, dG(e), \tag{40}$$

Figure 4: Entry choice



where we define

$$\epsilon^*_{it}(e) := \frac{\Upsilon\left(q^*_{it}(e)\right)}{\Upsilon'\left(q^*_{it}(e)\right) q^*_{it}(e)},$$

and

$$p^*_{it}(e) := \frac{\Upsilon'\left(q^*_{it}(e)\right)}{\int \Upsilon'(q^*_t(z)) q^*_t(z)\, dH^*_t(z)}.$$

The term $\epsilon^*_{it}(e)$ is the *inverse elasticity* of the Kimball aggregator $\Upsilon(q)$ evaluated at the planner's allocation for a particular variety $q^*_{it}(e)$. The term $p^*_{it}(e)$ is the social value of an additional unit of that variety, i.e., the planner's counterpart to the market price.

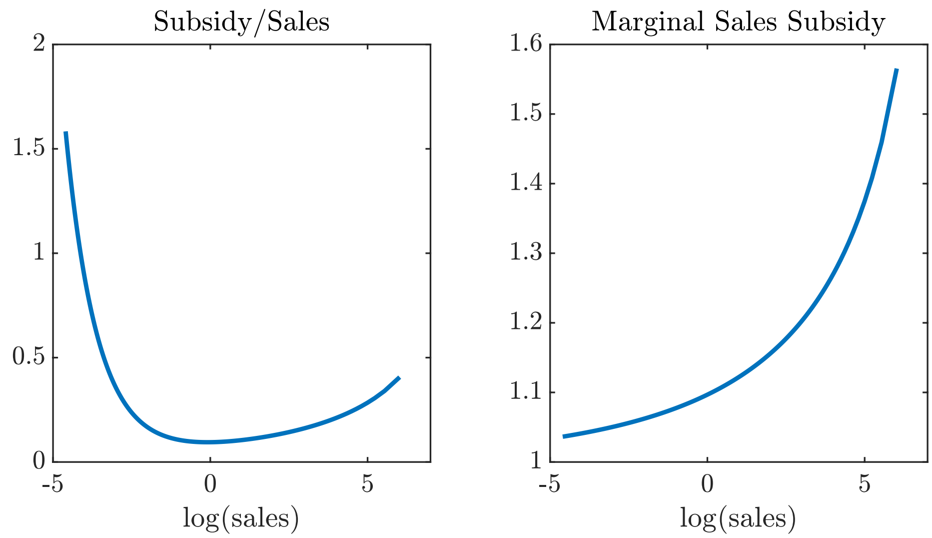
Comparing the free-entry condition in the decentralized equilibrium (19) to the planner's entry condition (40), we recover an important insight of Bilbiie, Ghironi and Melitz (2008), Zhelobodko, Kokovin, Parenti and Thisse (2012) and Dhingra and Morrow (2016): the planner's incentives to create new varieties are determined by the inverse elasticity $\epsilon(e)$ of the aggregator while the incentives for new firms to enter are determined by their markups $\mu(e)$. CES demand is the knife-edge special case where these incentives coincide, i.e., where $\mu(e) = \epsilon(e) = \sigma/(\sigma - 1)$ for all e. Figure 4 plots markups $\mu(e)$ and the inverse elasticity $\epsilon(e)$ against productivity. Low productivity firms have low markups and do not value entry as much as the planner values their entry. High productivity firms have high markups and value entry more than the planner does.

In steady state, the mass of varieties chose by the planner is given by

$$\frac{N^*}{Y^*} = \frac{1}{\frac{1}{\beta} - 1 + \delta} \frac{1}{\kappa \mathrm{MPL}^*} \frac{\int \left(\epsilon^*(e) - 1\right) \frac{q^*(e)}{e}\, dG(e)}{\int \frac{q^*(e)}{e}\, dG(e)},$$

where MPL^* denotes the marginal product of labor for the planner. Likewise, in steady state

Figure 5: Optimal size-dependent subsidy



the mass of firms in the decentralized equilibrium is given by

$$\frac{N}{Y} = \frac{1}{\frac{1}{\beta} - 1 + \delta} \frac{1}{\kappa \mathrm{MPL}} \frac{\int (\mu(e) - 1) \frac{q(e)}{e} \, dG(e)}{\int \frac{q(e)}{e} \, dG(e)}.$$

Whether the N/Y ratio is too low or too high compared to the efficient allocation is ambiguous and depends on precise details of the parameterization.

To summarize, misalignment between the planner's and the firms' incentives to enter is another source of inefficiency in this economy. The amount of entry in the decentralized equilibrium is determined by the firm's expected markups, which do not coincide with the planner's marginal valuation of new varieties except in the special case of CES demand.

Implementation. One way to implement the planner's allocations in the decentralized equilibrium is to subsidize production. Suppose that each firm receives a size-dependent subsidy $T(s)$ that depends on the amount the firm sells, $s = py$. It is straightforward to show that the planner's allocation can be implemented by setting a subsidy equal to

$$T(s) = \Upsilon\left(\frac{s}{p(s)Y}\right) DY - s$$

where D is the demand index in (4) and $p(s)$ is the firm's price. This subsidy ensures that the private incentives to produce, invest, and enter are aligned with those of the planner.

Figure 5 illustrates the shape of the subsidy function. The left panel shows the average subsidy, $T(s)/s$. Since $\Upsilon(0) > 0$, the optimal subsidy is positive even if the firm does not produce at all, $T(0) > 0$.[6] This lump-sum component of the subsidy ensures that the amount of entry is optimal and implies the average subsidy is U-shaped in the amount the firm sells. The right panel shows the marginal subsidy $T'(s)$ which, unlike the average subsidy, *is* strictly increasing in size. This is because the marginal subsidy is equal to the desired markup of a firm of that size, $T'(s) = \sigma/(\sigma - q(s)^{\varepsilon/\sigma})$, which increases in the firm's relative size.

4 Quantifying the model

In this section we first outline our calibration strategy and our model's implications for the cross-sectional distribution of markups. We then calculate the aggregate productivity losses due to misallocation in this economy. Our calibrated model features a considerable amount of markup dispersion and standard methods for estimating the losses due to misallocation based on a constant demand elasticity, such as Hsieh and Klenow (2009), would conclude that misallocation losses are large. But as we show, these standard methods based on a constant demand elasticity significantly overstate the actual misallocation losses in our model.

4.1 Calibration strategy

The level and dispersion of markups in our model depend crucially on three underlying parameters: (i) the average elasticity of demand σ, (ii) the sensitivity of a firm's demand elasticity to its relative size, as determined by the '*super-elasticity*' parameter ε, and (iii) the amount of productivity dispersion. For parsimony and as is standard in the literature we assume that the distribution of productivity $G(e)$ is Pareto with tail parameter ξ.

Intuitively, σ pins down the overall level of markups, ε pins down how markups μ_i and hence a firm's wage bill, $w_i l_i \sim p_i y_i/\mu_i$ vary with firm sales, and the Pareto tail parameter ξ pins down the concentration of firm sales. We choose these three parameters by simultaneously matching, for the 2012 US economy, an estimate of the aggregate markup, key moments of the distribution of sales for firms in 6-digit NAICS industries, and the relationship between a firm's wage bill (payroll) and sales for the 2012 US economy.[7]

4.1.1 Assigned parameters

We assume that a period is one year and set the discount factor $\beta = 0.96$ and exit rate $\delta = 0.1$. We set the inverse of the Frisch elasticity of labor supply to $\nu = 1$. We normalize the disutility from labor supply ψ and the entry cost κ to achieve a steady-state output of

[6] Since $\Upsilon'(0) < \infty$, and we assume constant returns to capital and variable inputs, there is a cutoff level of productivity $\underline{e}$ below which the firm does not produce since its price would be above the *choke price*. Whether there is indeed a mass of firms whose productivity falls below this cutoff depends on the exact parameterization of the model.

[7] See https://www.sba.gov/advocacy/firm-size-data, which contains data on total sales, the number of firms, and total wage bill for firms in about 15 revenue-based size classes.

Table 1: Parameterization

Panel A: Assigned parameters

β	discount factor	0.96
ν	labor supply elasticity	1
δ	exit rate	0.10
η	variable input elasticity	0.865
ϕ	weight on labor	0.676
θ	elasticity subs. between labor & materials	0.50

Panel B: Calibrated parameters

		benchmark	high ε/σ	high $\mathcal{M}$
σ	average elasticity	11.55	15.36	6.52
ε	super-elasticity	2.18	6.14	0.79
ξ	Pareto tail	6.66	6.69	4.17

$Y = 1$ and a steady-state total mass of firms $N = 1$ for our benchmark economy. We set the elasticity of substitution between materials and labor equal to $\theta = 0.5$, and set the weight on labor in production $\phi = 0.676$ to match the 45% share of materials in total sales in the US private business sector in 2012. Finally, we set the elasticity of the variable input, $\eta = 0.865$ to match the 0.15 ratio of private non-residential investment to private sector value added in 2012 in the US. We report all these parameter choices in Panel A of Table 1.

4.1.2 Calibrated parameters

In this section we explain how we choose values for the key parameters σ, ε, and ξ that determine the amount of concentration in sales and level and dispersion of markups.

Aggregate markup. First, we require that our model matches an aggregate markup of $\mathcal{M} = 1.15$, corresponding to the estimate of Barkai (2017) for 2012. Intuitively, this moment pins down the average elasticity σ.

Distribution of relative sales. Second, we require that our model matches the unweighted and weighted (by firm sales) distribution of *relative sales* of firms in each 6-digit industry. We define relative sales as the average sales of firms in a given size class and industry relative to the average sales of all firms in that industry. For brevity, from now on we refer to a group of firms in a given size class as *firms*. We pool observations of relative sales across all industries and report moments of this distribution in the left column of Table 2.

Table 2: Distribution of Relative Sales

Panel A: Unweighted

	US data	benchmark	high ε/σ	high $\mathcal{M}$
fraction of firms with				
relative sales ≤ 0.1	0.329	0.366	0.442	0.209
relative sales ≤ 0.5	0.761	0.747	0.692	0.738
relative sales ≤ 1	0.877	0.848	0.798	0.852
relative sales ≤ 2	0.942	0.916	0.888	0.923
relative sales ≤ 5	0.979	0.968	0.965	0.971
relative sales ≤ 10	0.990	0.987	0.991	0.988
relative sales ≤ 50	0.999	0.999	1.000	0.999
relative sales ≤ 100	1.000	1.000	1.000	1.000

Panel B: Sales-weighted

	US data	benchmark	high ε/σ	high $\mathcal{M}$
fraction of sales in firms with				
relative sales ≤ 0.1	0.019	0.026	0.014	0.020
relative sales ≤ 0.5	0.088	0.128	0.091	0.156
relative sales ≤ 1	0.154	0.211	0.183	0.250
relative sales ≤ 2	0.271	0.323	0.338	0.364
relative sales ≤ 5	0.507	0.509	0.630	0.537
relative sales ≤ 10	0.660	0.661	0.847	0.671
relative sales ≤ 50	0.951	0.928	1.000	0.910
relative sales ≤ 100	0.978	0.977	1.000	0.963

Now consider Panel A of Table 2 which summarizes the unweighted distribution of relative sales. In the data, 32.9% of all firms have average sales that are less then 1/10th of the industry average. The vast majority of firms, some 87.7%, sell less then their industry average. About 1% of all firms have sales that exceed 10 times the industry average and about 0.1% of all firms sell more than 50 times the industry average. Now consider Panel B of Table 2 which summarizes the sales-weighted distribution. The 32.9% smallest firms that have relative sales below 1/10th of their industry average account for a total of 1.9% of overall sales in the US. The 87.7% smallest firms that have relative sales below their industry average together account for 15.4% of overall sales. Finally, the 1% of the firms whose sales exceed 10 times their industry average account for a share of $1 - 0.66 = 0.34$ or 34% of total sales and the 0.1% of firms whose sales exceed 50 times their industry average account for a share of $1 - 0.951 = 0.049$ or 4.9% of total sales. We choose the Pareto tail ξ to minimize the distance between all these moments in the data and the model.

Relationship between relative wages and relative sales. We now discuss the set of moments that allow us to pin down the super-elasticity ε. We calculate, for each size class in each industry, the relative wage bill of firms in that size class, defined as the average wage bill of firms in that size class in that industry relative to the average wage bill of firms in that industry. The model implies that the relative wage bill of firm i depends on its relative sales and relative markup according to

$$\text{relative wage bill}_i = \frac{\text{relative sales}_i}{\text{relative markup}_i}.$$

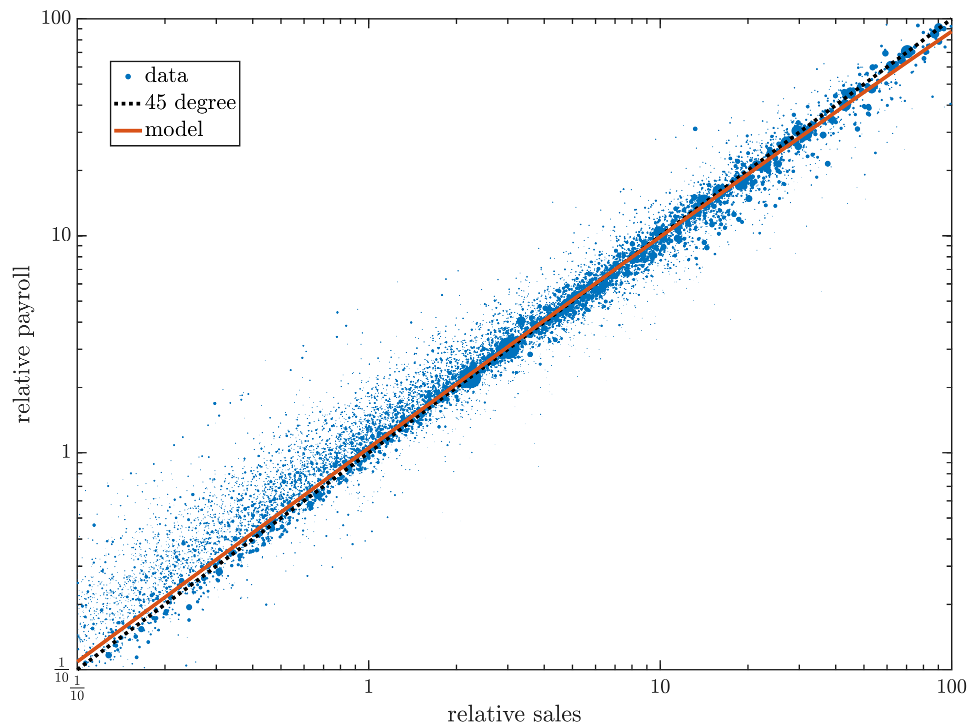
If ε is equal to zero, markups do not increase with firm sales and the relative wage bill would increase one-for-one with relative sales. But if ε is positive, markups increase with firms sales implying that the relative wage bill increases less than one-for-one with sales. The extent to which the relative wage bill increases with relative sales is therefore informative about the extent to which markups increase with firm size. By expressing both the wage bill and sales in relative terms we are effectively subtracting industry-specific differences in production functions (in say η or ϕ) and using within-industry variation to identify ε.

Figure 6 shows the relationship between the relative wage bill and relative sales in the data. Each circle corresponds to one size class in a given industry and the diameter of the circle indicates the total sales accounted for by firms in that particular size class. The dotted line in the figure is the 45-degree line, which corresponds to an economy with $\varepsilon = 0$ in which markups do not systematically increase with size. Though the pattern is difficult to see from simply eyeballing the data, the relative wage bill increases less than one-for-one with sales. The slope coefficient of a regression, weighted by firm sales, is equal to 0.964 when we restrict the sample to firms with relative sales greater than 1 (0.970 for the full sample). Larger firms thus have a smaller share of payments to labor, a pattern which our model interprets as evidence that markups increase with sales.

There are of course other plausible explanations for such a pattern. For example, a fixed (overhead) component to a firm's wage bill would also imply that larger firms have a disproportionately low wage bill.[8] Similarly, this pattern could arise if larger firms outsource

[8]See Autor, Dorn, Katz, Patterson and Reenen (2017b); Bartelsman, Haltiwanger and Scarpetta (2013).

Figure 6: Relative wage bill vs. relative sales



a larger fraction of their activities or have a larger capital share. In this sense we view our estimates as providing an *upper bound* on how rapidly markups increase with size.

We now explain how we use this evidence to estimate the super-elasticity ε. Our model implies a non-linear relationship between the relative wage bill and relative sales which is a function of ε and the other parameters

$$\log(\text{relative wage bill}_i) = F(\log(\text{relative sales}_i)\,;\varepsilon),$$

with a higher ε implying a flatter slope. We can use this relationship to calculate what the model predicts a firm's relative wage bill should be given its relative sales for any given ε in the steady state of the model. We then choose ε to minimize the distance between the model's prediction and the actual relative wage bill observed in the data

$$\sum_i \omega_i \left[\log\left(\text{relative wage bill}_i^{\text{data}}\right) - F\left(\log\left(\text{relative sales}_i^{\text{data}}\,;\varepsilon\right)\right)\right]^2,$$

where ω_i is the overall sales share of firms in each size class.

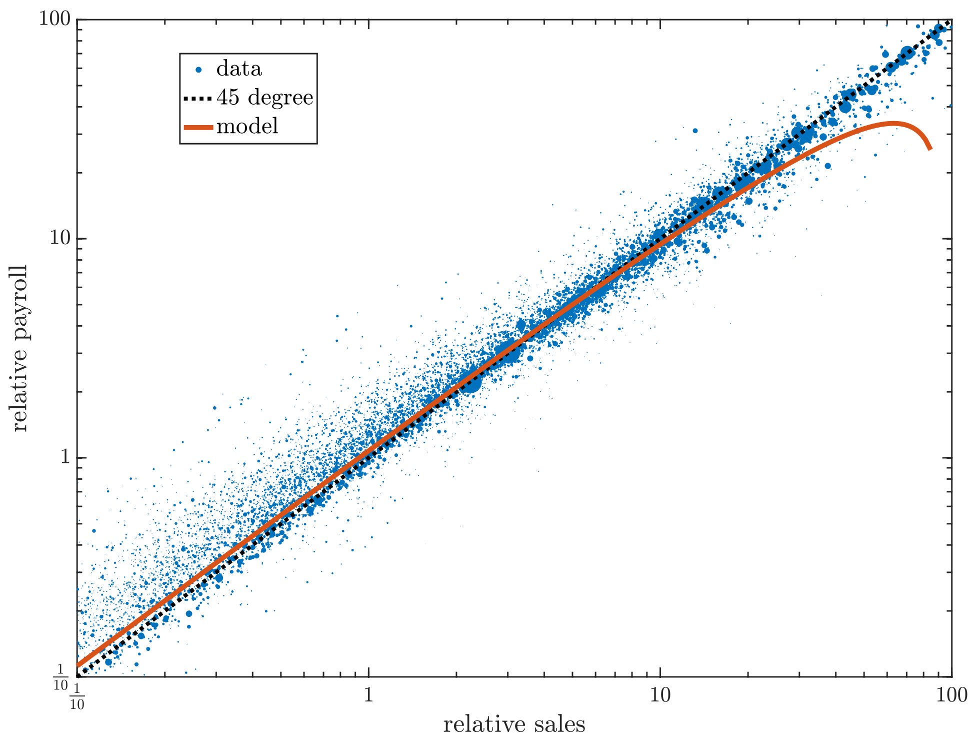
To summarize, we jointly choose the parameters σ, ε and ξ to (i) match a 15% aggregate markup, (ii) match the distribution of relative sales summarized in Table 2, and (iii) minimize the distance between the model's predictions for a firm's wage bill as a function of its relative sales and the corresponding observations in the data. For our benchmark calibration we pool together data from all industries. In our robustness section below we provide alternative estimates of ξ and ε based on various NAICS industries.

Model Fit. Panel B of Table 1 above reports the parameter values that minimize our objective function. The elasticity of substitution σ is equal to 11.55, the super-elasticity ε is equal to 2.18, while the Pareto tail coefficient ξ is equal to 6.66. Though our estimate of $\varepsilon/\sigma = 0.189$ is much lower than typically assumed in macro studies that attempt to match the response of prices to changes in monetary policy or exchange rates, it is in line with the micro estimates surveyed by Klenow and Willis (2016). In our robustness section below we present alternative estimates of ε/σ derived from more disaggregated product-level data on markups and sales for the panel of Taiwanese manufacturing firms studied by Edmond, Midrigan and Xu (2015). We find that a ratio of about ε/σ of about 0.15 best fits the cross-sectional relationship between markups and market size in that micro data, an estimate very close to that produced by our benchmark calibration strategy.

With these parameters the model matches the aggregate markup of 15% exactly. Table 2 shows that the model reproduces well the concentration in industry sales observed in the data, especially at the top. For example, in the data the fraction of firms that sell at least 10 times more than their industry average is equal to 1% and these firms account for 34% of all sales. In the model the fraction of firms that sell at least 10 times more than their industry average is equal to 1.3% and these firms account for 33.9% of all sales. Finally, the solid line in Figure 6 shows the model's predictions for how the relative wage bill varies with relative firm size. Recall that in the data the slope coefficient of a regression, weighted by firm sales, is equal to 0.964 when we restrict the sample to firms with relative sales greater than 1. The corresponding elasticity in the model is 0.965.

Alternative 'high ε/σ' calibration. The magnitude of ε/σ is critical for the model so we now provide some intuition as to how this ratio is identified. In particular, we report results for an alternative 'high ε/σ' calibration with $\varepsilon/\sigma = 0.4$, about twice as large as our benchmark. For this experiment we re-calibrate the Pareto tail parameter ξ to match the size distribution of firms and the elasticity parameter σ to match the 15% aggregate markup while continuing to assign all other parameters as before. We report the re-calibrated parameter values and the model's fit for the distribution of sales in Table 1 and Table 2 above. Importantly, this version of the model provides much worse fit to the concentration in sales at the top. Since high ε/σ means that markups are increasing rapidly with size, this version of the model predicts too few large firms compared to the data. Figure 7 shows this economy's predictions for how the relative wage bill changes with relative sales. The poor fit at the top

Figure 7: Alternative high ε/σ calibration provides worse fit



of the distribution is evident. With a higher ε the model implies much higher markups for the largest firms and therefore a much smaller wage bill relative to sales and is incapable of reproducing the relationship between the wage bill and sales in the data.[9]

4.2 Markup distribution

Our model's implications for the steady-state markup distribution are given in Panel A of Table 3. Here we report the aggregate markup $\mathcal{M}$, i.e., the cost-weighted average of individual markups, and the cost-weighted percentiles of the markup distribution for both our benchmark and the alternative high ε/σ calibration. We also compare our model's implications to estimates of markups from the publicly available Compustat data for the US for 2012. To calculate these, we follow the approach of De Loecker and Eeckhout (2017) using the ratio of sales to the cost of goods sold, scaled by estimates (at the 2-digit industry level) of the cost elasticity in the production function from Karabarbounis and Neiman (2018).[10]

[9] We have also used the methods proposed by Andrews, Gentzkow and Shapiro (2017) to examine this argument more formally. We find that the derivate of the moments for wage bill relative to sales and relative

Table 3: Steady state implications

Panel A: Distribution of markups (cost-weighted)

	Compustat	benchmark	high ε/σ	high $\mathcal{M}$
aggregate markup, $\mathcal{M}$	1.26	1.15	1.15	1.25
p25 markup	0.97	1.10	1.08	1.18
p50 markup	1.12	1.14	1.13	1.24
p75 markup	1.31	1.19	1.19	1.31
p90 markup	1.69	1.24	1.27	1.39

Panel B: Productivity losses from misallocation

	benchmark	high ε/σ	high $\mathcal{M}$
actual losses, $\log(E^*/E) \times 100$	1.2	2.1	1.3
losses with CES and $\bar{\sigma} = \frac{\mathcal{M}}{\mathcal{M}-1}$	8.4	16.7	8.7

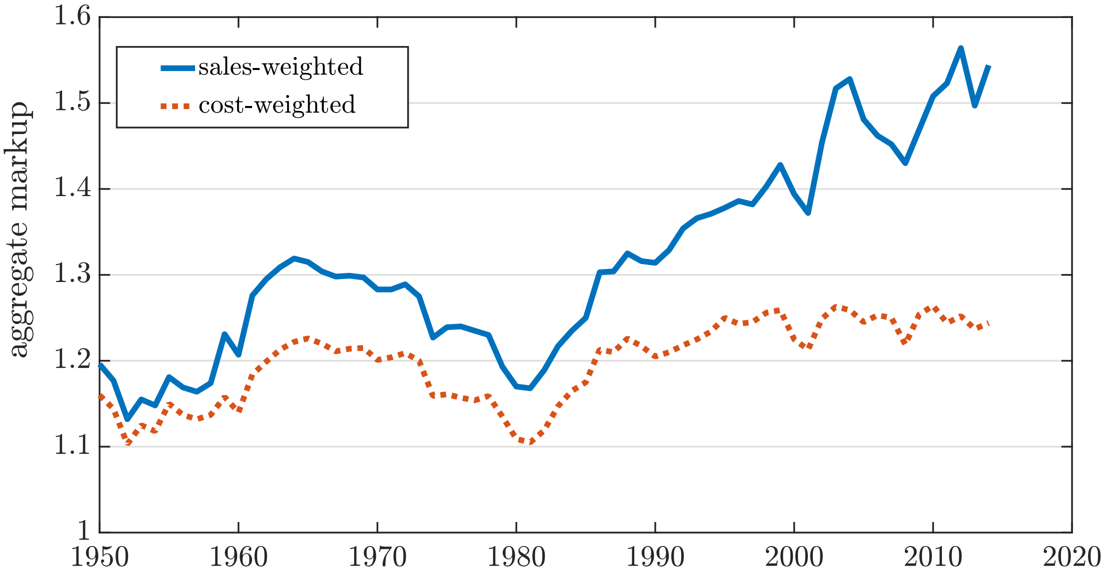
The distribution of markups in our benchmark model ranges from 1.1 at the 25th percentile to 1.24 at the 90th percentile. The dispersion of markups increases very little in the high ε/σ economy which implies a 25th percentile of 1.08 and a 90th percentile of 1.27. The Compustat data implies an aggregate markup of 1.26, larger than our calibration target of 1.15. The Compustat data also implies more dispersed markups than in our model. We do not find these discrepancies between the model and the data critical for two reasons. First, the sample of Compustat firms includes only a subset of the very largest firms in the US, those that are publicly traded. By contrast, our calibration uses the estimates of the aggregate markup from Barkai (2017) for the entire US private sector. Second, the ratio of sales to costs in the data may reflect distortions other than markups (for example credit constraints) or perhaps may vary across firms due to non-convexities, differences in technologies or costs of adjusting factors of production. Indeed, we find that most of the markup dispersion in the Compustat data is not systematically related to firm size.[11]

sales concentration are both very responsive to the value of ε/σ and in this sense ε/σ is locally identified.

[10] If we include selling, general and administrative expenses (SGA) in our measure of costs, as Traina (2018) does, then the cost-weighted average markup falls from 1.26 to 1.21.

[11] See also Hall (2018) who finds that the average US markup (weighted by value-added shares) increased from 1.12 in 1988 to 1.38 in 2015 in KLEMS data.

Figure 8: Cost-weighted vs. sales-weighted average markups, Compustat data



Solid blue line shows the sales-weighted average of firm-level markups, as in De Loecker and Eeckhout (2017). Dashed red line shows the cost-weighted average of firm-level markups. The former has increased by a larger amount, but the latter is the relevant measure of the aggregate distortion to first-order conditions that results in welfare losses.

Our observation that the aggregate markup in the 2012 Compustat data is about 1.26 may seem to contradict the findings of De Loecker and Eeckhout (2017) who report an aggregate markup of about 1.6. But there is in fact no contradiction. The measure of aggregate markup we construct is the cost-weighted average of individual markups (equivalently, the *harmonic* sales-weighted average), since this is the object that distorts the aggregate first-order conditions and results in welfare losses. By contrast, De Loecker and Eeckhout (2017) report the *arithmetic* sales-weighted average of markups. As Figure 8 shows, the latter has increased much more in the last several decades than the cost-weighted average, owing to an increase in markups at the top of the distribution.[12] Viewed through the lens of our model, the increase in the sales-weighted average markup overstates the increase in the distortions to inputs because it implicitly overstates the amount of inputs hired by high-markup firms.

Alternative 'high $\mathcal{M}$' calibration. Our benchmark calibration targets an aggregate markup of $\mathcal{M} = 1.15$ corresponding to the estimates from Barkai (2017) for the entire US private sector. To assess the sensitivity of our results to this target for the overall level of market power we also report results for an alternative 'high $\mathcal{M}$' calibration that targets $\mathcal{M} = 1.25$ corresponding to the cost-weighted average of markups in the Compustat data.

[12] See also Figure B.4(b) in De Loecker and Eeckhout (2017) which reports a very similar pattern.

We re-calibrate the parameters σ, ε and ξ to match this higher aggregate markup but keeping our other target moments the same as in our benchmark.[13] We report the re-calibrated parameter values and the model's fit for the distribution of sales in Table 1 and Table 2 above. This version of the model provide a similar fit to the concentration data as our benchmark, though it predicts slightly too many firms at the very top when compared to the data (or our benchmark model). As shown in Panel A of Table 3 above, this version of the model implies larger and more dispersed markups than our benchmark model but does not fully capture the dispersion in markups in the Compustat data.

4.3 Implications for misallocation

The markup dispersion generated by our model implies that there are aggregate productivity losses due to misallocation. How large are these losses due to misallocation? As shown in Panel B of Table 3 above, aggregate productivity E in the steady state of our benchmark economy is 1.2% below the level of aggregate productivity that could be achieved by a planner facing the same technology and resource constraints who could optimally reallocate all factors of production (including capital) across producers. Since the high ε/σ calibration implies larger and more dispersed markups, it implies a larger 2.1% loss from misallocation.

Our benchmark 1.2% loss from misallocation is an economically substantial effect but is much smaller than the losses from misallocation that have featured prominently in the literature, e.g., Restuccia and Rogerson (2008) and Hsieh and Klenow (2009).[14] We now show that we can reconcile our findings with these estimates in the literature by recognizing that existing estimates typically rely on the assumption of CES demand, whereas our calculations use the actual demand system implied by the Kimball aggregator. We show that incorrectly assuming CES demand implies losses from misallocation that are much larger.

Misallocation with a CES aggregator. Suppose a researcher uses data on the distribution of markups $\mu(e)$ generated by our model but incorrectly specifies the final goods aggregator to be of the CES, rather than the Kimball form:

$$Y = \left(\int y(e)^{\frac{\bar{\sigma}-1}{\bar{\sigma}}} \, dG(e) \right)^{\frac{\bar{\sigma}}{\bar{\sigma}-1}},$$

for some constant elasticity $\bar{\sigma} > 1$ where intermediate goods are produced with a constant returns technology in, say, labor and in which firms differ in their productivity e, so that

[13] We also adjust the weight on labor in production to $\phi = 0.610$ and the elasticity of the variable input to $\eta = 0.853$ so that we continue to match our targets for the share of materials in total sales and the share of investment in total value added.

[14] See also Baqaee and Farhi (2018) who proposes an alternative non-parametric approach to calculating the evolution of these misallocation losses over time.

$y(e) = el(e)$. Then aggregate labor productivity in the efficient allocation is given by

$$E^* = \left(\int e^{\bar{\sigma}-1} \, dG(e) \right)^{\frac{1}{\bar{\sigma}-1}},$$

while the actual level of productivity

$$E = \frac{\left(\int \left(\frac{e}{\mu(e)} \right)^{\bar{\sigma}-1} \, dG(e) \right)^{\frac{\bar{\sigma}}{\bar{\sigma}-1}}}{\int \left(\frac{e}{\mu(e)} \right)^{\bar{\sigma}} \frac{1}{e} \, dG(e)},$$

is below that under the efficient allocation whenever markups are dispersed.

In Panel B of Table 3 we report the loss implied by comparing E and E^* calculated using the CES formula. For our benchmark model we set $\bar{\sigma} = \frac{\mathcal{M}}{\mathcal{M}-1} = 7.67$, i.e., the constant elasticity that rationalizes our model's aggregate markup of $\mathcal{M} = 1.15$. Assuming a CES aggregator we would conclude that misallocation losses are 8.4%, almost 7 times larger than the actual loss of 1.2% implied by the Kimball aggregator. For our high $\mathcal{M} = 1.25$ economy we set $\bar{\sigma} = \frac{\mathcal{M}}{\mathcal{M}-1} = 5$. This gives similar results to our benchmark. The CES formula implies a loss of 8.7% whereas the true loss implied by the Kimball aggregator is 1.3%. For our high ε/σ economy markups are more dispersed so the CES formula implies a loss from misallocation of 16.7%, much larger than the 2.1% true loss implied by the Kimball aggregator.

Why does the CES measurement overstate the true misallocation losses? The CES measurement overstates the gains the planner could achieve by reallocating factors of production from low e firms to high e firms. To understand this, observe that the true demand elasticity with the Kimball technology is

$$-\frac{\Upsilon'(q)}{\Upsilon''(q)q} = \sigma q^{-\frac{\varepsilon}{\sigma}},$$

which implies that with the Kimball technology the planner encounters strongly diminishing marginal product from allocating more factors to firms that already have high q. Loosely speaking, it is as if the planner encounters a form of '*near-satiation*'. It is of course precisely this form of near-satiation that leads high e firms in the decentralized equilibrium to charge high markups. For high q firms lowering prices generates few additional sales so higher productivity simply translates to higher markups. The CES assumption interprets these high markups as a great potential source of gains from reallocation because it does not recognize that reallocating factors towards such firms will run into the same strongly diminishing marginal product that generates high markups in the first place.

The key point is that explicitly modeling the source of markup variation has important implications for inferring their costs. Dispersion in markups may not necessarily be as costly as implied by CES calculations which do not take an explicit stand on the underlying source of the distortions in the firms' optimality conditions. Of course, these results reflect a very specific source of markup variation, namely Kimball demand. But in our robustness section below we show that similar conclusions are reached in an alternative model of oligopolistic competition calibrated to match the same concentration facts.

5 How costly are markups?

We now present our main quantitative results on the welfare costs of markups. We answer two questions. First, how large are the total welfare costs of markups in our economy? Second, what is the relative importance of the three channels by which markups distort allocations?

We answer the first question by asking how much the representative consumer would benefit from a full implementation of the efficient allocation that eliminates markup distortions, taking all of the transitional dynamics into account. We find that the total welfare costs of markups are large. Implementing the efficient allocation results in a consumption equivalent welfare gain of about 7.5%.

We answer the second question by removing each of the three channels in isolation using offsetting subsidies. We show that a uniform output subsidy of 15% that exactly offsets the aggregate markup goes a long way towards achieving full efficiency, removing about two-thirds of the total costs of markups. We show that size-dependent taxes and subsidies that remove misallocation while keeping the aggregate markup unchanged remove about one-third of the total costs of markups. Although the channels are not strictly additive, this suggests there is not much scope for large gains from correcting distortions to the entry margin. Indeed we find that subsidizing entry removes at most about 1/10th of the total cost of markups.

5.1 Total cost of markups

We first contrast the distorted steady state in our decentralized equilibrium to that chosen by a planner, then calculate the transitional dynamics of the economy from the initial distorted steady state to the efficient steady state, and finally calculate the welfare gains from implementing the efficient steady state taking these transitional dynamics into account.

Steady state comparison. Table 4 compares the distorted steady state to the efficient steady state. In the efficient steady state output is higher by 37%, consumption by 30.5%, and employment by 16.7% relative to the distorted steady state. The efficient steady state also calls for more product variety, the mass of firms is higher by 15.7%. The capital stock is 50.9% higher. Aggregate efficiency is 3.8% higher. As discussed above, misallocation only

Table 4: Steady state allocations under alternative policies

	efficient	uniform output subsidy	remove misallocation	entry subsidy
log deviation from benchmark, $\times 100$				
output, Y	37.0	33.0	3.2	4.5
consumption, C	30.5	24.9	4.1	6.3
employment, L	16.7	15.5	0.6	3.4
mass of firms, N	15.7	6.4	3.2	20.3
capital, K	50.9	47.0	3.2	4.6
aggregate efficiency, E	3.8	1.1	1.7	3.4
welfare gains, CEV, %	7.5	4.9	2.5	0.7

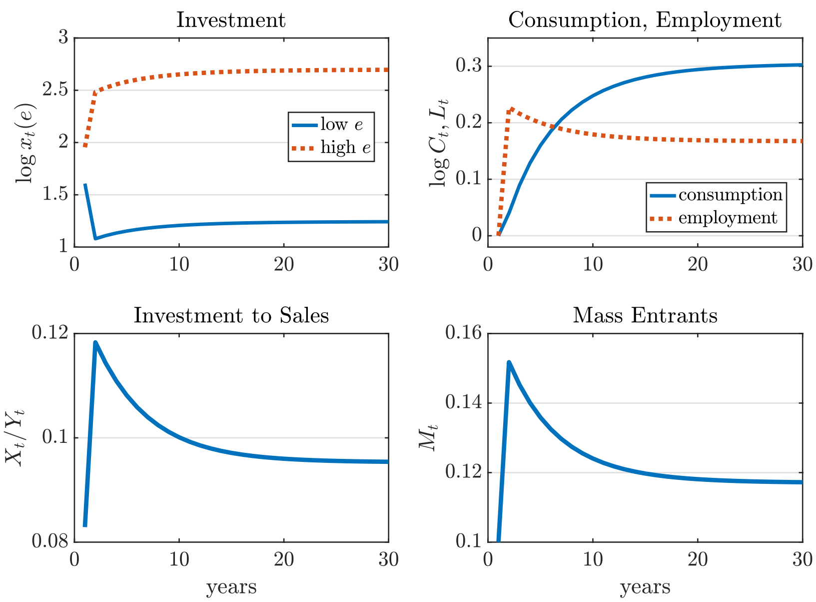
reduces efficiency in our benchmark economy by 1.2%, so the bulk of this increase in efficiency is due to the increase in product variety, not the removal of misallocation.

Welfare gains from implementing efficient allocation. We calculate the welfare gains by solving for the planner's optimal paths for investment, variety creation, labor supply, etc, starting from the distribution $H_0(z)$ in the distorted steady state. Both the mass of varieties and the amount of intangible capital for each firm are initially distorted, so the transitional dynamics are long-lasting, reflecting both the planner's desire to smooth consumption and the irreversibility of the initial intangible investment choices.

Figure 9 shows the planner's choices during the transition from the distorted steady state to the efficient one. The upper-left panel shows that the planner increases the amount of dispersion in investment across the low and high productivity firms. The upper-right panel of the figure shows that consumption increases gradually, owing to the representative consumer's preference for consumption smoothing, but employment increases on impact, owing to the increase in aggregate efficiency and the removal of the implicit output tax. Finally, the bottom two panels of Figure 9 show that investment in both varieties and physical capital overshoots initially, leading to a rapid increase in the economy's two types of capital.

The last row of Table 4 above reports the welfare gains for the representative consumer in consumption-equivalent units including the transition, i.e., these take into account the deferred increase in consumption as investment builds up and the overshooting of employment during the transition. We find that the representative consumer needs to be compensated with an additional 7.5% consumption in each period in order for her to be indifferent between the initial distorted steady state and the transition to the efficient steady state.

Figure 9: Transition to efficient allocation



5.2 Relative importance of each channel

We now decompose the overall gains into the three channels by which markups distort allocations, i.e., the aggregate markup, misallocation, and inefficient entry. We do this by removing each of the three channels in isolation using offsetting subsidies. These subsidies are financed by lump-sum taxes levied on the representative consumer. We view these experiments as simply isolating the role of each distortion and illustrating the welfare costs of markups. The actual welfare consequences of such schemes would of course be much more complex in economies with heterogeneous consumers and other frictions.

Removing aggregate markup. We first study the consequence of introducing a uniform output subsidy χ for all producers that eliminates the aggregate markup distortion.

A firm's profits in this environment are

$$\pi_t(z) = \max_{p_t(z)} \, (1+\chi)\, p_t(z) y_t(z) - P_{v,t} v_t(z),$$

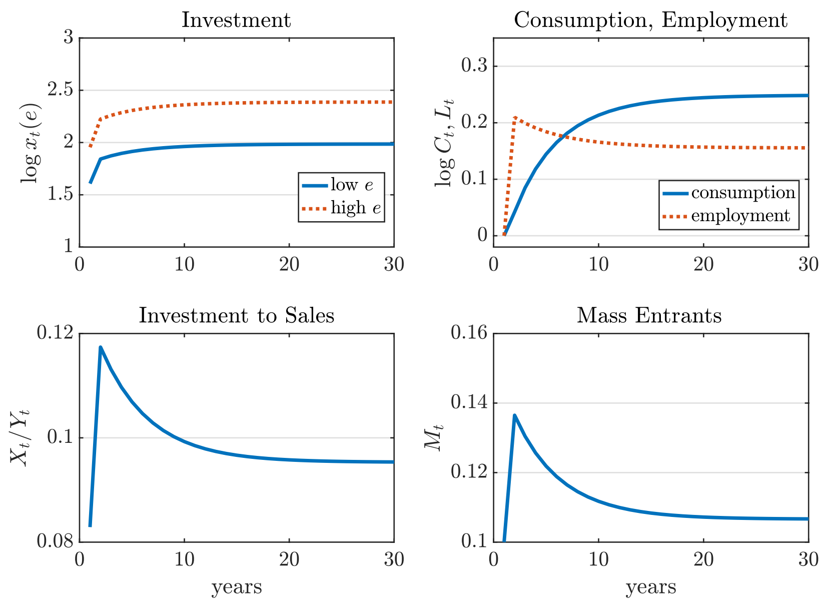
and its optimal price is

$$p_t\,(z) = \mu_t\,(z)\,\frac{1}{\eta}\frac{1}{1+\chi}P_{v,t}\frac{v_t\,(z)}{y_t\,(z)}.$$

The subsidy thus increases the steady state intangible capital to output ratio,

$$\frac{K}{Y} = \frac{1-\eta}{\frac{1}{\beta}-1+\delta}\frac{1+\chi}{\mathcal{M}},$$

Figure 10: Transitional dynamics with uniform output subsidy



the variable input cost to sales ratio,

$$\frac{P_v V}{Y} = \eta \frac{1+\chi}{\mathcal{M}},$$

as well as the mass of firms to output ratio,

$$\frac{N}{Y} = \frac{1+\chi}{\kappa W} \frac{1}{\frac{1}{\beta} - 1 + \delta} \left(1 - \frac{1}{\mathcal{M}}\right).$$

We set $1 + \chi = \mathcal{M} = 1.15$, to entirely eliminate the aggregate markup distortion.

Table 4 above reports the effect of introducing the output subsidy on the steady state of our benchmark model. The subsidy increases output by 33%, consumption by 25%, and employment by 16%. These increases are only slightly smaller than those from eliminating all markup distortions. The key difference between the efficient allocation and the economy with a uniform output subsidy is the lower efficiency E in the latter. This lower efficiency reflects the continued presence of misallocation and the somewhat smaller mass of firms.

Figure 10 shows the transitional dynamics after the introduction of the uniform subsidy. These transitions are very similar to when we remove all markups distortions, with one important exception. Under the efficient allocation the planner chooses to *increase* overall concentration, and does this by increasing the amount of investment in more productive firms and reducing the amount of investment in less productive firms. This outcome cannot be reproduced by a uniform output subsidy. Nevertheless, as the last row of Table 4 shows, a uniform output subsidy that eliminates the aggregate markup increases welfare by 4.9%, about two-thirds of the 7.5% total costs of markups.

Removing misallocation. We next consider size-dependent subsidies that equate the marginal product of factors across firms but leave the aggregate markup unchanged. Table 4 shows that such subsidies would have a modest impact, increasing output by 3.2%, consumption by 4.1%, and employment by 0.6%. Aggregate efficiency increases by 1.7%, reflecting both the removal of misallocation and the 3.2% increase in the mass of firms that results from subsidies that disproportionately benefit high productivity firms. Overall, removing misallocation increases welfare by 2.5%, about one-third of the total costs of markups.

Although the three channels we identify are not additive, taken together the aggregate markup and the misallocation channels seem to account for the bulk of the costs of markups. We now demonstrate that indeed the costs due to inefficient entry are quite small.

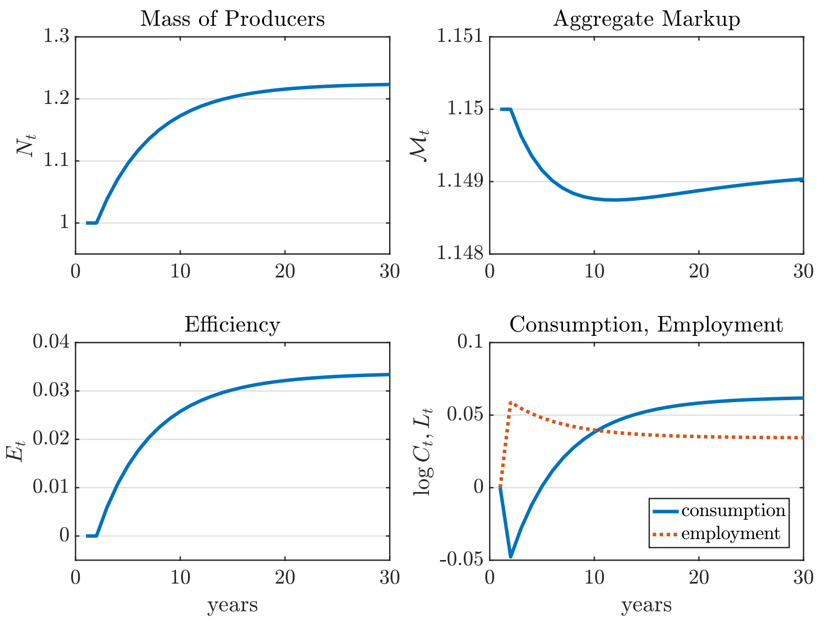
Subsidizing entry. We now consider a uniform subsidy χ that reduces the cost of creating a new variety to $\frac{1}{1+\chi}$ and increases the mass of firms to

$$\frac{N}{Y} = \frac{1+\chi}{\kappa W} \frac{1}{\frac{1}{\beta} - 1 + \delta} \left(1 - \frac{1}{\mathcal{M}}\right).$$

We calculate the gains from such a policy for many values of χ. We find that the largest gains from such a policy occur when the entry subsidy is $\chi = 0.297$ which causes the steady state mass of firms to increase by 20.3%. Table 4 shows that this subsidy has a modest effect on economic activity, increasing output by 4.5%, consumption by 6.3%, and employment by 3.4%. Aggregate efficiency increases by 3.4%, reflecting the increase in variety. But these increases in economic activity do not lead to similarly-sized welfare gains. The welfare gains from this entry subsidy are 0.7%, less then 1/10th of the total cost of markups. Larger entry subsidies can actually *backfire* and reduce welfare. For example, a subsidy of $\chi = 0.69$ that increases the mass of firms by 50% would lead to a welfare *loss* of 0.2%. Such losses occur when too much labor is allocated to variety creation as opposed to production.

Why are the welfare gains from entry subsidies so low? It turns out that increasing the number of firms has virtually no effect on the aggregate markup or losses from misallocation. Figure 11 illustrates the dynamics of the economy after the introduction of an entry subsidy $\chi = 0.297$ which increases the number of firms by 20.3%. The aggregate markup falls by a tiny amount, from 1.150 to 1.149. Though aggregate efficiency increases, it does so entirely due to love-of-variety effects, not due to a reduction in misallocation. Overall, the welfare gains from such an entry subsidy are small because consumption must fall and employment must rise to finance the increased investment in new firms.

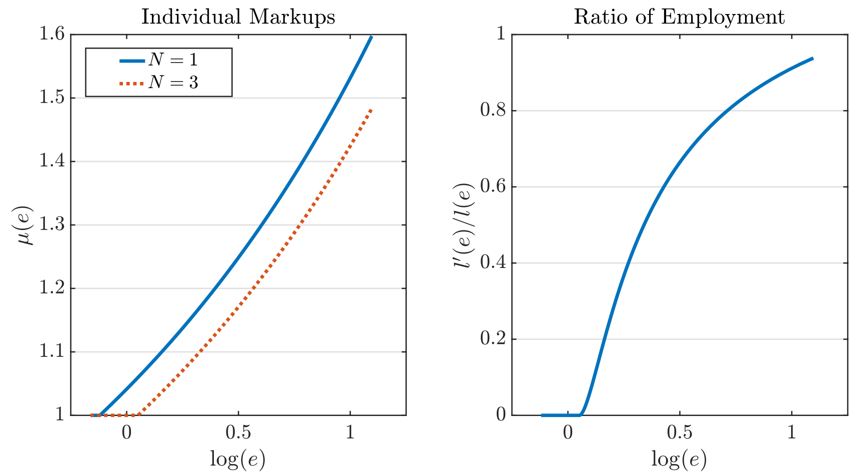
Figure 11: Transitional dynamics with entry subsidy



The result that *more competition* does not decrease the aggregate markup may appear counterintuitive but is, in fact, a robust result in a large class of models in the international trade literature which have shown that the removal of trade costs (which subjects domestic producers to more competition) leaves the markup distribution unchanged.[15] To understand this result, recall that the aggregate markup is a cost-weighted average of firm-level markups $\mu(q)$. An increase in the number of firms has two effects on this weighted average. The direct effect is a reduction in the relative quantity q and hence a reduction in the markups $\mu(q)$ of each firm. But there is also an important compositional effect. Recall that in our model, small firms face more elastic demand. This makes them more vulnerable to competition from entrants. By contrast large firms face less elastic demand and are less vulnerable to competition from entrants. A entry subsidy that increases the number of firms causes small, low markup firms to contract by more than large, high markup firms and the resulting reallocation means high markup firms get relatively more weight in the aggregate markup calculation. In our model, this offsetting compositional effect is almost exactly as large as the direct effect so that overall the aggregate markup falls by a negligible amount. We develop this argument more formally in Appendix C.

[15] See Bernard, Eaton, Jensen and Kortum (2003) and Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2017) who show that the markup distribution is invariant to changes in trade costs in models where variable markups arise due to limit pricing and monopolistic competition with non-CES demand, respectively.

Figure 12: Effect of entry subsidy on markups



We illustrate the two offsetting effects in Figure 12. For visual clarity, we consider an extreme parameterization in which we choose the entry subsidy large enough to triple the number of firms. Notice in the left panel that markups fall for all firms when the number of firms increases. The right panel shows that the most efficient firms lose only about 5% of their employment. By contrast, the least efficient firms contract their employment by a lot more and indeed some find it optimal to shut down altogether. Though we have derived this result in a quite specific model of monopolistic competition, in our robustness section below we show that similar conclusions are reached in an alternative model of oligopolistic competition calibrated to match the same concentration facts.

Welfare results for our alternative 'high $\mathcal{M}$' calibration. We have also calculated the welfare costs of markups in our alternative calibration that targets an aggregate markup of $\mathcal{M} = 1.25$, as in the Compustat data. Not surprisingly, since the aggregate markup is larger the distorted steady state is much further from the efficient steady state. As reported in Table 5, in the efficient steady state output is higher by 67.6%, consumption by 56.8%, and employment by 26.3% relative to the high $\mathcal{M}$ steady state. The total welfare costs of markups are correspondingly larger. The representative consumer needs to be compensated with an additional 18.6% consumption in each period in order for her to be indifferent between the high $\mathcal{M}$ steady state and the transition to the efficient steady state. In this sense, the total welfare costs of markups are more than twice as large in the $\mathcal{M}$ economy as in our benchmark. The high $\mathcal{M}$ economy has both larger and more dispersed markups than the benchmark, but we again find that it is the aggregate markup itself that accounts for most of the welfare costs. A 25% uniform output subsidy that eliminates the aggregate markup increases welfare by 15.2%, more than four-fifths of the 18.6% total costs of markups in this

Table 5: Steady state allocations in high $\mathcal{M}$ calibration

	efficient	uniform output subsidy	remove misallocation
log deviation from benchmark, $\times 100$			
output, Y	67.6	63.3	3.4
consumption, C	56.8	50.4	4.1
employment, L	26.3	25.1	0.3
mass of firms, N	15.9	9.7	0.2
capital, K	90.0	85.7	3.4
aggregate efficiency, E	5.6	2.6	1.3
welfare gains, CEV, %	18.6	15.2	3.0

alternative calibration. In other words, the aggregate markup channel is even more important in this alternative calibration than it is in our benchmark. Removing misallocation increases welfare by 3%, slightly higher than in our benchmark. Again we find that, taken together, the aggregate markup and misallocation channels account for almost all the costs of markups.

5.3 Explaining the rise in markups

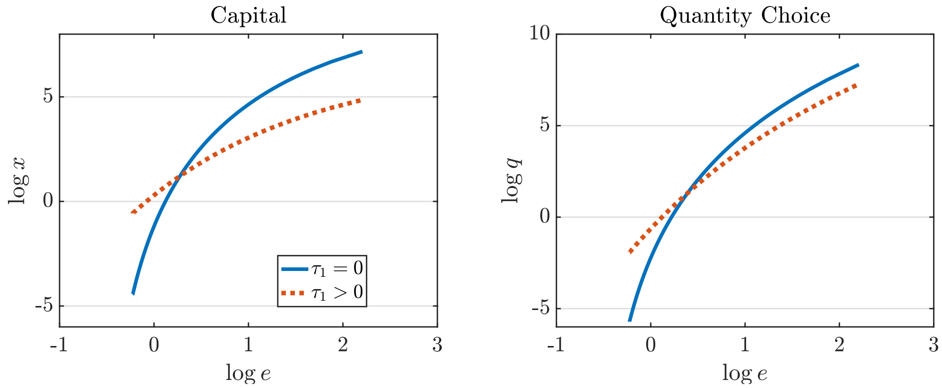
So far we have focused on our model's normative implications. But given the pronounced rise in markups observed in US data over recent decades, it is worth asking if our model can shed light on this phenomenon.

Problem with the barriers to entry story. One explanation for the rise in markups is a rise in entry barriers and a resulting decline in competition.[16] Our model casts doubt on this explanation. In our model, a rise in entry barriers of this kind would counterfactually *reduce concentration* by shifting production to less efficient firms that can now survive due to less intense competition. And moreover this reallocation of production towards less efficient firms would leave the aggregate markup essentially unchanged, despite an increase in firm-level markups, because of the compositional effects illustrated in Figure 12 above.

Given that an increase in entry barriers cannot explain the patterns in the data, what are the forces that could potentially rationalize the observed increase in markups? We briefly consider two other possible explanations: (i) a decline in antitrust enforcement, as emphasized

[16]See Grullon, Larkin and Michaely (2017) and Gutiérrez and Phillippon (2017) for discussion.

Figure 13: Effect of size-dependent investment taxes



by Peltzman (2014) and Grullon, Larkin and Michaely (2017) and others, and (ii) production technologies that are more intensive in intangible capital and hence more *scalable*, as emphasized by Haskel and Westlake (2017).

Decline in antitrust enforcement. A detailed model of mergers and acquisitions is beyond the scope of this paper. We model the effects of antitrust policy in a simple reduced-form way. Specifically, we view antitrust policy as a schedule of *size-dependent investment taxes*. Our interpretation of a reduction in the enforcement of antitrust policy is then a decline in the progressivity of these size-dependent investment taxes which disproportionately benefits large, high markup firms and increases markups and concentration.

To this end, let $T(x)$ denote the tax paid by a firm that wants to invest x with

$$T(x) = \tau_0 x^{1+\tau_1} - x.$$

The firm's tax-inclusive expenditure is then $\tau_0 x^{1+\tau_1}$. Here the parameter τ_1 determines the progressivity of the tax with a positive τ_1 implying higher marginal taxes for larger firms, while τ_0 determines the average tax rate. Clearly, a progressive tax schedule disproportionately hurts larger, more productive firms. Indeed, the investment choices in this economy are proportional to

$$x(e) \sim \left(\frac{q(e)}{e}\right)^{\frac{1}{1+\eta\tau_1}},$$

and therefore scale less with productivity than they would in the absence of taxes. Figure 13 shows that when τ_1 is positive, both capital and the amount firms sell become less dispersed.

To illustrate the effects of these taxes, we choose $\tau_1 = 0.80$ which reduces the sales share of the top 1% of firms from 0.30 in our benchmark to 0.18. This corresponds to the 60% increase in the top 1% sales share in the Compustat data from 1980 to 2012. We set $\tau_0 = 0.61$ to keep the capital-to-output ratio K/Y unchanged relative to our benchmark model.

Table 6: Effect of size-dependent investment taxes

	benchmark	size-dependent taxes
top 1% sales share	0.30	0.18
top 5% sales share	0.58	0.37
aggregate markup	1.15	1.12
losses from misallocation, %	1.16	9.28
log-deviation from benchmark, $\times 100$		
mass firms, N	–	22.0
aggregate efficiency, E	–	-4.4
output, Y	–	-11.4
consumption, C	–	-20.5
employment, L	–	8.9

Table 6 compares the steady states in the two economies, one with 1980 levels of concentration and size-dependent investment taxes, the other our benchmark model with 2012 levels of concentration but without size-dependent investment taxes. Returning the economy to 1980 levels of concentration in this way would reduce the aggregate markup, from 1.15 to 1.12. This is a much larger effect than the entry subsidy we discuss above. But the reduction in the aggregate markup comes at a considerable cost. The losses from misallocation are now 9.3%, much larger than the 1.2% in our benchmark. Consequently, aggregate efficiency in the economy falls by 4%, despite a 22% increase in the mass of firms. Output is lower by 11% and consumption is lower by 21% despite a 9% increase in employment.

Overall we find that policies to limit concentration can be costly. Even though they succeed in reducing the level of markups, especially at the top, they result in considerable misallocation across firms thereby generating large efficiency losses. Empirically, this also suggests that if the rise in concentration and markups observed in recent US data is due to a reduction in, say, antitrust enforcement, then it may be the case that the overall level of markups rose yet at the same time misallocation fell. This is admittedly speculative, but is consistent with Baqaee and Farhi (2018) who document that the increase in concentration and markups in the US has been accompanied by an improvement in allocative efficiency.

Intangible capital and scalability. Recently Haskel and Westlake (2017) have argued that the advent of technologies that are intensive in intangible capital — in design, branding, R&D, etc — have made production technologies more *scalable*. One simple interpretation of this idea is that production functions are beginning to exhibit less diminishing returns to scale. This means that a given amount of dispersion in productivity will generate a larger amount of dispersion in output and hence in markups. This argument suggest an alternative, technological, interpretation of the simultaneous rise in concentration and markups to go along side a decline in antitrust enforcement. Still, regardless of whether the simultaneous rise in concentration and markups is due to changes in technology, changes in regulation, or some mix of the two, our key point is that it may be costly to undo these changes. Size-dependent policies that reduce concentration in an attempt to bring down the overall level of markups do so at the cost of increasing misallocation and reducing aggregate productivity.

6 Robustness

In this section we consider a number of robustness checks designed to assess the sensitivity of our results to parameter choices and other assumptions.

6.1 Estimates of super-elasticity ε for narrower NAICS industries

In our benchmark model we assume the same productivity dispersion parameter ξ and super-elasticity ε for each industry. We now relax this assumption by estimating these two parameters for narrower sets of NAICS industries. Our first set excludes the finance, education, and religion industries. As Table 7 shows, excluding these industries has hardly any effect on our estimated parameters or the implied amount of misallocation. Our second set excludes in addition those industries that feature more "local" competition, namely health, accommodation and food. Again we find that the Pareto tail ξ is close to our benchmark but now the super-elasticity ε is slightly higher at 2.43. This slightly higher ε leads to slightly larger losses from misallocation, 1.3% up from our benchmark 1.2%. Our third set includes manufacturing only. Since the wage bill in manufacturing is much less concentrated than the top sales share for manufacturing, we now estimate a distinctly higher $\varepsilon = 4.54$ with losses from misallocation rising to 1.9%. Still, the bottom line is that the losses from misallocation implied by variable markups are relatively small.

6.2 Estimates of super-elasticity ε from Taiwanese micro data

We now estimate the ratio ε/σ using a rich product-level panel dataset from Taiwanese manufacturing industries that we previously studied in Edmond, Midrigan and Xu (2015).

Table 7: Narrower NAICS estimates and misallocation losses

NAICS industries	ξ	σ	ε	misallocation, %
benchmark	6.66	11.55	2.18	1.2
(1) exclude finance, real estate, education, religion	6.76	11.54	2.16	1.2
(2) exclude all (1), and health, accommodation, food	6.67	11.79	2.43	1.3
(3) just manufacturing	6.72	13.11	4.54	1.9

The Taiwanese manufacturing data is more detailed than the 6-digit NAICS data and allows us to control for any product-year specific effects that capture sectoral differences.

For this exercise, we use the Klenow-Willis specification of the Kimball aggregator to derive the relationship between a producer's markup μ_i and its sales $p_i y_i$, namely

$$\frac{1}{\mu_i} + \log\left(1 - \frac{1}{\mu_i}\right) = \text{constant} + \frac{\varepsilon}{\sigma}\log(p_i y_i) \tag{41}$$

The key idea is that if we had measures of the markups μ_i so that the LHS of this expression is known then the ratio ε/σ can be estimated as the slope coefficient in a regression on sales. To implement this, we follow the methodology of De Loecker and Warzynski (2012) to construct estimates of producer level markups μ_i. In particular, we estimate an industry-specific production function from which we can infer the markup from the producer's cost minimization problem based on one of the variable inputs. The inverse markup is then calculated as the variable input share adjusted for the estimated factor output elasticity.

We then estimate equation (41) above in two ways. In the first specification we simply exploit the cross-sectional variation of producers within a given product category by including product-year fixed effects. This gives an estimate of $\varepsilon/\sigma = 0.15$ that is tightly estimated with a standard error of 0.002. In the second specification we exploit the panel structure of the data and include a producer fixed effect, thus using the time-series comovement of a producer's sales and their markups to identify the super-elasticity. This gives an estimate of $\varepsilon/\sigma = 0.16$ with a standard error of 0.007. Reassuringly, both of these estimates are fairly close to our benchmark estimate of 0.189 from the 6-digit NAICS data.

6.3 Dotsey-King specification of the Kimball aggregator

In our benchmark model we use the specification of the Kimball aggregator proposed by Klenow and Willis (2016). A popular alternative is the specification proposed by Dotsey and King (2005) which can be written

$$\Upsilon(q) = \frac{1}{(1+\zeta)\varkappa}\big((1+\zeta)q-\zeta)\big)^{\varkappa} + 1 - \frac{1}{(1+\zeta)\varkappa}$$

where to ensure concavity of the aggregator we need the two parameters ζ and $\varkappa$ to be such that $(1+\zeta)(1-\varkappa) > 0$. Both specifications have two parameters, one which controls the average demand elasticity and the other which controls the super-elasticity. In slight abuse of notation, let $\sigma(q)$ denote the demand elasticity and let $\varepsilon(q)$ denote the super-elasticity as functions of relative size. For the Klenow-Willis specification these are $\sigma(q) = \sigma\, q^{-\varepsilon/\sigma}$ and $\varepsilon(q) = \varepsilon\, q^{-\varepsilon/\sigma}$ so that $\sigma(1) = \sigma$ and $\varepsilon(1) = \varepsilon$. For the Dotsey-King specification these are

$$\sigma(q) = \frac{1}{(1+\zeta)(1-\varkappa)}\,\frac{(1+\zeta)q-\zeta}{q}$$

and

$$\varepsilon(q) = -\frac{1}{(1+\zeta)(1-\varkappa)}\,\frac{\zeta}{q}$$

We calibrate the three key parameters, namely the Pareto tail parameter ξ and the Dotsey-King parameters ζ and $\varkappa$ using the same strategy as for our benchmark model. Matching our target moments requires $\xi = 6.69$, $\zeta = -0.54$ and $\varkappa = 0.82$. This value of the Pareto tail parameter $\xi = 6.69$ is very close to our benchmark estimate of 6.66, suggesting that the amount of productivity dispersion needed to match the concentration in the data is not very sensitive to the details of the aggregator. The values $\zeta = -0.54$ and $\varkappa = 0.82$ for the Dotsey-King aggregator give point demand elasticity $\sigma(1) = 11.64$, again very close to the point demand elasticity $\sigma = 11.55$ in our Klenow-Willis benchmark, suggesting that the average demand elasticity needed to match an aggregate markup of 1.15 is again not very sensitive to the details of the aggregator. By contrast, the point super-elasticity is given by $\varepsilon(1)/\sigma(1) = 0.54$, higher than our benchmark estimate of 0.189 and more in line with our alternative 'high ε/σ' calibration above.[17]

We report the results from this experiment in Table 8. The fit to the distribution of relative sales is slightly worse than our benchmark specification, in particular the match to the top end of the distribution is worse. This is one reason why we prefer our benchmark Klenow-Willis specification. The implied distribution of markups is very similar to our benchmark except at the very top. For example, the 99th cost-weighted percentile of markups is 1.37 in

[17] Note that these estimates of the ε/σ ratio from US concentration data are much lower than is typically used in macro models that assume a representative firm. For example, Sbordone (2010) uses a ε/σ ratio of 2 or 3 and Dotsey and King (2005) use a ratio of 6. By contrast, all of our estimates are substantially below 1 and imply much milder real rigidities at the firm level.

our benchmark but is 1.23 with the Dotsey-King specification. The losses from misallocation are now about 1.7%, up just slightly from our benchmark loss of 1.2% but somewhat lower than the 2.1% loss from our alternative 'high ε/σ' calibration of the Klenow-Willis specification that is arguably a better point of comparison. We conclude from this that our finding of misallocation losses that are small relative to the existing literature is not driven by the particular details of the Klenow-Willis specification and is a robust implication of the US concentration data viewed through the lens of a model with endogenously variable markups. We also find that doubling the number of competitors has negligible effects on the aggregate markup and the amount of misallocation, again consistent with our benchmark results.

6.4 Oligopolistic competition

We now present calculations based on the oligopolistic competition model of Atkeson and Burstein (2008) that we used in Edmond, Midrigan and Xu (2015). In this model there is a continuum of sectors aggregated by a CES technology with elasticity θ and then within each sector there is a finite N firms that are aggregated with another CES technology with elasticity $\gamma > \theta$ and these N firms engage in Cournot competition. We use a simplified version of Edmond, Midrigan and Xu (2015) with one country, symmetric sectors (no systematic productivity differences between sectors, an identical number of firms per sector), and without fixed operating costs. We assume that each firm receives an IID productivity draw from a Pareto distribution with tail parameter ξ. We fix $N = 1100$, the median number of firms in the 6-digit NAICS data, and calibrate the three key parameters γ, θ and ξ using the same strategy as for our benchmark model. Matching our target moments requires $\gamma = 10.25$, $\theta = 0.46$, and $\xi = 8.64$. The demand elasticity facing each firm is a (harmonic) weighted average of γ and θ, namely

$$\varepsilon(s_i) = \left(s_i \frac{1}{\theta} + (1 - s_i) \frac{1}{\gamma} \right)^{-1}$$

where s_i is the sales share of firm $i = 1, ..., N$ in its sector.

We report the results in Table 8. The model fits the lower end of the distribution of relative sales worse than our benchmark, but it matches the top of the distribution well.[18] The distribution of markups predicted by the two models is similar, except at the very top. For example, the 99th cost-weighted percentile of markups is 1.37 in our benchmark and 1.50 in the model with oligopolistic competition. Intuitively, with a finite number of producers in any given industry there is a small set of sectors in which the largest firm is much more productive than the remaining competitors and charges very high markups. Owing to these higher markups at the very top, the Atkeson-Burstein model predicts more misallocation than our benchmark, 2.89% as opposed to 1.2%.

[18] A richer specification of the productivity distribution allows us to fit the data nearly as well as in our benchmark model and implies nearly identical results, so we omit it for brevity.

Table 8: Alternative market structures

Monopolistic competition with Dotsey-King aggregator
and oligopolistic competition with Atkeson-Burstein nested CES aggregators

Panel A: Unweighted

	US data	Dotsey-King	Atkeson-Burstein
fraction of firms with			
relative sales ≤ 1	0.877	0.878	0.865
relative sales ≤ 2	0.942	0.955	0.931
relative sales ≤ 5	0.979	0.988	0.973
relative sales ≤ 10	0.990	0.996	0.987
relative sales ≤ 50	0.999	1.000	0.999
relative sales ≤ 100	1.000	1.000	1.000

Panel B: Sales-weighted

	US data	Dotsey-King	Atkeson-Burstein
fraction of firms with			
relative sales ≤ 1	0.154	0.376	0.286
relative sales ≤ 2	0.271	0.539	0.389
relative sales ≤ 5	0.507	0.709	0.533
relative sales ≤ 10	0.660	0.796	0.648
relative sales ≤ 50	0.951	0.912	0.910
relative sales ≤ 100	0.978	0.938	0.979

Panel C: Cost-weighted distribution of markups

	benchmark	Dotsey-King	Atkeson-Burstein
aggregate markup	1.15	1.15	1.15
p25 markup	1.10	1.10	1.11
p50 markup	1.14	1.16	1.12
p75 markup	1.19	1.21	1.15
p90 markup	1.24	1.23	1.23
p99 markup	1.37	1.23	1.50
loss from misallocation, %	1.16	1.66	2.89

Now consider the effect of increasing competition in this model. Doubling the number of firms in each sector only reduces the aggregate markup from 1.15 to 1.146 and only reduces the misallocation losses from 2.89% to 2.88%. Thus, as in our benchmark model, more competition does not alleviate the key source of losses from markups. We have also solved for equilibria in versions of the model with Bertrand competition in which goods sold by producers that belong to a given sector are perfect substitutes so that the most productive firm engages in limit pricing and charges a markup that depends on the second-best producer's costs. We found that our results are robust to this extension as well, with implied losses from misallocation on the order of 2%. We thus conclude that our results key are robust to these alternative settings with oligopolistic competition. Solving for the dynamic equilibrium and the welfare costs of markups in these settings is computationally impractical because of the very high dimensionality of the state-space, but it is reassuring that our key steady-state implications are robust to these other market structures.

7 Conclusion

We study the welfare costs of product market distortions in a dynamic model with heterogeneous firms and endogenously variable markups. We calibrate the model to match the amount of concentration observed in US industry in 2012. We find that the welfare costs of markups are large. For our benchmark calibration, the representative consumer would gain 7.5% in consumption-equivalent terms if all markup distortions were eliminated, once transitional dynamics are taken into account. In our model markups reduce welfare because the aggregate markup distortion acts like a uniform output tax and reduces employment and investment by all firms, because markup variation across firms causes misallocation of factors of production, and because there is an inefficiently low rate of entry due to the misalignment between private and social incentives to create new firms. We find that the aggregate markup accounts for about two-thirds of the total welfare costs in our benchmark model, misallocation accounts for about one-third, and the costs due to inefficient entry are negligible.

Although we focus on the normative implications of our model, our results also have clear empirical implications. One simple but important finding is that the overall level of markups is best measured as a *cost-weighted* average of firm-level markups. This is the relevant aggregate distortion to employment and investment decisions. By contrast a *revenue-weighted* average of firm-level markups, as used in the existing literature, overstates the rise in the overall level of market power. In addition, our results provide two reasons to be skeptical of explanations for the simultaneous rise in concentration and markups that focus on increasing barriers to entry. First, in our model increasing barriers to entry *reduce concentration*, because the resulting lack of competition makes it easier for small firms to survive. Second, in our model changes in entry have negligible effects on the overall level of markups because entry is associated with a reallocation of production towards high productivity, high markup firms.

Reductions in antitrust enforcement or increases in the scalability of production may provide better explanations for the rise in concentration and markups. But whatever the underlying cause, our key conclusion is that size-dependent policies aimed at reducing concentration and markups need to be viewed with caution. While such policies can reduce the overall level of markups, they can also greatly increase misallocation and thereby reduce aggregate productivity.

Appendix

A Cost-weighted vs. sales-weighted average markups

The aggregate markup can be written as either a cost-weighted arithmetic average of firm-level markups or a sales-weighted harmonic average of firm-level markups. To see this simply, consider a special case of our model where labor is the only variable input so that a firm with productivity z produces $y = zl^{\eta}$. Since price is a markup over marginal cost we have

$$p(z) = \mu(z)\, W \left(\frac{y(z)}{z}\right)^{1/\eta} \frac{1}{y(z)}$$

Hence we can write the firm's wage share as

$$\frac{Wl(z)}{p(z)y(z)} = \frac{\eta}{\mu(z)}$$

As in the main text, the aggregate markup $\mathcal{M}$ is implicitly defined by the aggregate labor share

$$\frac{W\tilde{L}}{Y} = \frac{\eta}{\mathcal{M}}$$

where $\tilde{L}$ denotes aggregate labor used in production. Hence we can write the sales shares

$$\frac{p(z)y(z)}{Y} = \frac{\mu(z)}{\mathcal{M}} \frac{l(z)}{\tilde{L}} \tag{A1}$$

Now let the distribution of productivity be $H(z)$. Since $Y = \int p(z)y(z)\, dH(z)$ we can integrate both sides of (A1) and rearrange to get

$$\mathcal{M} = \int \mu(z) \frac{l(z)}{\tilde{L}}\, dH(z)$$

which expresses the aggregate markup as an arithmetic average of the firm-level markups $\mu(z)$ with cost weights $l(z)/\tilde{L}$. Equation (28) in the main text is simply this formula but for our more general model where both labor and materials are variable inputs. Alternatively, re-write (A1) as

$$\frac{p(z)y(z)}{Y} \frac{\mathcal{M}}{\mu(z)} = \frac{l(z)}{\tilde{L}} \tag{A2}$$

Since $\tilde{L} = \int l(z)\, dH(z)$ we can integrate both sides and rearrange to get

$$\mathcal{M} = \left(\int \frac{1}{\mu(z)} \frac{p(z)y(z)}{Y}\, dH(z)\right)^{-1}$$

which expresses the aggregate markup as a harmonic average of $\mu(z)$ with sales weights $p(z)y(z)/Y$.

These calculations do not depend on the details of the demand system or market structure. In particular, they do not on the assumption of Kimball demand and monopolistic competition. In Edmond, Midrigan and Xu (2015) we obtained equivalent formulas in the Atkeson-Burstein model with oligopolistic competition.

B Planner's valuation of new varieties

In this appendix we provide more details on the planner's valuation of new varieties M_t^*. Recall that the planner's problem is to maximize (33) subject to (34) and (35). Let $\lambda_{1,t}^*$ and $\lambda_{2,t}^*$ denote the multipliers on these constraints. The planner's first order condition for M_t^* can then be written

$$\kappa\psi L_t^{*\nu} + \lambda_{1,t}^* \int x_t(e)\, dG(e) + \eta\beta \sum_{i=1}^{\infty} \left[\beta\left(1-\delta\right)\right]^{i-1} \lambda_{1,t+i} Y_{t+i}^* Z_{t+i}^{*\frac{1}{\eta}} \int \left(\frac{q_{t+i}^*(e)}{z_{t+i}^*(e)}\right)^{\frac{1}{\eta}} dG(e)$$
$$= \beta \sum_{i=1}^{\infty} \left[\beta\left(1-\delta\right)\right]^{i-1} \lambda_{2,t+i} \int \Upsilon\left(q_{t+i}^*(e)\right) dG(e).$$

The LHS of this expression gives the marginal cost of new varieties, i.e., the initial labor cost κ plus the cost of investment allocated to the new varieties plus the discounted variable input costs used by these varieties. The RHS gives the marginal benefit from the new varieties. Using $\lambda_{1,t}^* = 1/C_t^*$ and (37) and (38) and simplifying gives

$$\kappa\psi C_t^* L_t^{*\nu} = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}^*}{C_t^*}\right)^{-1} \frac{Y_{t+i}^* Z_{t+i}^{*\frac{1}{\eta}}}{A_{t+i}^*} \int \left[\Upsilon\left(q_{t+i}^*(e)\right) - \Upsilon'\left(q_{t+i}^*(e)\right) q_{t+i}^*(e)\right] dG(e).$$

As in the main text, we define the inverse elasticity

$$\epsilon_{t+i}^*(e) = \frac{\Upsilon\left(q_{t+i}^*(e)\right)}{\Upsilon'\left(q_{t+i}^*(e)\right) q_{t+i}^*(e)}$$

We can then simplify terms on the RHS to get

$$\kappa\psi C_t^* L_t^{*\nu} = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}^*}{C_t^*}\right)^{-1} \frac{Y_{t+i}^* Z_{t+i}^{*\frac{1}{\eta}}}{A_{t+i}^*} \int \left[\epsilon_{t+i}^*(e) - 1\right] \Upsilon'\left(q_{t+i}^*(e)\right) q_{t+i}^*(e)\, dG(e).$$

Next, integrate the static allocation (36) across all varieties available in period t and use the expression for aggregate productivity Z_t^* in (27) to obtain

$$\int \Upsilon'(q_t^*(z)) q_t^*(z)\, dH_t^*(z) = A_t^* Z_t^{*-\frac{1}{\eta}}.$$

This allows us to write the planner's optimal choice of new varieties as in the main text, namely

$$\kappa\psi C_t^* L_t^{*\nu} = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}^*}{C_t^*}\right)^{-1} \int \left[\epsilon_{t+i}^*(e) - 1\right] p_{t+i}^*(e) y_{t+i}^*(e) dG\left(e\right),$$

where $p_{t+i}^*(e)$ denotes the planner's valuation of an additional unit of output of that variety.

C Why entry has negligible effect on aggregate markup

In this appendix we show why increasing the number of competitors has a negligible effect on the aggregate markup in our model, a result that is analogous to findings in the trade literature, especially the work of Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2017). They study a model with monopolistic competition and variable markups with more general non-CES demand which nest the Kimball aggregator we use. We adopt their approach to calculating the response of the aggregate markup $\mathcal{M}$ to a marginal change in the number of firms. To this end, note that a firm's employment $l(e)$ is proportional to its relative quantity scaled by productivity, $q(e)/e$, so we can write the aggregate markup as

$$\mathcal{M} = \frac{\int_1^\infty \mu(q(e)) \frac{q(e)}{e}\, dG(e)}{\int_1^\infty \frac{q(e)}{e}\, dG(e)}$$

where the limits of the integral use our assumption that $G(e)$ is Pareto on $[1, \infty)$. Using the optimal steady state investment choice $x(e)$ we can express the optimality condition that determines a firm's relative size

$$\Upsilon'(q) = \mu(q)\frac{1}{Ae},$$

where $A > 0$ is a scalar that depends on the aggregate demand index D and the cost of the variable input P_v. Since the latter changes as we increase the number of producers, so does the scalar A. In particular, $A'(N) < 0$ so that competition effectively increases all firms' variable costs.

This optimality condition clearly shows that a firm's quantity choice is a function of the product Ae, not of e and A in isolation. We can then use a change of variables $\tilde{e} = Ae$ and the assumption that $G(e)$ is Pareto to write the aggregate markup as

$$\mathcal{M} = \frac{\int_A^\infty \mu(q(\tilde{e}))\frac{q(\tilde{e})}{\tilde{e}}\,dG(\tilde{e})}{\int_A^\infty \frac{q(\tilde{e})}{\tilde{e}}\,dG(\tilde{e})}.$$

Hence changes in the number of competitors, summarized by changes in A, only change the aggregate markup through their effect on the markups of the smallest firms. A direct calculation then gives

$$\mathcal{M}'(N) = -\left(\mu(q(A)) - \mathcal{M}\right)\,\frac{q(A)g(A)}{\int_A^\infty \frac{q(\tilde{e})}{\tilde{e}}\,dG(\tilde{e})}\,\frac{A'(N)}{A} \leq 0$$

Since the markups of the smallest firms, $\mu(q(A))$, are lower than the aggregate markup, $\mathcal{M}$, an increase in the number of firms reduces the aggregate markup. But this effect is quantitatively small in our calibration since $q(A)g(A)$ is relatively small because the smallest firms sell very little. Thus, even though we do not assume a choke price, as Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2017) do in deriving their exact neutrality result, our quantitative results are very similar.

References

Amiti, Mary, Oleg Itskhoki, and Josef Konings, "International Shocks and Domestic Prices: How Large Are Strategic Complementarities?," 2017.

Andrews, Isiah, Matthew Gentzkow, and Jesse M. Shapiro, "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, November 2017, *132* (4), 1553–1592.

Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare, "The Elusive Pro-Competitive Effects of Trade," *Review of Economic Studies*, 2017, *forthcoming.*

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

_ **and** _ , "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, June 2010, *118* (3), 433–484.

_ **and** _ , "The Aggregate Implications of Innovation Policy," *Journal of Political Economy*, 2018, *forthcoming.*

Autor, David, David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "Concentrating on the Fall of the Labor Share," *American Economic Review: Papers & Proceedings*, 2017, *107* (5), 180–185.

_ , _ , _ , _ , and _ , “The Fall of the Labor Share and the Rise of Superstar Firms,” 2017. MIT working paper.

Baqaee, David Rezza and Emmanuel Farhi, “Productivity and Misallocation in General Equilibrium,” 2018. LSE working paper.

Barkai, Simcha, “Declining Labor and Capital Shares,” 2017. LBS working paper.

Bartelsman, Eric, John Haltiwanger, and Stefano Scarpetta, “Cross-Country Differences in Productivity: The Role of Allocation and Selection,” *American Economy Review*, 2013, *103* (1), 305–334.

Behrens, Kristian, Giordano Mion, Yasusada Murata, and Jens Suedekum, “Quantifying the Gap between Equilibrium and Optimum Under Monopolistic Competition,” January 2018. UQAM working paper.

Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum, “Plants and Productivity in International Trade,” *American Economic Review*, September 2003, *93* (4), 1268–1290.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz, “Monopoly power and endogenous product variety: Distortions and remedies,” October 2008. NBER working paper 14383.

De Loecker, Jan and Frederic Warzynski, “Markups and Firm-Level Export Status,” *American Economic Review*, October 2012, *102* (6), 2437–2471.

_ and Jan Eeckhout, “The Rise of Market Power and the Macroeconomic Implications,” August 2017. NBER working paper.

Dhingra, Swati and John Morrow, “Monopolistic Competition and Optimum Product Diversity Under Firm Heterogeneity,” *Journal of Political Economy*, 2016, *forthcoming.*

Dixit, Avinash K. and Joseph E. Stiglitz, “Monopolistic Competition and Optimum Product Diversity,” *American Economic Review*, 1977, *67* (3), 297–308.

Dotsey, Michale and Robert G. King, “Implications of State-Dependent Pricing for Dynamic Macroeconomic Models,” *Journal of Monetary Economics*, January 2005, *52* (1), 213–242.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, “Competition, Markups, and the Gains from Internaitonal Trade,” *American Economic Review*, October 2015, *105* (10), 3183–3221.

Gopinath, Gita and Oleg Itskhoki, “Frequency of Price Adjustment and Pass-Through,” *Quarterly Journal of Economics*, 2010, *125* (2), 675–727.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, “Are U.S. Industries Becoming More Concentrated?,” August 2017. Rice University working paper.

Gutiérrez, Germán and Thomas Phillippon, "Declining Competition and Investment in the US," November 2017. NYU Stern working paper.

Hall, Robert E., "New Evidence on the Markup of Prices over Marginal Costs and the Role of Mega-Firms in the US Economy," May 2018. NBER working paper.

Haskel, Jonathan and Stian Westlake, *Capitalism without Capital. The Rise of the Intangible Economy*, Princeton University Press, 2017.

Hsieh, Chang-Tai and Peter J. Klenow, "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, November 2009, *124* (4), 1403–1448.

Jones, Charles I, "Intermediate goods and weak links in the theory of economic development," *American Economic Journal: Macroeconomics*, 2011, *3* (2), 1–28.

Karabarbounis, Loukas and Brent Neiman, "Accounting for Factorless Income," June 2018. NBER working paper.

Kehrig, Matthias and Nicolas Vincent, "Growing Productivity without Growing Wages: The Micro-Level Anatomy of the Aggregate Labor Share Decline," 2017. Duke University working paper.

Kimball, Miles S., "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit, and Banking*, 1995, *27* (4, Part 2), 1241–1277.

Klenow, Peter J. and Jonathan L. Willis, "Real Rigidities and Nominal Price Changes," *Economica*, July 2016, *83*, 443–472.

Lerner, Abba P., "The Concept of Monopoly and the Measurement of Monopoly Power," *Review of Economic Studies*, 1934, *1* (3), 157–175.

Peltzman, Sam, "Industrial Concentration under the Rule of Reason," *The Journal of Law and Economics*, 2014, *57* (S3).

Peters, Michael, "Heterogeneous Mark-Ups, Growth and Endogenous Misallocation," December 2016. Yale Unviersity working paper.

Restuccia, Diego and Richard Rogerson, "Policy Distortions and Aggregate Productivity with Heterogeneous Establishments," *Review of Economic Dynamics*, October 2008, *11* (4), 707–720.

Sbordone, Argia M., "Globalization and Inflation Dynamics: The Impact of Increased Competition," in Jordi Gali and Mark J. Gertler, eds., *International Dimensions of Monetary Policy*, University of Chicago Press, 2010, pp. 547–579.

Traina, James, "Is Aggregate Market Power Increasing? Production Trends Using Financial Statements," February 2018. University of Chicago working paper.

Zhelobodko, Evgeny, Sergey Kokovin, Mathieu Parenti, and Jacques-François Thisse, "Monopolistic Competition: Beyond the Constant Elasticity of Substitution," *Econometrica*, November 2012, *80* (6), 2765–2784.