How Costly Are Markups?

Chris Edmond[*] Virgiliu Midrigan[†] Daniel Yi Xu[‡]

June 2018

Abstract

We study the welfare costs of markups in a dynamic model with heterogeneous firms engaged in monopolistic competition. In our model more efficient producers have larger market shares, charge higher markups, and produce too little relative to the social optimum. We decompose the costs of markups into three sources: i) a uniform output tax levied on all producers, ii) misallocation of factors of production across producers, iii) inefficiently low entry. The uniform tax distortion is the largest source of losses in our economy. Losses from misallocation are relatively low because very efficient producers face strongly diminishing returns and the gains from reallocating factors of production to them are low. Policies that subsidize firm entry have a relatively modest impact because while competition reduces individual producers' markups, it also reallocates market shares towards the larger producers and consequently does not reduce the aggregate markup distortion. Size-dependent policies that reduce concentration can reduce the aggregate markup but greatly increase misallocation, causing large aggregate efficiency losses.

Keywords: concentration, misallocation, firm dynamics.

JEL classifications: D4, E2, L1, O4.

[*] University of Melbourne, cedmond@unimelb.edu.au.

[†] New York University and NBER, virgiliu.midrigan@nyu.edu.

[‡] Duke University and NBER, daniel.xu@duke.edu.

1 Introduction

How large are the welfare costs of product market distortions? What kinds of simple policy interventions can best alleviate these costs? We study these questions in a quantitative model with heterogeneous firms with endogenously variable markups. We show that one can decompose the welfare cost of markups into three channels. First, the *aggregate markup*, that is, the cost-weighted average of individual producers' markups, acts like a uniform tax on output levied on all producers. Second, markups are dispersed in our economy because larger producers face less competition and consequently charge higher markups. Third, there is too little entry in this economy. Our goal is to use the model and U.S. data on the size distribution of firms to evaluate the relative importance of these three channels, and evaluate what types of policies are most effective at alleviating the efficiency losses from markups.

Our model features heterogeneous firms engaged in monopolistic competition with a non-CES demand system, as in Kimball (1995). Within a given industry, more productive firms are, in equilibrium, larger and face endogenously less elastic demand and so charge higher markups than less productive firms. Because of this, changes in the environment that allow more productive firms to expand 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 producers with relatively high measured labor shares to producers with relatively low measured labor shares (Autor, Dorn, Katz, Patterson and Reenen, 2017a,b; Kehrig and Vincent, 2017). Critically, markups in our model are returns to past sunk investments, in both developing new products, as well as acquiring capital. Policies aimed at reducing markups may therefore have unintended consequences but distorting such investments and reducing welfare.

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, which can be implemented using a scheme of size-dependent output subsidies financed by lump-sum taxes. We calibrate the initial steady state to match the levels of concentration in sales in the U.S. data as well as the relationship between payments to labor and sales at the firm level. We find that the welfare costs of markups are sizable. In our baseline calibration, we find that 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 various channels by which markups reduce welfare.

In our model, markups reduce welfare through channels — (i) the aggregate markup acts as a tax that reduces employment and investment, just as in a representative firm model,

(ii) the distribution of markups implies that factors of production are *misallocated* so that aggregate TFP is low, and (iii) the distribution of markups changes the expected returns to entry.

We find that the aggregate markup distortion is the most important source of welfare losses. A simple uniform subsidy on all producers that offsets this distortion would eliminate two-thirds of the overal costs of markups. Misallocation itself accounts for the rest of these welfare losses, while the distortion associated with the entry margin is negligible.

While the losses from misallocation in our economy are sizable – efficient reallocation of factors of production would increase overall TFP by 1.2%, these losses are much smaller than existing estimates in the literature. We argue that our findings are not inconsistent with these estimates. The reason misallocation is relatively low in our setting is that large producers who charge higher markups also face low demand elasticities, that is, strong diminishing returns. These strongly diminishing returns are precisely the reason such firms find it optimal to restrict production and charge high markups. But this feature of the model also implies that a benevolent planner has little to gain by reallocating factors of production towards these producers. The losses from misallocation are therefore small in our model. In contrast, when we calculate the losses from misallocation under the incorrect assumption that demand elasticities are constant, as a large literature on misallocation does, we find much larger gains from reallocation. With constant demand elasticities a planner stands to gain a lot from reallocating production towards large producers.

We also demonstrate that policies that subsidies firm entry are an inefficient tool for correcting the markup distortions. Changes in the number of competitors within an industry have a negligible impact on the aggregate markup distortion or misallocation. This result, reminiscent of findings in the trade literature[1], implies that the gains from increasing the number of producers only accrue from love-for-variety effects and are therefore relatively small. The intuition for why an increase in the number of competitors leads to a negligible change in the aggregate markup is as follows. The aggregate markup is a cost-weighted average of the markups set by individual firms. The direct effect of a large increase in the amount of competition is to reduce the markups of individual firms. But there is an offsetting compositional effect. As the number of competitors increases, low-productivity firms contract significantly while high-productivity firms contract by a much smaller amount. This reallocation of factors from low-productivity to high-productivity firms mitigates the direct effect and is driven by the key mechanism in our model, endogenously variable demand elasticities. Small firms face elastic demand and are vulnerable to more competition from entrants. Large firms face relatively inelastic demand and are less vulnerable to competition. Hence, even though more competition reduces all producer markups, it does not change

[1] See Bernard et al. (2003) and Arkolakis et al. (2017).

the cost-weighted average of markups because of the reallocation of production towards the larger, higher markup firms.

We finally evaluate the effect of size-dependent policies aimed at reducing within-industry concentration and the markups of the larger producers. We show that although such policies can indeed succeed in reducing the aggregate markup distortion, they come at a considerable cost. Intuitively, the decentralized allocations feature *too little* concentration relative to what is socially optimal, so further reducing the amount of concentration grealy increases the TFP losses from misallocation, output and welfare. Our model therefore suggests that if the rise in concentration and markups observed in recent years was indeed due to less restrictive anti-trust enforcement, overall efficiency went up despite the increase in markups. This hypothesis is indeed consistent with the evidence in Baqaee and Farhi (2018) who document that the increase in concentration and markups in the U.S. has been accompanied by an improvement in allocative efficiency, and the work of Peltzman (2014) and Grullon et al. (2017) who document a significant decline in antitrust enforcement in the US.

We conduct most of our analysis using a specific model of monopolistic competition with non-CES demand as in Kimball (1995). We show, however, that our results are robust to an alternative model in which variable markups arise due to oligopolistic competition among a finite number of heterogeneous producers in an industry (Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015)). In particular, in a model of oligopolistic competition calibrated to match the same US concentration facts we find that the losses from misallocation are relatively small and that even large increases in the number of competitors have small effects on the aggregate markup and misallocation.

Related Work Our paper is related to a number of papers that study the cost of markups. Biblbiie Melitz does dynamics but no heterogeneity. A bunch of IO guiys (ZHelobodko, Dhingra-Morrow, French guys do heterogeneity but no dynamics). We do both. This is important because markups are returns to past investments and heterogeneity changes many insights of models with representative firms (eg markup doesn't change with competition)

Markups and misallocation. Our model with variable markups endogenously generates a form of misallocation in the sense of Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). Here we study a specific form of misallocation, namely markups that increase with firms size. We do not argue misallocation due to markups is very small. More markup dispersion due to competition in small markets, e.g. a firm may have multiple locations and operate in different product markets and charge different markups in each dependeing on how much competiton it faces. The key is that policies that subsidies firms by size are unlikely to yeild gains much more then 2-3% because there isn't systematic strong relationship between size and labor productivity in the data. If anything relationship could be due to fixed costs or increase in capital share in production so our results are an upper bound.

The remainder of the paper proceeds as follows. Section 2 presents the model. Section 3 presents the corresponding planner's problem and characterizes the efficient allocations against which we assess the welfare costs of markups. Section 4 explains how we quantify the model and in particular how we calibrate the model to match the US concentration facts. Section 5 presents our main results on the welfare costs of markups. Section 6 conducts a number of robustness checks. Section 7 concludes.

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 allocations and the transition dynamics after one-time policy reforms.

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 final consumption C_t, investment in intangible capital X_t, or as a source of materials B_t, so that

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

The use of final goods as materials gives the model a simple *roundabout* production structure and, as in Jones (2011) and Baqaee and Farhi (2018), amplifies the distortions due to markups.

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 at date t. 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 $D_t = \frac{\sigma}{\sigma-1}$ so that (2) reduces to the familiar $p_t(\omega) = (y_t(\omega)/Y_t)^{-\frac{1}{\sigma}}$.

Klenow-Willis specification. We use throughout most of this paper 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.

Gains from variety. This specification of the production function implies *gains from 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 stock L of labor available for production, in a symmetric equilibrium $y = L/N$, so that the total amount of the final output in the economy is given by $N\Upsilon(y/Y) = N\Upsilon(L/(NY)) = 1$. Aggregate productivity $A = Y/L$ is implicitly defined by $N\Upsilon(1/(NA)) = 1$. In the CES special case $\varepsilon = 0$ we get the familiar expression $A = N^{\frac{1}{\sigma-1}}$. When $\varepsilon > 0$, aggregate productivity A 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 $e \sim G(e)$. We will 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 $i = 0, 1, 2, ...$ is

$$k_{t+i,i}(e) = x_t(e).$$

Intermediate producers are forced to exit with exogenous probability δ per period. 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 allows us to interpret our model as also capturing investments in intangible capital, whose resale value is much lower than that of tangible capital.[2]

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

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

[2] See Haskel and Westlake (2017).

where $v_{t,i}$ 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.

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 $v_{t,i}(e)$ units of the composite variable input will allocate that amongst labor and materials according to

$$l_{t,i}(e) = \phi^{\theta} \left(\frac{W_t}{P_{v,t}}\right)^{-\theta} v_{t,i}(e),$$

and

$$b_{t,i}(e) = (1-\phi)^{\theta} \left(\frac{1}{P_{v,t}}\right)^{-\theta} v_{t,i}(e)$$

where $P_{v,t}$ is the unit price of the composite variable input

$$P_{v,t} = \left[\phi W_t^{1-\theta} + (1-\phi)\right]^{\frac{1}{1-\theta}}.$$

Each firm maximizes profits taking as given the production function (6) and the demand curve (2). Letting $z := e\, k_{t,i}(e)^{1-\eta}$ denote the firm's effective productivity, we can write the static problem of a firm of type z as

$$\pi_t(z) = \max_{y_t(z)} \left[P_t \Upsilon'\Big(\frac{y_t(z)}{Y_t}\Big) y_t(z) - P_{v,t}\left(\frac{y_t(z)}{z}\right)^{\frac{1}{\eta}}\right]. \tag{8}$$

The solution to this problem implies that the optimal price $p_t(z)$ can be written as a markup $\mu_t(q_t(z))$ over the firm's marginal cost,

$$p_t(z) = \mu_t\left(q_t(z)\right) \times P_{v,t}\, \frac{1}{\eta} \left(\frac{y_t(z)}{z}\right)^{\frac{1}{\eta}} \frac{1}{y_t(z)}, \tag{9}$$

where $q_t(z) = y_t(z)/Y_t$ is the relative quantity supplied by the producer. Using the Klenow-Willis specification in (5) gives

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

which implies the markup function

$$\mu(q) = \frac{\sigma}{\sigma - q^{\frac{\varepsilon}{\sigma}}}. \tag{10}$$

When $\varepsilon = 0$, this reduces to the familiar CES markup $\mu = \frac{\sigma}{\sigma - 1}$. 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 $\frac{\varepsilon}{\sigma}$. The ratio of these two parameters is therefore critical in shaping how markups and quantities change with productivity and competition.

Figure 3 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.

We also note that the firm's quantity choice is bounded. A profit maximizing firm would not increase production to the point where the elasticity of demand is less than unity. With the Klenow-Willis specification (5) this requires that the demand elasticity be

$$-\frac{\Upsilon'(q)}{\Upsilon''(q)q} = \sigma q^{-\frac{\varepsilon}{\sigma}} > 1,$$

which implies a bound on the relative quantity equal to

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

The model therefore implies a threshold level of productivity $\bar{z}_t$ above which all producers 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

$$\pi_t(z) = p_t(z)y_t(z) - P_{v,t}v_t(z) \tag{11}$$

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

$$\frac{P_{v,t}v_t(z)}{p_t(z)y_t(z)} = \frac{\eta}{\mu(q_t(z))}. \tag{12}$$

Since markups are increasing in relative size $q_t(z)$ 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 $q_t(z)$.

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 $i = 1, 2, \ldots$. Hence the benefit of choosing an investment $x_t(e)$ at entry is the stream of flow profits $\pi_{t+i}(ex_t(e)^{1-\eta})$ for $i = 1, 2, \ldots$ and so the firm chooses $x_t(e)$ to

$$\max \; -x_t(e) + \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \pi_{t+i}\left(ex_t(e)^{1-\eta}\right), \tag{13}$$

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

$$x_t(e) = \frac{1-\eta}{\eta} \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} P_{v,t+i} v_{t+i}(e x_t(e)^{1-\eta}), \tag{14}$$

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 (14) gives the firm's optimal investment choice $x_t(e)$. Using (12) we can also write this as

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

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. We let M_t denote the mass of entrants in period t and κ the amount of labor required to create a new firm. 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, firms discount future flows at rate $\beta^i C_t / C_{t+i}$ and exit at exogenous rate δ, we have

$$\kappa W_t = \int \left(\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}}{C_t}\right)^{-1} \pi_{t+i}\left(e x_t(e)^{1-\eta}\right) - x_t(e) \right) dG(e), \tag{16}$$

which, using (11), (12) and (14) implies

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

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 $z = e x_t(e)^{1-\eta}$ 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.

We now highlight a few key equilibrium conditions. 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

$$H_{t+1}(z) = (1-\delta)H_t(z) + M_t \int \mathbb{I}\{\, e x_t(e)^{1-\eta} \leq z \,\}\, dG(e), \tag{18}$$

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

Labor market clearing requires

$$L_t = \int l_t(z)\, dH_t(z) + M_t \kappa. \tag{19}$$

Similarly, goods market clearing requires

$$Y_t = C_t + M_t \int x_t(e)\, dG(e) + \int b_t(z) dH_t(z), \tag{20}$$

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 'wedges' in the firms' input choices relate to the cross-sectional distribution of markups. These aggregation results motivate a two-step approach that we use to compute the equilibrium of this economy. 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:

$$Y_t = Z_t V_t^{\eta}. \tag{21}$$

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

$$V_t = \left[\phi \left(L_{p,t}\right)^{\frac{\theta-1}{\theta}} + (1-\phi) B_t^{\frac{\theta-1}{\theta}}\right]^{\frac{\theta}{\theta-1}}, \tag{22}$$

where $L_{p,t} = \int l_t(z) dH_t(z)$ denotes the quantity of labor *used in production.*

Similarly, let $\mathcal{M}_t$ denote the *aggregate markup* of this economy, implicitly defined as the solution to

$$\frac{P_{v,t} V_t}{Y_t} = \frac{\eta}{\mathcal{M}_t}. \tag{23}$$

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. This wedge also reduces the share of production labor in output:

$$\frac{W_t L_{p,t}}{Y_t} = \frac{\eta}{\mathcal{M}_t} \times \phi^{\theta} \left(\frac{W_t}{P_{v,t}} \right)^{1-\theta}, \tag{24}$$

and materials.

Some algebra shows that the aggregate productivity Z_t relates to the individual productivities z according to

$$Z_t = \left(\int \left(\frac{q_t\left(z\right)}{z} \right)^{\frac{1}{\eta}} dH_t\left(z\right) \right)^{-\eta}, \tag{25}$$

while the aggregate markup is simply equal to a *cost-weighted* average[3] of individual producers' markup:

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

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

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

denote the aggregate capital stock in the economy. We can then write

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

where

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

is the aggregate efficiency in this economy, a harmoninc weighted average of individual producer's productivity, with weights given by each producer's relative quantity $q(e)$.

Solution Algorithm. We briefly outline the algorithm we use to solve the model. We use the aggregation results above to calculate the aggregate production function and evaluate the 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

[3] Or equivalently, the sales-weighted *harmonic* average of individual markups. See Edmond et al. (2015).

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

$$\tilde{\pi}_t(z) = \max_{q_t(z)} \Upsilon'(q_t(z))q_t(z) - A_t \left(\frac{q_t(z)}{z}\right)^{\frac{1}{\eta}}, \tag{26}$$

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

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

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

$$\Upsilon'(q(z,A))q(z,A) = \mu(q(z,A))\frac{A}{\eta}\left(\frac{q(z,A)}{z}\right)^{\frac{1}{\eta}}. \tag{27}$$

We can then solve for the equilibrium A_t using the definition of the Kimball aggregator

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

which allows us to recover the equilibrium relative quantities

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

demand index

$$D_t = \int \Upsilon'(q_t(z))q_t(z)dH_t(z),$$

individual firm prices,

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

and finally the aggregate markup and productivity.[4]

Given an initial conjecture for how the measure $H_t(z)$ evolves over time, we can therefore compute the aggregate prices and quantities at each date and then use these, together with the free entry condition (17) and an entrant's optimal investment choice (15) to update this conjecture until the sequence of measures $H_t(z)$ for each date t during the transition converges.

Steady State Entry and Intangible Investment. To build intuition, we now briefly characterize the steady-state amount of capital $K = N \int x(e)dG(e)$ and measure of firms N. Using (15) and aggregating across all firms, we have

$$\frac{K}{Y} = \frac{1-\eta}{\frac{1}{\beta} - 1 + \delta}\frac{1}{\mathcal{M}}, \tag{28}$$

[4] See also the work of Gopinath and Itskhoki (2010) and Amiti et al. (2017) who describe how to solve for the equilibrium in this setting in more detail.

so that the stock of capital is distorted by the aggregate markup $\mathcal{M}$, just as all static choices are.

Similarly, evaluating (17) at the steady-state allocations allows us to write

$$\frac{N}{Y}=\frac{1}{\kappa W}\frac{1}{\frac{1}{\beta}-1+\delta}\left(1-\frac{1}{\mathcal{M}}\right), \tag{29}$$

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 Allocations

We now derive the efficient allocations in this economy and decompose the losses from markups into three components. First, the aggregate markup acts like a uniform sales tax on all producers. Second, heterogeneity in markups generates *misallocation* in that more productive firms produce too little relative to what is socially optimal, reducing total factor productivity. Third, markups distort the amount of entry. We illustrate these three sources of inefficiency by comparing the equilibrium outcomes to the allocations chosen by a planner that faces the same technological and resource constraints.

Planner's Problem Given an initial distribution of productivities $H_0(z)$, the planner maximizes the representative consumer's utility (1) by choosing how many varieties to create each period, how to allocate variable inputs across different productive units, how much to invest, consume, and work, subject to the labor and goods resource constraints (19) and (20), the law of motion for the distribution of productivity (18), the individual producers' production functions described by (6) and (7) and the aggregate production function implied by (2).

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 seeks to maximize

$$\sum_{t=0}^{\infty}\beta^t\left(\log C_t^*-\psi\frac{\left(L_{p,t}^*+\kappa M_t^*\right)^{1+\nu}}{1+\nu}\right) \tag{30}$$

subject to the resource constraint for goods

$$C_t^*+X_t^*+B_t^*=\left(\sum_{i=1}^{\infty}(1-\delta)^{i-1}M_{t-i}^*\int\left(\frac{q_{t,i}^*(e)}{ex_{t-i}^*(e)^{1-\eta}}\right)^{\frac{1}{\eta}}dG(e)\right)^{-\eta}V(L_{p,t}^*,B_t^*)^{\eta} \tag{31}$$

and the Kimball aggregator

$$\left(\sum_{i=1}^{\infty}(1-\delta)^{i-1}M_{t-i}^*\int\Upsilon\left(q_{t,i}^*(e)\right)dG(e)\right)=1, \tag{32}$$

where $q^*_{t,i}(e)$ 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 in equation (25).

As with the equilibrium allocations, we find it useful to solve the problem of the planner in two stages, by first calculating the relative quantity allocated to each variety, and then the other choices.

Planner's Static Choice. Let $\mu^*_{1,t}$ denote the multiplier on the planner's resource constraint (31) and $\mu^*_{2,t}$ denote the multiplier on the Kimball aggregator (32). The first order condition that pins down $q^*_{t,i}(e)$ (or equivalently $q^*_t(z)$, since age only matters through the initial choice of investment which is summarized by z) requires that

$$\Upsilon'(q^*_t(z))q^*_t(z) = A^*_t \left(\frac{q^*_t(z)}{z}\right)^{\frac{1}{\eta}}, \tag{33}$$

where

$$A^*_t = \frac{\mu^*_{1,t}}{\mu^*_{2,t}} Y^*_t {Z^*_t}^{\frac{1}{\eta}}. \tag{34}$$

As with the decentralized allocations, the distribution of producer productivities only affects the choice of the relative quantity of an individual variety through the term A^*_t. We can therefore solve (33) for an arbitrary value of A^* and then find A^*_t to satisfy the Kimball aggregator (32).

Misallocation. Comparing the equilibrium allocation in (27) and the planner's allocation in (33) reveals the *misallocation* among existing producers in the decentralized equilibrium. Since more productive producers have higher markups, they choose to produce too little compared to what a planner would optimally choose. Figure 4 illustrates this point by comparing the amount of labor chosen by individual producers to that chosen by the planner. Notice that the planner's choice of labor is not a log-linear function of productivity, in contrast to what would be optimal with a CES technology, reflecting the strongly diminishing marginal product of intermediate inputs as their relative quantity increases. As we discuss below, this feature of the model implies that the gains from reallocating factors of production among producers are not as high (for a given distribution of markups) in this economy as would be the case in an economy with a CES technology.

Planner's Initial Investment Choice. Consider next the planner's choice of how much capital to allocate to each individual entrant and how many new varieties to create in a given period. Recognizing that $\mu^*_{1,t} = 1/C^*_t$ is simply the marginal utility of consumption and that

$X^*_t = M^*_t \int x^*_t(e)\, dG(e)$, the planner's first order condition for $x_t(e)$ can be written as

$$x^*_t(e) = (1-\eta)\beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C^*_{t+i}}{C^*_t}\right)^{-1} Y^*_{t+i} Z^{*\frac{1}{\eta}}_{t+i} \left(\frac{q^*_{t+i}(e)}{z^*_{t+i}(e)}\right)^{\frac{1}{\eta}}, \tag{35}$$

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

This expression implies that the steady-state stock of capital chosen by the planner satisfies

$$\frac{K^*}{Y^*} = \frac{1-\eta}{\frac{1}{\beta} - 1 + \delta}, \tag{36}$$

so the capital-output ratio is higher than in the decentralized equilibrium, as shown in (28). The equilibrium allocations thus imply too little investment because of the aggregate markup distortion.

Planner's Choice of Variety Creation. Consider next the planner's first order condition for M_t. We have

$$\kappa\psi \left(L^*_t\right)^{\nu} + \mu^*_{1,t} \int x_t(e)\, dG(e) + \eta\beta \sum_{i=1}^{\infty} \left[\beta(1-\delta)\right]^{i-1} \mu_{1,t+i} Y^*_{t+i} \left(Z^*_{t+i}\right)^{\frac{1}{\eta}} \int \left(\frac{q^*_{t+i}(e)}{z^*_{t+i}(e)}\right)^{\frac{1}{\eta}} dG(e)$$
$$= \beta \sum_{i=1}^{\infty} \left[\beta(1-\delta)\right]^{i-1} \mu_{2,t+i} \int \Upsilon\left(q^*_{t+i}(e)\right) dG(e)$$

The left-hand side of the expression gives the overall cost of creating a new variety: the initial labor cost κ, the cost of the investment allocated to the new varieties, as well as the discounted variable input costs used by these varieties. The right-hand side of the expression gives the benefit of the additional varieties.

Using $\mu^*_{1,t} = 1/C^*_t$, as well as (34) and (35) we can simplify this expression to

$$\kappa\psi C^*_t L^{*\nu}_t = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C^*_{t+i}}{C^*_t}\right)^{-1} \frac{Y^*_{t+i} Z^{*\frac{1}{\eta}}_{t+i}}{A^*_{t+i}} \int \left[\Upsilon\left(q^*_{t+i}(e)\right) - \Upsilon'\left(q^*_{t+i}(e)\right) q^*_{t+i}(e)\right] dG(e).$$

To contrast this expression with the market allocations, let us define

$$\epsilon^*_{t+i}(e) = \frac{\Upsilon\left(q^*_{t+i}(e)\right)}{\Upsilon'\left(q^*_{t+i}(e)\right) q^*_{t+i}(e)}$$

as the *inverse elasticity* of the Kimball aggregator Υ evaluated at the planner's optimal quantity choice. We then have

$$\kappa\psi C^*_t L^{*\nu}_t = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C^*_{t+i}}{C^*_t}\right)^{-1} \frac{Y^*_{t+i} Z^{*\frac{1}{\eta}}_{t+i}}{A^*_{t+i}} \int \left[\epsilon^*_{t+i}(e) - 1\right] \Upsilon'\left(q^*_{t+i}(e)\right) q^*_{t+i}(e) dG(e).$$

Next, integrate (33) across all varieties available in period t and use the expression for aggregate productivity Z_t^* in (25) to write

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

which allows us to write the planner's optimal choice of new varieties as

$$\kappa\psi C_t^* L_t^{*\nu} = \beta \sum_{i=1}^{\infty} (\beta(1-\delta))^{i-1} \left(\frac{C_{t+i}^*}{C_t^*}\right)^{-1} \int \left[\epsilon_{t+i}^*(e) - 1\right] p_{t+i}^*(e) y_{t+i}^*(e) dG\left(e\right), \tag{37}$$

where

$$p_t^*(e) = \frac{\Upsilon'\left(q_t^*(e)\right)}{\int \Upsilon'(q_t^*(z))q_t^*(z)dH_t^*(z)}$$

is the planner's marginal valuation of an additional unit of output of a particular variety.

A comparison of (37) and (17) reveals many similarities between the planner's entry choice and the free entry condition in the decentralized equilibrium. In particular, it is clear, as pointed out by Bilbiie et al. (2008), Zhelobodko et al. (2012) and Dhingra and Morrow (2016), that the incentives to enter in the decentralized equilibrium are determined by the producer's markups μ while the planner's incentives to create varieties are determined by the inverse elasticity of the production function ϵ). Importantly, these incentives are aligned in the special case of CES in which $\mu = \epsilon = \sigma/(\sigma - 1)$.

This can be seen most easily by contrasting the steady state mass of producers in the decentralized and planner allocations. Since in steady state $x(e) \sim q(e)/e$ in both sets of allocations, equations (17) and (37) reduce to

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

and

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

where $MPL = \partial Y/\partial L_p$ is the marginal product of labor.

The left panel of Figure 5 plots the markup $\mu(q(e))-1$ and the inverse elasticity $\epsilon(q(e))-1$ against productivity. Low productivity producers have relatively low markups and thus value entry too little compared to how the planner values entry by low-productivity varieties. In contrast, high productivity producers earn high markups and value entry more than the planner does. Whether the N/Y ratio is too low or too high compared to the efficient allocations is therefore ambiguous and depends on the exact details of the parameterization.

To summarize, the overall amount of entry in the economy with markups is inefficient, as it is determined by the firm's expected markups which do not coincide with the planner's marginal valuation of new varieties except for in the special case of CES technology.

Misalignment between the planner's and the firms' incentives to enter is another source of inefficiency in this economy.

Planner's Remaining Choices. The planner's optimal allocations of labor, capital and materials are similar to those in the decentralized equilibrium, except for the absence of the aggregate markup. We have

$$\psi C_t^* \left(L_t^*\right)^{\nu} = MPL_t^*,$$

and

$$1 = MPB_t^* = \frac{\partial Y_t^*}{\partial B_t^*}.$$

A comparison of these expressions to those in the decentralized allocations in, say, (23) and (24) illustrates the third source of inefficiency due to markups: the aggregate markup $\mathcal{M}$ acts like a uniform tax on sales, depressing employment and investment choices.

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 $T(s_t)$ that depends on the amount the firm sells, $s_t = p_t y_t$. It is straightforward to show that a subsidy equal to

$$T_t(s_t) = D_t Y_t \Upsilon\left(\frac{s_t}{p_t(s_t) Y_t}\right) - s_t,$$

where D_t is the demand index in (4) and $p_t(s_t)$ is the firm's price, restores efficiency as it aligns the private incentives to produce, invest and enter with those of the planner.

Figure 6 illustrates the shape of the subsidy function. The left panel shows the average subsidy, $T_t(s_t)/s_t$. Since $\Upsilon(0) > 0$, the optimal subsidy is positive even if the firm does not produce at all, $T_t(0) > 0$.[5] 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 marginal sales subsidy (right panel) increases, in contrast, in a firm's sales since the optimal marginal subsidy is simply equal to the optimal markup of the firm, $T_t'(s_t) = \sigma/(\sigma - q_t(s_t)^{\varepsilon/\sigma})$, 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

[5] 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.

due to misallocation in this economy. Our calibrated model features a considerable amount of markup dispersion and conventional methods for estimating the aggregate productivity losses due to misallocation based on a constant elasticity of demand, such as Hsieh and Klenow (2009), would conclude that misallocation losses are large. But as we show, these conventional methods based on a constant elasticity of demand 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 shape parameter ξ.

Intuitively, σ pins down the overall level of markups; ε pins down how markups μ_t and therefore a firm's wage bill, $w_t l_t \sim p_t y_t / \mu_t$, change with firm sales; and ξ pins down the concentration of firm sales. We thus choose these three parameters by requiring that the model reproduces the U.S. data on the distribution of sales for firms in narrowly-defined industries, the relationship between a firm's wage bill (payroll) and the firm's sales, as well as estimates of the aggregate level of markups in the U.S. We use the 2012 U.S. data for 6-digit NAICS industries on the payroll and sales of firms in different size classes[6] to calculate a number of moments that we require our model to match.

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 parameter ψ and the entry cost κ to achieve a steady-state output of $Y = 1$ 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 U.S. 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 U.S. We report all these parameter choices in Panel A of Table 1.

[6] https://www.sba.gov/advocacy/firm-size-data. The dataset contains information on total sales, number of firms and total wage bill for firms in about 15 size classes, where a class is defined based on the firm's revenues.

Calibrated Parameters. We next describe how we have chosen values for the key parameters ξ, σ and ε that determine the amount of concentration in sales and level and dispersion of markups. We choose these to minimize the distance between three sets of moments in the model and in the data. First, we require our model to match a 1.15 value for the aggregate markup $\mathcal{M}$, corresponding to the estimate of Barkai (2017) for 2012. Intuitively, this moment pins down the average elasticity σ.

Second, we require the model to match 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 on relative sales across all industries and report moments of this distribution in the left column of Table 2.

Consider first Panel A which summarizes the unweighted distribution of relative sales. About one-third of all firms in the data have average sales that are less then 1/10th of the industry average. The majority of firms (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. Consider next 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 U.S. 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 34% of (1 - 0.66) of overall sales and the 0.1% of firms whose sales exceed 50 times their industry average account for about 4.9% (1 - 0.951) of total sales. We require our model to match all these statistics by choosing the Pareto tail ξ to minimize the distance between these moments in the data and the model.

We finally discuss the set of targets that allow us to pin down the super-elasticity ε. We calculate, for each class size in each industry, the relative wage bill of firms in that class size, defined as the average wage bill of firms in that class size 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 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 would not increase with firm sales, and the relative wage bill would increase one-for-one with relative sales. 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 substracting industry-specific differences in production functions (say η or ϕ) and using within-industry variation to identify ε.

Figure 7 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, more so for larger firms.

Of course, there are alternative plausible explanations that would give rise to such a pattern. For example, a fixed (overhead) component to a firm's wage bill would also imply that larger firms pay their workers disproportionaly less.[7] Similarly, such a pattern can arise if larger firms outsource a greater fraction of their activities, or have a larger capital share. We thus think of our estimates as providing an *upper bound* on how rapidly markups increase with firm size.

We next explain how we have used this evidence to identify 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 rest of the parameters:

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

wigh 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 thus 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 ξ, σ and ε to jointly match i) a 15% aggregate markup, ii) 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. In our baseline exercise we pool together data from all industries. In our robustness section below we extendthe analysis to

[7] See Autor et al. (2017b) and Bartelsman et al. (2013).

allow for heterogeneity in ξ and ε across industries, by explicitly estimating these parameters for each industry and incorporating this heterogeneity into the model.

Model Fit. Panel B of Table 1 shows 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 is equal to $\xi = 6.66$. Though our estimate of ε/σ of 0.189 is much less than what is 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-economic estimates surveyed by Klenow and Willis (2016). As a robustness check, we present below 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 data, an estimate very close to that produced by our calibration strategy above.

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 7 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.

Since the ratio ε/σ is critical for the model's implications, we next provide some intuition for how this ratio is identified by reporting results for a 'high ε/σ' economy with $\varepsilon/\sigma = 0.4$), about twice larger than in our benchmark. For this alternative economy we continue to assign the other parameters as before and adjust the Pareto shape parameter ξ to match the size distribution of firms and the elasticity parameter σ to match the 15% aggregate markup. We report the re-calibrated parameter values and the model's fit for the distribution of sales in Tables 1 and 2. Notice that this version of the model fits the overall concentration in sales at the top much worse than our benchmark model. Since markups are increasing rapidly with size here, the model predicts too few large firms compared to the data.

Figure 8 shows this economy's predictions for how the relative wage bill changes with relative sales. The poor fit at the top 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.

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. We do so for our benchmark specification and the alternative parameterization with high ε/σ. We also compare our model's implications to estimates of markups from the publicly available Compustat data for the U.S. 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).

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. Our model predicts an aggregate markup that is smaller than that in the Compustat data (1.15 vs. 1.26). The model also predicts much less dispersion in markups, which in the data ranges from a 25th percentile of 0.97 to a 90th percentile of 1.69. 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 U.S., those that are publicly traded. In contrast, our calibration uses the estimates of the aggregate markup from Barkai (2017) for the entire U.S. 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. Hence, when we re-estimate our parameters to match the relationship between variable costs and sales in the Compustat data we find estimates of ε that are, if anything, smaller than those in our benchmark model.

Our observation that the aggregate markup in the U.S. Compustat data is equal to 1.26 may seem to contradict the findings of De Loecker and Eeckhout (2017) who report numbers in the neighborhood of 1.60. 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 efficiency losses. De Loecker and Eeckhout (2017), in contrast, report the sales-weighted *arithmetic* average of markups. The latter has increased much more in the last several decades, as Figure 9 shows[8], than the cost-weighted average, owing to an

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

increase in markups at the top of the distribution. Interpreted through the lens of our model, this increase in markups at the top does not distort input choices too much because of the relatively low amount of inputs hired by these producers.

4.3 Implications for misallocation

The markup dispersion generated by the model implies that there are aggregate productivity losses due to misallocation. We next ask: how large are these misallocation losses? As shown in Panel B of Table 3, aggregate productivity E in the steady state of our benchmark economy is 1.16% 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 2.13% loss from misallocation.

Our benchmark 1.16% 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), Hsieh and Klenow (2009)).[9] We now show that we can easily reconcile our findings with these estimates in the literature by recognizing that existing estimates rely on the assumption of a CES demand system, whereas our calculations use the actual demand system implied by the Kimball aggregator. We show that incorrectly assuming a CES demand system implies losses from misallocation that are much larger.

Misallocation with a CES technology. 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 producers differ in their productivity e, so that $y(e) = el(e)$. 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

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

[9] See also Baqaee and Farhi (2018) who proposes an alternative non-parametric approach to calculating how these losses evolve over time.

is below that under the efficient allocations 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 these calculations 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 15%. Assuming a CES technology we would conclude that misallocation losses are 8.4%, almost 7 times larger than the actual loss of 1.16% implied by the Kimball technology. For our high ε/σ economy markups are more dispersed so the CES technology implies a loss from misallocation of 16.7%, much greater than the 2.13% true loss implied by the Kimball technology.

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

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

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 producers lowering prices generates few additional sales so higher productivity simply translates to higher markups. The assumption of a CES technology 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 wedges in the firms' optimality conditions. Of course, these results reflect a very specific source of markup variation, namely Kimball demand. But as we show in our robusness section below, similar conclusions obtain in an alternative economy, similar to that studied by Atkeson and Burstein (2008), in which markups arise due to oligopolistic competition among a finite number of producers in a given industry.

5 How Costly Are Markups?

We next ask two questions. First, how large are the overall efficiency losses due to markups in our economy? Second, what is the relative importance of the three distortions: i) the uniform output tax, ii) entry distortion and iii) misallocation? We answer the first question by comparing the equilibrium allocations to those chosen by a benevolent planner and computing the welfare gains from implementing the planner's allocations taking the transition dynamics into account. We find that markups are quite costly overall: implementing the efficient allocations results in a consumption equivalent welfare gain of 7.49%.

We answer the second question by removing each of the three distortions in isolation using offseting subsidies. We show that a uniform sales subsidy of 15% that exactly offsets the aggregate markup goes a long way towards restoring efficiency, removing two-thirds of the overall losses from markups. Policies that subsidize entry, in contrast, have a relatively modest impact and would only offset about 1/10th of the overall cost of markups. The reason entry subsidies have a modest impact is that the resulting increase in the number of producers does not change the aggregate markup, a result that echoes findings in the trade literature.[10] Finally, removing misallocation by using size-dependent subsidies, but keeping the aggregate markup unchanged, would eliminate about one-thrid of the cost of markups, resulting in a 2.5% consumption-equivalent welfare gain.

5.1 Efficient Allocations

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

Steady State Comparison. We compare the steady state allocations in the market equilibrium with the efficient allocations in Table 4. We note that output and consumption would be about 37% and 31% greater under the efficient allocations, while employment would be about 17% higher. The efficient allocations call for more product variety in steady state: the mass of producers N is about 16% greater than under the decentralized allocations. The capital stock would be about 51% greater, while aggregate efficiency, E, would be about 3.8% greater. As discussed earlier, misallocation only reduces efficiency in our benchmark economy by 1.16%, so the bulk of this increase in efficiency is due to the increase in the number of varieties, not the removal of misallocation.

Welfare Gains From Removing Markups. We calculate these gains by first computing how the planner chooses the paths for investment, variety creation, labor supply etc.

[10] See Bernard et al. (2003) and Arkolakis et al. (2017) who demonstrate the invariance of the markup distribution to changes in trade barriers in models in which variable markups arise due to limit pricing and monopolistic competition with non-CES demand, respectively.

to maximize the representative consumer's utility starting from the steady state distribution $H_0(z)$ in the decentralized equilibrium. Both the mass of varieties and the amount of intangible capital of each producer is distorted in the decentralized steady state, so the transitions are long-lasting, reflecting the planner's desire to smooth consumption, as well as the irreversibility of the initial intangible investment choices.

Figure 10 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 producers. 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 overall efficiency and the removal of the implicit output tax. Finally, the bottom two panels of Figure 10 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 reports the welfare gains, expressed in consumption equivalent units, the representative consumer experiences by transiting from the distorted economy to the efficient steady state. These gains take into account the gradual increase in consumption and the overshooting of employment during the transition. We find that the consumer would be willing to give up 7.49% of her consumption in each period, in order to be indifferent between the status quo and the removal of the markup distortions.

We next decompose these gains into the three margins by considering simpler subsidy schemes that remove each distortion in isolation. These subsidies are financed by lump-sum taxes levied on the representative consumer. We emphasize that we think of these experiments as simply isolating the role of each distortion and illustrating the efficiency losses from markups. Clearly, the welfare consequences of such schemes would change in richer economies with heterogeneous consumers and frictions.

Uniform Output Subsidy. 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

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

We set $1+\chi = \mathcal{M} = 1.15$, to entirely eliminate the aggregate markup distortion. The subsidy thus increases the steady state intangible capital to output ratio,

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

the variable input cost to sales ratio,

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

as well as the number of producers to output ratio,

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

Table 4 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 altogether. The key difference between the efficient allocations and those in an economy with a uniform sales subsidy is the lower efficiency E in the latter. This lower level of efficiency reflects the presence of misallocation, as well as the somewhat smaller number of varieties.

Figure 11 shows the transition dynamics after the introduction of the uniform subsidy. These transitions are very similar to those arising when we remove all markups distortions, with one exception. Under the efficient allocations the planner chooses to increase the overall concentration in the economy, by increasing the amount of investment in the more productive firms and reducing it in the less productive firms, outcomes which the uniform output subsidy cannot reproduce. Nevertheless, as the last row of Table 4 shows, the uniform output subsidy increases welfare by 4.86%, a sizable amount that eliminates nearly two-thirds (4.86 out of 7.49) of the overall costs of markups.

Uniform Entry Subsidy. We next consider the consequence of an alternative policy, a uniform subsidy χ that reduces the cost of creating a new variety to $\frac{1}{1+\chi}$ and increases the number of producers to

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

It turns out that the largest gains accrue from a subsidy $\chi = 0.297$ under which the steady state measure of producers increases by 22.5%. Table 4 compares the steady state allocations of this economy with those in our benchmark model. Output would increase by about 4.5%, consumption by 5.3% and employment by 3.4%. Aggregate efficiency would increase as well, by about 3.4%. The welfare gains from such a subsidy are equal to 0.67% consumption equivalent units, or less then 1/10th of the overall cost of markups. Much larger entry subsidies would be costly. For example, a subsidy of $\chi = 0.69$ that increase the number of producers by 50% would lead to a consumption-equivalent welfare loss of 0.19% by reallocating too much labor 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 producers has virtually no effect on the aggregate markup or losses from misallocation, the two key sources of inefficiency in our economy. Figure 12 illustrates the dynamics of the economy after the introduction of an entry subsidy that increases the number of competitors by 50%. The aggregate markup falls only a little, from 1.150 to 1.149. Though overall efficiency E increases, it does so entirely due to love-for-variety, not due to a reduction in misallocation which actually increases slightly in the new steady state (from 1.16% to 1.19%). Overall, the welfare gains from such an entry subsidy are relatively small because consumption must fall and employment must increase to finance the increased investment in new varieties.

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[11] which have shown that the removal of trade barriers (which subjects domestic producers to more competition) leaves the markup distribution unchanged. Intuitively, more competition has two offsetting effects. On one hand, the relative quantity $q = y/Y$ sold by each individual producer falls since each producer becomes smaller relative to the aggregate when the number of firms increases. Since the optimal markup $\mu(q)$ is an increasing function of one's relative output, all individual markups fall. On the other hand, with firm heterogeneity this effect is largely offset by a countervailing compositional effect arising due to a reallocation of production from smaller to larger producers. This reallocation arises because more efficient, larger producers have lower demand elasticities. These producers therefore lose a relatively smaller market share compared to smaller, high elasticity producers. Employment and all other factors thus reallocate towards the larger firms. Since the aggregate markup is a weighted average of individual markups, with weights given by each firm's cost share, the reallocation of factors towards higher markup producers offsets the overall decline in individual markups, leaving the aggregate markup unchanged.

We illustrate the two offsetting effects Figure 13. For visual clarity, we consider an extreme parameterization in which we choose the entry subsidy large enough to triple the number of competitors. Notice in the left panel that markups fall for all producers when the number of competitors increases. The right panel shows that the most efficient producers lose only about 5% of their employment. In contrast, the least efficient producers contract their employment by a lot more and indeed some find it optimal to shut down altogether. Employment thus reallocates towards higher markup firms and so the employment-weighted average of markups does not change. This result once again arises due to the very ingredient that gives rise to variable markups, namely variable demand elasticities. Low productivity producers sell little, have high demand elasticities and experience large declines in the demand

[11] See Bernard et al. (2003) and Arkolakis et al. (2017).

for their goods when faced with additional competition. High productivity producers have high market shares, low demand elasticities and are thus relatively immune to competition. Once again, though we have derived this result in the context of a particular model of variable markups, our robustness section below shows that a similar result obtains in the Atkeson and Burstein (2008) model of oligopolistic competition.

Overall, we conclude that entry subsidies are an inefficient tool for dealing with the markup distortions.

Removal of Misallocation. We finally consider the implications of size-dependent subsidies that equate the marginal product of factors across producers, but leave the aggregate markup unchanged. The last column of Table 4 shows that such policies would have a modest impact on output and consumption: these would increase by 3.2% and 4.1%, respectively. This increase in consumption comes with virtually no cost, however, since employment would only increase by 0.6%. Overall efficiency would increase by about 1.7%, reflecting the removal of misallocation, as well as a small increase in the number of producers (3.2%) that results from subsidies that disproportionately benefits higher-productivity firms. Overall, the representative agent's welfare would go up by the equivalent of 2.5% consumption units, about one-third of the overall gains from removing markups.

Size-Dependent Investment Taxes. We have focused so far on the model's normative implications. Our results, however, also have clear positive implications. For example, our model implies that it is difficult to attribute the rise in markups observed in the past few decades to an increase in entry barriers and a reduction in the number of competitors. Such changes would reduce concentration by shifting production to less efficient producers that can now survive due to lack of competition. This reallocation would leave the aggregate markup unchanged, despite an increase in firm-level markups, as explained above.

Given that an increase in entry barriers cannot explain the patterns in the data, what are the forces that could potentially rationalize the increase in markups that a number of researchers have documented? One possibility that we consider here is the decline in anti-trust enforcement starting from the 1980s.[12] One simple way of capturing anti-trust enforcement in our model is to assume that the government levies a tax on investment which depends on the amount of capital the firm acquires. Such a progressive tax policy disproportionately hurts the more efficient, higher markup producers and reduces markups and concentration.

Specifically, suppose that a firm that would like to purchase x units of investment pays a

[12] See Peltzman (2014) and Grullon et al. (2017) who document a significant decline in antitrust enforcement in the U.S.

tax of

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

The firm's overall expenditure, including the tax, 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 producers, while τ_0 determines the average tax rate. Clearly, a progressive tax schedule disproportionately hurts more efficient producers. 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 in the absence of taxes. Figure 14 shows that when τ_1 is positive, both capital and the amount firms sell become less dispersed.

We illustrate the role such taxes play by choosing $\tau_1 = 0.80$ so as to reduce the sales share of the largest 1% of producers from 30% in our benchmark to 18%. This corresponds roughly to the 60% increase in the top 1% sales share in the Compustat data from 1980 to 2012. We then set $\tau_0 = 0.61$ so as to keep the capital-to-output ratio K/Y unchanged relative to the benchmark model.

Table 5 compares the steady states in the two economies, with and without size-dependent investment taxes. Clearly, a progressive tax schedule reduces concentration at the top and the aggregate markup, from 1.15 to 1.12. The cost of the lower markup, however, is a much greater misallocation of factors across producers: the difference between the level of aggregate efficiency E and its efficient level E^* is now 9.3%. Consequently, overall efficiency in the economy falls by 4%, despite a 22% increase in the number of producers. Overall, a policy that reduces concentration at the top is quite costly: it reduces output by 0.21 log-points, consumption by 0.21 log-points, despite an increase in employment of about 9%.

We thus conclude that policies that limit concentration can be quite costly. Even though they succeed in reducing the overall level of markups, especially at the top, they result in a great deal of misallocation across producers, generating large efficiency losses. Absent additional micro-level evidence, it is premature for us to speculate whether the increase in concentration in the U.S. observed in the last few decades was indeed due to less anti-trust enforcement and thus less progressive implicit producer taxes. We note, however, that this interpretation is not without merit since, as Baqaee and Farhi (2018) document, the increase in concentration and markups in the U.S. has also been associated with an improvement in allocative efficiency.

6 Robustness

We next discuss a number of robustness checks we have considered.

6.1 Estimates of Super-Elasticity for each 2-digit NAICS Industry

In our benchmark model we have assumed a representative industry characterized by a single productivity dispersion parameter ξ and super-elasticity ε. Here we relax this assumption by estimating these two parameters for each 2-digit NAICS sector. As Table 6 shows, industries differ quite a bit in the Pareto tail parameter ξ which ranges from 3.53 in 'arts and entertainment' to 7.03 in 'retail'. These differences reflect differences in the degree of concentration across sectors. In contrast, we find that the ratio ε/σ is very similar across industries, with most estimates in the 0.10 – 0.20 neighborhood. Given these estimates, we also report in the table the models implications for the aggregate markup and the losses from misallocation. The aggregate markup ranges from 1.15 to 1.28, with most estimates close to the 1.21 in our benchmark model. The losses from misallocation range from 0.3% to 3.7%, with most estimates around 1%. We thus conclude that allowing for sectoral heterogeneity does not greatly change our model's implications.

6.2 Estimates of Super-Elasticity from Taiwanese Micro Data

Here we exploit a rich product-level dataset from Taiwanese manufacturing industries that we have previously studied in Edmond et al. (2015) in order to estimate the ratio ε/σ. We do so by recognizing that with the Klenow-Willis specification of the Kimball aggregator the following relationship between a firm's markups and sales holds:

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

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

To implement our estimation, we first follow De Loecker and Warzynski (2012) to construct a markup measure for each producer. Specifically, we first estimate an industry-specific production function and then infer the markup from the producer's cost minimization problem based on one of the variable inputs. The inverse markup is calculated as the variable input share adjusted for the estimated factor output elasticity.

We estimate the equation 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 effect. We obtain an estimate of ε/σ of 0.145 that is tightly estimated with a standard error of 0.002. In a 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 markups to identify the super-elasticity. We obtain an estimate of ε/σ of 0.161 with a standard error of 0.007. Both of these estimates a very close to the 0.16 estimate we have obtained using the U.S. data in our benchmark parameterization.

6.3 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 i.i.d productivity draw from a Pareto distribution with shape parameter ξ. We fix $N = 1100$, the median number of firms in a given U.S. 6-digit sector, and choose the three parameters γ, θ and ξ to reproduce the same statistics we have targeted in our benchmark model. Matching these requires $\gamma = 10.25$, $\theta = 0.46$, and $\xi = 8.64$. The demand elasticity facing each producer 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 from this experiment in Table 7. Notice that the model fits the lower end of the distribution of relative sales worse than our Benchmark model, but it matches the top of the distribution well.[13] The distribution of markups predicted by the two models is very 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: the efficiency losses are equal to 2.89%. Most of this misallocation is within industries: equating the marginal product of factors of production within sectors only would increase efficiency by 2.84%.

Consider finally the effect of increasing competition in this model. Doubling the number of producers in each sector reduces the aggregate markup to only 1.146 from its baseline value of 1.150 and the losses from misallocation to only 2.88% from its baseline value of 2.89%. Thus, as in our benchmark model, more competition does not alleviate the key source of losses from markups. We thus conclude that our results key are robust in this alternative setting with oligopolistic competition. Unfortunately solving for the dynamic equilibrium

[13] 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.

and the welfare costs of markups in this economy is computationally impractical owing to the large dimensionality of the state-space in each industry, but it is reassuring that our key steady-state implications are robust to this alternative popular class of models. We have also solved for equilibria in versions of the model with price competition (Bertrand) 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 producers' costs. We found that our results are robust to this extension as well, with implied losses from misallocation in a calibrated version of the model on the order of 2%.

7 Conclusion

We study the welfare costs of product market distortions in a dynamic model with heterogeneous firms with endogenously variable markups. When calibrated to match the amount of concentration observed in the US industry in 2012 the model implies large welfare costs of markups. In our baseline calibration the representative consumer would experience gains of about 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 acts as a uniform tax on output and because the markup distribution causes misallocation and distorts entry decisions. We find that two-thirds of the welfare costs of markups comes from the aggregate markup distortion and can thus be offset using a uniform output tax.

References

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

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.

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.

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.

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," 2017.

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

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?," 2017. 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," 2018.

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.

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

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

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.

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 labor & materials	0.50

Panel B: Calibrated Parameters

	Benchmark	High ε/σ
σ, average elasticity	11.55	15.36
ε, super-elasticity	2.18	6.14
ξ, Pareto tail	6.66	6.69

Table 2: Distribution of Relative Sales

Panel A: Unweighted

	U.S. Data	Benchmark	High ε/σ
fraction of firms with			
relative sales ≤ 0.1	0.329	0.366	0.442
relative sales ≤ 0.5	0.761	0.747	0.692
relative sales ≤ 1	0.877	0.848	0.798
relative sales ≤ 2	0.942	0.916	0.888
relative sales ≤ 5	0.979	0.968	0.965
relative sales ≤ 10	0.990	0.987	0.991
relative sales ≤ 50	0.999	0.999	1.000
relative sales ≤ 100	1.000	1.000	1.000

Panel B: Sales-Weighted

	U.S. Data	Benchmark	High ε/σ
fraction of sales in firms with			
relative sales ≤ 0.1	0.019	0.026	0.014
relative sales ≤ 0.5	0.088	0.128	0.091
relative sales ≤ 1	0.154	0.211	0.183
relative sales ≤ 2	0.271	0.323	0.338
relative sales ≤ 5	0.507	0.509	0.630
relative sales ≤ 10	0.660	0.661	0.847
relative sales ≤ 50	0.951	0.928	1.000
relative sales ≤ 100	0.978	0.977	1.000

Table 3: Steady State Implications

Panel A: Distribution of Markups (cost-weighted)

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

Panel B: Productivity Losses from Misallocation

	Benchmark	High ε/σ
actual losses, $\log(E^*/E) \times 100$	1.16	2.13
losses with CES and $\bar{\sigma} = \frac{\mathcal{M}}{\mathcal{M}-1} = 7.67$	8.4	16.7

Table 4: Steady State Allocations Under Alternative Policies

	efficient	unif. sales subsidy	entry subsidy	remove misallocation
log deviation from benchmark, ×100				
output, Y	37.0	33.0	4.5	3.2
consumption, C	30.5	24.9	6.3	4.1
employment, L	16.7	15.5	3.4	0.6
number producers, N	15.7	6.4	20.3	3.2
capital, K	50.9	47.0	4.6	3.2
aggregate efficiency, E	3.8	1.1	3.4	1.7
welfare gains, CEV, %	7.49	4.86	0.67	2.50

Table 5: 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		
number producers, N	–	0.22
aggregate efficiency, E	–	-0.04
output, Y	–	-0.11
consumption, C	–	-0.21
employment, L	–	0.09

Table 6: Sectoral Estimates and Productivity Losses

NAICS sector	ξ	ε/σ	Agg. Markup	Misallocation, %
mining	4.04	0.17	1.25	2.10
utilities	7.13	0.05	1.14	0.30
construction	5.96	0.15	1.17	1.11
manufacturing	4.85	0.21	1.21	2.00
wholesale	3.78	0.25	1.26	2.99
retail	8.03	0.05	1.13	0.21
transportation	5.66	0.13	1.18	1.09
information	5.35	0.09	1.19	0.90
prof. services	6.40	0.09	1.16	0.65
adm. services	6.75	0.10	1.15	0.62
education	5.99	0.20	1.17	1.37
health care	6.79	0.14	1.15	0.82
arts and enter.	3.53	0.30	1.28	3.72
accom./food services	6.97	0.08	1.15	0.48

Figure 1: Gains From Variety with Kimball Aggregator

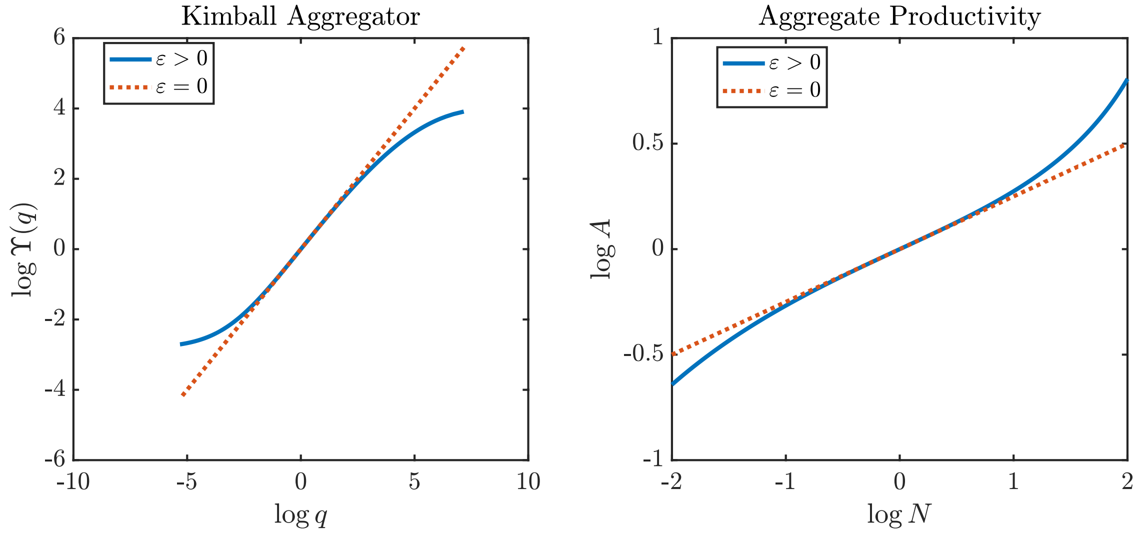


Table 7: Comparison with Atkeson-Burstein Model

Panel A: Unweighted

	U.S. Data	Atkeson-Burstein
fraction of firms with		
relative sales ≤ 1	0.877	0.865
relative sales ≤ 2	0.942	0.931
relative sales ≤ 5	0.979	0.973
relative sales ≤ 10	0.990	0.987
relative sales ≤ 50	0.999	0.999
relative sales ≤ 100	1.000	1.000

Panel B: Sales-Weighted

	U.S. Data	Atkeson-Burstein
fraction of sales in firms with		
relative sales ≤ 1	0.154	0.286
relative sales ≤ 2	0.271	0.389
relative sales ≤ 5	0.507	0.533
relative sales ≤ 10	0.660	0.648
relative sales ≤ 50	0.951	0.910
relative sales ≤ 100	0.978	0.979

Panel C: Cost Weighted Distribution of Markups

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

Figure 2: Demand Function

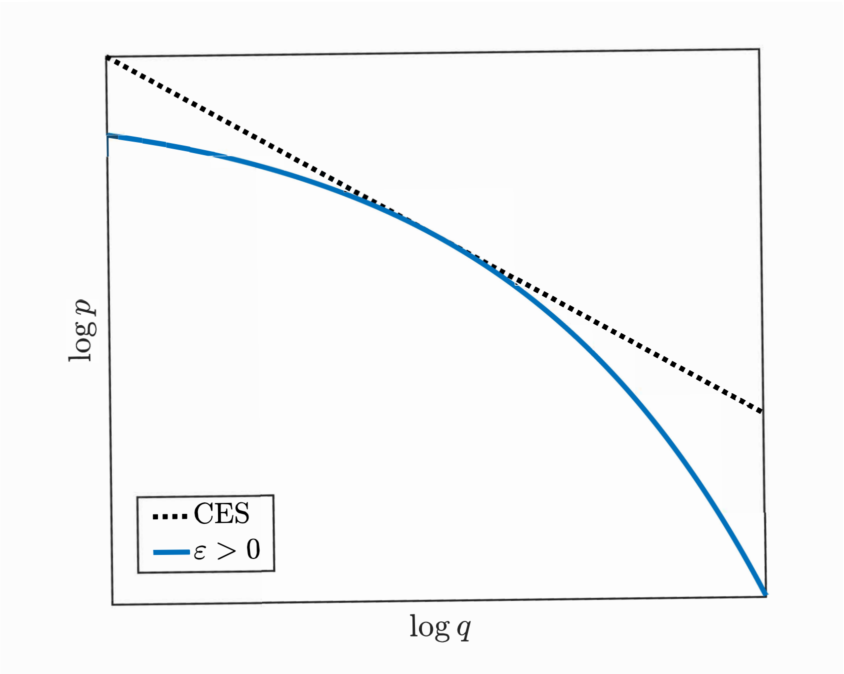


Figure 3: Static Choices

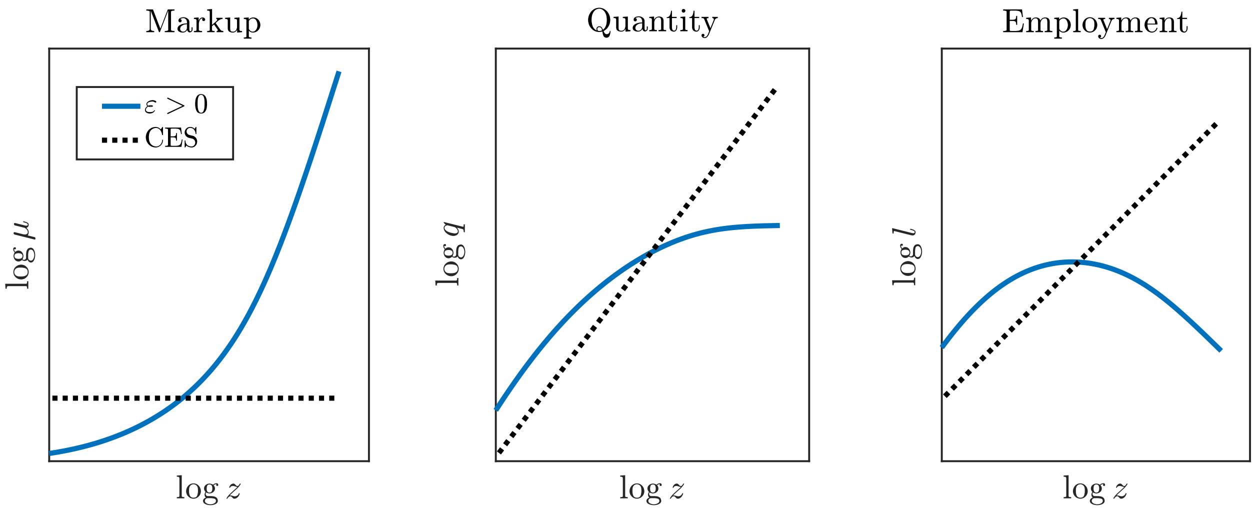


Figure 4: Equilibrium and Planner's Allocations Compared

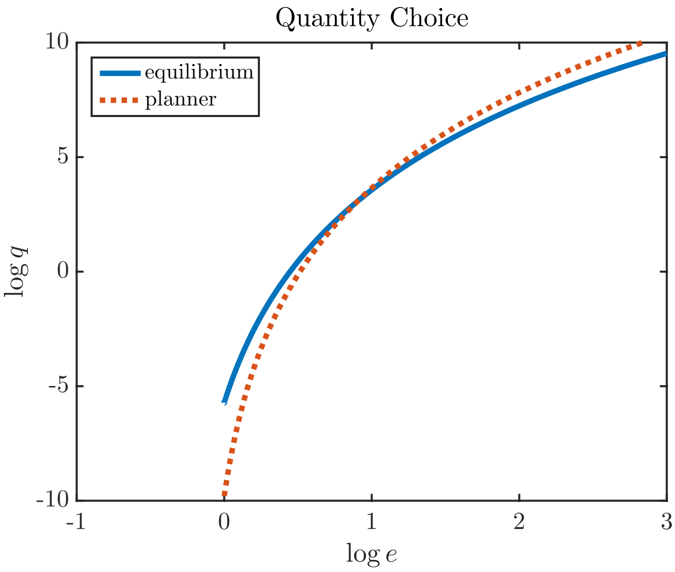


Figure 5: Entry Choice

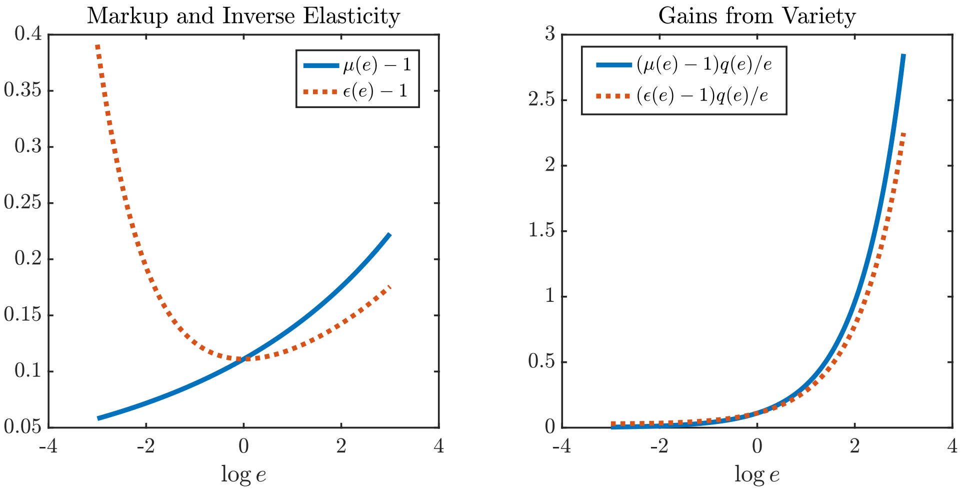


Figure 6: Optimal Size-Dependent Subsidy

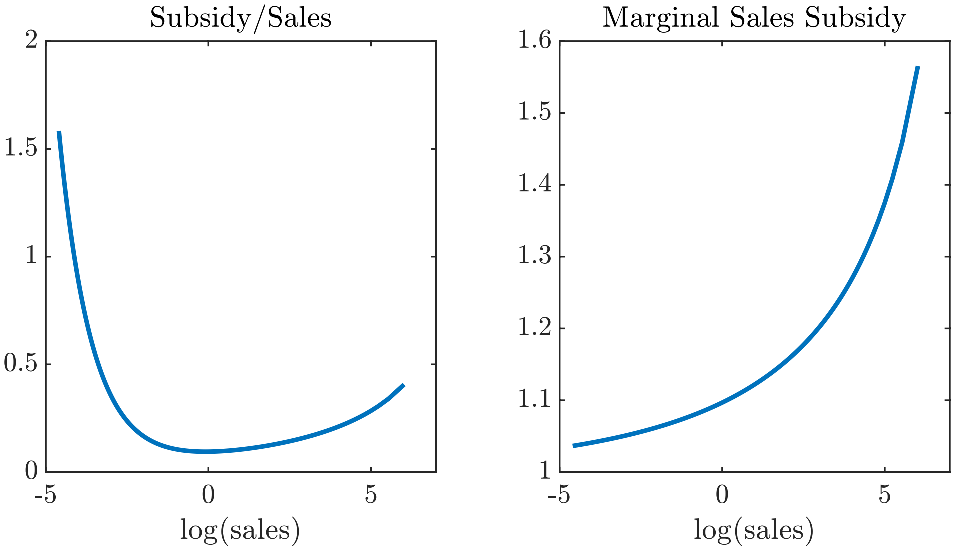


Figure 7: Relative Wage Bill vs. Relative Sales

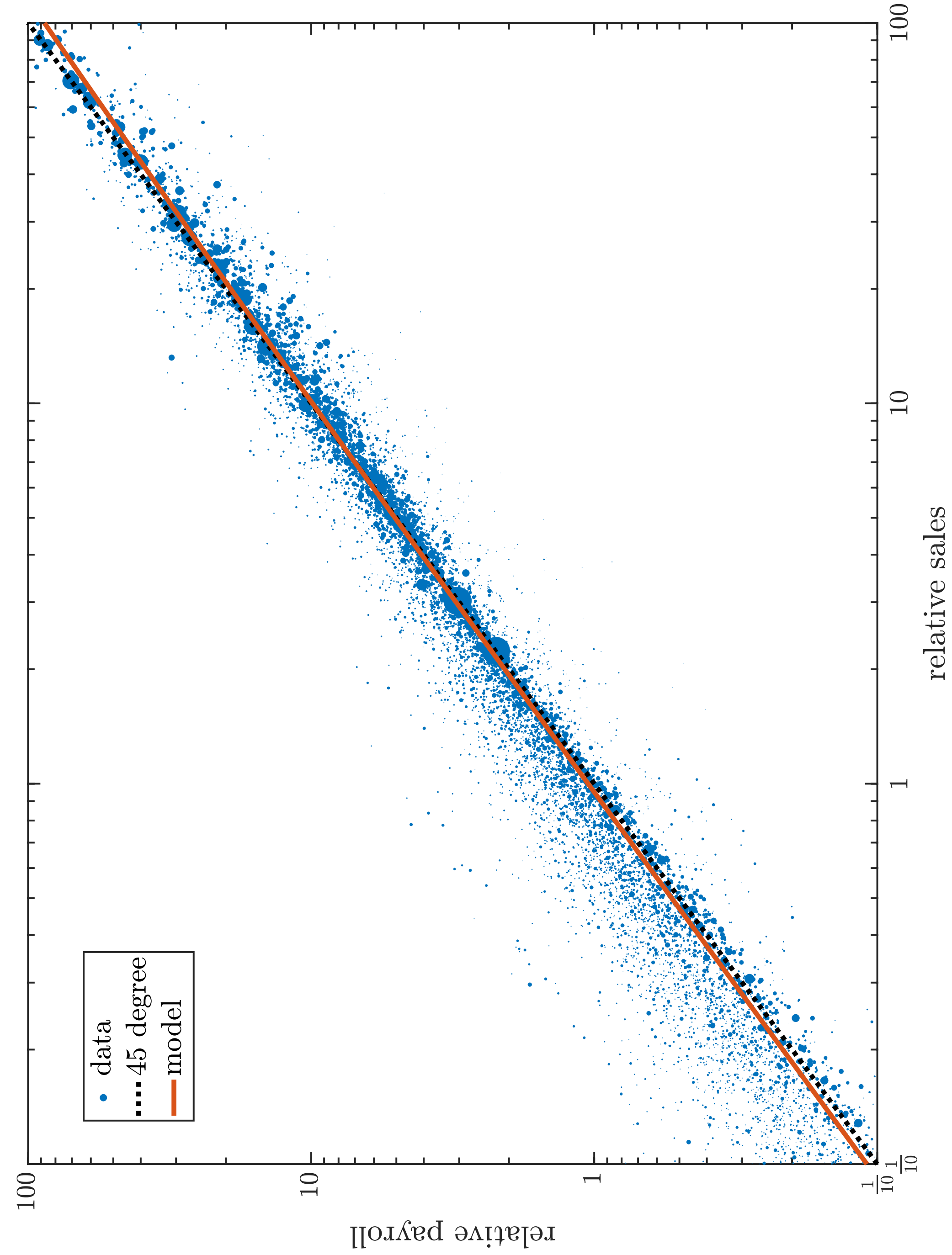


Figure 8: Relative Wage Bill vs. Relative Sales. High ε/σ

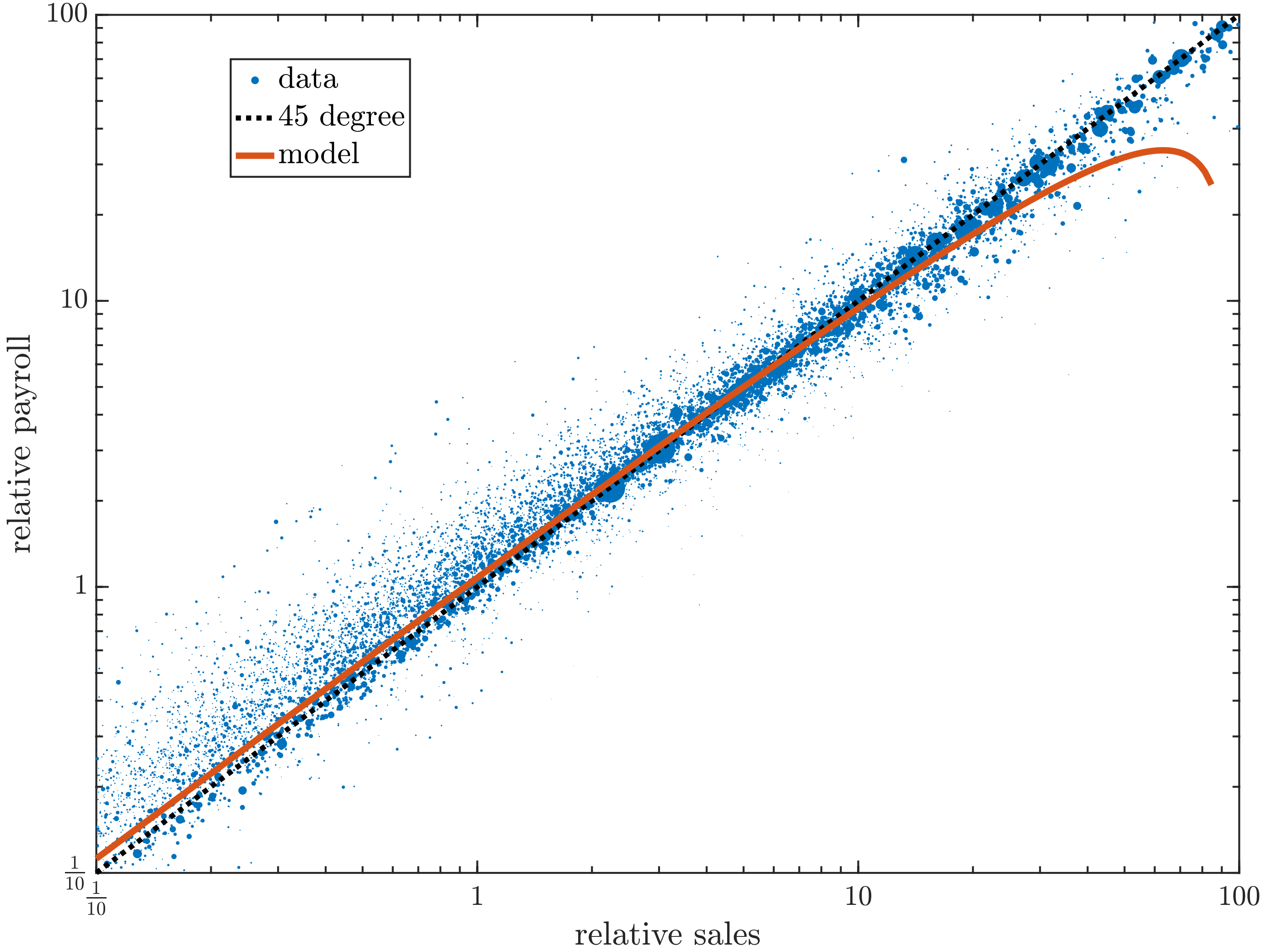


Figure 9: Cost vs. Sales-Weighted Average Markups, Compustat

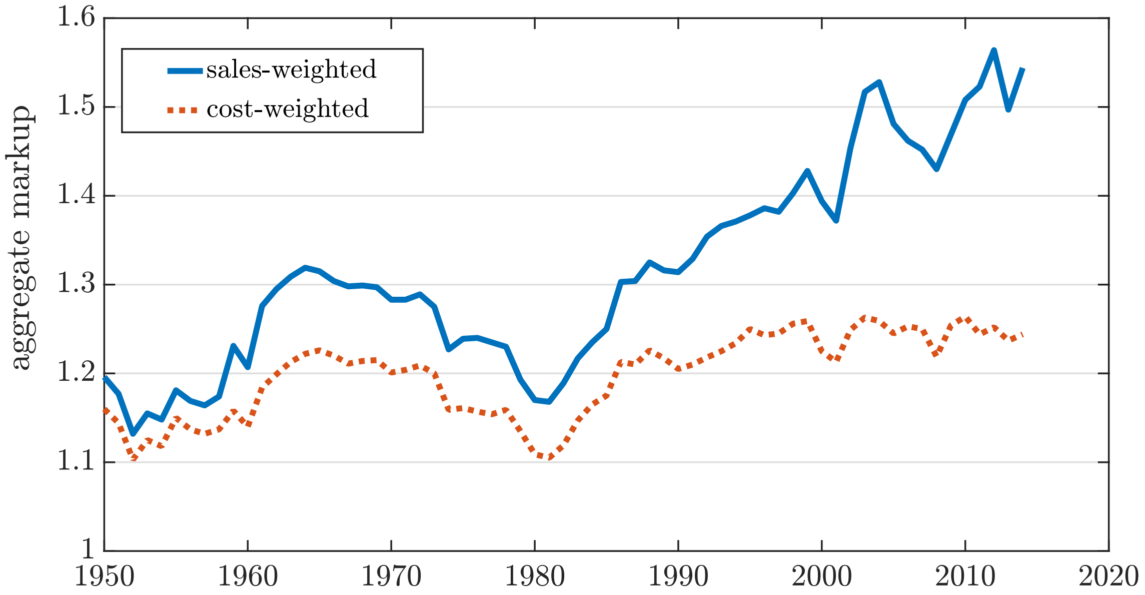


Figure 10: Transition to Efficient Allocations

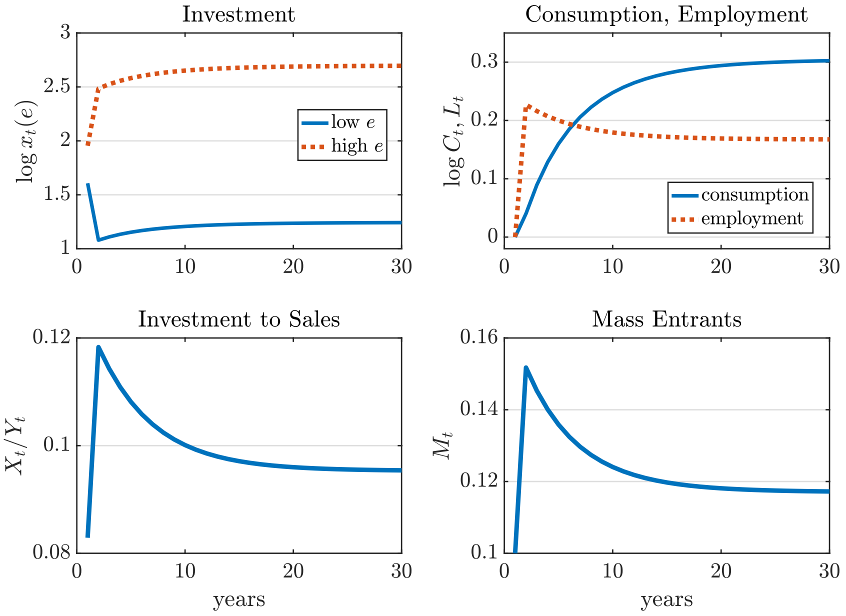


Figure 11: Transition Dynamics with Uniform Sales Subsidy

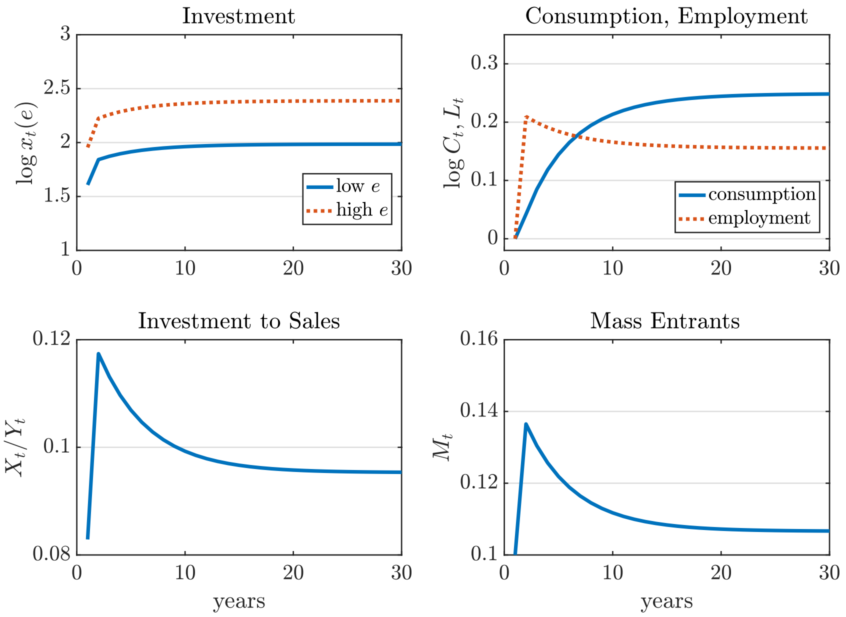


Figure 12: Transition Dynamics with Entry Subsidy

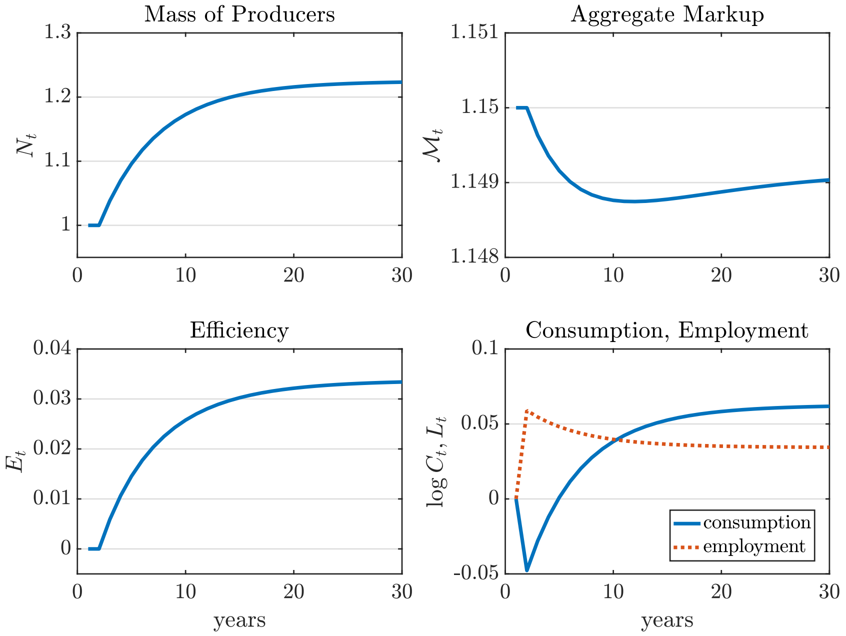


Figure 13: Effect of Entry Subsidy on Markups

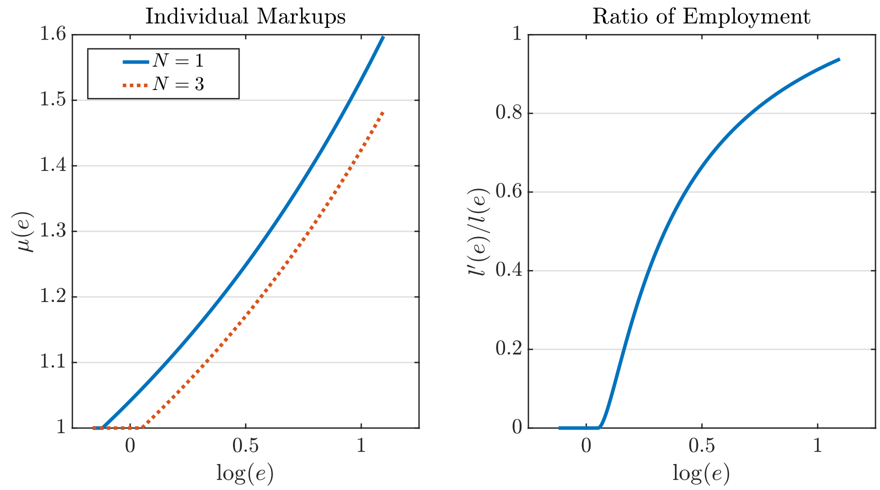
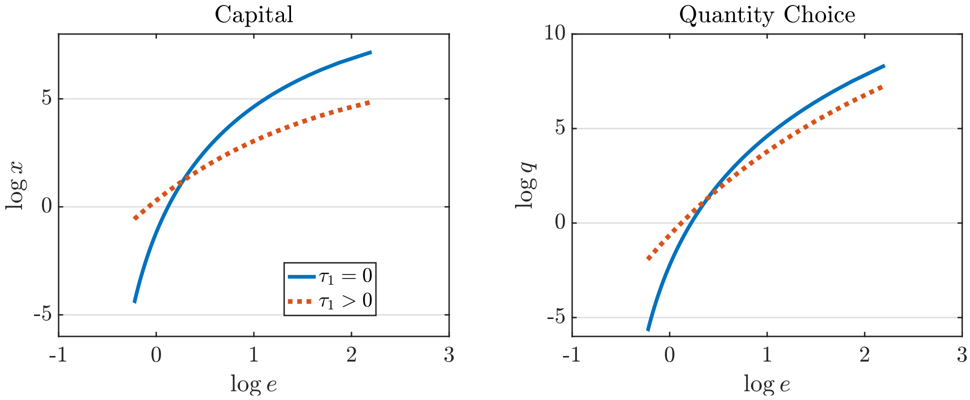


Figure 14: Effect of Size-Dependent Policies


Capital
$\log x$
$\log e$
$\tau_1 = 0$
$\tau_1 > 0$
Quantity Choice
$\log q$
$\log e$