NBER WORKING PAPER SERIES

ENTRY VS. RENTS

David Baqaee
Emmanuel Farhi

Working Paper 27140
http://www.nber.org/papers/w27140

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2020

Emmanuel Farhi acknowledges research financial support from the Ferrante fund at Harvard University. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Entry vs. Rents
David Baqaee and Emmanuel Farhi
NBER Working Paper No. 27140
May 2020
JEL No. E0,E1,E3,O0,O11,O21,O25,O3,O4,O41

ABSTRACT

We show that the tension between entry and rents lies at the core of a general theory of aggregation with scale effects. We characterize the responses of macro aggregates to micro shocks in disaggregated economies with general forms of entry, internal or external returns to scale, input-output linkages, and distortions. In particular, we decompose changes in aggregate productivity into changes in technical and allocative efficiency, and show that the latter depend on changes in rents and quasi-rents across markets. In addition, we give formulas for the social costs of distortions. Finally, we prove that while first-best industrial policy is network-independent, second-best policy supports the more "networked" parts of the economy by boosting the backward linkages of markets with high forward linkages and returns to scale. As an application, we quantify the misallocation from markups in the U.S.: accounting for entry raises the aggregate efficiency loss from 20% to 40% This number depends sensitively on how entry is modeled, in ways that we make precise.

David Baqaee
Department of Economics
University of California at Los Angeles
Bunche Hall
Los Angeles, CA 90095
and NBER
baqaee@econ.ucla.edu

Emmanuel Farhi
Harvard University
Department of Economics
Littauer Center
Cambridge, MA 02138
and NBER
emmanuel.farhi@gmail.com

A data appendix is available at http://www.nber.org/data-appendix/w27140

1 Introduction

In macroeconomics, exogenous aggregate productivity is god of the gaps, bridging the distance between model and data and explaining inexplicable variations in output. In this paper, we consider how scale economies, shaped by entry-exit and accompanying distortions, endogenously determine aggregate productivity and output. We analyze how microeconomic shocks translate into macroeconomic effects, characterize the efficiency losses from misallocation, and derive optimal first- and second-best industrial policies.

The class of models that we consider is much broader than the family of macro models typically used to analyze scale effects, for example, Dixit and Stiglitz (1977), Krugman (1979), Romer (1987), Aghion and Howitt (1992), Murphy et al. (1989), Hopenhayn (1992), and Melitz (2003). In particular, we allow for an arbitrary pattern of distorting wedges and technological heterogeneity. We allow for increasing, decreasing, or constant internal and external returns to scale, as well as within and cross-industry heterogeneity. Finally, we study disaggregated production structures accommodating input-output linkages in both production and entry.

We decompose changes in aggregate productivity into changes in technical and allocative efficiency via an aggregation equation:

$$\Delta \log TFP = \Delta \log TFP^{\text{tech}} + \Delta \log TFP^{\text{alloc}}.$$

Technical efficiency measures the direct impact of technology shocks, holding fixed the allocation of resources, and allocative efficiency measures the indirect effect of shocks due to the reallocation of resources.[1]

We show that changes in technical efficiency are given by

$$\Delta \log TFP^{\text{tech}} = \sum_i \lambda_i^{\text{F}} \Delta \log A_i,$$

where $\Delta \log A_i$ is an appropriately normalized productivity shock to the ith producer and λ_i^{F} is a measure of forward linkages from i to the household. The weight λ_i^{F} depends on expenditure shares and is related but not exactly equal to the size of i as measured by its sales as a share of GDP. It has the same intuition as Hulten (1978)'s theorem.

Under conditions that guarantee efficiency of the equilibrium, the logic of the envelope theorem implies that reallocation effects can be ignored to a first order $\Delta \log TFP^{\text{alloc}} = 0$, leaving only the direct effect of the shocks. Once we stray from efficiency, changes in

[1] There are different notions of changes in allocative efficiency. In this paper, we define them as changes in output due to reallocations of resources. See Baqaee and Farhi (2019a) for a detailed discussion.

allocative efficiency play a dominant role in determining the aggregate consequences of disturbances. These indirect reallocation effects depend on which markets expand and shrink. They also depend on whether these adjustments in market sizes occur through changes in the size of existing producers or through changes in the number of producers. We show that the resulting changes in allocative efficiency can be summarized by changes in rents and quasi-rents.

To be specific, we define rents to be income accruing to proprietors after variable costs have been deducted from revenues. Proprietors earn rents because of non-constant returns to scale (Ricardian rents) and because of markups (monopoly rents).[2] We define the quasi-rents associated with a given market as the expenditures on entry that were paid by the producers who entered that market.

Our treatment of entry is novel. In a broad class of models, entrants pay an entry cost to obtain a, perhaps random, production technology. In such models, the entry of new producers, and the quasi-rents associated with that entry, can be represented using linear projections. Let λ_π denote the vector of rents as a share of GDP in each market. We show that changes in quasi-rents associated with each market is the projection of the vector $\Delta \log \lambda_\pi$ on the space spanned by the linear entry technology

$$\widehat{\Delta \log \lambda_\pi},$$

with residual

$$\Delta \log \lambda_\pi - \widehat{\Delta \log \lambda_\pi}.$$

The projection determines the amount of entry into the different markets and interacts with increasing external economies. The residual measures the imperfect ability of entry to keep up with rents and interacts with decreasing internal economies. Therefore, in a least-squares sense, entry minimizes rents claimed by existing producers, and the projection and residuals from a regression summarize reallocation effects in general equilibrium.

In particular, in response to productivity shocks, changes in allocative efficiency are

$$\Delta \log TFP^{\text{alloc}} = -\sum_i \lambda_i^{\text{F}} \mathcal{E}_i^{\text{int}} \left(\Delta \log \lambda_{\pi,i} - \widehat{\Delta \log \lambda_{\pi,i}} \right) + \sum_i \lambda_i^{\text{F}} \mathcal{E}_i^{\text{ext}} \widehat{\Delta \log \lambda_{\pi,i}},$$

where $\mathcal{E}_i^{\text{int}}$ is an internal scale elasticity (of market output to the variable inputs of producers), $\mathcal{E}_i^{\text{ext}}$ is an external scale elasticity (of market output to the number of producers), and the sum is over all markets i (including primary factors). We derive similar formulas for

[2]Here, monopoly rents also includes all the revenues collected by distortionary wedges (since other distorting wedges, say taxes, can be represented as markups).

the response to changes in markups and other distortions.

There are two terms in this expression. The first term depends on decreasing internal returns to scale ($\mathcal{E}_i^{\text{int}} > 0$). A positive residual in market i means that quasi-rents are failing to keep up with rents. This implies that producers in market i are scaling up and running into diminishing returns. This raises the shadow price of their producer-specific fixed factor and lowers output. If weighted sum of residuals is negative, that means beneficial reallocations, by making better use of resources, have made factors of production less scarce. The second term depends on external increasing returns to scale. With increasing external returns to scale ($\mathcal{E}_i^{\text{ext}} > 0$), allocative efficiency further improves if reallocations increase entry in markets with high external economies on balance across markets. The sum of these terms is always zero in efficient economies, but not in inefficient economies.[3]

It is often believed that entry when there are decreasing internal returns, like in Hopenhayn (1992), and entry when there are increasing external returns, like in Dixit and Stiglitz (1977) or Melitz (2003), give similar results. However, the equation above shows that this folk wisdom is generally incorrect when there are inefficiencies. Reallocation effects in Dixit-Stiglitz/Melitz-type models depend on the projection, whereas in the Hopenhayn-type models they depend on the residual. This distinction matters and leads to differences in macroeconomic behavior.

We complete this new perspective by providing propagation equations which show how changes in these different categories of rents are determined in equilibrium as a function of the microeconomic primitives and the shocks. These propagation equations, which capture backward and forward propagation through supply chains, also characterize how every price and quantity responds to a shock in equilibrium. The aggregation and propagation equations fully characterize the model's positive properties to a first order.

From a normative perspective, we also characterize optimal industrial policy as well as the gains from implementing it. We show that while first-best policy is network-independent, second-best policies do depend very much on the network structure. In particular, for economies with increasing returns, we rationalize and revise Hirschman (1958)'s influential argument that policy should encourage those sectors with the most forward and backward linkages, and we give precise formal definitions for these loose concepts. We show that the optimal marginal intervention aims to boost backward linkages for producers that have relatively high forward linkages and returns to scale.

Finally, we show that the social cost of inefficiencies is approximately the sales-

[3]Both terms in this formula can be interpreted as changes in the shadow value of fixed-factors associated with non-constant-returns-to-scale. There are fixed factors associated with decreasing internal returns, and fixed factors associated with increasing external returns. The former have positive shadow prices, and the latter have negative shadow prices.

weighted sum of a series of Harberger triangles, some associated with production and some associated with entry. We also characterize these Harberger triangles in terms of microeconomic primitives.

Although our main contribution is theoretical, we also provide an example application by quantifying the social costs of markups using micro data for the U.S. We decompose the losses into losses arising from misallocation of resources in production (due to dispersion in markups) and misallocation of resources in entry (due to bias in average markups). One might imagine that since markups incentivize entry, models with endogenous entry would assign smaller losses to markups than models without an entry margin. On the contrary, we find that distortions on the entry margin, caused by the markups, are quantitatively as important as distortions on the production margin. To use a concrete example, without entry we find that markups estimated by a production-function approach à la De Loecker et al. (2019) reduce aggregate productivity by around 20%.[4] Accounting for entry can double these losses. Furthermore, the specific number one attaches to these losses depends sensitively on knowledge of the production structure, including the strength of external economies, the extent to which entry is targeted, the resources used for entry, and the view one takes on the presence of entry restrictions in the data. While these features are critical on a theoretical and quantitative level, little is know about them in practice, and more empirical work is needed to convincingly measure them.

The structure of the paper is as follows. In Section 2, we set up the general model and define the equilibrium notion. In Section 3, we prove conditions under which the equilibrium is efficient and derive comparative statics for the efficient case. In Section 4, we specialize the model and introduce notation necessary to analyze inefficient equilibria. In Section 5, we provide and discuss the aggregation formula for how shocks affect aggregate output. Section 6 contains backward and forward propagation equations that determine how rents respond to shocks as a function of primitives. In Section 8, we analyze the normative properties of the economy, including first- and second-best optimal policy and the social costs of distortions. Finally, Section 9 is a quantitative application where we use a calibrated model to compute and dissect the social costs of markups and the benefits of industrial policy in the U.S. using firm-level data on markups.

Related Literature. Our results apply to a broad range of popular models in the macro, trade, and growth literatures. For instance, our framework encompasses and generalizes

[4]We also use alternative approaches for estimating markups: an alternative implementation of the production-function (PF) approach with different categories of costs, the user-cost approach (UC), and the accounting-profits (AP) approach. Although the numbers depend on the specification, the qualitative message remains the same.

models of entry like Dixit and Stiglitz (1977) or (a finite-horizon version of) Hopenhayn (1992), the closed economy version of Melitz (2003), and finite-horizon versions of models of endogenous growth with lab-equipment like Romer (1987) and Grossman and Helpman (1991). It also nests multi-sector and production network models like Hulten (1978), Long and Plosser (1983), and much of the subsequent literature like Gabaix (2011), Acemoglu et al. (2012), Carvalho and Gabaix (2013), and Baqaee and Farhi (2019b), amongst others.

This paper is most closely related to Baqaee (2018) and Baqaee and Farhi (2019a) which establish aggregation and propagation results for inefficient production networks with and without entry. Baqaee (2018) considers production networks with external economies, entry, and distortions. This paper builds on that framework using a more general model, allowing for a more sophisticated handling of the entry condition, returns to scale, production functions, and network linkages in both production and entry. Furthermore, unlike Baqaee (2018), this paper also characterizes reallocation, misallocation, and optimal policy.

On the other hand, Baqaee and Farhi (2019a) analyze reallocation and misallocation but, unlike this paper, abstract from entry. This paper nests and provides a new angle on that paper. In particular, Baqaee and Farhi (2019a) show that a reduction in factor shares can imply an improvement in allocative efficiency due to reallocation. In this paper, the factor shares are rents not offset by entry, and so a decline in the price of these factors represents beneficial reallocation (i.e. beneficial reallocation makes factors cheaper/less scarce).

This paper also relates to other papers on cross-sectional misallocation and industrial policy, with or without externalities, like Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Epifani and Gancia (2011), Edmond et al. (2018), Liu (2017), Osotimehin and Popov (2017), Behrens et al. (2016), and Bartelme et al. (2019). By showing that even in non-neoclassical models losses can be approximated using Harberger triangles, the paper also extends the insights of Harberger (1954) and Harberger (1964).

2 General Framework

The model consists of a representative household, a set of producers, and a set of entrants. In this section, we describe the model, and define the equilibrium. A circular flow diagram of the economy is depicted in Figure 1. Each rectangle represents a type of agent in the model. Loosely speaking, entrants buy resources to enter. After paying the entry costs, entrants are assigned (perhaps randomly) to produce. Meanwhile, producers produce using intermediate materials they purchase from other producers. The representative

household owns all resources in the economy and purchases consumption goods using national income. We begin by describing the problem each agent is faced with, starting with the producers.

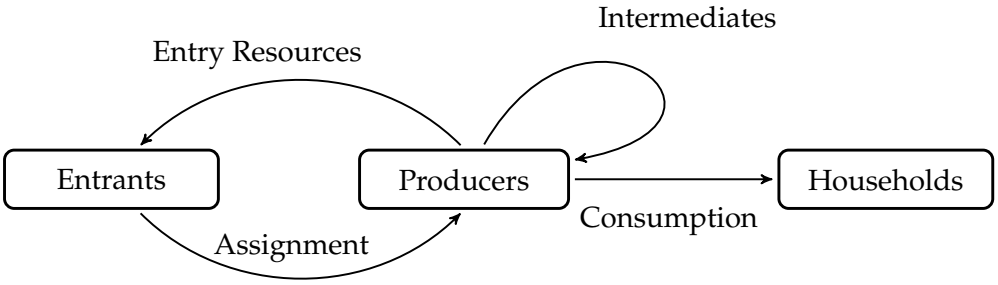


Figure 1: Circular flow schematic of the economy showing the flow of resources.

2.1 Markets and Producers

There is a set of *markets* indexed by $i \in \mathcal{N}$. Each market i is populated by an endogenous mass M_i of identical producers with output

$$y_i = f_i\left(\left\{x_{ij}\right\}_{j\in\mathcal{N}}, A_i\right),$$

where f_i is a neoclassical production function, A_i is some scalar indexing productivity, x_{ij} is the input quantity of market good j (including primary factors). Each producer minimizes costs and sets its price p_i^y equal to its marginal cost times an exogenous markup μ_i.

The output good of market i is given by

$$Y_i = F_i\left(M_i y_i\right),$$

where the market aggregator F_i may have constant, decreasing, or increasing returns to scale in the producer-level output y_i. The price of the market good P_i^Y is equal to the marginal cost of producing Y_i times an exogenous wedge μ_i^Y. Unlike the producer-level markup μ_i, revenues generated by the market-level wedge μ_i^Y are *not* rebated to the owner of i and instead go directly to the household. The market-level wedge μ_i^Y therefore acts like an output tax.

To understand the versatility of this modeling block, consider the following examples. Let x_i denote a bundle of inputs and ignore productivity by setting $A_i = 1$ as an argument.

Assume that $f_i(x) = x^{1-\varepsilon_i}$. Suppose that $F_i(x) = x^{\frac{1}{1-\varepsilon_i}}$. Then $Y_i = (M_i x_i^{1-\varepsilon_i})^{\frac{1}{1-\varepsilon_i}}$ captures a CES market structure with an elasticity of substitution $1/\varepsilon_i$ between differentiated varieties produced under constant returns to scale. Suppose instead that $F_i(x) = x$. Then $Y_i = M_i x_i^{1-\varepsilon_i}$ captures a market structure with perfectly substitutable varieties produced under decreasing returns to scale.

Primary Factors. A subset of markets $\mathcal{F} \subset \mathcal{N}$ correspond to *primary factors*. These markets are populated by an exogenous mass M_f of producers whose production functions f_f have zero returns to scale. We also assume that the market aggregator has constant returns to scale $F_f(M_f y_f) = M_f y_f$. In addition, we assume that there are no markups/wedges $\mu_f = \mu_f^Y = 1$. Basically, there is no entry into the market, each producer produces a fixed amount of output, and producer outputs are aggregated linearly, so that total market output is also fixed. This allows us to capture endowments of primary factors such as labor, land, or the initial capital stock.

2.2 Entrants

There is an infinite supply of potential entrants who are grouped into types indexed by $j \in E$. Entrants pay fixed costs and enter subject to a zero-profit condition.

Fixed Costs. To enter, potential entrants must pay a fixed cost

$$g_j\left(\left\{x_{E,ji}\right\}_{i\in\mathcal{N}}\right), \tag{1}$$

where g_j has constant returns, and $x_{E,ji}$ is the input quantity of market good i. A simple example is when firms pay entry costs in units of labor if they choose to enter, as in Hopenhayn (1992) or Melitz (2003).

The entry matrix ζ is an $|E| \times |\mathcal{N} - \mathcal{F}|$ positive-valued matrix. Type-j entrants who pay the sunk cost are randomly assigned, according to $\zeta(j, i)$, the ability to produce in market $i \in \mathcal{N} - \mathcal{F}$. Without loss of generality, assume that the rows of ζ are linearly independent.[5] A simple example is that there is only one type of entrant and technology is assigned randomly, as in Hopenhayn (1992) or Melitz (2003). We denote by $M_{E,j}$ the endogenous mass of type-j entrants who pay the entry cost.

If there is no way to enter market $i \in \mathcal{N}$, which occurs when $\zeta(j, i) = 0$ for all $j \in E$, then we allow for an exogenous mass M_i of incumbents who operate in the market without

[5] If the rows of ζ are not linearly independent, then some entry types are redundant (can be replicated by playing a mixed entry strategy).

having to enter.

We refer to markets where entry is not possible as *uncontested markets* and denote their collection by $\mathcal{N}^c$. We also sometimes simply call them *incumbents*, since each of these markets operate like a representative incumbent. We refer to markets where entry is possible as *contested markets* and denote their collection by $\mathcal{N}^u$. Note that since we are flexible in the way we define and combine markets, we can capture a situation where incumbents and entrants coexist by having them operate in different markets with highly-substitutable market goods.

Sunk vs. Overhead Costs. The entry matrix ζ can capture sunk and overhead costs simultaneously. To capture sunk costs, suppose that $\zeta(j,i)$ has positive support for a range of different i's. In this case, once the entry cost j has been paid, the entrant will always choose to operate all of its technologies i since the entry cost is sunk. At the other extreme, suppose that $\zeta(j,i) = 1$ for one specific i and zero otherwise. In this case, entrant j will only choose to pay the cost if operating technology i is worth paying the fixed cost. In other words, the fixed cost is not sunk.

We can also consider intermediate situations in which entrant j pays a sunk cost and draws a mixture of zero-returns technologies j'. Other entrants j'' can purchase the output of j' and combine it with another fixed cost to enter with certainty into producing i. This structure mimics the entry decision in standard models such as Hopenhayn (1992) and Melitz (2003) where potential entrants first pay a sunk cost and then decide whether or not to pay an additional overhead cost before operating.[6]

Zero-Profit Conditions. The zero-profit condition for type-j entrants is

$$\sum_i \frac{\zeta(j,i)M_{E,j}}{M_i}\lambda_{\pi,i} = M_{E,j}\sum_{k\in\mathcal{N}} P_k^Y x_{E,jk},$$

where

$$\lambda_{\pi,i} = M_i p_i^y y_i - M_i \sum_{j\in\mathcal{N}} P_j^Y x_{ij}$$

is the total *rent* or *variable profit* (we use the two terms interchangeably) earned by all the producers of market i. The left-hand side of the zero-profit condition is the expected

[6]The difference between our treatment of overhead costs and that in Hopenhayn (1992) and Melitz (2003) is that we assume divisibility and that they assume non-divisibility. We could capture non-divisibility by letting $g_j(\{x_{E,ji}\}_{i\in\mathcal{N}})$ have variable (possibly increasing) returns to scale (for example, by making it a step function). This would not affect Theorems 1 or 2. See Appendix A for more details and for an alternative set-theoretic formalization which sidesteps these issues.

total rent earned by type-j entrants and the right-hand side is the total cost of entry. This condition ensures that the rents earned by type-j entrants are *quasi-rents* rather than *pure rents*.

2.3 Households

There is a representative household whose preferences are given by a homothetic utility function over market goods

$$Y = \mathcal{D}\left(\{C_i\}_{i \in \mathcal{N}}\right).$$

To avoid corners, we require that $Y \leq 0$ whenever $C_i = 0$ for any $i \in \mathcal{N}$.

The budget constraint of this representative household requires total final expenditure to equal total income defined as revenues net of expenditures

$$\sum_{i \in \mathcal{N}} P_i^Y C_i = \sum_{i \in \mathcal{N}} P_i^Y Y_i - \sum_{j \in \mathcal{N}} P_j^Y x_{ij} - \sum_{j \in E} M_{E,j} \sum_{k \in \mathcal{N}} P_k^Y x_{E,jk}.$$

Note that payments to primary factors are included as the revenues of zero-returns-to-scale incumbents in markets $\mathcal{F} \subset \mathcal{N}$.

2.4 Resource Constraints

The resource constraint for market good $i \in \mathcal{N}$ is

$$Y_i = C_i + \sum_{j \in \mathcal{N}} M_j x_{ji} + \sum_j M_{E,j} x_{E,ji},$$

where the mass of entrants for contested markets $i \in \mathcal{N}^c$ is given by

$$M_i = \sum_{j \in E} \zeta(j,i) M_{E,j}.$$

This reflects the fact that market good i is used by households, producers (as intermediate inputs), and entrants (as fixed costs), and that non-incumbent producers in market i are entrants of different types.

2.5 Equilibrium

The decentralized equilibrium is an allocation of resources and collection of prices which clears markets and solves each agents' decision problem.

Definition 1. A *decentralized equilibrium* is a collection of prices $\left\{P_i^Y\ p_i^y\right\}$ and quantities $\left\{C_i, Y_i, y_i, x_{ij}, x_{E,ij}, M_{E,j}, M_i\right\}$, such that given productivities $\{A_i\}$ and markups/wedges $\left\{\mu_i, \mu_i^Y\right\}$: (i) the representative household maximizes utility; (ii) each price is equal to marginal cost times the markup; (iii) entrants earn zero profits; (iv) prices clear all markets.

We treat markups/wedges as exogenous and provide comparative statics with respect to changes in technology and in markups/wedges. Endogenizing markups/wedges requires additional assumptions and results in additional equations for changes in markups. Those equations can then be combined with our comparative statics, using the chain rule, to generate comparative statics.

Since we allow for reduced-form wedges, this means that many types of distortions like taxes, financial frictions, or nominal rigidities, are nested as special cases. For instance, to capture a financial friction on i's ability to purchase inputs, add a fictitious incumbent producer to the model who buys inputs on behalf of i. An output wedge on this fictitious producer can then implement the same allocation as a financial friction on i.

Similarly, since we allow for productivity shocks to producers, we can capture productivity shocks to the entry or overhead costs of operation by adding fictitious incumbents who buy inputs to be used for entry or overhead.

Finally, using the Arrow-Debreu trick of indexing commodities by dates and states of the world, we can capture dynamic stochastic models.

Going forward, the primary object of interest is the response of aggregate output $\mathrm{d}\log Y$ to shocks. Since the supply of primary factors is fixed, changes in aggregate output also coincides with changes aggregate productivity $\mathrm{d}\log TFP$ as well as with changes consumer welfare.[7] In Appendix F.2 we generalize our results to the case where factor supply is not perfectly inelastic .

2.6 Noteworthy Special Cases

At this level of abstraction, with appropriately defined markups, the model nests most general equilibrium models with entry, including models where goods are perfectly substitutable and firms have diminishing returns, as well as models where goods are imperfectly substitutable and firms have constant marginal cost. For example, it nests models of industry dynamics like (a finite-horizon version of) Hopenhayn (1992), the closed-economy

[7] We abstract away from the well-understood issues related the treatment of new goods in the measurement of aggregate output. However, we note that they depend on the extent to which new goods appear directly in final demand or as intermediates. To the extent that new goods appear as intermediates, they are adequately captured in real GDP as it is constructed in the data.

version of Melitz (2003), models with product variety like Dixit and Stiglitz (1977), Krugman (1979), and Dhingra and Morrow (2019), (finite-horizon versions of) growth models with lab-equipment, like Romer (1987) and Grossman and Helpman (1991), and models of production networks without entry like Baqaee and Farhi (2019a) or with entry like Baqaee (2018).

3 Marginal-Cost-Pricing Benchmark

In this section, we consider the marginal-cost pricing benchmark defined as follows.

Definition 2. A *marginal-cost pricing equilibrium* is a decentralized equilibrium where $\mu_i = \mu_i^Y = 1$ for all $i \in \mathcal{N}$.

We prove two theorems. Theorem 1 shows that the marginal cost-pricing benchmark is efficient. This normative theorem is also important from a positive perspective since it ensures that the response of aggregate output to shocks can easily be obtained by applying the envelope theorem. Theorem 2 uses this insight to derive comparative statics.

Theorem 1 (First Welfare Theorem)**.** *The marginal-cost pricing equilibrium is Pareto-efficient.*

Theorem 1, which generalizes the first welfare theorem to an environment with fixed and sunk costs of operation is interesting for both normative and positive reasons. From a normative perspective, it immediately clarifies how the optimal allocation can be implemented using linear taxes, and we use this implementation in Section 8 when we approximate the decentralized economy's distance from the Pareto-efficient frontier. From a positive perspective, Theorem 1 implies the following result.

Theorem 2 (Comparative Statics under Efficiency)**.** *In the marginal-cost pricing equilibrium, the response of aggregate output to a Hicks-neutral productivity shock* $d \log A_i$ *is given by*

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_i} = \frac{M_i p_i^y y_i}{GDP},$$

which is the total sales of market i as a share of GDP. Similarly, the response of aggregate output to an entry productivity shock $\mathrm{d} \log \zeta(j,i)$ *is given by*

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log \zeta(j,i)} = \frac{\lambda_{\pi,i} \zeta(i,j) M_{E,j}}{GDP},$$

which is the rents earned by type-j entrants from producing in market i as a share of GDP.

Theorem 2 is an envelope theorem which extends Hulten (1978) to economies with selection, fixed costs, increasing returns, and an extensive margin of product creation and destruction. In particular, it shows that, for marginal-cost-pricing equilibria, simple and readily observable sufficient statistics like the sales or profit shares summarize the macroeconomic impact of microeconomic disturbances in general equilibrium.[8]

Extending Theorem 2 to cover biased technical change, for example factor-augmenting shocks, or shocks to the entry or overhead costs of operation is trivial. To model these shocks, say a shock to i's ability to use input k, simply introduce a new producer who buys from k and sells to i. A Hicks-neutral shock to this new producer is the same as a biased shock in the original model. This trick allows us to restrict attention to Hicks-neutral shocks without loss of generality. In the next section, we derive comparative statics for Hicks-neutral shocks when the economy is inefficient.

4 Inefficient Framework

Comparative statics in efficient models are easy to derive because, following the logic of the envelope theorem, reallocation effects can be ignored. Comparative statics in inefficient models are harder to obtain because, since the envelope theorem no longer applies, reallocation effects can no longer be ignored. To emphasize our mechanisms of interest, we specialize our general framework by making some simplifying assumptions.

4.1 Capturing IRS and DRS

We split the production Y_i of market good i into three steps.

Assumption 1. For each $i \in \mathcal{N}$, there is γ_i and ε_i in $[0,1]$ such that[9]

$$Y_i = (M_i y_i)^{\frac{1}{\gamma_i}}, \qquad y_i = q_i^{1-\varepsilon_i}, \quad \text{and} \quad q_i = A_i^{\frac{\gamma_i}{1-\varepsilon_i}} f_i\left(\left\{x_{ij}\right\}_{j\in\mathcal{N}}\right),$$

where f_i has constant returns to scale.

In effect, this assumption splits the production of every market good i into three distinct steps each with a homothetic production function.[10] Inputs are first combined together to

[8] For the proof in Appendix A, we also allow for non-divisible overhead costs.

[9] Our choice of notation for the exponents $1-\varepsilon_i$ and γ_i reflects a compromise to make the specializations of the formulas to both the DRS case and the CRS case natural.

[10] In Appendix F.1 we relax Assumption 1, and show that our results extend to the case where internal economies are non-isoelastic allowing for variable returns to scale at the producer level, and where external economies are non-isoelastic along the lines of Kimball (1995).

form producer-level q_i. These are then passed through a decreasing returns to scale (DRS) production function to make producer-level y_i. Finally, the mass M_i of y_i's are linearly combined and passed through an aggregator function to make the aggregated good Y_i. The aggregator F_i may be subject to increasing returns to scale (IRS) $\gamma_i < 1$. The parameters ε_i and γ_i control internal and external returns to scale (on the margin) respectively. The exponent on A_i is a convenient normalization made without loss of generality to ensure that Y_i is unit-elastic in the productivity shock.[11]

We also allow at each step for markups/wedges over marginal costs μ_i^q, μ_i^y, and μ_i^Y. Letting p_i^q and p_i^y be the prices of q_i and y_i, we can derive the following relationships between the revenues of the different steps and the costs of the first step

$$P_i^Y Y_i = \gamma_i \mu_i^Y M_i p_i^y y_i = \frac{\gamma_i}{1-\varepsilon_i} \mu_i^Y \mu_i^y M_i p_i^q q_i = \frac{\gamma_i}{1-\varepsilon_i} \mu_i^Y \mu_i^y \mu_i^q M_i \left(\sum_{j \in \mathcal{N}} P_j^Y x_{ij} \right). \tag{2}$$

Although it complicates the notation, we introduce both q_i and y_i because it makes it more straightforward to map the model to the rest of the literature. In practice, there are two common approaches for setting up entry models. By explicitly tracking q_i, y_i, and Y_i, and assigning markups/wedges μ_i^q, μ_i^y, and μ_i^Y, for each, we can easily switch back and forth between these two interpretations.

The first approach exemplified by Hopenhayn (1992), imagines that the sales data should be mapped to $p_i^y y_i$, with goods that are perfect substitutes and produced with diminishing returns. It is obtained by assuming that $1 - \varepsilon_i < \gamma_i = 1$ and that $\mu_i^q = \mu_i^Y = 1$. We then have decreasing internal returns to scale and constant external returns to scale. We refer to this approach as the DRS benchmark.

The second approach, exemplified by Dixit and Stiglitz (1977), imagines that the sales data should be mapped to $p_i^q q_i$, with goods that are imperfect substitutes a la CES and produced with constant returns. This case is obtained by assuming that $1 - \varepsilon_i = \gamma_i < 1$ and that $\mu_i^y = 1 - \varepsilon_i$ and $\mu_i^Y = 1/\gamma_i$. We then have constant internal returns to scale and increasing external returns to scale (due to love of varieties). We refer to this approach as the IRS benchmark. Note that under the IRS benchmark, there are offsetting (implicit) markups and markdowns $\mu_i^y = 1/\mu_i^Y$ separate from the usual markup μ_i^q.[12]

Proposition 1. *The DRS and IRS benchmarks can be obtained as special cases of the model. They are attained by setting for each $i \in \mathcal{N} - \mathcal{F}$:*

[11] For comparison, note that we did not impose this unit-elasticity normalization in Section 2.

[12] Intuitively, when solving models with product differentiation, we implicitly assume that the curvature from the CES aggregator itself does not generate any income, or in other words that y_i and Y_i are sold at average rather than marginal, cost.

1. $\gamma_i = 1$, $\mu_i^q = 1$, *and* $\mu_i^Y = 1$ *for the DRS benchmark;*

2. $\gamma_i = 1 - \varepsilon_i$, $\mu_i^y = 1 - \varepsilon_i$, *and* $\mu_i^Y = 1/\gamma_i$, *for the IRS benchmark.*

A corollary of Proposition 1 and Theorem 1 is that the competitive equilibrium under perfect competition ($\mu_i^y = 1$) in Hopenhayn-style models are generally efficient, even if such models are generalized to include multiple sectors and input-output linkages.

Define $\mu_i = \mu_i^y \mu_i^q$. We assume that the revenues generated by μ_i^q and μ_i^y both go to the entrant, so it is only their product that matters for the determination of equilibrium outcomes. Therefore, we refer to μ_i as *the* markup of market i. On the other hand, we assume that the revenues generated by μ_i^Y go directly to the household and are not earned by entrants, therefore, we refer to μ_i^Y as the *wedge* of market i (isomorphic to an output tax).

The next assumption rules out corners in M_i by ensuring that markups are not so low that producer i always makes negative profits.

Assumption 2. For each $i \in \mathcal{N}^c$, $\mu_i \geq 1 - \varepsilon_i$.

4.2 Notation and Other Preliminaries

We normalize nominal GDP to one throughout. This means that all prices are quoted in the nominal GDP numeraire and that all sales, revenues, expenditures, and costs are expressed as shares of GDP.

We also represent the final demand function $Y = \mathcal{D}(C_1, \ldots, C_N)$ as the first producer in $\mathcal{N}$. In other words, we represent real GDP as the output of some incumbent producer standing in for the household. To emphasize the unique role the household plays in the economy, we index it by the number 0, to remind the reader that the zero-th producer is the household.

All the objects introduced below are defined at the initial equilibrium (around which we provide first-order and second-order approximations). We normalize the mass of entrants $M_{E,j}$ to one at the initial equilibrium.

The Entry Matrix

Define the $|\mathcal{E}| \times |\mathcal{N}|$ normalized entry matrix $\tilde{\zeta}$ by

$$\tilde{\zeta}(j,i) = \frac{\zeta(j,i) M_{E,j}}{\sum_{k \in E} \zeta(k,i) M_{E,k}}$$

whenever market i is contested and zero otherwise. This matrix gives the fraction of producers in market i who are type-j entrants.

Let λ_i^{B} be the $|\mathcal{N}| \times 1$ vector of sales $P_i^Y Y_i$. The superscript B anticipates that sales are the natural measure of *backward* linkages in this model. Let λ_π and λ_E be the $|\mathcal{N}| \times |\mathcal{N}|$ and $|E| \times |E|$ diagonal matrices of total rents (variable profits) and total expenditures on entry. The zero-profit condition for entry implies that for each entrant type j, entry costs exactly offset expected rents

$$\lambda_{E,j} = \sum_{i \in \mathcal{N}} \tilde{\zeta}(j,i) \lambda_{\pi,i},$$

where the total rent of market i is

$$\lambda_{\pi,i} = \lambda_i^{\mathrm{B}} \pi_i, \quad \text{where} \quad \pi_i = \left(1 - \frac{1-\varepsilon_i}{\mu_i}\right) \frac{1}{\gamma_i \mu_i^Y}. \tag{3}$$

Here π_i is the share of market i's sales that are claimed as profits. The profit margin π_i consists of the rents due to market power $1 - 1/\mu_i$ and the rents due to diminishing returns ε_i/μ_i. Revenues generated by μ_i^Y are not paid out to the entrants, and so the profit margin is decreasing in μ_i^Y.

Let $\mathrm{d}\log\lambda_\pi$ be the $|\mathcal{N}|\times 1$ vector of changes in rents, $\mathrm{d}\log M$ the $|\mathcal{N}|\times 1$ vector of changes in masses of producers, and $d\log P_E$ the $|E| \times 1$ vector of changes in entry prices.[13,14] Then we can state the following key lemma, which will prove very useful in characterizing equilibrium outcomes.

Lemma 1. *In equilibrium,*

$$\begin{aligned} \mathrm{d}\log M &= \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi\,\mathrm{d}\log\lambda_\pi - \lambda_E\,\mathrm{d}\log P_E\right), \\ &= \widehat{\mathrm{d}\log\lambda}_\pi - \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\,\mathrm{d}\log P_E. \end{aligned}$$

Holding fixed entry costs ($\mathrm{d}\log P_E = 0$), Lemma 1 shows how entry responds to changes in rents across the economy: entry changes to match the changes in rents $\mathrm{d}\log\lambda_\pi$ to the extent possible. The normalized entry matrix $\tilde{\zeta}$ acts like the data matrix in a regression, and the response of the entrants to a change in rents is the linear projection of the changes in rents $\mathrm{d}\log\lambda_\pi$ onto the space spanned by $\tilde{\zeta}$. Therefore, new entry acts to

[13]The entry price $P_{E,j}$ of the jth entrant is the marginal cost associated with the production function in equation (1).

[14]We are purposefully defining λ_π as an $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix and $d\log\lambda_\pi$ be the $|\mathcal{N}| \times 1$ vector in order to streamline the matrix expressions for projections below. Throughout the draft, and in order to lighten the notation, we often use the same symbol to denote vectors and their counterparts as diagonalized matrices.

minimize the rents going to existing producers.

The ith component of this projection, denoted by $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$, measures the changes in quasi-rents in market i. In other words, it is the change in the amount of resources spent by those entrants who go on to become producers of type i. Changes in quasi-rents in market i can depend on the changes in rents $\mathrm{d}\log\lambda_{\pi,j}$ in all markets $j \in \mathcal{N}$.[15]

We can decompose changes in overall rents into its projection (changes in quasi-rents) and its residual

$$\underbrace{\mathrm{d}\log\lambda_\pi}_{\Delta\text{Rents}} = \underbrace{\widehat{\mathrm{d}\log\lambda}_\pi}_{\text{Projection}} + \underbrace{(\mathrm{d}\log\lambda_\pi - \widehat{\mathrm{d}\log\lambda}_\pi)}_{\text{Residual}}.$$

This projection and residual will turn out to neatly summarize reallocation effects in the presence of inefficiencies.

Holding fixed entry prices ($\mathrm{d}\log P_E = 0$), if there are as many entrant types as there are markets $|E| = |\mathcal{N} - \mathcal{F}|$, then a change in profits in a given market maps, one for one, into a change in the mass of entrants in that market. We call this situation fully directed entry, because in this case, changes in rents are captured entirely by new entrants as quasi-rents. When entry is fully-directed, the residual is zero and existing producers can be perfectly replicated through entry.

Definition 3. Entry is *fully-directed* if there are as many entrant types as there are markets $|E| = |\mathcal{N} - \mathcal{F}|$.

If there are fewer entrant types than markets $|E| < |\mathcal{N} - \mathcal{F}|$, entry into a particular product type may be restricted, or even impossible. When entry into a product type i is impossible, $\zeta(j,i) = 0$ for every $j \in E$, product i is either not produced, or if it is produced, then it is produced by incumbents. In this case, increases in i's rents $\mathrm{d}\log\lambda_{\pi,i}$ will not affect entry into i at all, since $\widehat{\mathrm{d}\log\lambda}_{\pi,i} = 0$.

We say entry is non-overlapping when multiple entrant types cannot enter into the same market i. We impose non-overlapping entry without loss of generality.[16]

Assumption 3. Entry is *non-overlapping*. That is, for each $i \in \mathcal{N}$, there is at most one entrant type $j \in E$ that can produce product i: $\zeta(j,i) \neq 0$.

[15] More generally, for any $|\mathcal{N}| \times 1$ vector X we will write $\widehat{X}_i$ to denote its ith projection.

[16] To see why we can impose this without loss of generality, consider a situation where entrants 1 and 2 enter into the same market, so that $M = M_{E,1} + M_{E,2}$ with $Y = (Mq^{1-\varepsilon})^{1/\gamma}$. To turn this into a model with non-overlapping entry, create two fictitious markets $Y_i = (M_i q_i^{1-\varepsilon})^{1/\gamma}$ with $M_i = M_{E,i}$ for $i \in \{1,2\}$. Now create a third fictitious market, with no entry, where $Y_3 = (Y_1^\gamma + Y_2^\gamma)^{1/\gamma}$. Note that $Y_3 = Y$, which means that we have recast a model with overlapping entry into an equivalent model with non-overlapping entry. We impose this assumption throughout.

IO Matrices

We introduce the *forward* and *backward* Input-Output (IO) matrices Ω^{B} and Ω^{F} and their accompanying Leontief inverses Ψ^{B} and Ψ^{F}. Intuitively, the backward matrix encodes the endogenous transmission of sales backward from downstream customers to their upstream suppliers, whereas the forward matrix captures the transmission of prices forward from upstream suppliers to their downstream customers.[17]

Shocks to productivities and markups/wedges work through a linear system of equations for changes in sales $d \log \lambda_i^{\mathrm{B}}$ and prices $d \log P_i^Y$ with forcing terms given by the shocks. The elementary building blocks used to derive these equations describe: how an autonomous change in the sales of i affects the sales of the other j's, directly and indirectly through the network, and holding prices and shocks constant; and how an autonomous change in the price of j affects the prices of the other i's, directly and indirectly through the network, and holding sales and shocks constant. The backward and forward Leontief inverses Ψ^{B} and Ψ^{F} encode precisely this information.

Backward IO Matrix. Let Ω^V be the $|\mathcal{N}| \times |\mathcal{N}|$ matrix whose ijth element is equal to i's variable expenditures on inputs from j as a share of revenues

$$\Omega_{ij}^V \equiv \frac{M_i P_j^Y x_{ij}}{P_i^Y Y_i}.$$

Let Ω^E be the $|E| \times |\mathcal{N}|$ matrix whose ijth element is equal to entrant i's expenditures on inputs from j as a share of the total entry costs

$$\Omega_{ij}^E \equiv \frac{P_j^Y x_{E,ij}}{\sum_{k \in \mathcal{N}} P_k^Y x_{E,ik}}.$$

Finally let π be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix of profit shares.

The *backward* IO matrix combines variable and fixed expenditures

$$\Omega^{\mathrm{B}} = \Omega^V + \pi \tilde{\zeta}' \Omega^E.$$

Its ijth element Ω_{ji}^{B} is the fraction of the revenues of j directly paid out to i for variable

[17] In Baqaee (2018), these are referred to as the supply- and demand-side matrices. In Baqaee and Farhi (2019a), these are referred to as the revenue- and cost-based matrices.

production and entry. The associated backward Leontief inverse is

$$\Psi^{\mathrm{B}} = \left(I - \Omega^{\mathrm{B}}\right)^{-1} = I + \Omega^{\mathrm{B}} + \left(\Omega^{\mathrm{B}}\right)^2 + \cdots .$$

Its ijth element Ψ^{B}_{ij} is the fraction of the revenues of i directly and indirectly (through the network) paid out to j for variable production and entry.

The sales of j can be broken down into to its sales to the the different i's according to $\lambda^{\mathrm{B}}_j = \sum_i \lambda^{\mathrm{B}}_i \Omega^{\mathrm{B}}_{ij}$. By implication, the ijth element of the backward IO matrix therefore encodes the the elasticity of the sales of j to the sales of i, so that $\Omega^{\mathrm{B}}_{ij} = \partial \log \lambda^{\mathrm{B}}_j / \partial \log \lambda^{\mathrm{B}}_i$, where the partial derivative indicates that prices and shocks as well as other sales are held constant. The ijth element of the backward Leontief inverse therefore encodes the elasticity of the sales of j to the sales of i, so that $\Psi^{\mathrm{B}}_{ij} = \partial \log \lambda^{\mathrm{B}}_j / \partial \log \lambda^{\mathrm{B}}_i$, where the partial derivative indicates that prices and shocks are held constant.[18]

Forward IO Matrix. Let μ, μ^Y, and ε/γ be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrices of μ_i, μ^Y_i, and ε_i/γ_i. The forward IO matrix is

$$\Omega^{\mathrm{F}} = \mu^Y \mu \Omega^V + \frac{\varepsilon}{\gamma} \tilde{\zeta}' \Omega^E.$$

Its ijth element Ω^{F}_{ij} is the fraction of the cost of i directly attributable to the price of j through variable production and entry. The associated forward Leontief inverse is

$$\Psi^{\mathrm{F}} = \left(I - \Omega^{\mathrm{F}}\right)^{-1} = I + \Omega^{\mathrm{F}} + \left(\Omega^{\mathrm{F}}\right)^2 + \cdots .$$

Its ijth element Ω^{F}_{ij} is the fraction of the cost of i directly and indirectly (through the network) attributable to the price of j through variable production and entry.

By Shepard's lemma, the $ij$the element of the forward IO matrix encodes the elasticity of the price of i to the price of j, so that $\partial \log P^Y_i / \partial \log P^Y_j = \Omega^{\mathrm{F}}_{ij}$, where the partial derivative indicates that sales and shocks as well as other prices are held constant. By repeated applications of Shepard's lemma, the ijth element of the forward Leontief therefore encodes the elasticity of the price of j to the price of i, so that $\Psi^{\mathrm{F}}_{ij} = \partial \log P^Y_i / \partial \log P^Y_j$, where the partial derivative now indicates that sales and shocks are held constant but that other prices are allowed to vary.

[18] As we shall see, that prices and shocks are held constant implies that Ω^{B} is also held constant.

Domar Weights. Following Domar (1961), the *Domar weight* of market i is

$$\lambda_i^{\mathrm{B}} = \frac{P_i^Y Y_i}{GDP} = P_i^Y Y_i,$$

where the last equality follows from our choice of numeraire. Theorem 1 implies that for the efficient benchmark, Domar weights are key sufficient statistics.

As a matter of accounting the Domar weight of i coincides with its *backward Domar weight* defined as the ith element of the zero-th row of the backward Leontief inverse

$$\lambda_i^{\mathrm{B}} = \sum_j \Omega_{0j}^{\mathrm{B}} \Psi_{ji}^{\mathrm{B}} = \Psi_{0i}^{\mathrm{B}}.$$

It captures the household's exposure to i via backward linkages or equivalently i's centrality in demand.

The *forward Domar weight* of product i is the ith element of the zero-th row of the forward Leontief inverse

$$\lambda_i^{\mathrm{F}} = \Psi_{0i}^{\mathrm{F}} = \sum_j \Omega_{0j}^{\mathrm{F}} \Psi_{ji}^{\mathrm{F}}.$$

It captures the household's exposures to i via forward linkages or equivalently i's centrality in supply.[19]

In the efficient marginal-cost pricing benchmark, the forward and backward Domar weights of market i coincide $\lambda_i^{\mathrm{B}} = \lambda_i^{\mathrm{F}}$, so that the supply centrality (forward Domar weight) of the market is equal to its demand centrality (backward Domar weight), and both are equal to its sales share. By contrast, with inefficiencies, in general, the backward and forward Domar weights of market i differ $\lambda_i^{\mathrm{B}} \neq \lambda_i^{\mathrm{F}}$ and their ratio $\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}}$ measures the wedge between the supply and demand centralities of the market, or equivalently the reduction in the size of the market caused by the cumulated distortions in its downstream supply chain.

5 Aggregation

We now proceed to generalizing Theorem 2 to inefficient economies. We provide our comparative statics in two successive steps. First, in this section , we provide an aggregation equation which gives the response to shocks of aggregate output as a function of changes in sales, rents, and quasi-rents. Second, in the next section, we provide propa-

[19] The backward and forward Domar weight generalize the revenue- and cost-based Domar weights in Baqaee and Farhi (2019a).

gation equations which give changes in sales (or rents) and quasi-rents, as a function of microeconomic primitives. Putting the two steps together yields our result. The shocks that we consider are shocks to all productivities, markups, and wedges which we write in vector form as $(\mathrm{d}\log A, \mathrm{d}\log \mu, \mathrm{d}\log \mu^Y)$.

5.1 The Aggregation Equation

Theorem 3 (Comparative Statics with Inefficiencies)**.** *The response of aggregate output (welfare) to shocks* $(\mathrm{d}\log A, \mathrm{d}\log \mu, \mathrm{d}\log \mu^Y)$ *is given by*

$$\begin{aligned} d\log Y = &\sum_{i\in\mathcal{N}} \lambda_i^F d\log A_i - \sum_{i\in\mathcal{N}} \lambda_i^F \frac{1-\varepsilon_i}{\gamma_i} d\log \mu_i^Y - \sum_{i\in\mathcal{N}} \lambda_i^F \frac{1-\varepsilon_i}{\gamma_i} d\log \mu_i \\ &- \sum_{i\in\mathcal{N}} \lambda_i^F \left(1-\frac{1-\varepsilon_i}{\gamma_i}\right)\left(d\log \lambda_i^B - \widehat{\mathrm{d}\log \lambda}_{\pi,i}\right) + \sum_{i\in\mathcal{N}} \lambda_i^F \left(\frac{1}{\gamma_i}-1\right)\widehat{\mathrm{d}\log \lambda}_{\pi,i}. \end{aligned} \tag{4}$$

Theorem 3 is the key result of the paper, and we spend the rest of this section unpacking the intuition and working through some examples. Since aggregate nominal output is normalized to one, changes in aggregate output are the opposite of changes in final-demand prices $\mathrm{d}\log Y = -\mathrm{d}\log P_0^Y$. And indeed, there is a simple *dual* interpretation of the formula tracking changes in final-demand prices. The first line captures changes in final-demand prices when sales and quasi-rents are held constant, and the second line accounts for changes in sales and quasi-rents. Intuitively, the first term on second line captures how for each market i, changes in the scale of operation of individual producers affect the price of the market good because of decreasing internal returns to scale. The second term on the second line captures how, for each market i, changes in entry affect the price of the market good by stimulating external economies. In both cases, what matters is then how, for each market i, the change in the price of the good affects the price of final-demand.

We can re-express Theorem 3 in different useful ways using the relationship between rents, sales, and profit margins $\mathrm{d}\log \lambda_{\pi,i} = \mathrm{d}\log \lambda_i^{\mathrm{B}} + \mathrm{d}\log \pi_i$. For example, we can write

$$\begin{aligned} \mathrm{d}\log Y = &\sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} d\log A_i - \sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} d\log \mu_i^Y - \sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} d\log \mu_i - \sum_{i} \lambda_i^{\mathrm{F}} \left(1-\frac{1-\varepsilon_i}{\gamma_i}\right) d\log\left(\frac{\frac{\varepsilon_i}{\mu_i}}{1-\frac{1-\varepsilon_i}{\mu_i}}\right) \\ &- \sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} \left(1-\frac{1-\varepsilon_i}{\gamma_i}\right)\left(d\log \lambda_{\pi,i} - \widehat{d\log \lambda}_{\pi,i}\right) + \sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} \left(\frac{1}{\gamma_i}-1\right)\widehat{d\log \lambda}_{\pi,i}. \end{aligned} \tag{5}$$

Intuitively, the first line captures the effect of shocks on prices when rents and quasi-rents

are held constant, and the second line accounts for changes in rents and quasi-rents. The fact that we now prevent or allow rents and quasi-rents to adjust, rather than sales and quasi-rents as in the previous expression, introduces two differences. First, the last term on the first line is new and accounts for the fact that, holding fixed rents, changes in markups change the ratio of Ricardian to monopoly rents. Increases in markups in market i cause individual producers to scale down which, in the presence of decreasing internal returns to scale, mitigates the increase in the price of the market good caused by the increase in markups, and hence also the increase in final-demand prices. Second, the first term on the second line now features the residuals of the projection of rents on entry. If rents outpace quasi-rents in market i, then the price of the market good increases because of decreasing internal returns to scale, and hence so do final-demand prices.

5.2 The Role of Reallocation

We now show that Theorem 3 also provides an interpretable decomposition of changes in output into changes in technical and allocative efficiency along the lines of Baqaee and Farhi (2019a) and thereby offer an alternative *primal* interpretation of the formula.

Let $\mathcal{X}$ denote the $(|\mathcal{N}| + |E|) \times |\mathcal{N}|$ *allocation matrix* of the economy, where $\mathcal{X}_{ij}$ records the fraction of good j used by a producer or entrant $i \in \mathcal{N} + E$. It is defined in *physical* units. Together with the vector of productivity shifters A, the allocation matrix pins down the whole allocation, and hence aggregate output $\mathcal{Y}(A, \mathcal{X})$.

In particular, equilibrium aggregate output is obtained by using the equilibrium allocation matrix $\mathcal{X}(A, \mu, \mu^Y)$ where μ and μ^Y are the vectors of markups/wedges. Changes in equilibrium aggregate output in response to shocks can therefore be written, in matrix notation, as

$$\mathrm{d}\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log A}\,\mathrm{d}\log A}_{\Delta\text{Technical Efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}}\,\mathrm{d}\mathcal{X}}_{\Delta\text{Allocative Efficiecny}},$$

where the first term is the direct effect of changes in technology, holding the allocation of resources constant, and the second term is the indirect effect of equilibrium reallocations

$$\mathrm{d}\mathcal{X} = \frac{\partial \mathcal{X}}{\partial \log A}\,\mathrm{d}\log A + \frac{\partial \mathcal{X}}{\partial \log \mu}\,\mathrm{d}\log \mu + \frac{\partial \mathcal{X}}{\partial \log \mu^Y}\,\mathrm{d}\log \mu^Y.$$

Proposition 2 (Decomposition with Inefficiencies)**.** *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log \mu, \mathrm{d}\log \mu^Y)$,

changes in aggregate output can be decomposed in changes in technical efficiency

$$\frac{\partial \log \mathcal{Y}}{\partial \log A} \, \mathrm{d} \log A = \sum_{i \in \mathcal{N}} \lambda_i^F \, \mathrm{d} \log A_i,$$

and changes in allocative efficiency

$$\begin{aligned} \frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} \, \mathrm{d} \mathcal{X} = & - \sum_{i \in \mathcal{N}} \lambda_i^F \frac{1 - \varepsilon_i}{\gamma_i} d \log \mu_i^Y - \sum_{i \in \mathcal{N}} \lambda_i^F \frac{1 - \varepsilon_i}{\gamma_i} d \log \mu_i \\ & - \sum_{i \in \mathcal{N}} \lambda_i^F \left(1 - \frac{1 - \varepsilon_i}{\gamma_i}\right) \left(\mathrm{d} \log \lambda_i^B - \widehat{\mathrm{d} \log \lambda}_{\pi,i}\right) + \sum_{i \in \mathcal{N}} \lambda_i^F \left(\frac{1}{\gamma_i} - 1\right) \widehat{\mathrm{d} \log \lambda}_{\pi,i}. \end{aligned}$$

Changes in technical efficiency are a Hulten-like weighted sum of changes in productivities. The weights are forward Domar weights rather than traditional Domar weights. This is because when the allocation of resources is kept constant, productivity shocks are pushed forward through supply chains to the household.

Changes in allocative efficiency can be traced back to reductions in prices (shares) of specific fixed factors associated with individual producers and with entry. Focus on productivity shocks for simplicity, so that the first line of the expression for changes in allocative efficiency is zero. There are two terms on the second line.

The first term depends on decreasing internal returns to scale $1 - (1 - \varepsilon_i)/\gamma_i$. When $\mathrm{d} \log \lambda_i^B - \widehat{\mathrm{d} \log \lambda}_{\pi,i} > 0$, this means that individual producers in market i are scaling up and running into diminishing returns. This raises the shadow price of their producer-specific fixed factor and contributes negatively to changes in allocative efficiency in proportion to the forward Domar weight $\lambda_i^F(1 - (1 - \varepsilon_i)/\gamma_i)$ of these specific fixed factors.[20,21]

The second term depends on increasing external returns to scale $1/\gamma_i - 1$. When $\widehat{\mathrm{d} \log \lambda}_{\pi,i} > 0$, this means that entry is increasing in market i and triggering external economies from love of variety. This reduces the (negative) shadow price of the specific fixed factor associated with entry and contributes positively to changes in allocative efficiency in proportion to the forward Domar weight $\lambda_i^F(1/\gamma_i - 1)$ of these specific fixed factors.

[20] When we refer to the price of producer-specific fixed factors, we rely on Lionel McKenzie's insight that any non-CRS production function $h(x)$ can be represented by a CRS technology $\tilde{h}(x, z) = zh(x/z)$ where z is a producer-specific fixed factor with supply $z = 1$. The marginal cost of $h(x)$ coincides with the marginal cost of $\tilde{h}(x, z)$, where the effect of scale in the former is captured by the (shadow) price of the fixed factor in the latter.

[21] Recall that primary factors $f \in \mathcal{F} \subset \mathcal{N}$ are captured as producer-specific fixed factors in factor markets with zero-returns-to-scale individual producers $(1 - (1 - \varepsilon_f)/\gamma_f = 1)$ aggregated linearly $(1/\gamma_f = 1)$ with no entry $(\widehat{\mathrm{d} \log \lambda}_{\pi,f} = 0)$.

Improvements in allocative efficiency can be measured by a forward-weighted sum of reductions in the shadow prices of fixed factors. Beneficial equilibrium reallocations, by using more resources more efficiently, reduce the shadow prices of fixed factors on balance across markets. This can only occur when the economy is inefficient. When the economy is efficient, reductions in the shadow prices of some specific fixed factors are exactly compensated by increases in others.

Corollary 1 (Decomposition under Efficiency)**.** *In the marginal-cost pricing equilibrium, as long as* $\varepsilon_i < 1$ *for all* $i \in \mathcal{N}$*, changes in technical and allocative efficiency are given by*[22]

$$\frac{\partial \log \mathcal{Y}}{\partial \log A} \, \mathrm{d} \log A = \sum_{i \in \mathcal{N}} \lambda_i^F \, \mathrm{d} \log A_i \qquad \textit{and} \qquad \frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} \, \mathrm{d} \mathcal{X} = 0,$$

with $\lambda_i^F = \lambda_i^B$.

In the efficient benchmark, technology shocks only have direct effects and not indirect reallocation effects. Of course, this does not mean that reallocations do not occur in efficient models, but merely that their impact is irrelevant to a first order.

5.3 Useful Special Cases

We build additional intuition by considering different specializations of Theorem 3. We consider a univariate productivity shock $\mathrm{d} \log A_i$ (holding constant other productivities, wedges, and markups). The intuition for a shock to markups/wedges is similar, but for brevity, we relegate this discussion to Appendix G.

We use the reformulation in equation (5) to get

$$\mathrm{d} \log Y = \lambda_i^{\mathrm{F}} \, \mathrm{d} \log A_i - \sum_{j \in \mathcal{N}} \lambda_j^{\mathrm{F}} \left(\frac{1 - \varepsilon_j}{\gamma_j} - 1 \right) \left(\mathrm{d} \log \lambda_{\pi,j} - \mathrm{d} \widehat{\log \lambda}_{\pi,j} \right) + \sum_{j \in \mathcal{N}} \left(\frac{1}{\gamma_j} - 1 \right) \mathrm{d} \widehat{\log \lambda}_{\pi,j}.$$

In a Cobb-Douglas economy where all the productions functions f_j and g_j are Cobb-Douglas, all sales are constant ($\mathrm{d} \log \lambda_j^B = 0$). All profit margins are also constant ($\mathrm{d} \log \pi_j = 0$) because there are no shocks to markups and wedges. Hence rents and quasi-rents are constant ($\mathrm{d} \log \lambda_{\pi,j} = \mathrm{d} \widehat{\log \lambda}_{\pi,j} = 0$). The expression then simplifies to $\mathrm{d} \log Y = \lambda_i^{\mathrm{F}} \, \mathrm{d} \log A_i$. Intuitively, the productivity shock is pushed forward through the supply chain to the household.

[22] The extra assumption ensures that entry is not socially wasteful. When it is violated, equilibrium reallocations affecting entry can reduce (but not increase) aggregate output to a first order. See Appendix E for more details.

When the economy is not Cobb-Douglas, productivity shocks induce changes in the sizes $\mathrm{d}\log\lambda_j^B$ of the different markets, and hence also changes in rents $\mathrm{d}\log\lambda_{\pi,j}$ and quasi-rents $\mathrm{d}\widehat{\log\lambda_{\pi,j}}$. The response of aggregate output is then complicated by the endogenous response of entry to changes in profits. We analyze productivity shocks in three benchmarks, corresponding respectively to constant, decreasing, and increasing returns to scale. That is, we consider the following cases: (CRS) $1-\varepsilon_j=\gamma_j=1$; (DRS) $1-\varepsilon_j<\gamma_j=1$; and (IRS) $1-\varepsilon_j=\gamma_j<1$ for every $j\in\mathcal{N}-\mathcal{F}$.

Productivity shocks with CRS. When $1-\varepsilon_j=\gamma_j=1$ for all $j\in\mathcal{N}-\mathcal{F}$, Theorem 3 reduces to

$$\mathrm{d}\log Y=\lambda_i^{\mathrm{F}}\,\mathrm{d}\log A_i-\sum_{f\in\mathcal{F}}\lambda_f^{\mathrm{F}}\,\mathrm{d}\log\lambda_f^{\mathrm{B}}.$$

There are neither decreasing internal economies nor increasing external economies. We therefore only need to track changes in primary factors prices. Here we have used the fact primary factor markets $f\in\mathcal{F}$ satisfy $1-(1-\varepsilon_f)/\gamma_f=1$, $1/\gamma_f=1$, and $\mathrm{d}\log\lambda_f^{\mathrm{B}}-\widehat{\mathrm{d}\log\lambda}_{\pi,f}=\mathrm{d}\log\lambda_f^{\mathrm{B}}$, because of respectively zero returns to scale, linear aggregation, and no entry.

The content of this equation is that equilibrium reallocations from the productivity shock are beneficial if they reduce the forward-weighted average of changes in primary factor prices.

When there is no entry ($\tilde{\zeta}=0$), we recover the result of Baqaee and Farhi (2019a). If there is entry, then it is socially wasteful. The clearest example is when entry is possible in all non-primary-factor markets and there is a single primary factor. We then get

$$\mathrm{d}\log Y=\lambda_i^{\mathrm{F}}\,\mathrm{d}\log A_i,$$

so that socially wasteful entry absorbs or exudes resources so that there are no changes in allocative efficiency, even though there are reallocations and the economy is inefficient.

Productivity shocks with DRS. When $1-\varepsilon_j<\gamma_j=1$ for all $j\in\mathcal{N}-\mathcal{F}$, Theorem 3 becomes

$$\mathrm{d}\log Y=\lambda_i^{\mathrm{F}}\,\mathrm{d}\log A_i-\sum_{f\in\mathcal{F}}\lambda_f^{\mathrm{F}}\,\mathrm{d}\log\lambda_f^{\mathrm{B}}-\sum_{j\in\mathcal{N}-\mathcal{F}}\lambda_j^{F}\varepsilon_j\left(d\log\lambda_{\pi,j}-\widehat{\mathrm{d}\log\lambda}_{\pi,j}\right).$$

There are decreasing internal economies but no increasing external economies. This means that we no longer only need to track changes in the prices of primary factors as in the

CRS case, but also of changes in the specific fixed factor prices associated with individual producers. The key sufficient statistic for these two sorts of specific fixed factors is the residual from the projection of rents on entry.

If for some market $j \in \mathcal{N}$, entry cannot keep up with rents so that $\mathrm{d}\log\lambda_{\pi,j} - \widehat{\mathrm{d}\log\lambda}_{\pi,j} > 0$, then individual producers in this market scale up and run into diminishing returns. As a result, the prices of their producer-specific factors increase. This reallocation contributes to reducing aggregate output in proportion to the forward Domar weight $\lambda_i^F \varepsilon_i$ of these specific fixed factors.

The total effect of reallocations is obtained by summing over all markets (non-primary factor markets and primary factor markets). Reallocations lead to a more efficient use of resources when they reduce the scarcity of fixed factors by making them cheaper.

Such improvements in allocative efficiency cannot occur when the economy is efficient. In this case factor prices cannot go down on balance across markets. To see this, note that, in the efficient case, $\lambda_{\pi,j} = \lambda_j \varepsilon_j$ and so $\sum_{j\in\mathcal{N}} \lambda_j^F \varepsilon_j (d\log\lambda_{\pi,j} - \widehat{\mathrm{d}\log\lambda}_{\pi,j}) = 0$ by definition of the residual.

Changes in allocative efficiency can also be zero even the economy is inefficient and when there are equilibrium reallocations. The clearest example is when entry not only possible in all non-primary-factor markets but is also fully directed, and when in addition there is a single primary factor. We then get

$$\mathrm{d}\log Y = \lambda_i^{\mathrm{F}}\, \mathrm{d}\log A_i.$$

There are only changes in technical efficiency, and no changes in allocative efficiency. Intuitively, in this case, changes in the prices of market goods are determined independently from changes in their sales because changes in sales are accommodated entirely through changes in entry. In other words, reallocations happen entirely on the extensive margin of entry and exit and offset one another.

Productivity shocks with IRS. When $\gamma_j = 1 - \varepsilon_j < 1$ for all $j \in \mathcal{N} - \mathcal{F}$ so that all non-primary factor markets are of the CES kind, Theorem 3 becomes

$$\mathrm{d}\log Y = \lambda_i^{\mathrm{F}} d\log A_i - \sum_{f\in\mathcal{F}} \lambda_f^{\mathrm{F}}\, \mathrm{d}\log\lambda_f^{\mathrm{B}} + \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}} \left(\frac{1}{\gamma_j} - 1\right) \widehat{\mathrm{d}\log\lambda}_{\pi,j}.$$

There are constant internal returns to scale and increasing external returns to scale. Hence we still need to keep track of the changes int he prices of primary factors, but we no longer need to track of changes in the prices of specific fixed factors associated with individual

producers as in the DRS case. Instead, we now need to keep track of changes in the entry margin itself — that is, the projection of changes in rents on entry.

If in some market j, quasi-rents increase so that $\widehat{\mathrm{d}\log\lambda}_{\pi,j} > 0$, then entry in the market increases and triggers external economies from love of variety. This reduces the negative price of the associated specific fixed factor. This reallocation contributes to increasing aggregate output in proportion to the forward Domar weight $\lambda_i^F(1/\gamma_i - 1)$ of these specific fixed factors.

The total effect of reallocations is obtained by summing over all markets (non-primary factor markets and primary factor markets). Reallocations lead to a more efficient use of resources when they reduce the scarcity of specific fixed factors by making them cheaper.

The clearest example is once again when entry is possible in all non-primary-factor markets and there is a single primary factor. We then get

$$\mathrm{d}\log Y = \lambda_i^{\mathrm{F}} d\log A_i + \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j} - 1\right)\widehat{\mathrm{d}\log\lambda}_{\pi,j}.$$

6 Propagation

The aggregation equation in the previous section gives changes in aggregate output as a function of changes in sales (or rents) and quasi-rents. In this section, we complete the theory by deriving propagation equations for the changes in sales (or rents) and quasi-rents. We do this in two steps: forward and backward propagation.[23] Forward propagation establishes how changes in prices feed forward from suppliers to consumers, and backward propagation describes how changes in sales feed backward from consumers to their suppliers. Together, they pin down changes in sales, rents, and quasi-rents, as well as all other disaggregated variables such as prices, quantities, etc.

6.1 Forward Propagation

We start by describing the response of prices to shocks.

Proposition 3 (Forward Propagation)**.** *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log\mu, \mathrm{d}\log\mu^Y)$*, changes in prices are given by*

$$d\log P_i^Y = -\sum_{j\in\mathcal{N}} \Psi_{ij}^F d\log A_j + \sum_{j\in\mathcal{N}} \Psi_{ij}^F \frac{1-\varepsilon_j}{\gamma_j} d\log\mu_j^Y + \sum_{j\in\mathcal{N}} \Psi_{ij}^F \frac{1-\varepsilon_j}{\gamma_j}\,\mathrm{d}\log\mu_j$$

[23] This generalizes the changes in consumer and producer centrality introduced in Baqaee (2018).

$$+\sum_{j\in\mathcal{N}}\Psi^F_{ij}\left(1-\frac{1-\varepsilon_j}{\gamma_j}\right)\left(d\log\lambda^B_j-\widehat{\mathrm{d}\log\lambda}_{\pi,j}\right)-\sum_{j\in\mathcal{N}}\Psi^F_{ij}\left(\frac{1}{\gamma_j}-1\right)\widehat{\mathrm{d}\log\lambda}_{\pi,j}$$

Proposition 3 is similar to Theorem 3. In fact, since aggregate nominal output is normalized to one, changes in aggregate output are just the opposite of the changes in final-demand prices $\mathrm{d}\log Y = -\mathrm{d}\log P_0$. Therefore, Proposition 3 can be specialized to yield Theorem 3.[24] The general intuition is similar.

6.2 Backward Propagation

We continue by describing the responses of sales, rents, and quasi-rents to shocks.

Lemma 2 (Profit Shares)**.** *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log\mu, \mathrm{d}\log\mu^Y)$*, changes in sales, rents, and profit margins are related through*

$$\mathrm{d}\log\lambda_{\pi,i}=\mathrm{d}\log\lambda^B_i+\mathrm{d}\log\pi_i,\quad \textit{where}\quad \mathrm{d}\log\pi_i=-\mathrm{d}\log\mu^Y_i+\frac{\frac{1-\varepsilon_i}{\mu_i}}{1-\frac{1-\varepsilon_i}{\mu_i}}\mathrm{d}\log\mu_i.$$

This lemma implies that it is enough to characterize changes in sales $\mathrm{d}\log\lambda^{\mathrm{B}}_i$ since we can then immediately obtain changes in rents $\mathrm{d}\log\lambda_{\pi,i}$ and quasi-rents $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$.

For simplicity, we assume that all production and entry functions in the economy f_i and g_i are of a nested-CES form. By relabelling each CES aggregator to be a new incumbent (by which we mean a representative producer in an uncontested market with an exogenous unit mass of producers) and linking them together via the input-output network, we can, without loss of generality, assume that each production function in variable production f_i corresponds to one nest with a single elasticity of substitution θ_i.[25] Furthermore, by introducing a new incumbent transforming the different inputs of g_i into a single customized input, we can, without loss of generality, assume that each production function in entry g_i uses a single input.

In order to state our results, we use the *input-output covariance operator*:

$$Cov_m(X,\Psi^{\mathrm{B}}_{(i)})=\sum_{k\in\mathcal{N}}\mu^Y_m\mu_m\Omega^V_{mk}x_k\Psi^{\mathrm{B}}_{ki}-\left(\sum_{k\in\mathcal{N}}\mu^Y_m\mu_m\Omega^V_{mk}\Psi^{\mathrm{B}}{}_{ki}\right)\left(\sum_{k\in\mathcal{N}}\mu^Y_m\mu_m\Omega^V_{mk}x_k\right),$$

where $\Psi^{\mathrm{B}}{}_{(i)}$ is the ith column of the backward Leontief inverse Ψ^{B}. It is the covariance

[24] We call Theorem 3 a theorem and Proposition 3 a proposition because of their economic, rather than mathematical, significance.

[25] See the discussion of standard-form economies in Baqaee and Farhi (2019b) for more information.

between the vector X and the ith column of Ψ^{B}, using the mth row of $\mu^{Y}\mu\Omega^{V}$ as the probability distribution, where we rely on the fact that $\sum_{k\in\mathcal{N}}\mu_{m}^{Y}\mu_{m}\Omega_{mk}^{V}=1$.

Proposition 4 (Backward Propagation)**.** *In response to shocks* $(\mathrm{d}\log A,\mathrm{d}\log\mu,\mathrm{d}\log\mu^{Y})$*, absolute changes in sales are given by*

$$\mathrm{d}\,\lambda_{i}^{B}=-\sum_{m\in\mathcal{N}}\lambda_{m}^{B}\frac{1-\varepsilon_{m}}{\gamma_{m}\mu_{m}\mu_{m}^{Y}}\sum_{k\in\mathcal{N}}\Omega_{mk}^{V}\Psi_{ki}^{B}\,\mathrm{d}\log(\mu_{m}\mu_{m}^{Y})+\sum_{m\in\mathcal{N}}\lambda_{m}^{B}\frac{1-\varepsilon_{m}}{\gamma_{m}\mu_{m}\mu_{m}^{Y}}\sum_{j\in E}\tilde{\zeta}_{jm}\sum_{k\in\mathcal{N}}\Omega_{jk}^{E}\Psi_{ki}^{B}d\log\mu_{m}$$
$$-\sum_{m\in N}\lambda_{m}^{B}\left(1-\frac{1-\varepsilon_{m}}{\mu_{m}}\right)\frac{1}{\gamma_{m}\mu_{m}^{Y}}\sum_{j\in E}\tilde{\zeta}_{jm}\sum_{k\in N}\Omega_{jk}^{E}\Psi_{ki}^{B}d\log\mu_{m}^{Y}-\sum_{m\in\mathcal{N}}\lambda_{m}^{B}\frac{1-\varepsilon_{m}}{\gamma_{m}\mu_{m}\mu_{m}^{Y}}(\theta_{m}-1)Cov_{m}\left(-d\log P^{Y},\Psi_{(i)}^{B}\right).$$

Proportional changes in sales can be deduced using $\mathrm{d}\log\lambda_{i}^{B}=\mathrm{d}\,\lambda_{i}^{B}/\lambda_{i}^{B}$.

The intuition is familiar from the structure of neoclassical input-output models (see Baqaee and Farhi, 2019b). The first term is the mechanical effect of changes in markups/wedges on the demand for i via the intensive margin: and increase in m's markup/wedge $\mathrm{d}\log(\mu_{m}\mu_{m}^{Y})>0$ reduces m's demand for each of its input k in proportion to $\lambda_{m}^{\mathrm{B}}(1-\varepsilon_{m})/(\gamma_{m}\mu_{m}\mu_{m}^{Y})\Omega_{mk}^{V}$, and this in turn reduces the demand for i in proportion to the exposure Ψ_{ki}^{B} of k to i.

The second term is the effect of changes in markups on the demand for i via the extensive margin: an increase in m's markup $\mathrm{d}\log\mu_{m}$ stimulate entrant type j's expenditures on k in proportion to $\lambda_{m}^{\mathrm{B}}(1-\varepsilon_{m})/(\gamma_{m}\mu_{m}\mu_{m}^{Y})\tilde{\zeta}_{jm}\Omega_{jk}^{E}$, and this in turn increases the demand for i in proportion to the exposure Ψ_{ki}^{B} of k to i.

The third term is the mechanical effect changes in wedges on the demand for i via the extensive margin: an increase in m's output tax on $d\,\mathrm{d}\log\mu_{m}^{Y}$ discourages entrant type j's expenditure on k in proportion to $\lambda_{m}^{\mathrm{B}}(1-(1-\varepsilon_{m})/\mu_{m})/(\gamma_{m}\mu_{m}^{Y})\tilde{\zeta}_{jm}\Omega_{jk}^{E}$, and this in turn reduces the demand for i in proportion to the exposure Ψ_{ki}^{B} of k to i.

The fourth and final term captures the effect of substitutions on the intensive margin. Changes in relative prices $\mathrm{d}\log P$ caused by the shocks lead individual producer in every market $m\in\mathcal{N}$ to shift their expenditures on their inputs. This in turn changes expenditures on i in proportion to $\lambda_{m}^{\mathrm{B}}(1-\varepsilon_{m})/(\gamma_{m}\mu_{m}\mu_{m}^{Y})(\theta_{m}-1)Cov_{m}(\mathrm{d}\log P,\Psi_{(i)}^{\mathrm{B}})$.

7 Illustrative Examples

In this section, we provide three fully worked-out illustrative examples: the one-sector heterogenous-firm economy, the multi-sector economy, and the roundabout-entry economy. All three examples have a unique primary factor in exogenous supply which we

refer to as labor (indexed by L). We explain how different assumptions about returns to scale and entry shape the comparative statics for aggregate output or equivalently for aggregate TFP since labor supply is exogenous. For brevity, we only discuss shocks to productivities. We provide illustrative examples with shocks to markups/wedges in Section 8 in the context of a broader analysis of optimal policy.

We use the following notation throughout. Given three vectors U, V, and W with $\sum_k U_k = 1$, we write $\mathbb{E}_U(V) = \sum_k U_k V_k$ and $Cov_U(V, W) = \sum_k U_k(V_k W_k) - (\sum_k U_k V_k)(\sum_k U_k W_k)$. We also sometimes use overlines to signal initial values when there is an ambiguity , but we drop them when there is none: for example, we alternatively write $\overline{\lambda}_i^{\mathrm{B}}$ or λ_i^{B} depending on the context.

7.1 One-Sector Heterogenous-Firm Economy

In this example, we consider a one-sector heterogenous-firm economy. We focus on a "superstar" shock: a shock to productivities that disproportionately increases the productivity of high-markup firms. Here we are motivated by recent evidence from the U.S. showing that the rise in average markups is largely a within-sector and across-firms phenomenon driven by the rise of superstar firms with high markups (Autor et al., 2017; Vincent and Kehrig, 2017; Baqaee and Farhi, 2019a; De Loecker et al., 2019).

In Baqaee and Farhi (2019a), we argued that the reallocations of resources towards high-markup firms associated with this shock explain almost 50% of aggregate TFP growth over the past 20 years. Here we explain how different views on returns to scale (IRS vs. DRS) and entry (no entry vs. free entry) shape the response of aggregate output.

7.1.1 IRS à la Melitz (2003)

Consider a single-sector economy where aggregate output

$$Y = \left(\sum_k q_k^{\frac{\theta-1}{\theta}} \right)^{\frac{\theta-1}{\theta}}$$

is a CES aggregate of differentiated varieties indexed by k with an elasticity of substitution $\theta > 1$ and associated gross external returns to scale $1/\gamma = \theta/(\theta - 1)$. Each variety k is produced from labor with constant returns and a productivity A_k by a single firm and sold at a constant exogenous markup $\mu_k^q > 1$ over marginal cost

$$q_k = A_k l_k, \quad p_k^q = \mu_k^q mc_k.$$

To mimic the trends in the data, we submit this economy to a vector of firm-level productivity shocks $\mathrm{d}\log A$ which disproportionately increase the productivity of high-markup firms so that $Cov_{\lambda^B}(\mathrm{d}\log A, 1-1/\mu_q) > 0$.

To apply our formulas, we will use the backward Domar weight (sales share) and the forward Domar weight of each firm, which are equal to each other $\lambda_k^{\mathrm{B}} = \lambda_k^{\mathrm{F}}$. When there is no entry, we will also use the backward Domar weight (income share) and the forward Domar weight of labor which are then given by $\lambda_L^{\mathrm{B}} = 1/\bar{\mu}^q < 1$ and $\lambda_L^{\mathrm{F}} = 1$, where $\bar{\mu}^q = 1/\mathbb{E}_{\lambda^{\mathrm{B}}}(1/\mu^q)$ is the (harmonic) average markup. Finally, when there is free entry, we will use the fact that the backward Domar weight (income share) of labor is one $\lambda_L^{\mathrm{B}} = 1$. These identities will continue to hold in the DRS case.

No Entry. We first consider the case where there is no entry. We take the set of firms as exogenously given. Changes in aggregate output are given by

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) - \mathrm{d}\log \lambda_L^{\mathrm{B}},$$

where

$$\mathrm{d}\log \lambda_L^{\mathrm{B}} = -(\theta-1)(\bar{\mu}^q - 1)\, Cov_{\lambda^{\mathrm{B}}}\left(\frac{1-\frac{1}{\mu^q}}{1-\frac{1}{\bar{\mu}^q}}, \mathrm{d}\log A\right).$$

The term $\mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A)$ is Hulten-like and captures the effect of the shock on technical efficiency. The term $-\mathrm{d}\log \lambda_L^{\mathrm{B}}$ is the reduction in the labor share and captures the effect of the shock on allocative efficiency. Since the shock disproportionately increase the productivity of high-markup firms ($Cov_{\lambda^B}(\mathrm{d}\log A, 1-1/\mu_q) > 0$), and since firms are substitutes ($\theta > 1$), it reallocates labor towards high-markup firms and reduces the labor share (rents earned by labor). This reallocation improves allocative efficiency, because high-markup firms were too small to begin with from a social perspective, and boosts aggregate output.

Free Entry. Now consider a long-run steady-state version of the same model with free entry instead of no entry. The set of firms is endogenous. Potential entrants pay a fixed entry cost in units of labor and draw a k at random from a fixed distribution, which determines their productivity A_k and markup μ_k^q. Now Theorem 3 implies that

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) + \frac{1}{\theta-1}\,\mathrm{d}\log\left(\mathbb{E}_{\lambda^B}\left(1-\frac{1}{\mu^q}\right)\right),$$

where, from Proposition 4,

$$\mathrm{d}\log\left(\mathbb{E}_{\lambda^B}\left(1-\frac{1}{\mu^q}\right)\right)=(\theta-1)Cov_{\lambda^B}\left(\frac{1-\frac{1}{\mu^q}}{1-\frac{1}{\bar{\mu}^q}},\mathrm{d}\log A\right).$$

Unlike in the no-entry case, the labor share is one $\lambda_L^B=1$ even though the variable profit share is positive $\mathbb{E}_{\lambda^B}(1-1/\mu^q)>0$. Changes in the labor share (rents earned by labor) $\mathrm{d}\log\lambda_L^B=0$ no longer play any role in the determination of changes in allocative efficiency. Instead, we must now track changes in the variable profit share $\mathrm{d}\log(\mathbb{E}_{\lambda^B}(1-1/\mu^q))$ (quasi-rents). Similarly, the elasticity of substitution θ no longer plays any role because its roles in substitution and love for variety cancel out.

There are changes in technical efficiency captured by the Hulten-like term $\mathbb{E}_{\lambda^B}(\mathrm{d}\log A)$ and changes in allocative efficiency are $Cov_{\lambda^B}\left(\frac{1-\frac{1}{\mu^q}}{1-\frac{1}{\bar{\mu}^q}},\mathrm{d}\log A\right)$. Average profits increase because the shock triggers reallocations towards high-markup firms. This in turn increases entry and generates improvements in allocative efficiency by enabling external economies arising from love for variety.

The improvements in technical efficiency are given by the same Hulten-like term as in the no entry-case. By contrast, the changes in allocative efficiency, which are still positive, are different. As before, improvements in efficiency brought about by reallocation of labor to high-markup firms economizes on labor. However, unlike the no-entry case, the labor saved is no longer used towards the variable production of incumbents, but instead is used for the entry and variable production of new firms.[26]

7.1.2 DRS à la Hopenhayn (1992)

We slightly modify the model and instead assume that aggregate output

$$Y=\sum_k y_k,$$

is a linear aggregate of undifferentiated varieties. The varieties are perfectly-substitutable and can be thought of as being the same good. Each variety k is produced from labor with decreasing returns $1-\varepsilon<1$ and productivity A_k by a single firm and sold at a constant

[26]Comparing the free-entry and no-entry model shows that increases in aggregate output are larger when there is no entry if, and only if, there is too much entry to begin with. That is, if $(\theta-1)(\bar{\mu}^q-1)>1$. In this case, releasing labor towards variable production by incumbents dominates releasing labor towards entry and variable production of new firms. Somewhat surprisingly, this means that the rise of markups may be less bad if incumbents are able to restrict (excessive) entry through lobbying or deterrence.

exogenous markup μ_k^y over marginal cost

$$y_k = A_k l_k^{1-\varepsilon}, \quad p_k^y = \mu_k^y mc_k.$$

It turns out that the response $\mathrm{d}\log Y$ of aggregate output to the shock $\mathrm{d}\log A$ is exactly the same as in the IRS case if we set $\varepsilon = 1/\theta < 1$ and $\mu_k^y = \mu_k^q(\theta-1)/\theta$. This is true regardless of whether there is no entry or free entry. The easiest way to see this is that the DRS model is equivalent to a power transformation to the utility function of the household with an offsetting inverse power transformation of the productivity shocks.

It is tempting to draw from this example the comforting lesson that the differences between IRS models à la Melitz (2003) and DRS models à la Hopenhayn (1992) are cosmetic, and that fundamentally the two approaches can be used interchangeably in applications based on convenience. In fact, as we shall see with the next example, this lesson is highly specific to the one-sector heterogenous-firm setting and does not generalize.

7.2 Multi-Sector Economy

In this section, we consider a multi-sector economy with homogenous firms within sectors, targeted free entry in all sectors with entry costs paid in units of labor. We focus on a shock to sector-level productivities. We explain how different views on returns to scale (IRS vs. DRS) now lead to very different responses of aggregate output.

7.2.1 IRS with Free Entry à la Dixit and Stiglitz (1977)

Consider a multi-sector economy where aggregate output

$$Y = \left(\sum_k Y_k^{\frac{\theta_0-1}{\theta_0}} \right)^{\frac{\theta_0-1}{\theta_0}}$$

is a CES aggregate of differentiated sectors indexed by k with an elasticity of substitution θ_0. Each sector k's output

$$Y_k = \left(M_k q_k^{\frac{\theta_k-1}{\theta_k}} \right)^{\frac{\theta_k}{\theta_k-1}}$$

is itself a CES aggregate of an endogenous mass M_k of differentiated varieties with an elasticity of substitution $\theta_k > \min\{\theta_0, 1\}$ and associated gross external returns to scale $1/\gamma_k = \theta_k/(\theta_k - 1)$. Each variety in sector k is produced from labor with constant returns

and productivity A_k by a single firm and sold at a markup $\mu_k^q > 1$ over marginal cost

$$q_k = A_k l_k, \quad p_k^q = \mu_k^q mc_k.$$

We are interested in a long-run steady-state with targeted free entry in all sectors where potential entrants choose which sector to enter into and pay a sector-specific fixed entry cost in units of labor. We submit this economy to a vector of sector-level productivity shocks $\mathrm{d}\log A$ and denote by θ the vector of within-sector elasticities of substitution.

To apply our formulas, we will use the backward Domar weight (sales share) and the forward Domar weight of each sector, which are equal to each other $\lambda_k^{\mathrm{B}} = \lambda_k^{\mathrm{F}}$. We will also use the fact that the backward Domar weight (income share) of labor is one $\lambda_L^{\mathrm{B}} = 1$. These identities continue to hold in the DRS case.

From Theorem 3, changes in aggregate output are given by

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) + \sum_k \frac{1}{\theta_k - 1}\lambda_k^{\mathrm{B}}\,\mathrm{d}\log\left(\lambda_k^{\mathrm{B}}\left(1 - \frac{1}{\mu_k^q}\right)\right),$$

and from Proposition 4,

$$\sum_k \frac{1}{\theta_k - 1}\lambda_k^{\mathrm{B}}\,\mathrm{d}\log\left(\lambda_k^{\mathrm{B}}\left(1 - \frac{1}{\mu_k^q}\right)\right) = \frac{Cov_{\lambda^{\mathrm{B}}}\left(\frac{\theta-1}{\theta-\theta_0}, d\log A\right)}{\mathbb{E}_{\lambda^{\mathrm{B}}}\left(\frac{\theta-1}{\theta-\theta_0}\right)}.$$

The Hulten-like term $\mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A)$ is the change in technical efficiency. The term $\sum_k 1/(\theta_k - 1)\lambda_k^{\mathrm{B}}\,\mathrm{d}\log(\lambda_k^{\mathrm{B}}(1-1/\mu_k^q))$, reflecting changes in allocative efficiency, is the weighted sum of changes in variable profits with weights reflecting the size and external economies in each sector.

Assume that sectors are substitutes with $\theta_0 > 1$ and that the shock disproportionately increases the productivity of sectors with high external economies (low elasticities of substitution). Then $Cov_{\lambda^{\mathrm{B}}}((\theta-1)/(\theta-\theta_0), d\log A) > 0$ and so the shock leads to improvements in allocative efficiency. Basically, the shock triggers beneficial reallocations of labor towards sectors with high external economies which were too small to begin with from a social perspective. These forces operate in reverse when sectors are complements with $\theta_0 < 1$.

The correlation of productivity shocks and markups, which was key in the single-sector heterogenous-firm economy, is now completely irrelevant. This is because now labor reallocations happen purely on the extensive margin via changes in entry in the different sectors, while the intensive margin remains unchanged as individual producers

in the different sectors keep operating at the same scale. Instead, the key is now the correlation of productivity shocks and returns to scale.

7.2.2 DRS with Free Entry

We now show that changes in aggregate output are very different under DRS. Consider the same multi-sector model but now assume that each sector k's output

$$Y_k = M_k y_k$$

is a linear aggregate of an endogenous mass M_k of undifferentiated varieties. Each variety in sector k is produced from labor with decreasing returns $1 - \varepsilon_k$ and productivity A_k by a single firm and sold at a markup μ_k^y over marginal cost

$$y_k = A_k l_k^{1-\varepsilon_k}, \quad p_k^y = \mu_k^y mc_k.$$

Like in the IRS case, we are again interested in how a long-run steady-state with targeted free entry in all sectors responds to a vector of sector-level productivity shocks $\mathrm{d}\log A$.

Changes in aggregate output are given by

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^B}(\mathrm{d}\log A).$$

Changes in technical efficiency are captured by the same Hulten-like term as in the IRS case. By contrast, there are no longer any changes in allocative efficiency. This occurs despite the fact that there are equilibrium reallocations and that the model is inefficient. Basically, the adjustment in the sizes of the different sectors happen entirely on the extensive margin via changes in entry. Individual producers in the different sectors keep operating at the same scale so that there is no change on the intensive margin. Since in addition there are no external economies, changes in prices only reflect exogenous changes in productivities with no changes in the shadow prices of specific fixed factors. Reallocations therefore do not save on specific fixed factors and are, therefore, neutral on efficiency grounds. This example clarifies that the IRS and DRS models are, in general, very different.

7.3 Roundabout-Entry Economy

In this section, we consider a roundabout-entry economy with one product sector populated by homogenous firms and free entry with entry costs paid in units of an entry good produced with labor and products. We focus on a shock to the productivity of the product

sector. We explain how different views regarding the costs of entry (labor vs. products) and returns to scale (IRS vs. DRS) shape the response of aggregate output. [27]

IRS with Free Entry. Consider a roundabout-entry economy where aggregate output

$$Y = Y_1 - x_{21}$$

is the output of the product sector which is not used for entry. The output of the product sector

$$Y_1 = \left(M_1 q_1^{\frac{\theta_1 - 1}{\theta_1}} \right)^{\frac{\theta_1 - 1}{\theta_1}}$$

is a CES aggregate of an endogenous mass M_1 of differentiated varieties with an elasticity of substitution $\theta_1 > 1$ and associated external returns to scale $1/\gamma_1 = \theta_1/(\theta_1 - 1)$. Each variety is produced from labor with constant returns and productivity A_1 by a single firm and sold at a markup μ_1^q over marginal cost

$$q_1 = A_1 l_1, \quad p_1^q = \mu_1^q mc_1.$$

We are interested in a long-run steady-state with free entry where potential entrants pay a fixed entry cost that relies on both goods and factors. In particular, suppose the entry good is produced from labor and products and sold at marginal cost

$$Y_2 = \left(\overline{\Omega}_{2L}^V \left(\frac{l_2}{\overline{l}_2} \right)^{\frac{\theta_2 - 1}{\theta_2}} + \overline{\Omega}_{21}^V \frac{x_{21}}{\overline{x}_{21}}^{\frac{\theta_2 - 1}{\theta_2}} \right)^{\frac{\theta_2}{\theta_2 - 1}}, \quad P_2^Y = \mu_2^Y mc_2.$$

Consider a positive productivity shock $\mathrm{d} \log A_1 > 0$.

To apply our formulas, we will use the backward Domar weight (sales share) and Domar weight of products $\lambda_1^{\mathrm{B}} = 1/(1-1/(1-1/\mu_1^q)\Omega_{21}^V) \geq 1$ and $\lambda_1^{\mathrm{F}} = 1/(1-1/(\theta_1-1)\Omega_{21}^V) \geq 1$. We will also use the fact that the backward Domar weight (income share) of labor is one $\lambda_L^{\mathrm{B}} = 1$.

Consider first the special case where entry uses only labor $\Omega_{21}^V = 0$. Changes in aggregate output are then given by

$$\mathrm{d} \log Y = \mathrm{d} \log A_1.$$

[27] Note that there is a non-trivial input-output structure in entry but not in variable production. We refer the reader to Section 8 for some illustrative examples with a non-trivial input structure in entry.

There are only changes in technical efficiency, given by a Hulten-like term, and no changes in allocative efficiency.

Consider next the general case where entry also uses products $\Omega_{21}^V > 0$. From Theorem 3, changes in aggregate output are now given by

$$\mathrm{d}\log Y = \lambda_1^{\mathrm{F}}\,\mathrm{d}\log A_1 + \frac{\frac{\theta_2-1}{\theta_1-1}\lambda_1^{\mathrm{F}}(\lambda_1^{\mathrm{B}}-1)\left(1-\Omega_{21}^V\right)}{1-\frac{\theta_2-1}{\theta_1-1}\lambda_1^{\mathrm{F}}(\lambda_1^{\mathrm{B}}-1)\left(1-\Omega_{21}^V\right)}\lambda_1^{\mathrm{F}}\,\mathrm{d}\log A_1.$$

The first term on the right-hand side is Hulten-like and captures changes in technical efficiency, and the second term captures changes in technical efficiency.

When labor and products are Cobb-Douglas in entry $\theta_2 = 1$, there are only changes in technical efficiency and no changes in allocative efficiency. The effects of productivity shocks on technical efficiency are amplified, the more so, the higher is λ_1^{F}, or equivalently the greater the roundaboutness Ω_{21}^V of the economy. This is because the productivity shock hits products which are then used to make more products via entry.

When labor and products are not Cobb-Douglas in entry $\theta_2 \neq 1$, there are also changes in allocative efficiency. Focus on the substitutes case $\theta_2 > 1$. Changes in allocative efficiency are then positive and are responsible for a second source of amplification. An increase in the productivity of products leads to a reduction in their price. This leads to a substitution towards products and away from labor in entry, and generates more entry. This in turn further reduces the price of products because of love for variety, etc. ad infinitum. In fact changes in aggregate output become arbitrarily large when θ_2 increases towards a critical threshold.

This example shows that with IRS, the denomination of the entry costs influences the comparative statics of the model in important ways. We now show that this dependence, while still present, is lessened with DRS.

DRS with Free Entry. Consider now a version of the same model but where the product sector, instead of being modeled as IRS sector, is modeled as a DRS sector which linearly aggregates a mass of undifferentiated varieties each produced under decreasing returns $1-\varepsilon_1$ with productivity A_1 and sold at a markup μ_1^y over marginal cost. Changes in aggregate output are then given by

$$\mathrm{d}\log Y = \lambda_1^{\mathrm{F}} d\log A_1,$$

where $\lambda_1^F = 1/(1-\varepsilon_1\Omega_{21}^V) \geq 1$. There are only changes in technical efficiency, given by a Hulten-like term, and no changes in allocative efficiency. Unlike in the IRS case, this

property holds no matter whether products are used in entry or not.

Together with the multi-sector example above, this example underscores that the DRS and IRS models are only isomorphic when the input-output structure is trivial (a single sector and entry in labor). Away from this special case, they behave very differently.

8 Optimal Policy and Misallocation

In this section, we turn to optimal policy. We describe first-best policies when instruments are unrestricted as well as second-best policies when only limited instruments are available. We also characterize the gains from optimal policy by computing the economy's distance to the efficient production possibility frontier.

8.1 First-Best Policy

Theorem 1 implies that the first best is attained when $\mu_i^Y = \mu_i = 1$ for all $i \in \mathcal{N}$. The first best can be implemented in different ways using different instruments. Starting from a second-best equilibrium when markups/wedges are not at their first-best values, this can be achieved by applying, in each market i, a combination of (gross) output taxes $\tau_i^Y = 1/\mu_i^Y$ and $\tau_i = 1/\mu_i$, with the revenues collected paid respectively to the household and to the producers of the good.[28] This is immediate since the taxes and markups/wedges matter only through $\tau_i^Y \mu_i^Y$ and $\tau_i \mu_i$. Alternatively, the first best can be implemented by keeping τ_i^Y and replacing τ_i by markup regulations that ensure $\mu_i = 1$.

For example, in the DRS benchmark where $1 - \varepsilon_i < \gamma_i = 1$ and $\mu_i^Y = 1$, optimal policy can be obtained through perfect competition $\mu_i^y = 1$. In the IRS benchmark where $1 - \varepsilon_i = \gamma_i < 1$, $\mu_i^y = 1 - \varepsilon_i$, and $\mu_i^Y = 1/\gamma_i$, optimal policy can be obtained through monopolistic markups $\mu_i^q = 1/(1 - \varepsilon_i) > 1$ and output subsidies $\tau_i^Y = \gamma_i < 1$.

An important observation is that first-best policy is independent of the input-output network. The policy intervention in each market depends only on the markups/wedges and the returns to scale of that market.

8.2 Second-Best Policy

Whereas the first-best policy is network-independent, second-best policies do depend on the details of the network. In this section, we deliver simple bang-for-buck formulas to assess and compare the merits of different marginal interventions. These formulas revive

[28] The revenues collected by a tax are negative if the gross tax is below one and acts like a gross subsidy.

and revise the informal policy recommendations of Hirschman (1958), who argued in favor of encouraging sectors with increasing returns that had the most backward and forward linkages. The analysis reveals the extent to which details matter: effective policy depends crucially on the nature of the intervention and on subtle features of the economy.

We focus on the IRS benchmark. We assume that there is only one primary factor which we call labor. We also assume that entry is possible in all markets, or, in other words, that all markets are contested. We consider marginal interventions at the no-intervention equilibrium. We investigate markup regulation and entry subsidization, which can loosely be thought of as capturing respectively competition and industrial policy. These two types of interventions neatly illustrate two very different ways in which forward and backward linkages can matter.

Markup Regulation. To start with, consider a budget-neutral intervention reducing the markups $\mathrm{d}\log\mu_i^q < 0$ of the producers of market i. The response of aggregate output, normalized by the revenues $-\lambda_i^{\mathrm{B}}\,\mathrm{d}\log\mu_i^q > 0$ transfered away from the producers by the associated implicit subsidy, is

$$-\frac{1}{\lambda_i^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_i^q} = \frac{\lambda_i^{\mathrm{F}}}{\lambda_i^{\mathrm{B}}}\frac{1}{\gamma_i}\left(\frac{\mu_i^q - 1/\gamma_i}{\mu_i^q - 1}\right) + \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j} - 1\right)\left(-\frac{1}{\lambda_i^B}\frac{\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}}{\mathrm{d}\log\mu_i^q}\right).$$

The first term is the direct effect of the markup reduction, holding sales constant. It captures two opposing effects on the price of market good i and in turn on final-demand prices. On the one hand, the policy reduces the price of each individual producer in market i, making the good cheaper for the household. On the other hand, the policy also dis-incentivizes entry into market i, which increases the effective price of i due to reduced variety. Overall, whether the sign of the direct effect is positive or negative depends on whether there is too little or too much entry in market i to begin with, which in turn depends on whether the initial markup μ_i^q is lower or higher than the infra-marginal surplus created by new varieties $1/\gamma_i$.

Under monopolistic competition, the direct effects exactly cancel $\mu_i^q = 1/\gamma_i$, leaving us the second term

$$-\frac{1}{\lambda_i^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_i^q} = \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j} - 1\right)\left(-\frac{1}{\lambda_i^B}\frac{\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}}{\mathrm{d}\log\mu_i^q}\right). \tag{6}$$

The bang-for-buck impact of the intervention is measured by a simple sufficient statistic: a

forward-weighted sum across markets j of the changes in backward-linkages (i.e. market size and hence entry) resulting from the intervention $-(1/\lambda_i^{\mathrm{B}})\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}/\,\mathrm{d}\log\mu_i^q$ interacted with increasing returns to scale $1/\gamma_j - 1$.[29]

Entry Subsidies. Now consider marginal entry subsidies to type-i entrants at the no-intervention equilibrium. Without loss of generality, we treat the entry production function $g_i(x_{E,ij})$ of i as though it were operated by an incumbent producer assembling the resources needed to enter and selling them at marginal cost $\mu_{E,i}^Y = 1$. These entry-good producers play a special role and we will to denote their backward and forward Domar weights as $\lambda_{E,i}^{\mathrm{B}}$ and $\lambda_{E,i}^{\mathrm{F}}$. The backward Domar weight is equal to the profits earned by type-i entrants $\lambda_{E,i}^{\mathrm{B}} = \sum_{j\in\mathcal{N}-\mathcal{F}}\tilde{\zeta}_{ij}\lambda_{\pi,j}$. The forward Domar weight is captures the impact of type-i entry on final-demand prices $\lambda_{E,i}^{\mathrm{F}} = \sum_{j\in\mathcal{N}-\mathcal{F}}\tilde{\zeta}_{ji}\lambda_j^{\mathrm{F}}(1/\gamma_j - 1)$.

Introducing an entry subsidy on type-i entrants is equivalent to reducing the markup $\mathrm{d}\,log\mu_{E,i} < 0$ of the producer of entry good i. At the no-intervention equilibrium the budgetary impact is just $-\lambda_{E,i}^{\mathrm{B}}\,\mathrm{d}\log\mu_{E,i}^Y > 0$. The response of aggregate output, normalized by its budgetary impact to allow bang-for-buck comparisons, is

$$-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_{E,i}^Y} = \left(\frac{\lambda_{E,i}^{\mathrm{F}}}{\lambda_{E,i}^{\mathrm{B}}} - \frac{\lambda_L^{\mathrm{F}}}{\lambda_L^{\mathrm{B}}}\right) + \sum_{j\in\mathcal{N}-\mathcal{F}}\lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j} - 1\right)\left(-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}}{\mathrm{d}\log\mu_{E,i}^Y}\right), \tag{7}$$

where, at the no-intervention equilibrium, $\lambda_L^{\mathrm{B}} = 1$ since all markets are contested, but $\lambda_L^{\mathrm{F}} \neq 1$ in general since there are inefficiencies.

The bang-for-buck impact of the intervention depends on two simple sufficient statistics corresponding to the two terms in this expression. The second term is exactly the same as that for measuring the bang-for-buck impact of markup regulations in equation 6 and it has the same intuition.

By contrast, the first term is specific to entry subsidies. It depends on the difference between two ratios of forward to the backward Domar weights: that of the entry condition $\lambda_{E,i}^{\mathrm{F}}/\lambda_{E,i}^{\mathrm{F}}$ where the intervention takes place, and that of labor $\lambda_L^{\mathrm{F}}/\lambda_L^{\mathrm{B}}$.[30] The ratio of forward to backward Domar weights i measures the reduction in the size i caused by cumulated markups downstream. Hence, the first term boils down to a comparison of the cumulated distortions downstream from entry good i compared to labor. Holding sales constant, the

[29] When there are intermediate goods, it is actually possible for a markup reduction to increase the sales of all contested markets $-\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}/\,\mathrm{d}\log\mu_i^q$ simultaneously, further increasing the room for policy improvements by encouraging intermediate-input use.

[30] This first term is reminiscent of Liu (2017), who studied marginal interventions around the decentralized equilibrium of a production network economy without variable returns to scale or entry.

entry subsidy stimulates entry by type-i entrants, which reduces final demand prices; but also absorbs more resources into entry, which increases the real price of labor (the labor share) and in turn final-demand prices.

Illustrative Example: Cobb-Douglas Economies. To make this more concrete, specialize the analysis to the case where all the productions functions f_i and g_i are Cobb-Douglas. In addition, assume that entry only uses labor. As we shall see, the optimal intervention formulas then take a simple form.

Under our assumptions the backward and forward Leontief inverses restricted to the non-primary-factor-markets (indicated by superscript $\mathcal{N}-\mathcal{F}$) are given by

$$\begin{aligned}\Psi^{\mathrm{B},\mathcal{N}-\mathcal{F}} &= \left(I-\Omega^{V,\mathcal{N}-\mathcal{F}}\right)^{-1} \quad = I+\Omega^{V,\mathcal{N}-\mathcal{F}}+\left(\Omega^{V,\mathcal{N}-\mathcal{F}}\right)^2+\cdots,\\ \Psi^{F,\mathcal{N}-\mathcal{F}} &= (I-\mu^q\Omega^{V,\mathcal{N}-\mathcal{F}})^{-1} = I+\mu^q\Omega^{V,\mathcal{N}-\mathcal{F}}+\left(\mu^q\Omega^{V,\mathcal{N}-\mathcal{F}}\right)^2+\cdots.\end{aligned}$$

For non-primary-factor-markets i, backward and forward Domar weights are given by $\lambda_i^{\mathrm{B}}=\Psi_{0i}^{\mathrm{B},\mathcal{N}-\mathcal{F}}$ and $\lambda_i^{\mathrm{F}}=\Psi_{0i}^{\mathrm{F},\mathcal{N}-\mathcal{F}}$. This shows that the ratio of forward to backward Domar weights $\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}}$, captures the extent to which market i is shrunk by multiple marginalization downstream. If in addition markups are at their monopolistic-competition levels $\mu_q^j=1/\gamma_j$, then $\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}}$ becomes a measure of the cumulated gross increasing returns in the downstream supply chain (excluding itself) of market i. The adjusted ratio $(\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}})(1/\gamma_i)$ also includes the gross increasing returns of market i rather than only those of markets strictly downstream from i.

This means that the bang-for-buck impact of a marginal entry subsidy for type-i entrants is

$$-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_{E,i}^{Y}}=\frac{\lambda_{E,i}^{\mathrm{F}}}{\lambda_{E,i}^{\mathrm{B}}}-\frac{\lambda_L^{\mathrm{F}}}{\lambda_L^{\mathrm{B}}}\quad\text{where}\quad\frac{\lambda_{E,i}^{\mathrm{F}}}{\lambda_{E,i}^{\mathrm{B}}}=\sum_{j\in\mathcal{N}-\mathcal{F}}\frac{\tilde{\zeta}_{ij}\lambda_j^{\mathrm{B}}\left(1-\frac{1}{\mu_j^q}\right)}{\sum_{j'\in\mathcal{N}-\mathcal{F}}\tilde{\zeta}_{ij'}\lambda_{j'}^{\mathrm{B}}\left(1-\frac{1}{\mu_{j'}^q}\right)}\left(\frac{\lambda_j^{\mathrm{F}}}{\lambda_j^{\mathrm{B}}}\frac{\frac{1}{\gamma_j}-1}{1-\frac{1}{\mu_j^q}}\right).$$

Entry subsidies are most effective when they target entrants i with the highest ratio of forward to backward Domar weights $\lambda_{E,i}^{\mathrm{F}}/\lambda_{E,i}^{\mathrm{B}}$, or equivalently entrants entering on average in markets j with the highest ratios of forward to backward Domar weights $\lambda_j^{\mathrm{F}}/\lambda_j^{\mathrm{B}}$ and net returns to scale to profit margins $(1/\gamma_j-1)/(1-1/\mu_j^q)$.[31]

If markups are at their monopolistic-competition levels $\mu_q^j=1/\gamma_j$, this boils down to

[31]The weights in the average across markets reflect the probability $\tilde{\zeta}_{ij}$ of entering market j and the associated profits $\lambda_j^{\mathrm{B}}\left(1-\frac{1}{\mu_j^q}\right)$. They reflect the relative importance of market j in incentivizing type-i entrants.

targeting entrants that enter on average in those markets j that have the most cumulated gross increasing returns in their downstream supply chains (including themselves) as captured by the adjusted ratio $(\lambda_j^{\mathrm{F}}/\lambda_j^{\mathrm{B}})(1/\gamma_j)$. If in addition, all markets have the same increasing returns to scale, then this tends to favor upstream interventions.

We next consider markup regulation. We assume for simplicity that at the no-intervention equilibrium, markups are their monopolistic-competition levels $\mu_q^j = 1/\gamma_j$, and that entry is perfectly targeted. The changes in market sizes triggered by the intervention take the simple form $-\mathrm{d}\log\lambda_j^{\mathrm{B}}/\mathrm{d}\log\mu_i^q = \Psi_{ij}^{\mathrm{B}}(\lambda_i^{\mathrm{B}}/\lambda_j^{\mathrm{B}})$ when $j \neq i$ and $-\mathrm{d}\log\lambda_j^{\mathrm{B}}/\mathrm{d}\log\mu_i^q = 0$ when $j = i$. The bang-for-buck impact of a marginal increase in the markups of market i is then

$$-\frac{1}{\lambda_i^{\mathrm{B}}}\frac{d\log Y}{d\log\mu_i^q} = \sum_{j\in\mathcal{N}-\mathcal{F}-\{i\}}\frac{\lambda_j^{\mathrm{F}}}{\lambda_j^{\mathrm{B}}}\left(\frac{1}{\gamma_j}-1\right)\Psi_{ij}^{\mathrm{B}}.$$

Therefore, markup regulations are most effective when they reduces markups in markets i that are downstream from supply chains Ψ_{ij}^{B} composed of markets j with the most net increasing returns $(1/\gamma_j - 1)$ and the most cumulated gross increasing returns in their own downstream supply chains (excluding themselves) as captured by the ratio of their forward to backward Domar weights $\lambda_j^{\mathrm{F}}/\lambda_j^{\mathrm{B}}$. If in addition, all markets have the same increasing returns to scale, then this tends to favor downstream interventions.

In Appendix H, we fully work out an example with three sectors: two final-good sectors and an intermediate-good sector. One of the final-good sectors uses only labor in variable production while the other one uses only the intermediate good. There is free entry using labor in all sectors. To maximize bang for buck, interventions should leverage the larger cumulated increasing returns along the vertical supply chain: markup reductions should be targeted to the final-good sector which is downstream from the intermediate good sector, and entry subsidies should be targeted to the upstream intermediate-good sector.

8.3 Social Costs of Distortions

In this section, we characterize the gains from optimal policy, which coincide with the social costs of distortions, the distance from the efficient frontier, or the amount of misallocation. We show that even with non-neoclassical ingredients like entry, non-convexities, and external economies, the distance to the frontier can be approximated via a Domar-weighted sum of Harberger triangles associated with variable production and entry. We then specialize the result and work through a series of examples to emphasize how seemingly minor details can drastically alter one's view about the extent of misallocation.

By Theorem 2, the marginal-cost pricing equilibrium with markups/wedges $\mu_i = \mu_i^Y = 1$ is efficient. We consider nearby equilibria associated with close-to-efficient markups/wedges. For any equilibrium variable X, we denote by $\mathrm{d}\log X$ the log-deviation of X from its value at the marginal-cost pricing equilibrium, which can also be thought of as the change in X caused by the deviations $\mathrm{d}\log\mu_i$ and $\mathrm{d}\log\mu_i^Y$ of the markups/wedges from their marginal-cost pricing values. We provide a second-order approximation in these deviations $(\mathrm{d}\log\mu, \mathrm{d}\log\mu^Y)$ of the associated aggregate efficiency loss $\mathcal{L} = -(1/2)\,\mathrm{d}^2\log Y$.

Proposition 5 (Deadweight-Loss)**.** *The efficiency loss can be approximated as the sum of Harberger triangles associated with variable production and entry*

$$\mathcal{L} \approx \frac{1}{2}\sum_{i\in\mathcal{N}-\mathcal{F}} \lambda_i^B \frac{1}{\gamma_i}\,\mathrm{d}\log y_i\,\mathrm{d}\log\left(\mu_i\mu_i^Y\right) + \frac{1}{2}\sum_{i\in\mathcal{N}-\mathcal{F}} \lambda_i^B \frac{1}{\gamma_i}\,\mathrm{d}\log M_i\,\mathrm{d}\log\mu_i^Y.$$

This expression is best suited for the DRS case because it emphasizes the change $\mathrm{d}\log y_i$ in the quantity of each undifferentiated variety in each market i. For the IRS benchmark, it is useful to rewrite the loss function to emphasize the change $\mathrm{d}\log q_i$ in the quantity of each differentiated variety in each i. We get

$$\mathcal{L} = \frac{1}{2}\sum_{i\in\mathcal{N}-\mathcal{F}} \lambda_i^B d\log q_i d\log\left(\mu_i\mu_i^Y\right) + \frac{1}{2}\sum_{i\in\mathcal{N}-\mathcal{F}} \lambda_i^B \frac{1}{\gamma_i} d\log M_i d\log\mu_i^Y.$$

In conjunction with the forward and backward propagation equations in Propositions 3 and 4, we can rewrite these loss functions in terms of microeconomic primitives. We relegate this general formula to Appendix A, and focus on a few prominent examples obtained by considering a special class of models with a sectoral structure.

8.3.1 Sectoral Models

To generate examples, we will use *sectoral* models defined by the following conditions:

1. every non-primary-factor-market $i \in \mathcal{N}-\mathcal{F}$ can be assigned to a unique sector $\mathcal{I}$, with common returns to scale $\varepsilon_i = \varepsilon_{\mathcal{I}}$ and $\gamma_i = \gamma_{\mathcal{I}}$, and so that its output matters only through sectoral output

$$Y_{\mathcal{I}} = \left(\sum_{i\in\mathcal{I}} Y_i^{\gamma_{\mathcal{I}}}\right)^{\frac{1}{\gamma_{\mathcal{I}}}} = \left(\sum_{i\in\mathcal{I}} M_i y_i\right)^{\frac{1}{\gamma_{\mathcal{I}}}} = \left(\sum_{i\in\mathcal{I}} M_i q_i^{1-\varepsilon_{\mathcal{I}}}\right)^{\frac{1}{\gamma_{\mathcal{I}}}};$$

2. individual producers in the markets market i in sector $\mathcal{I}$, have the same production

function $q_i = A_i f_{\mathcal{I}}(\{x_{i\mathcal{J}}\})$, where $x_{i\mathcal{J}}$ indicates that inputs are sectoral aggregates, but have different productivities A_i;

3. there is one type of entrant for each sector $\mathcal{I}$, and entrants are randomly assigned to markets $i \in \mathcal{I}$ according to some fixed distribution;

4. individual producers in the markets market i in sector $\mathcal{I}$ charge different markups $\mu_i = \mu_i^q \mu_i^y$ but share common output wedges $\mu_i^Y = \mu_{\mathcal{I}}^Y$.

Sectoral models, common in the literature, are worth singling out because their within-sector heterogeneity can be aggregated. We can then break the problem of computing the distance to the frontier into two recursive blocks, within and across sectors. See Appendix I for detailed derivations.

Throughout the following examples, we define the sales share of sector $\mathcal{I}$ to be $\lambda_{\mathcal{I}}^{\mathrm{B}} = \sum_{j \in \mathcal{I}} \lambda_j^{\mathrm{B}}$, and producer i's share of sector $\mathcal{I}$ to be $\lambda_i^{\mathcal{I},\mathrm{B}} = \lambda_i^{\mathrm{B}} / \lambda_{\mathcal{I}}^{\mathrm{B}}$. We will denote by $\mathbb{E}_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\mu)$ and $Var_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\mu)$ the within-sector weighted expectations and variances of changes in markups/wedges $\mathrm{d}\log\mu_i$ of producers $i \in \mathcal{I}$ with weights $\lambda_i^{\mathcal{I},\mathrm{B}}$.

8.3.2 Sectoral DRS Example

For sectoral models, we can provide a straightforward characterization of the loss function with DRS. That is, where $1 - \varepsilon_{\mathcal{I}} < \gamma_{\mathcal{I}} = 1$, $\mu_{\mathcal{I}}^Y = 1$, and $\mu_i^q = 1$. We proceed under the additional assumptions that there is only one primary factor, that entry paid in that factor, and that there are no deviations of output wedges from their efficient benchmarks $\mathrm{d}\log\mu_{\mathcal{I}}^Y = 0$.

Proposition 6. *In sectoral versions of the DRS benchmark with only one primary factor, with entry paid in that factor, and with no deviations of output wedges from their efficient benchmarks* $\mathrm{d}\log\mu_{\mathcal{I}}^Y = 0$*, the loss function is given by*

$$\mathcal{L} = \frac{1}{2}\sum_{\mathcal{I}} \lambda_{\mathcal{I}} \left(\frac{1}{\varepsilon_{\mathcal{I}}} - 1\right) Var_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\mu) + \frac{1}{2}\sum_{\mathcal{I}} \lambda_{\mathcal{I}} \left(\frac{1}{\varepsilon_{\mathcal{I}}} - 1\right) \left(\mathbb{E}_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\mu)\right)^2.$$

Because there are no output wedges, we know from Proposition 5 that there are no Harberger triangles associated with entry and only Harberger triangles associated with variable production. Of course, this does not mean that the entry margin is irrelevant, but changes in entry only matter through the impact on variable production.

The first term in the loss function captures misallocation arising from distortions in relative producer sizes driven by dispersed markups/wedges within sectors. The second

term captures misallocation arising from distortions in the average size of firms within sectors, or equivalently, distortions driven by an inappropriate average levels of markups within sectors. The losses increase with the returns to scale: they go to zero in the zero-returns to scale limit where $\varepsilon_{\mathcal{I}}$ goes to one, and they go to infinity in the constant-returns limit where $\varepsilon_{\mathcal{I}}$ goes to zero.

Proposition 6 is surprising if one is familiar with the misallocation literature. Normally, elasticities of substitution are key pieces of information. In this case, this information is not relevant since there is no misallocation across sectors.[32]

8.3.3 Sectoral IRS Examples

Now consider a sectoral version of the IRS benchmark where $1-\varepsilon_{\mathcal{I}} = \gamma_{\mathcal{I}} < 1$, $\mu^{Y}_{\mathcal{I}} = 1/\gamma_{\mathcal{I}}$, and $\mu^{y}_{i} = 1/(1-\varepsilon_{\mathcal{I}})$. We denote by $\theta_{\mathcal{I}} = 1/\varepsilon_{\mathcal{I}} = 1/(1-\gamma_{\mathcal{I}})$ the CES elasticity of substitution associated with every sector $\mathcal{I}$.

The behavior of sectoral IRS models is substantially more complicated than that of sectoral DRS models. In particular, whereas cross-sector elasticities of substitution are irrelevant under DRS, they remain very much relevant under IRS. Rather than providing the complicated general formula, we instead focus on some simple examples to give intuition. In each case, seemingly small changes in the assumptions about the nature of entry make the welfare costs of distortions quite different.

One-Sector Heterogenous-Firm Economy. We start with the one-sector model heterogenous-firm economy described in Section 7.1 with IRS and free entry. The aggregate efficiency loss is

$$\mathcal{L} = \frac{1}{2}\theta Var_{\lambda^{\text{B}}}\left(d\log\mu^{q}\right) + \frac{1}{2}\theta E_{\lambda^{\text{B}}}\left(d\log\mu^{q}\right)^{2}.$$

The first term, which captures misallocation on the intensive margin from the fact that high-markup firms are too small and low-markup firms are too big, depends on the elasticity of substitution and on the dispersion of markups. This term is standard in the literature (see e.g. Hsieh and Klenow, 2009; Baqaee and Farhi, 2019a). The second term, which captures misallocation on the extensive margin arising from the fact that there is too much or too little entry, depends on the elasticity of substitution and on the level of markups, and is new to the literature.

[32] In this model, changes in sectoral markups do not change relative sectoral prices to a first-order. An increase in a sector's markup increases the prices of producers in that sector but reduces their scale. At the efficient point, these effects cancel exactly. Therefore, sectoral prices do not change to a first-order, which means that to a first-order, the elasticity of substitution across sectors is not relevant for how quantities adjust. Since Harberger triangles are products of first-order changes in quantities and first-order changes in the markups, the cross-sectoral elasticity of substitution is irrelevant to a second-order.

If instead of inefficient markups μ_i^q, we considered inefficient output wedges μ_i^Y instead, then the extensive margin would be unaffected and we would only have misallocation on the intensive margin, leading to

$$\mathcal{L} = \frac{1}{2}\theta Var_{\lambda^B}\left(d\log\mu^Y\right).$$

This example underscores an important difference between models with and without free entry. In the latter markups and output taxes are equivalent (see e.g. Baqaee and Farhi, 2019a). However, in models with free entry, higher markups incentivize entry whereas higher output taxes do not.

Multi-Sector Economy. We continue with the multi-sector economy described in Section 7.1 with IRS and free entry. The aggregate efficiency loss is

$$\mathcal{L} = \frac{1}{2}\sum_{\mathcal{I}} \lambda_{\mathcal{I}}\theta_{\mathcal{I}}\left(\mathbb{E}_{\lambda^{\mathcal{I},B}}\left(d\log\mu\right)\right)^2.$$

Note that while the elasticities of substitution within sectors $\theta_{\mathcal{I}}$ matter, the elasticity of substitution in consumption across sectors θ_0 does not. This is because, at the efficient marginal-cost pricing equilibrium, changes in markups distort the allocation of resources within a sector between the extensive and intensive margins but these distortions have offsetting effects on the price of the sector good. Basically, there is only misallocation within sectors but no misallocation across sectors.

By contrast, with no entry and instead an exogenous mass M_i of incumbents in each market, the aggregate efficiency loss function becomes

$$\mathcal{L} = \frac{1}{2}\theta_0 Var_{\lambda^B}\left(\mathbb{E}_{\lambda^{\mathcal{I},B}}(d\log\mu)\right),$$

where the variance is the weighted variance of sectoral markups with weights given by sectoral sales shares $\lambda_{\mathcal{I}}^B$. In contrast to the free-entry case, the elasticity of substitution across sectors θ_0 is now relevant but the elasticities of substitution within sectors $\theta_{\mathcal{I}}$ are not, and there is only misallocation across sectors but not within sectors.

This examples illustrates that allowing for entry changes which elasticities of substitution are even relevant to misallocation.

Roundabout Economy. We finish with the roundabout economy described in Section 7.3. Suppose first that entry uses only labor. Then the aggregate efficiency loss is

$$\mathcal{L} = \frac{1}{2}\frac{1}{1-\gamma_1}(d\log\mu_1)^2 = \frac{1}{2}\theta_1(d\log\mu_1)^2.$$

The loss is increasing in the elasticity of substitution across products θ_1 since the love-of-variety effect is declining in θ_1, and goes to infinity as θ_1 goes to infinity since entry becomes socially wasteful.

Next, suppose that entry uses only products. Then the aggregate efficiency loss becomes

$$\mathcal{L} = \frac{1}{2}\frac{(\theta_1-1)^3}{(\theta_1-2)^2}(d\log\mu_1)^2.$$

Once again, the losses goes to infinity as θ_1 goes to infinity and for similar reasons. However, the loss is no longer increasing in θ_1, but is instead U-shaped, and also goes to infinity as θ_1 goes to 2 from above, since love of variety becomes so strong that output becomes linear in the mass of entrants. This example breaks the long-standing intuition in the misallocation literature that efficiency losses are increasing in the elasticity of substitution. In Appendix D.2.1, we show that this U-shaped pattern also arises with input-output linkages in variable production rather than entry.

This example illustrates how changing the input-output structure of entry can transform the losses from misallocation.

9 Quantitative Application

In this section, we quantify the social cost of distortions, or equivalently the gains from optimal policy. We also compute the social bang for a marginal buck of competition or industrial policy. We calibrate the model to fit U.S. data and provide a brief account of how we proceed; the details of how we map the model to data are in Appendix B.

9.1 Description of Quantitative Model

Our quantitative model has a sectoral structure with heterogenous firms within sectors and one primary factor capturing value-added. We merge firm-level data from Compustat with industry-level data from the BEA. We use annual input-output tables from the BEA, and assign each firm in the our Compustat sample to a BEA industry. In the data, we observe industry-level sales shares for industries $\mathcal{I}$; input-output entries for industries

$\mathcal{I}$ and $\mathcal{J}$; the sales shares of the Compustat firms i in industry $\mathcal{I}$; and the markup μ_i of Compustat firm i.

We adopt the baseline estimates of De Loecker et al. (2019) to obtain firm-level markups using a production-function (PF) approach. In Appendix C, we perform robustness checks by recomputing our results using three alternative methods for estimating markups: an alternative implementation of the production-function approach with different categories of costs, the user-cost approach (UC), and the accounting-profits (AP) approach. Although the numbers depend on the specific approach, the qualitative message remains the same.

The model has a nested CES structure where each firm i in industry $\mathcal{I}$ has a CES production function combining value-added and intermediate inputs with an elasticity of substitution θ_1. The intermediate input component is itself a CES aggregator of inputs from other industries with an elasticity of substitution θ_2. Finally, we have the within-sector elasticities $\varepsilon_{\mathcal{I}}$ and $\gamma_{\mathcal{I}}$ for each BEA industry. This means that at the sector level, we can have DRS or IRS.

Drawing on estimates from Atalay (2017), Herrendorf et al. (2013), and Boehm et al. (2014), we set the elasticity of substitution across sectors in consumption to be $\theta_0 = 0.9$, between value-added and intermediates to be $\theta_1 = 0.5$, and across sectors in intermediates to be $\theta_2 = 0.2$. Our results are not particularly sensitive to these choices.

We use the same within-sector elasticities for all sectors: $\varepsilon_{\mathcal{I}} = \varepsilon$ and $\gamma_{\mathcal{I}} = \gamma$. We consider two benchmarks corresponding to different returns to scale: IRS with $1 - \varepsilon = \gamma$ and DRS with $1 - \varepsilon < \gamma = 1$.

We consider two different scale elasticities, $1 - \varepsilon = 0.875$ and $1 - \varepsilon = 0.75$, which are equivalent in the IRS benchmark to using CES aggregators with elasticities of substitution given by $\theta = 8$ and $\theta = 4$ respectively.

Finally, we experiment with different ways of modeling entry: no entry, entry using primary factors, and entry using primary factors and goods (in the same way as variable production). The model without entry can be thought of as a short-run model, and the model with entry as a long-run model.

Our quantitative model is, of course, highly stylized. We therefore caution the reader from over-interpreting our quantitative results, which are meant only to be suggestive. Our analysis emphasizes just how sensitive the quantitative results are to difficult-to-measure parameters such as external economies, elasticities of substitution, input-output linkages in variable production and in entry, markups, and barriers to entry.

9.2 Social Costs of Distortions

We solve the model nonlinearly and compute the efficiency loss from misallocation. We report the numbers as the percentage gain in welfare achieved by implementing optimal policy starting from the decentralized equilibrium outcome. The results are in Table 1 for different combinations of assumptions regarding entry and returns to scale.

Across the board, the losses from inefficiency are higher when we allow entry than we do not, refuting the notion that endogenizing entry necessarily reduces the social cost of markups.

The "Level only" row eliminates the dispersion of markups within each sector by setting all markups within each sector equal to the harmonic average of markups in that sector. The "Dispersion only" row proportionately rescales the markups in the data so that their harmonic average within each sector is equal to one (this means average markups are equal to the CES markups when we adopt the IRS benchmark and equal to one when we adopt the DRS benchmark).

IRS, $1-\varepsilon=0.875$	No Entry	Entry uses Factors	Entry uses Goods and Factors
Level only	4.6%	14%	10%
Dispersion only	30%	30%	30%
Benchmark	36%	50%	41%
IRS, $1-\varepsilon=0.75$			
Level only	4.6%	17%	20%
Dispersion only	22%	23%	20%
Benchmark	19%	32%	37%
DRS, $1-\varepsilon=0.875$			
Level only	1.5%	7.8%	7.6%
Dispersion only	23%	23%	23%
Benchmark	26%	35%	32%
DRS, $1-\varepsilon=0.75$			
Level only	0.8%	9.5%	10%
Dispersion only	9.2%	9.2%	9.2%
Benchmark	9.6%	19%	20%

Table 1: Efficiency losses from misallocation. Firm-level returns to scale $1-\varepsilon=0.875$ under DRS correspond to elasticities of substitution across firms within sectors $\theta=8$ under IRS. Firm-level returns to scale $1-\varepsilon=0.75$ under DRS correspond to elasticities of substitution across firms within sectors $\theta=4$ under IRS.

When there is no entry, almost the entirety of the losses are explained by the dispersion

effect. The losses due to the dispersion effect are due to misallocation across firms within sectors, and are large because markups are very dispersed within sectors and because the relevant elasticities within sectors are large. The losses due to the level effect, when there is no entry, are entirely due to misallocation across sectors, and are small because markups are not so dispersed across sectors and because the cross-sectoral elasticities of substitution are low.

When there is entry, the level effect becomes comparable to the dispersion effect. The losses due to the level effect now also reflect misallocation between entry and variable production within sectors, and these losses are large because markups are in general too large and because the relevant elasticities are large.

Whether entry only uses primary factors or also intermediates has ambiguous effects. Depending on the scale elasticities, the relative size of the gains can go either way. When the entry margin is more important (ε is larger), the gains tend to be higher when entry also uses intermediates.

The efficiency losses are uniformly higher in the IRS benchmark than in the DRS benchmark. To understand this, it is useful to think about the limit where ε goes to one, which corresponds to a within-sector across-firm elasticity of one under IRS and a firm-level return to scale of zero under DRS. In this limit, under IRS, the efficiency losses become infinite if there are non-trivial input-output linkages because love of variety becomes extreme and so do the inefficiencies in entry. By contrast, under DRS, the efficiency losses go to zero as made clear by Propositions 5 and 6. This is because there are no Harberger triangles associated with variable production since firms have a fixed scale $d \log y_i = 0$, and there are no Harberger triangles associated with entry since there are no output wedges $d \log \mu_i^Y = 0$.

In Table 2, we compare the results of the benchmark model to versions of the model that employ some commonly used shortcuts: ignoring intermediate goods in production or entry (assuming no input-output); using a single-sector economy but allowing for intermediates (round-about economy); ignoring firm-level heterogeneity within sectors (no firm heterogeneity). We discuss each of these strawmen in turn.

The no-input-output-economy assumes away intermediates, and calibrates the size of each industry to be equal to its value-added share. Without entry, this economy undershoots the benchmark model for reasons discussed at length by Baqaee and Farhi (2019a). The undershooting becomes even more extreme once we allow for entry, underscoring even more strongly the need to model input-output linkages.

The round-about economy assumes that all firms in the economy belong to a single sector. The output of this sector is used both as the consumption good and as an in-

IRS	No Entry	Entry uses Factors	Entry uses Goods/Factors
Benchmark	36%	50%	40%
No Input-Output	16%	20%	–
Round-about Economy	139%	182%	133%
No Firm Heterogeneity	4.6%	14%	10%
DRS			
Benchmark	26%	35%	32%
No Input-Output	13%	18%	–
Round-about Economy	91%	123%	108%
No Firm Heterogeneity	1.0%	7.8%	7.6%

Table 2: Efficiency losses from misallocation when different disaggregated aspects of the economy are trivialized. We use firm-level returns to scale $1-\varepsilon = 0.875$ under DRS, which correspond to elasticities of substitution across firms within sectors $\theta = 8$ under IRS.

termediate input into production. The one-sector round-about economy overshoots the benchmark by a large amount. This is to be expected since the round-about economy aggregates all firms in the economy into a single sector. This means cross-sectoral dispersions in markups (which are less costly than within-sectoral dispersions) are treated as if they are within-sectors. Intuitively, dispersed markups now distort input choices across producers by more, since firms in two different industries are treated as if they are highly substitutable.

Finally, the no-firm-heterogeneity economy assumes that all firms in a sector are identical, with the same productivity shifter and the same markup equal to the sectoral markup. The homogeneous sectors economy undershoots the benchmark by a large amount because even though it accounts for cross-sectoral distortions, it abstracts away from within-sector misallocation.

All in all, the sensitivity of these numbers underscores the importance of theoretically unpacking and precisely measuring the details. Our results highlight the need for more empirical guidance on issues such as the strength of external economies and input-output linkages in variable production and in entry.

Role of the Elasticity of Substitution Across Firms Within Sectors. In many models of misallocation without entry, for example (e.g. Hsieh and Klenow, 2009; Baqaee and Farhi, 2019a), the distance to the frontier increases with the elasticity across firms within sectors. As discussed in Section 8.3, this intuition also holds in our model with entry under DRS but not under IRS.

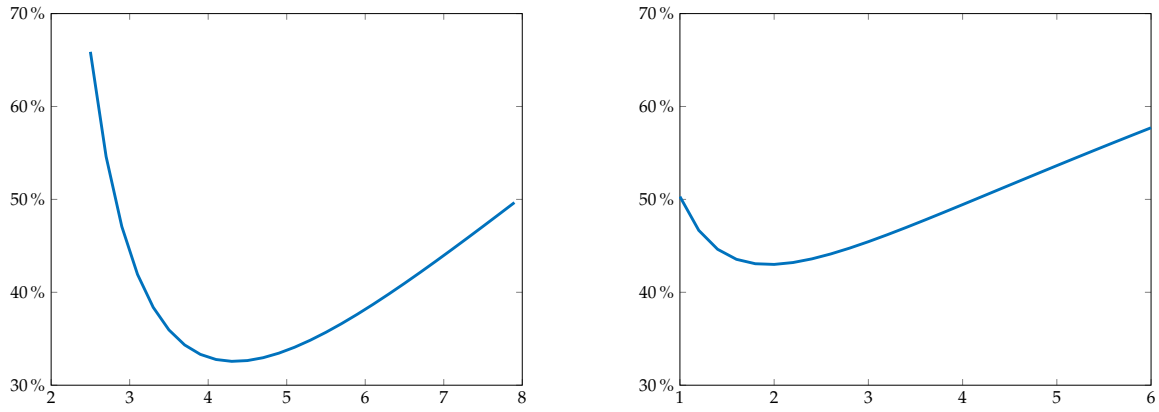
Figure 2a shows that for the IRS becnhmark, the distance to the frontier is U-shaped as a function of the elasticity of substitution across firms within sectors θ. For instance, the losses are 50% when $\theta = 8$. This number falls to 32% when the elasticity is lowered to $\theta = 4$, before rising to close to 65% when the elasticity is lowered further to $\theta = 2.5$. This is consistent with the theoretical discussion in the last example of Section 8.3. Intuitively, with non-trivial input-output linkages, a lower elasticity reduces misallocation along the intensive margin, but magnifies misallocation along the extensive margin. In the limit where θ goes to one, misallocation along the extensive margin becomes infinitely costly.

Role of Barriers to Entry. In our benchmark specifications with entry, we assume that that no friction interferes with free entry. In other words, we assume that all rents are quasi-rents rather than pure rents. However, it is plausible that, even in the long run, profits are not entirely offset by the costs of entry. For example, it may be that resources spent on entry are less than profits due to barriers to entry from regulations or due to anti-competitive strategic deterrence. We capture these barriers to entry in reduced form by introducing an entry tax/wedge.

Figure 2b displays the estimated distance to the frontier as a function of the view that one takes on the size of entry barriers in the data, where the size of entry barriers are measured by the size of the implicit entry tax/wedge (a value of one means that there are no barriers to entry). Perhaps surprisingly, the efficiency losses are non-monotonic in the size of entry barriers. Intuitively, whether barriers to entry increase or decrease the estimated distance to the frontier depends on whether there is too little or too much entry in the equilibrium with no entry barriers. Our estimated markups are relatively high, which implies that there is too much entry in the equilibrium with no entry taxes/wedges. As a result, if one takes the view that there are entry barriers in the data, then one is lead to a lower estimate of the distance to the frontier.

9.3 Bang for Buck of Marginal Policy Interventions

We turn to the the effect of a marginal policy intervention in the decentralized equilibrium. Figure 3 shows the bang-for-buck elasticity of aggregate output with respect to a marginal entry subsidy (a form of industrial policy) or markup reduction (a form of competition policy) in different industries. The elasticity is scaled by the revenues associated with the intervention as in Section 8.2. For this exercise, we focus on the IRS case where $\gamma_{\mathcal{I}} = 1 - \varepsilon_{\mathcal{I}} = 0.875$. We consider two alternative calibrations: one where we set markups equal to their CES monopolistic values, and one where we set markups equal to their



(a) Elasticity of substitution

(b) Entry wedges

Figure 2: Efficiency losses for the benchmark IRS model when entry uses factors.

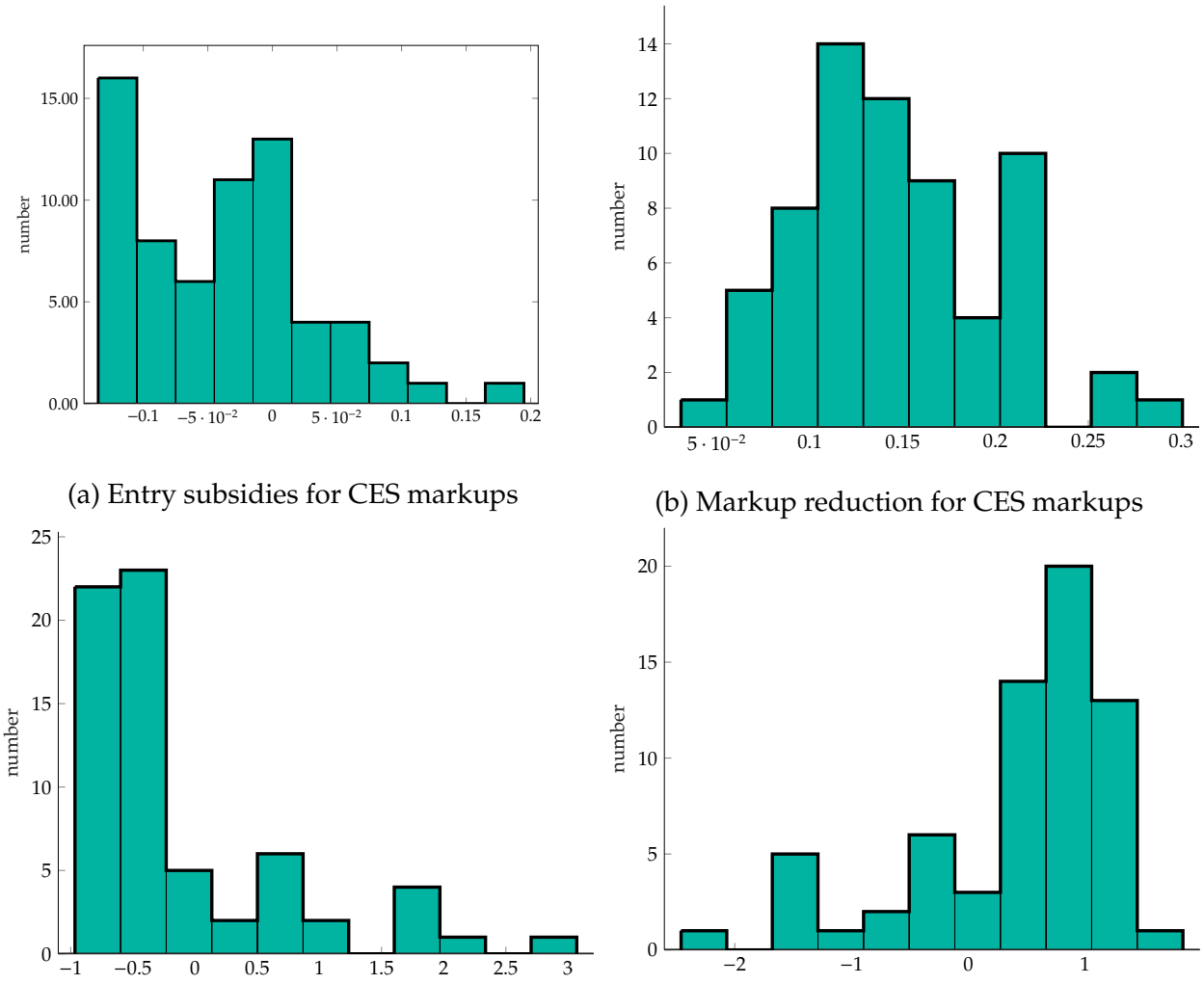
estimated values.

The monopolistic-markups calibration is a useful starting point for understanding the results, since it helps isolate the role played by the input-output network. In this case, as could be anticipated from the monopolistic-markup Cobb-Douglas example in Section 8, markup reductions are always beneficial. Because we have imposed the same increasing returns to scale from love of variety in all sectors, the greatest bang-for-buck are those with more complex supply chains, namely manufacturing industries like motor vehicles, metals, and plastics. The smallest gains come from those industries with the simplest supply chains, mostly service industries like housing or legal services but also primary industries like oil extraction or forestry.

For entry subsidies, the biggest gains, as expected, come from subsidizing those industries which are upstream in complex supply chains, namely primary industries like forestry, oil, and mining, whereas subsidizing entry into relatively downstream industries, like nursing, hospitals, or social assistance, is actually harmful.

When we move to the estimated markups, the shape of the input-output network is not the only determinant of the relative ranking of different industries, as now we must also consider whether each sector's markups are too high or too low on average relative to its external economies. Since we have imposed the same external economies from love of variety in all sectors but we have estimated markups, we do not read too much into the exact relative ranking of the different industries.

However, we can conclude that as we move farther away from the efficient frontier, as we do when we go from monopolistic markups to estimated markups, the potency of second-best policies increases dramatically. To see this, compare the magnitude of the elasticities in the top row to the bottom row of Figure 6.



(a) Entry subsidies for CES markups

(b) Markup reduction for CES markups

(c) Entry subsidies for estim. markups

(d) Markup reduction for estim. markups

Figure 3: The elasticity of output with respect to reductions in markups or an entry subsidy to different sectors normalized by the cost of the intervention. The top row uses CES markups, whereas the bottom row uses estimated PF markups.

But the larger effect sizes are a mixed blessing. Once we are far away from the frontier, the scope for policy having unintended consequences also increases. Although there appear to be many free lunches available to policy makers, many of them are poisoned since policy interventions can have large positive or negative signs. In other words, as implied by the theory of the second-best, interventions that seem sensible in isolation, like reducing markups, can reduce output once we are deep inside the frontier.

10 Conclusion

Traditional theories of aggregation, by relying on aggregate envelope theorems, imply that the aggregate production function can be treated like a black-box machine whose

contents are irrelevant to a first order. The only causal ingredient from the production side of the economy is, therefore, exogenous changes to the shape of this function. These exogenous changes, or total factor productivity shocks, are highly volatile in the data and responsible for a large fraction of the trend and cyclical components of aggregate output.

This approach is untenable for economies that are inside the Pareto-efficient frontier, and disaggregated economies, where many different margins can be misallocated, are likely to be deep inside the frontier. In these economies, total factor productivity is an endogenous object and affected, to a first order, by reallocation effects. These reallocation effects can be large enough to account for the bulk of variations in aggregate output. Reallocation not only amplifies exogenous micro-level productivity shocks, it can even replace exogenous productivity as a causal mechanism. And unlike changes in technology, which are likely gradual, always positive, and related to the physical act of production, changes in reallocation can be abrupt, beneficial or harmful, and related to the economic choices about the allocation of scarce resources.

This paper shows that these reallocation forces are especially potent in the presence of non-convexities and entry, where competitive markets can become a non-starter. Beneficial reallocations, by making better use of scarce resources, reduce their shadow price, and these prices are given by rents and quasi-rents. This provides a useful framework for studying the macroeconomics of scale.

References

Acemoglu, D., V. M. Carvalho, A. Ozdaglar, and A. Tahbaz-Salehi (2012). The network origins of aggregate fluctuations. *Econometrica 80*(5), 1977–2016.

Ackerberg, D. A., K. Caves, and G. Frazer (2015). Identification properties of recent production function estimators. *Econometrica 83*(6), 2411–2451.

Aghion, P. and P. Howitt (1992). A model of growth through creative destruction. *Econometrica: Journal of the Econometric Society*, 323–351.

Atalay, E. (2017). How important are sectoral shocks? *American Economic Journal: Macroeconomics (Forthcoming)*.

Autor, D., D. Dorn, L. Katz, C. Patterson, and J. Van Reenen (2017). The fall of the labor share and the rise of superstar firms.

Baqaee, D. R. (2018). Cascading failures in production networks. *Econometrica 86*(5), 1819–1838.

Baqaee, D. R. and E. Farhi (2019a). Productivity and misallocation in general equilibrium. *The Quarterly Journal of Economics 135*(1), 105–163.

Baqaee, D. R. and E. Farhi (2019b). The Macroeconomic Impact of Microeconomic Shocks: Beyond Hulten's Theorem. *Econometrica 87*(4), 1155–1203.

Bartelme, D. G., A. Costinot, D. Donaldson, and A. Rodriguez-Clare (2019, August). The Textbook Case for Industrial Policy: Theory Meets Data. NBER Working Papers 26193, National Bureau of Economic Research, Inc.

Behrens, K., G. Mion, Y. Murata, and J. Suedekum (2016). Distorted monopolistic competition.

Boehm, C., A. Flaaen, and N. Pandalai-Nayar (2014). Complementarities in multinational production and business cycle dynamics. Technical report, Working paper, University of Michigan.

Carvalho, V. and X. Gabaix (2013). The Great Diversification and its undoing. *The American Economic Review 103*(5), 1697–1727.

Ciccone, A. and K. Matsuyama (1996). Start-up costs and pecuniary externalities as barriers to economic development. *Journal of Development Economics 49*(1), 33–59.

Claus, J. and J. Thomas (2001). Equity premia as low as three percent? evidence from analysts' earnings forecasts for domestic and international stock markets. *The Journal of Finance 56*(5), 1629–1666.

De Loecker, J., J. Eeckhout, and G. Unger (2019). The rise of market power and the macroeconomic implications. Technical report.

Dhingra, S. and J. Morrow (2019). Monopolistic competition and optimum product diversity under firm heterogeneity. *Journal of Political Economy 127*(1), 196–232.

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, 297–308.

Domar, E. D. (1961). On the measurement of technological change. *The Economic Journal 71*(284), 709–729.

Edmond, C., V. Midrigan, and D. Y. Xu (2018). How costly are markups? Technical report, National Bureau of Economic Research.

Epifani, P. and G. Gancia (2011). Trade, markup heterogeneity and misallocations. *Journal of International Economics 83*(1), 1–13.

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

Georgescu-Roegen, N. (1951). Some properties of a generalized leontief model. *Activity Analysis of Allocation and Production. John Wiley & Sons, New York*, 165–173.

Grossman, G. M. and E. Helpman (1991). *Innovation and growth in the global economy*. MIT press.

Gutiérrez, G. and T. Philippon (2016). Investment-less growth: An empirical investigation. Technical report, National Bureau of Economic Research.

Harberger, A. C. (1954). Monopoly and resource allocation. In *American Economic Association, Papers and Proceedings*, Volume 44, pp. 77–87.

Harberger, A. C. (1964). The measurement of waste. *The American Economic Review 54*(3), 58–76.

Herrendorf, B., R. Rogerson, and A. Valentinyi (2013). Two perspectives on preferences and structural transformation. *American Economic Review 103*(7), 2752–89.

Hirschman, A. O. (1958). *The strategy of economic development*, Volume 58. Yale University Press New Haven.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*, 1127–1150.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing TFP in China and India. *The quarterly journal of economics 124*(4), 1403–1448.

Hulten, C. R. (1978). Growth accounting with intermediate inputs. *The Review of Economic Studies*, 511–518.

Kimball, M. S. (1995). The quantitative analytics of the basic neomonetarist model. *Journal of Money, Credit, and Banking 27*(4).

Krugman, P. R. (1979). Increasing returns, monopolistic competition, and international trade. *Journal of international Economics 9*(4), 469–479.

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

Liu, E. (2017). Industrial policies and economic development. Technical report.

Long, J. B. and C. I. Plosser (1983). Real business cycles. *The Journal of Political Economy*, 39–69.

McKenzie, L. W. (1959). On the existence of general equilibrium for a competitive market. *Econometrica: journal of the Econometric Society*, 54–71.

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

Murphy, K. M., A. Shleifer, and R. W. Vishny (1989). Industrialization and the big push. *Journal of political economy 97*(5), 1003–1026.

Olley, G. S. and A. Pakes (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica: Journal of the Econometric Society*, 1263–1297.

Osotimehin, S. and L. Popov (2017). Misallocation and intersectoral linkages. Technical report, Mimeo.

Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics 11*(4), 707–720.

Romer, P. M. (1987). Growth based on increasing returns due to specialization. *The American Economic Review 77*(2), 56–62.

Samuelson, P. A. (1951). Abstract of a Theorem Concerning Substitutability in Open Leontief Models. In T. Koopmans (Ed.), *Activity Analysis of Production and Allocation*, New York. Wiley.

Vincent, N. and M. Kehrig (2017). Growing productivity without growing wages: The micro-level anatomy of the aggregate labor share decline. In *2017 Meeting Papers*, Number 739. Society for Economic Dynamics.

Appendix A Proofs

We prove Theorem 1 using a slightly more general framework that allows for non-divisible overhead costs. This allows us to prove Theorems 1 and 2 for environments that nest Hopenhayn (1992) and Melitz (2003).

Non-Divisible Overhead

To augment the model with non-divisible overhead costs, suppose that in order to produce i, each producer must also pay an overhead/fixed-cost

$$h_i\left(\{x_{O,ik}\}_{k\in\mathcal{N}}\right),$$

where h_i has constant returns to scale and $x_{O,ik}$ are inputs used for overheads. A simple example is when firms must pay an overhead cost in units of labor if they choose to produce, as in Hopenhayn (1992) or Melitz (2003). When this overhead cost is zero, we recover the set up in the main paper.

The set of goods j can produce is $\operatorname{supp}\zeta(j) = \{i \in \mathcal{N} : \zeta(i,j) \neq 0\}$, let $\mathfrak{B}(j)$ be a set of subsets of $\operatorname{supp}_i \zeta(j,i) \subset \mathcal{N}$. For each $B \in \mathfrak{B}(j)$, entrant j can choose whether or not to produce the products in B. We say $B \subset \mathcal{N}$ is active if its expected profits are greater than zero

$$\left(\sum_{i\in B} \zeta(j,i)\frac{\lambda_{\pi_i}}{M_i}\right)\mathbf{1}\{B \text{ is active}\} \geq 0.$$

Entrant j does not produce i if i is ever a member of an inactive set. Hence, define the active set for j as

$$B^*(j) = \bigcap_{B\in\mathfrak{B}(j)} \{B \text{ is active}\}. \tag{8}$$

Let the mass of type-j entrants who have paid the sunk cost be $M_{E,j}$. Then, if $i \in \mathcal{N}$ is not produced by incumbents, then the mass M_i of products i is

$$M_i = \sum_{j\in E} \zeta(j,i) M_{E,j} \mathbf{1}\{i \in B^*(j)\}.$$

In words, M_i is the mass of products controlled by those entrants who paid both the sunk and overhead cost. This way of modelling the entrants' decision to operate is very general. At one extreme, we may imagine that entrant j can choose, one by one, whether or not to operate each technology i. In this case, $\mathfrak{B}(j)$ is the set of all singleton elements of $\operatorname{supp}_i \zeta(j,i)$. On the other extreme, we might imagine that entrant j may be constrained

to either operate all of its technologies or none of them. In this case, $\mathfrak{B}(j)$ contains only a single element $\{\text{supp}_i \zeta(j,i)\}$. Using $\mathfrak{B}(j)$, we can model intermediate situations when j can only choose whether or not to operate certain subsets of its technologies jointly.

A simple example is a case where firms can only operate at a given point in time if they have chosen to operate in every previous point in time, as in Hopenhayn (1992) or Melitz (2003). Of course, to model time in this fashion, we would simply index goods by time.

To prove Theorem 1, we require the following lemma.

Lemma 3. *At the prices in a decentralized equilibrium, the marginal cost pricing equilibrium maximizes aggregate profits.*

Proof. For each $i \in \mathcal{N}$,define

$$F_i^Z(M_i y_i, Z_i) = Z_i F_i\left(\frac{M_i y_i}{Z_i}\right)\mathbf{1}(M_i y_i > 0), \tag{9}$$

and note that $Z_i = 1$ gives $F_i^Z(M_i y_i, Z_i) = F_i(M_i y_i)$. By assumption, since F_i has constant or increasing returns to scale, $\frac{\partial F_i^Z}{\partial Z_i} \leq 0$.

When i sets prices according to marginal cost, we have $P_i^Y Y_i - p_i^y M_i y_i = \partial F_i/\partial Z_i$. Equivalently, we can think of the producer of Y_i as purchasing inputs $y_i M_i$ and Z_i at price $p_i^y > 0$ and $P_i^Z = \partial F_i/\partial Z_i \leq 0$ to maximize profits. Since every good is essential for consumption, we can assume that every good must be produced and $Z_i = 1$. In other words, we can treat Z_i as a fixed factor in inelastic supply $Z_i = 1$ with a negative price.

Aggregate profits are then given by

$$\begin{aligned}
\Pi(p,w) = P^Y \cdot Y - P^Z \cdot Z &- \sum_{j\in E}\sum_{i\notin B(j)} \zeta(j,i) M_{E,j}\left(\sum_{j\in\mathcal{N}-\mathcal{F}}\left(x_{ij} + \frac{x_{O,ij}}{M_i}\right) + \left(\sum_{f\in\mathcal{F}} x_{if} + \frac{x_{O,if}}{M_i}\right)\right) \\
&- \sum_{k\in\mathcal{N}-\mathcal{F}} p_k x_{E,jk} - \sum_{f\in\mathcal{F}} p_f x_{E,jf} \\
&= \sum_{j\in E}\sum_{i\notin B(j)} \zeta(j,i) M_{E,j}\left(p_i A_i f_i(x_{ij}, x_{if})\right) - \sum_{f\in\mathcal{F}} p_f x_{E,jf} \\
&- \sum_{j\in E}\sum_{i\notin B(j)} \zeta(j,i) M_{E,j}\left(\sum_{j\in\mathcal{N}-\mathcal{F}}\left(x_{ij} + \frac{x_{O,ij}}{M_i}\right) - \left(\sum_{f\in\mathcal{F}} x_{if} + \frac{x_{O,if}}{M_i}\right)\right) - \sum_{k\in\mathcal{N}} p_k x_{E,jk}
\end{aligned}$$

Since all producers have increasing or constant marginal cost, profits generated at this allocation must be lower than if those producers produced at the profit-maximizing,

cost-minimizing allocation chosen by the marginal-cost pricing equilibrium. Hence,

$$\begin{aligned}\Pi(p,w) &\leq \sum_{i\notin B(j)} \zeta(j,i) M_{E,j} \lambda_{\pi_i} - \sum_{k\in\mathcal{N}} p_k x_{E,jk} - \sum_{f\in\mathcal{F}} w_f l_{E,jf} \\ &\leq \sum_{j\in E} \left(\sum_{i\notin B^*(j)} \zeta(j,i)\lambda_{\pi_i} - C_j \right) M_{E,j},\end{aligned}$$

where C_j is the unit cost of the jth entrant. We know that for the prices prevailing in the equilibrium, $\left(\sum_{i\notin B^*(j)} \zeta(j,i)\lambda_{\pi_i} - C_j\right) = 0$, for every $j \in E$. Hence, at the decentralized prices,

$$\Pi(p,w) \leq 0, \tag{10}$$

for any allocation. ■

Proof of Theorem 1. Suppose that there exists some feasible allocation (denoted by primes) which is preferred to the decentralized one. Call this allocation c', and since it is strictly preferred, there must be some $i \in N$ such that $c'_i > c_i$. This means that c' is unaffordable at the equilibrium prices, hence $P^Y \cdot c' > P^Y \cdot c$. The allocation c' is delivered by some production plan, so that summing over all markets using the decentralized prices, we must have

$$\begin{aligned}\sum_{i\in\mathcal{N}} P^Y c'_i =& P^Y \cdot Y' - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} p_i \left(x'_{ji} M'_j + x'_{E,ji} \right) - \sum_{i\in\mathcal{N}}\sum_{j} p_i x'_{E,ji} \\ >& P^Y \cdot Y - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} p_i \left(x_{ji} M_j + x_{O,ji} \right) - \sum_{i\in\mathcal{N}}\sum_{j} p_i x_{E,ji}.\end{aligned}$$

Denote the total supply of factor $f \in \mathcal{F}$ by L_f. Now, subtract $\sum_{f\in\mathcal{F}} p_f L_F + P^Z \cdot Z'$ from both sides

$$\begin{aligned}& P^Y \cdot Y' - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} p_i \left(x'_{ji} M'_j + x'_{O,ji} \right) - \sum_{i\in\mathcal{N}}\sum_{j\in E} p_i x'_{E,ji} \\ & - P^Z \cdot Z' - \sum_{f\in F} p_f \left(\sum_{i\in\mathcal{N}} x'_{if} M_i \sum_{i\in E} x'_{E,if} + \sum_{i\in\mathcal{N}} x'_{O,if} \right) \\ > \\ & P^Y \cdot Y - \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} p_i \left(x_{ji} M_j + x_{O,ji} \right) - \sum_{i\in\mathcal{N}}\sum_{j\in E} p_i x_{E,ji} \\ & - P^Z \cdot Z - \sum_{f\in\mathcal{F}} p_f \left(\sum_{i\in\mathcal{N}} x_{if} M_i \sum_{i\in E} x_{E,if} + \sum_{i\in\mathcal{N}} x_{O,if} \right).\end{aligned}$$

The last line follows from factor market clearing for every $f \in \mathcal{F}$ and the fact that $Z_i = Z_i' = 1$.

Rewrite this as

$$\sum_{i\in\mathcal{N}} \lambda'_{\pi_i} M_i - \sum_{i\in\mathcal{N}-\mathcal{F}} \sum_{j\in E} p_i x'_{E,ji} - \sum_{f\in\mathcal{F}} p_f \sum_{i\in E} x'_{E,if} >$$
$$\sum_{i\in\mathcal{N}} \lambda_{\pi_i} M_i - \sum_{i\in\mathcal{N}-\mathcal{F}} \sum_{j\in E} p_i x_{E,ji} - \sum_{f\in\mathcal{F}} p_f \sum_{i\in E} x_{E,if}.$$

Substitute in the values of M_i to get

$$\sum_{j\in E}\sum_{i\in\mathcal{N}} \lambda'_{\pi_i}\left(\zeta(j,i) M'_{E,j} \mathbf{1}(i \notin \mathcal{B}^{*\prime}(j))\right) - \sum_{i\in\mathcal{N}-\mathcal{F}} \sum_{j\in E} p_i x'_{E,ji} - \sum_{f\in\mathcal{F}} p_f \sum_{i\in E} x'_{E,if} >$$
$$\sum_{j\in E}\sum_{i\in\mathcal{N}} \lambda_{\pi_i}\left(\zeta(j,i) M_{E,j} \mathbf{1}(i \notin \mathcal{B}^{*}(j))\right) - \sum_{i\in\mathcal{N}-\mathcal{F}} \sum_{j\in E} p_i x_{E,ji} - \sum_{f\in\mathcal{F}} p_f \sum_{i\in E} x_{E,if}$$

This contradicts the lemma, hence such an allocation cannot exist. ■

Proof of Theorem 2. An application of the envelope theorem to the planning problem decentralized by the marginal-cost pricing equilibrium. ■

Proof of Proposition 1. The perfect substitutes version is straightforward. Under the imperfect substitutes assumption, we assume differentiated products sell at prices p_i^q and quantity q_i. These differentiated varieties are then purchased by different agents j, so that the total quantity of good i purchased by j is defined to be

$$x_{ji} = \left(M_i q_{ji}^{1-\varepsilon_i}\right)^{\frac{1}{1-\varepsilon_i}} = M_i^{\frac{1}{1-\varepsilon_i}} q_{ji},$$

where q_{ji} is the quantity of variety i purchased by j. The budget j spends on purchasing variety i is then

$$P_i^Y x_{ij} = M_i p_i^q q_{ji}.$$

Recall that P_i^Y is the marginal cost of producing x_{ij} so we can write

$$P_i^Y x_{ij} = \mu_i^Y \mu_i^y \frac{\gamma_i}{1-\varepsilon_i} M_i p_i^q q_{ji} = \mu_i^Y \mu_i^y M_i p_i^q q_{ji}.$$

So we require that $\mu_i^Y \mu_i^y = 1$.

On the other hand, under the differentiated products interpretation, the profits accruing to each variety must be

$$(1 - 1/\mu_i^q) p_i q_i.$$

In general, the profits accruing to each producer are

$$\left(1-\frac{1-\varepsilon_i}{\mu_i^q\mu_i^y}\right)\frac{p_i^q q_i}{\mu_i^Y\gamma_i}.$$

Letting $\mu_i^y = (1-\varepsilon_i)$ and $\mu_i^Y = 1/\gamma_i = 1/(1-\varepsilon_i)$ gives us the correct amount of rents.

■

Proof of Lemma 1. We assume that the rows of ζ are linearly independent, otherwise there are trivial entry types. Initialize the quilibrium where all M_E have been normalized to unity, we have the zero-profit conditions

$$\begin{aligned}\lambda_{E,i} &= \sum_{j\in N}\left(\frac{\zeta_{ij}}{\sum_{k\in E}\zeta_{kj}}\right)\lambda_{\pi_j},\\ &= \sum_{j\in N}\tilde{\zeta}_{ij}\lambda_{\pi_j},\end{aligned}$$

where $\tilde{\zeta}_{ij} = \zeta_{ij}M_{E,i}/\left(\sum_{k\in E}\zeta_{kj}M_{E,k}\right)$. Using the fact that

$$M_i = \sum_j \zeta_{ji}M_{E,j}. \tag{11}$$

loglinearize to get the zero-profit condition

$$\sum_j \tilde{\zeta}_{ij}\lambda_{\pi_j}d\log\lambda_{\pi_j} - \left(\sum_j \tilde{\zeta}_{ij}\lambda_{\pi_j}\right)\left(\sum_j \Omega_{ij}^E d\log P_j\right) = \sum_j \tilde{\zeta}_{ij}\lambda_{\pi_j}d\log M_j, \tag{12}$$

or in matrix notation, where λ_E and λ_π are diagonal matrices:

$$\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P = \tilde{\zeta}\lambda_\pi\tilde{\zeta}' d\log M_E. \tag{13}$$

If ζ has linearly independent columns then

$$(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\right) = d\log M_E, \tag{14}$$

and

$$\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\right) = d\log M. \tag{15}$$

From constant returns to scale, we know that $\Omega^E d\log P = d\log P_E$. ■

Proof of Theorem 3. The aggregation equation is

$$d\log Y = -\Omega'_{(0)} d\log P. \tag{16}$$

Hence we can write, for any individual firm,

$$\begin{aligned}
d\log p_i^q &= d\log \mu_i^q + \sum_j \mu_i^q \Omega_{ij}^V d\log P_j^Y - \left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log A_i^q \\
d\log p_i^y &= d\log p_i^q + \frac{\varepsilon_i}{1-\varepsilon_i} d\log y_i + d\log \mu_i^y, \\
d\log P_i^Y &= d\log \mu_i^Y + d\log p_i^y + (\gamma_i - 1)\, d\log Y_i, \\
d\log P_i^Y &= d\log \mu_i^Y + \frac{\varepsilon_i}{1-\varepsilon_i} d\log y_i + d\log \mu_i^y + \\
&\qquad d\log \mu_i^q + \sum_j \mu_i^q \Omega_{ij}^V d\log P_j^Y + (\gamma_i - 1)\, d\log Y_i - \left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log A_i^q, \\
&= d\log \mu_i^Y + \frac{\varepsilon_i}{1-\varepsilon_i} (\gamma_i d\log Y_i - d\log M_i) + d\log \mu_i^y + d\log \mu_i^q \\
&\quad + \sum_j \mu_i^q \Omega_{ij}^V d\log P + (\gamma_i - 1)\, d\log Y_i - \left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log A_i^q, \\
&= d\log \mu_i^Y + \left(\frac{\varepsilon_i}{1-\varepsilon_i}\gamma_i + \gamma_i - 1\right) d\log Y_i - \frac{\varepsilon_i}{1-\varepsilon_i} (d\log M_i) + d\log \mu_i^y \\
&\quad + d\log \mu_i^q + \sum_j \mu_i^q \Omega_{ij}^V d\log P_j^Y - \left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log A_i^q, \\
&= d\log\left(\mu_i^Y \mu_i^y \mu_i^q\right) + \left(\frac{\gamma_i}{1-\varepsilon_i} - 1\right) d\log \lambda_i^B - \left(\frac{\gamma_i}{1-\varepsilon_i} - 1\right) d\log P_i \\
&\quad - \frac{\varepsilon_i}{1-\varepsilon_i} (d\log M_i) + \sum_j \mu_i^q \Omega_{ij}^V d\log P - d\log A_i^q, \\
\left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log P_i &= d\log\left(\mu_i^Y \mu_i^y \mu_i^q\right) + \left(\frac{\gamma_i}{1-\varepsilon_i} - 1\right) d\log \lambda_i^B - \frac{\varepsilon_i}{1-\varepsilon_i} (d\log M_i) \\
&\quad + \sum_j \mu_i^q \Omega_{ij}^V d\log P_j - \left(\frac{\gamma_i}{1-\varepsilon_i}\right) d\log A_i^q, \\
d\log P_i &= \left(\frac{1-\varepsilon_i}{\gamma_i}\right) d\log\left(\mu_i^Y \mu_i^y \mu_i^q\right) + \left(1 - \left(\frac{1-\varepsilon_i}{\gamma_i}\right)\right) d\log \lambda_i^B \\
&\quad - \frac{\varepsilon_i}{\gamma_i} (d\log M_i) + \left(\frac{1-\varepsilon_i}{\gamma_i}\right) \sum_j \mu_i^q \Omega_{ij}^V d\log P_j^Y - d\log A_i^q,
\end{aligned}$$

We know that

$$d\log M = \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi \tilde{\zeta}')^{-1} \left(\tilde{\zeta}\lambda_\pi d\log \lambda_\pi - \lambda_E \Omega^E d\log P\right). \tag{17}$$

Substitute our expression for entry to get, in matrix notation,

$$\begin{aligned}
d\log P =& \frac{1-\varepsilon}{\gamma} d\log\mu - d\log A^q + \left(1 - \frac{1-\varepsilon}{\gamma}\right) d\log\lambda^Y + \frac{(1-\varepsilon)}{\gamma}\mu^q\Omega^V d\log P \\
&+ \frac{\varepsilon}{\gamma}\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\lambda_E\Omega^E d\log P\right) - \frac{\varepsilon}{\gamma} d\log\hat{\lambda}_\pi, \\
=& \Psi^F\left(\frac{1-\varepsilon}{\gamma} d\log\mu - d\log A^q + \left(1 - \frac{1-\varepsilon}{\gamma}\right) d\log\lambda - \frac{\varepsilon}{\gamma}\left(d\log\hat{\lambda}_\pi + d\log\hat{z}\right)\right), \\
=& \Psi^F\left(\frac{1-\varepsilon}{\gamma} d\log\mu - d\log A^q\right) + \Psi^F\left(\frac{\varepsilon}{\gamma}\left(d\log\lambda - d\log\hat{\lambda}_\pi\right) + \left(1 - \frac{1}{\gamma}\right) d\log\lambda\right),
\end{aligned}$$

where

$$\Psi^F = \left(I - \frac{(1-\varepsilon)}{\gamma}\mu^q\Omega^V - \frac{\varepsilon}{\gamma}\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\Omega^E\right)^{-1}. \tag{18}$$

■

Proof of Proposition 2. Without loss of generality, impose no-overlapping entry. Then $\tilde{\zeta} \in \{0,1\}^{E\times N}$, hence

$$(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E = I_{E\times E}. \tag{19}$$

Therefore,

$$\Psi^F = \left(I - \mu^Y\mu^q\Omega^V - \frac{\varepsilon}{\gamma}\tilde{\zeta}'\Omega^E\right)^{-1}. \tag{20}$$

The feasibility condition is

$$d\log Y_i = \frac{C_i}{Y_i} d\log C_i + \sum_{j\in N}\frac{M_j x_{ji}}{Y_i} d\log(x_{ji}M_j) + \sum_{j\in E}\frac{x_{E,ji}}{Y_i} d\log x_{E,ji}. \tag{21}$$

Now consider the feasible allocation rule which holds the allocation matrix $\mathcal{X}$constant:

$$d\log Y_i = d\log C_i = d\log x_{ji}M_j = d\log x_{E,ji}. \tag{22}$$

Under this allocation rule we have

$$d\log y_i = \gamma_i d\log A_i + (1-\varepsilon_i)\sum_{j\in N}\mu_i^q\Omega_{ij}^V d\log x_{ij} = \gamma_i d\log A_i + (1-\varepsilon_i)\sum_{j\in N}\mu_i^q\Omega_{ij}^V\left(d\log Y_j - d\log M_i\right). \tag{23}$$

Hence

$$d\log Y_i = d\log A_i + \sum_{j\in N}\mu_i^Y\mu_i^q\Omega_{ij}^V\left(d\log Y_j - d\log M_i\right) + \frac{1}{\gamma_i} d\log M_i$$

$$
\begin{aligned}
&= d\log A_i + \mu_i^Y \sum_{j\in N} \mu_i^q \Omega_{ij}^V d\log Y_j + \frac{\varepsilon_i}{\gamma_i} d\log M_i \\
&= d\log A_i + \mu_i^Y \sum_{j\in N} \mu_i^q \Omega_{ij}^V d\log Y_j + \frac{\varepsilon_i}{\gamma_i} \sum_{m\in E} \tilde{\zeta}_{mi} d\log M_{E,m} \\
&= d\log A_i + \mu_i^Y \sum_{j\in N} \mu_i^q \Omega_{ij}^V d\log Y_j + \frac{\varepsilon_i}{\gamma_i} \sum_{m\in E}\sum_{j\in N} \tilde{\zeta}_{mi} \Omega_{mj}^E d\log Y_j
\end{aligned}
$$

In matrix form

$$
\begin{aligned}
d\log Y &= d\log A + \mu^Y \mu^q \Omega^V d\log Y + \frac{\varepsilon}{\gamma} \tilde{\zeta}' \Omega^E d\log Y \\
&= (I - \mu^Y \mu^q \Omega^V d\log Y - \frac{\varepsilon}{\gamma} \tilde{\zeta}' \Omega^E)^{-1} d\log A \\
&= \Psi^F d\log A.
\end{aligned}
$$

■

Proof of Proposition 3. The proof for this is the same as that of Proposition 3. ■

Proof of Proposition 4. Define the notation $\lambda_i^Y = P_i Y_i$, $\lambda_i^y = p_i^y y_i$, and $\lambda_i^q = p_i^q q_i$. Note that $\lambda_i^Y = \lambda_i^B$. Now recall

$$
\lambda_{\pi_i} = \left(1 - \frac{1-\varepsilon_i}{\mu_i^q \mu_i^y}\right) \frac{\lambda_i^B}{\mu_i^Y \gamma_i}.
$$

$$
d\log \lambda_{\pi_i} = d\log \lambda_i^B - d\log \mu_i^Y + \frac{\frac{1-\varepsilon_i}{\mu_i^q \mu_i^y}}{\left(1 - \frac{1-\varepsilon_i}{\mu_i^q \mu_i^y}\right)} d\log \mu_i^q \mu_i^y \tag{24}
$$

In other words,

$$
\lambda_E = diag\,(M_E)\,\zeta \left(1 - (1-\varepsilon)\,(\mu^q \mu^y)^{-1}\right) diag(M)^{-1} diag(\mu^Y \gamma)^{-1} \lambda^B, \tag{25}
$$

So,

$$
\begin{aligned}
\lambda^{B'} &= \lambda^{B'} \Omega^V + (\lambda_E)'\, \Omega^E, \\
&= \lambda^{B'} \Omega^V + \lambda^{B'} \left(1 - (1-\varepsilon)\,(\mu^q \mu^y)^{-1}\right) diag(M)^{-1} diag(\mu^Y \gamma)^{-1} \zeta' diag\,(M_E)\,\Omega^E.
\end{aligned}
$$

Therefore, using the fact that $d\varepsilon = 0$,

$$
d\lambda^{B'} = \lambda^{B'} d\Omega^V + \lambda^{B'}\,(1-\varepsilon)\, diag(d\log(\mu^q \mu^y))(\mu^q \mu^y)^{-1} diag(M)^{-1} diag(\mu^Y \gamma)^{-1} \zeta' diag\,(M_E)\,\Omega^E
$$

$$
\begin{aligned}
&- \lambda^{B'}\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(d\log M + d\log \mu^{Y}) diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right)\Omega^{E}\\
&+ \lambda^{B'}\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right) d\Omega^{E}\\
&+ \lambda^{B'}\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right) diag(d\log M_{E}) d\Omega^{E}\\
&+ d\lambda^{B'}\left(\Omega^{V}+\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta'\Omega^{E}\right),\\
&= \lambda^{B'}\left(d\Omega^{V}+(1-\varepsilon)\, diag(d\log\left(\mu^{q}\mu^{y}\right)) diag\left(\mu^{q}\mu^{y}\right)^{-1} diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right)\Omega^{E}\right)\Psi^{B}\\
&- \lambda^{B'}\left(\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(d\log M + d\log \mu^{Y}) diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right)\Omega^{E}\right)\Psi^{B}\\
&+ \lambda^{B'}\left(\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right) diag\left(d\log M_{E}\right)\Omega^{E}\right)\Psi^{B}\\
&+ \lambda^{B'}\left(\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right) d\Omega^{E}\Psi^{B}\right),
\end{aligned}
\tag{26}
$$

where, using the fact that in the initial equilibrium $\zeta diag(M)^{-1} = \tilde{\zeta}$

$$
\begin{aligned}
\Psi^{B} &= \left(I-\Omega^{V}-\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(M)^{-1} diag(\mu^{Y}\gamma)^{-1}\zeta' diag\left(M_{E}\right)\Omega^{E}\right),\\
&= \left(I-\Omega^{V}-\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}'\Omega^{E}\right).
\end{aligned}
$$

The constituent parts of (26) are:

$$
d\Omega^{V}_{ij} = -\Omega^{V}_{ij} d\log\left(\mu^{q}_{i}\mu^{Y}_{i}\mu^{y}_{i}\right) + \left(\mu^{q}_{i}\right)^{-1}(1-\theta_{i}) Cov_{i}\left(d\log P, I_{(j)}\right), \tag{27}
$$

$$
\left[(1-\varepsilon)\, diag(d\log\mu^{q}\mu^{y}) diag\left(\mu^{q}\mu^{y}\right)^{-1} diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}' diag\left(M_{E}\right)\Omega^{E}\right]_{ij} = \sum_{k}\frac{(1-\varepsilon_{k})}{\gamma_{k}}\frac{d\log\left(\mu^{q}_{k}\mu^{y}_{k}\right)}{\mu^{q}_{k}\mu^{y}_{k}\mu^{Y}_{k}}\tilde{\zeta}_{ik}\Omega^{E}_{ij} \tag{28}
$$

$$
\begin{aligned}
&\left[\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(d\log M + d\log\mu^{Y}) diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}' diag\left(M_{E}\right)\Omega^{E}\right]_{ij}\\
&\quad = \left[\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(d\log M + d\log\mu^{Y}) diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}' diag\left(M_{E}\right)\Omega^{E}\right]_{ij}
\end{aligned}
\tag{29}
$$

$$
\begin{aligned}
&\left[\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}' diag\left(M_{E}\right) diag\left(d\log M_{E}\right)\Omega^{E}\right]_{ij}\\
&\qquad = \sum_{k}\left(1-(1-\varepsilon_{i})\left(\mu^{q}_{i}\mu^{y}_{i}\right)^{-1}\right)\left(\mu^{Y}_{i}\gamma_{i}\right)^{-1}\tilde{\zeta}_{ki}\Omega^{E}_{kj} d\log M_{E,i}
\end{aligned}
\tag{30}
$$

$$
\left[\left(1-(1-\varepsilon)\left(\mu^{q}\mu^{y}\right)^{-1}\right) diag(\mu^{Y}\gamma)^{-1}\tilde{\zeta}' diag\left(M_{E}\right) d\Omega^{E}\right]_{ij} = 0, \tag{31}
$$

where the last line follows from the fact that we have assumed (without loss of generality)

that Ω^E is degenerate. Combining all this, we get

$$
\begin{aligned}
d\lambda_i^B = &-\sum_{m\in N}\lambda_m^B\sum_{k\in N}\Omega_{mk}^V\Psi_{ki}^B d\log\left(\mu_m^q\mu_m^y\mu_m^\Upsilon\right)+\sum_m\lambda_m^B\left(\mu_m^q\right)^{-1}(1-\theta_m)Cov_m\left(d\log P,\Psi_{(i)}^B\right)\\
&+\sum_{j\in E}\sum_{m\in N}\sum_{k\in N}\lambda_m^B\frac{(1-\varepsilon_m)}{\gamma_k}\frac{d\log\left(\mu_m^q\mu_m^y\right)}{\mu_k^q\mu_k^y\mu_k^\Upsilon}\tilde{\zeta}_{jm}\Omega_{jk}^E\Psi_{ki}^B\\
&-\sum_{k\in N}\sum_{m\in N}\sum_{j\in E}\lambda_m^B\left(1-(1-\varepsilon_m)\left(\mu_m^q\mu_m^y\right)^{-1}\right)\frac{1}{\mu_m^\Upsilon\gamma_m}\tilde{\zeta}_{jm}\Omega_{jk}^E\left(d\log M_m+d\log\mu^\Upsilon\right)\Psi_{ki}^B\\
&+\sum_{k\in N}\sum_{m\in N}\lambda_m^B\sum_{j\in E}\left(1-(1-\varepsilon_m)\left(\mu_m^q\mu_m^y\right)^{-1}\right)\frac{1}{\mu_m^\Upsilon\gamma_m}\tilde{\zeta}_{jm}\Omega_{jk}^E d\log M_{E,j}\Psi_{ki}^B.
\end{aligned}
$$

With non-overlapping entry, we use the following identity

Lemma 4. *Under non-overlapping entry, the following identity holds:*

$$
\begin{aligned}
&\sum_{k\in N}\sum_{m\in N}\sum_{j\in E}\lambda_m^B(1-(1-\varepsilon_m)\left(\mu_m^q\mu_m^y\right)^{-1})\frac{1}{\mu_m^\Upsilon\gamma_m}\tilde{\zeta}_{jm}\Omega_{jk}^E d\log M_m\Psi_{ki}^B\\
&\qquad=\sum_{k\in N}\sum_{m\in N}\sum_{j\in E}\lambda_m^B(1-(1-\varepsilon_m)\left(\mu_m^q\mu_m^y\right)^{-1})\frac{1}{\mu_m^\Upsilon\gamma_m}\tilde{\zeta}_{jm}\Omega_{jk}^E d\log M_{E,j}\Psi_{ki}^B. \quad (32)
\end{aligned}
$$

Proof. Rearrange the left-hand side to be:

$$
\begin{aligned}
\sum_{k\in N}\sum_{m\in N}\sum_{j\in E}\tilde{\zeta}_{jm}\lambda_{\pi_m}\Omega_{jk}^E\left(d\log M_m-d\log M_{E,j}\right)\Psi_{ki}^B&=\sum_{k\in N}\sum_{j\in E}\Omega_{jk}^E\Psi_{ki}^B\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m-d\log M_{E,j}\right)\\
&=\sum_{k\in N}\sum_{j\in E}\Omega_{jk}^E\Psi_{ki}^B\left(\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m\right)-\lambda_{E,j}d\log M_{E,j}\right)\\
&=\sum_{k\in N}\sum_{j\in E}\Omega_{jk}^E\Psi_{ki}^B\left(\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m\right)-\lambda_{E,j}d\log M_{E,j}\right).
\end{aligned}
$$

The free-entry condition is

$$
P_{E,j}=\sum_{k\in N}\zeta_{jk}\lambda_{\pi_k}\frac{1}{M_k}. \quad (33)
$$

$$
\begin{aligned}
\lambda_{E,j}d\log P_{E,j}&=\sum_{k\in N}\tilde{\zeta}_{jk}\lambda_{\pi_k}d\log\lambda_{\pi_k}-\sum_{k\in N}\tilde{\zeta}_{jk}\lambda_{\pi_k}d\log M_k,\\
\sum_{k\in N}\tilde{\zeta}_{jk}\lambda_{\pi_k}d\log M_k&=\sum_{k\in N}\tilde{\zeta}_{jk}\lambda_{\pi_k}d\log\lambda_{\pi_k}-\lambda_{E,j}d\log P_{E,j}
\end{aligned}
$$

$$\tilde{\zeta}\lambda_\pi d\log M = \tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E d\log P_E.$$

On the other hand,

$$\lambda_E d\log\lambda_E = \lambda_E d\log M_E + \lambda_E d\log P_E. \tag{34}$$

Finally, note that, free entry requires that

$$\begin{aligned}\lambda_E &= \tilde{\zeta}\lambda_\pi,\\ \lambda_E d\log\lambda_E &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi + \tilde{\zeta}d\log\tilde{\zeta}\lambda_\pi.\end{aligned}$$

If there is non-overlapping entry, then

$$d\log\tilde{\zeta} = 0. \tag{35}$$

Hence,

$$\begin{aligned}\lambda_E d\log M_E &= \lambda_E d\log\lambda_E - \lambda_E d\log P_E\\ &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E d\log P_E\\ &= \tilde{\zeta}\lambda_\pi d\log M.\end{aligned}$$

Therefore,

$$\sum_{k\in N}\sum_{j\in E}\Omega^E_{jk}\Psi^B_{ki}\left(\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m\right) - \lambda_{E,j}d\log M_{E,j}\right) = 0, \tag{36}$$

as needed. In general,

$$\begin{aligned}\lambda_E &= \tilde{\zeta}\lambda_\pi,\\ \lambda_E d\log\lambda_E &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi + d\tilde{\zeta}\lambda_\pi\\ &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi + d\log M_E\tilde{\zeta}\lambda_\pi - \tilde{\zeta}d\log M\lambda_\pi\end{aligned}$$

Hence

$$\begin{aligned}\lambda_E d\log M_E &= \lambda_E d\log\lambda_E - \lambda_E d\log P_E\\ &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi + d\log M_E\tilde{\zeta}\lambda_\pi - \tilde{\zeta}d\log M\lambda_\pi - \lambda_E d\log P_E\\ &= \tilde{\zeta}\lambda_\pi d\log M + d\log M_E\tilde{\zeta}\lambda_\pi - \tilde{\zeta}d\log M\lambda_\pi\\ &= d\log M_E\tilde{\zeta}\lambda_\pi\end{aligned}$$

In other words,

$$\sum_{k\in N}\sum_{j\in E}\Omega^E_{jk}\Psi^B_{ki}\left(\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m - d\log M_{E,j}\right)\right) = 0. \tag{37}$$

Simplify it a bit

$$\begin{aligned}\tilde{\zeta}\lambda_\pi d\log M &= \tilde{\zeta}\lambda_\pi\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\right)\\ &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\end{aligned}$$

$$\begin{aligned}\lambda_E d\log M_E &= \lambda_E\left((\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^F d\log P\right)\right),\\ &= diag(\tilde{\zeta}\lambda_\pi\mathbf{1})(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^F d\log P\right).\end{aligned}$$

Hence

$$\begin{aligned}\tilde{\zeta}\lambda_\pi d\log M - \lambda_E d\log M_E &= \tilde{\zeta}\lambda_\pi d\log\lambda_\pi\\ &\quad - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})\Omega^F d\log P\\ &\quad - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})\Omega^F d\log P\right)\\ &= \left(I_{E\times E} - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\right)\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})\Omega^F d\log P\right)\end{aligned}$$

where we use the fact that

$$(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\right) = d\log M_E, \tag{38}$$

and

$$\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - \lambda_E\Omega^E d\log P\right) = d\log M. \tag{39}$$

Hence, in general we have

$$\begin{aligned}\sum_{k\in N}\sum_{j\in E}\Omega^E_{jk}\Psi^B_{ki}\left(\sum_{m\in N}\tilde{\zeta}_{jm}\lambda_{\pi_m}\left(d\log M_m\right) - \lambda_{E,j}d\log M_{E,j}\right) &= \Big[\left(I_{E\times E} - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\right)\\ &\quad \left(\tilde{\zeta}\lambda_\pi d\log\lambda_\pi - diag(\tilde{\zeta}\lambda_\pi\mathbf{1})\Omega^F d\log P\right)\Big]'\Omega^E\Psi^B.\end{aligned} \tag{40}$$

■

Having defined

$$d\log\hat{P} = \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\Omega^F d\log P, \tag{41}$$

with the aid of the lemma above, if we have non-overlapping entry, we get the simpler expressions

$$\begin{aligned}
d\lambda_i^B = &-\sum_{m\in N}\lambda_m^B\sum_{k\in N}\Omega_{mk}^V\Psi_{ki}^B d\log\left(\mu_m^q\mu_m^y\mu_m^Y\right) + \sum_{j\in E}\sum_{m\in N}\sum_{k\in N}\lambda_m^B\frac{(1-\varepsilon_m)}{\gamma_m}\frac{d\log\left(\mu_m^q\mu_m^y\right)}{\mu_m^q\mu_m^y\mu_m^Y}\tilde{\zeta}_{jm}\Omega_{jk}^E\Psi_{ki}^B\\
&-\sum_{k\in N}\sum_{m\in N}\sum_{j\in E}\lambda_m^B\left(1-(1-\varepsilon_m)\left(\mu_m^q\mu_m^y\right)^{-1}\right)\frac{1}{\mu_m^Y\gamma_m}\tilde{\zeta}_{jm}\Omega_{jk}^E d\log\mu_m^Y\Psi_{ki}^B\\
&+\sum_m\lambda_m^B{\mu_m^Y}^{-1}(1-\theta_m)Cov_{\tilde{\Omega}^{V,m}}\left(d\log P,\Psi_{(i)}^B\right).
\end{aligned}$$

■

Proof of Proposition 5. We start with

$$\begin{aligned}
d\log Y &= \sum_i b_i d\log C_i\\
C_i d\log C_i &= Y_i d\log Y_i - \sum_{j\in N}x_{ji}d\log x_{ji}M_j - \sum_{j\in N}x_{ji}M_j d\log M_j - \sum_{j\in E}x_{E,ji}d\log x_{E,ji}\\
P_iC_i d\log C_i &= P_iY_i d\log Y_i - \sum_{j\in N}P_ix_{ji}M_j d\log x_{ji} - \sum_{j\in N}P_ix_{ji}M_j d\log M_j - \sum_{j\in E}P_ix_{E,ji}d\log x_{E,ji}\\
b_i d\log C_i &= \lambda_i^B d\log Y_i - \sum_{j\in N}\frac{P_ix_{ji}}{PY}M_j\left(d\log x_{ji}+d\log M_j\right) - \sum_{j\in E}\frac{P_ix_{E,ji}}{PY}d\log x_{E,ji}\\
d\log Y &= \sum_{i\in N}\left(\lambda_i^B d\log Y_i - \sum_{j\in N}\frac{P_ix_{ji}}{PY}M_j\left(d\log x_{ji}+d\log M_j\right) - \sum_{j\in E}\frac{P_ix_{E,ji}}{PY}d\log x_{E,ji}\right),\\
&= \sum_{i\in N}\left(\lambda_i^B d\log Y_i - \sum_{j\in N}\frac{P_jx_{ij}}{PY}M_i\left(d\log x_{ij}+d\log M_i\right) - \sum_{j\in E}\frac{P_ix_{E,ji}}{PY}d\log x_{E,ji}\right),
\end{aligned}$$

We also have

$$P_iY_i = \mu_i^Y\mu_i^y\frac{\gamma_i}{1-\varepsilon_i}M_ip_i^qq_i^q = \mu_i^Y\gamma_iM_ip_i^yy_i.$$

Meanwhile, letting mc^q be the marginal cost of producing q,

$$mc_i^q\frac{\partial q_i}{\partial x_{ij}} = P_i,$$

Hence

$$\frac{\partial \log q_i}{\partial \log x_{ij}} = \mu_i^q \frac{P_i x_{ij}}{p_i^q q_i} = \mu_i^Y \mu_i^y \mu_i^q \frac{\gamma_i}{(1-\varepsilon_i)} M_i \frac{P_i x_{ij}}{P_i Y_i} = \mu_i^Y \mu_i^y \mu_i^q \frac{\gamma_i}{(1-\varepsilon_i)} M_i \Omega_{ij}^V, \tag{42}$$

Furthremore,

$$d \log q_i = \sum_j \mu_i^q \frac{p_j x_{ij}}{p_i q_i} d \log x_{ij}. \tag{43}$$

So,

$$\begin{aligned}
d \log Y_i &= \frac{1}{\gamma_i} \left(d \log M_i + (1-\varepsilon_i) d \log q_i \right) \\
&= \frac{1}{\gamma_i} \left(d \log M_i + (1-\varepsilon_i) \sum_j \frac{\partial \log q_i}{\partial \log x_{ij}} d \log x_{ij} \right) \\
&= \frac{1}{\gamma_i} \left(d \log M_i + (1-\varepsilon_i) \sum_j \mu_i^Y \mu_i^y \mu_i^q \frac{\gamma_i}{(1-\varepsilon_i)} M_i \Omega_{ij}^V d \log x_{ij} \right).
\end{aligned}$$

We can write

$$\begin{aligned}
d \log Y &= \sum_{i \in N} \left(\lambda_i^B d \log Y_i - \sum_{j \in N} \frac{P_j x_{ij}}{PY} M_i \left(d \log x_{ij} + d \log M_i \right) - \sum_{j \in E} \frac{P_i x_{E,ji}}{PY} d \log x_{E,ji} \right), \\
&= \sum_{i \in N} \left(\lambda_i^B d \log Y_i - \sum_{j \in N} \lambda_i^B \frac{P_j x_{ij}}{P_i Y_i} M_i \left(d \log x_{ij} + d \log M_i \right) - \sum_{j \in E} \frac{P_i x_{E,ji}}{PY} d \log x_{E,ji} \right), \\
&= \sum_{i \in N} \left(\lambda_i^B d \log Y_i - \lambda_i^B \sum_{j \in N} \Omega_{ij}^V \left(d \log x_{ij} + d \log M_i \right) - \sum_{j \in E} \frac{P_i x_{E,ji}}{PY} d \log x_{E,ji} \right), \\
&= \sum_{i \in N} \left(\lambda_i^B d \log Y_i - \lambda_i^B \frac{1}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} \left[\gamma_i d \log Y_i - d \log M_i + (1-\varepsilon_i) d \log M_i \right] - \sum_{j \in E} \frac{P_i x_{E,ji}}{PY} d \log x_{E,ji} \right), \\
&= \sum_{i \in N} \left(\lambda_i^B d \log Y_i - \lambda_i^B \frac{1}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} \left[\gamma_i d \log Y_i - \varepsilon_i d \log M_i \right] - \sum_{j \in E} \frac{P_i x_{E,ji}}{PY} d \log x_{E,ji} \right), \\
&= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{i \in N} \lambda_i^B \frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} d \log M_i - \sum_{j \in E} \sum_{i \in N} \Omega_{ji}^E \lambda_{E,j} d \log x_{E,ji}, \\
&= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{i \in N} \lambda_i^B \frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} d \log M_i - \sum_{j \in E} \lambda_{E,j} d \log M_{E,j}, \\
&= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{i \in N} \lambda_i^B \frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} \sum_{j \in E} \tilde{\zeta}_{ij} d \log M_{E,j} - \sum_{j \in E} \lambda_{E,j} d \log M_{E,j},
\end{aligned}$$

$$= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{i \in N} \lambda_i^B \frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} \sum_{j \in E} \tilde{\zeta}_{ij} d \log M_{E,j} - \sum_{j \in E} \lambda_{E,j} d \log M_{E,j},$$

Finally, note that

$$\begin{aligned} \lambda_{E,j} &= \sum_i M_i \lambda_i^y \left(1 - \frac{1 - \varepsilon_i}{\mu_i^y \mu_i^q} \right) \tilde{\zeta}_{ij} \\ &= \sum_i \frac{\lambda_i^B}{\mu_i^Y \gamma_i} \left(1 - \frac{1 - \varepsilon_i}{\mu_i^y \mu_i^q} \right) \tilde{\zeta}_{ij} \end{aligned}$$

Hence,

$$\begin{aligned} d \log Y &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) \\ &\quad + \sum_{j \in E} \sum_{i \in N} \lambda_i^B \frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} \tilde{\zeta}_{ij} d \log M_{E,j} - \sum_{j \in E} \sum_{i \in N} \frac{\lambda_i^B}{\mu_i^Y \gamma_i} \left(1 - \frac{1 - \varepsilon_i}{\mu_i^y \mu_i^q} \right) \tilde{\zeta}_{ij} d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{j \in E} \sum_{i \in N} \lambda_i^B \tilde{\zeta}_{ij} \left(\frac{\varepsilon_i}{\mu_i^q \mu_i^Y \mu_i^y \gamma_i} - \frac{1}{\gamma_i \mu_i^Y} \left(1 - \frac{1 - \varepsilon_i}{\mu_i^y \mu_i^q} \right) \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) + \sum_{j \in E} \sum_{i \in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} \left(\frac{\varepsilon_i}{\mu_i^q \mu_i^y} - 1 + \frac{1}{\mu_i^y \mu_i^q} - \frac{\varepsilon_i}{\mu_i^y \mu_i^q} \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) - \sum_{j \in E} \sum_{i \in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} \left(1 - \frac{1}{\mu_i^y \mu_i^q} \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log Y_i \right) - \sum_{j \in E} \sum_{i \in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} \left(1 - \frac{1}{\mu_i^y \mu_i^q} \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) \frac{1}{\gamma_i} \left(d \log y_i + d \log M_i \right) \right) - \sum_{j \in E} \sum_{i \in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} \left(1 - \frac{1}{\mu_i^y \mu_i^q} \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\lambda_i^B \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) \frac{1}{\gamma_i} d \log y_i \right) + \sum_{j \in E} \sum_{i \in N} \left(\lambda_i^B \frac{1}{\gamma_i} \tilde{\zeta}_{ij} \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) - \frac{1}{\mu_i^Y} \left(1 - \frac{1}{\mu_i^y \mu_i^q} \right) \right) d \log M_{E,j}, \\ &= \sum_{i \in N} \left(\frac{\lambda_i^B}{\gamma_i} \left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y} \right) d \log y_i \right) + \sum_{j \in E} \sum_{i \in N} \left(\frac{\lambda_i^B}{\gamma_i} \tilde{\zeta}_{ij} \left(1 - \frac{1}{\mu_i^Y} \right) \right) d \log M_{E,j}. \end{aligned}$$

Diffrentiate this expression a second time with respect to $d \log \mu$and $d \log \mu^Y$and evaluate

it at the efficient point to get

$$d^2 \log Y = \frac{1}{2}\left[\sum_{i\in N} \lambda_i^B d\log Y_i d\log\left(\mu_i^q \mu_i^y \mu_i^Y\right) - \sum_{j\in E}\sum_{i\in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} d\log\left(\mu_i^y \mu_i^q\right) d\log M_{E,j}\right]$$
$$= \frac{1}{2}\left[\sum_{i\in N} \lambda_i^B d\log Y_i d\log\left(\mu_i^q \mu_i^y \mu_i^Y\right) - \sum_{j\in E}\sum_{i\in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} d\log\left(\mu_i^y \mu_i^q\right) d\log M_{E,j}\right]$$

Another way to write this is as

$$2d^2 \log Y = \sum_{i\in N} \lambda_i^B d\log Y_i d\log\left(\mu_i^q \mu_i^y \mu_i^Y\right) - \sum_{j\in E}\sum_{i\in N} \frac{\lambda_i^B \tilde{\zeta}_{ji}}{\mu_i^Y \gamma_i} d\log\left(\mu_i^y \mu_i^q\right) d\log M_{E,j}$$
$$= \sum_{i\in N} \lambda_i^B d\log Y_i d\log \mu_i^Y + \sum_{i\in N} \lambda_i^B d\log\left(\mu_i^y \mu_i^q\right)\left(d\log Y_i - \frac{1}{\mu_i^Y \gamma_i}\sum_{j\in E} \tilde{\zeta}_{ji} d\log M_{E,j}\right)$$
$$= \sum_{i\in N} \lambda_i^B d\log Y_i d\log \mu_i^Y + \sum_{i\in N} \lambda_i^B d\log\left(\mu_i^y \mu_i^q\right)\left(d\log Y_i - \frac{1}{\mu_i^Y \gamma_i} d\log M_i\right)$$
$$= \sum_{i\in N} \lambda_i^B d\log Y_i d\log \mu_i^Y + \sum_{i\in N} \frac{\lambda_i^B}{\gamma_i} d\log\left(\mu_i^y \mu_i^q\right)\left(d\log y_i\right).$$

Where we use the fact that

$$d\log Y_i = \frac{1}{\gamma_i} d\log y_i + \frac{1}{\gamma_i} d\log M_i \tag{44}$$

and

$$d\log M_i = \sum_j \tilde{\zeta}_{ji} d\log M_{E,j}. \tag{45}$$

Yet another representation is

$$d\log Y = \sum_{i\in N}\left(\frac{\lambda_i^B}{\gamma_i}\left(1 - \frac{1}{\mu_i^q \mu_i^Y \mu_i^y}\right) d\log y_i\right) + \sum_{j\in E}\sum_{i\in N}\left(\frac{\lambda_i^B}{\gamma_i}\tilde{\zeta}_{ij}\left(1 - \frac{1}{\mu_i^Y}\right)\right) d\log M_{E,j}.$$
$$2d^2 \log Y = \sum_{i\in N}\left(\frac{\lambda_i^B}{\gamma_i} d\log(\mu_i^q \mu_i^Y \mu_i^y) d\log y_i\right) + \sum_{j\in E}\sum_{i\in N}\left(\frac{\lambda_i^B}{\gamma_i}\tilde{\zeta}_{ij} d\log \mu_i^Y\right) d\log M_{E,j}.$$
$$= \sum_{i\in N} \frac{\lambda_i^B}{\gamma_i}\left(d\log(\mu_i^q \mu_i^Y \mu_i^y) d\log y_i + d\log \mu_i^Y d\log M_i\right)$$

In the IRS case, this is

$$2d^2 \log Y = \sum_{i \in N} \lambda_i^B d \log(\mu_i^q \mu_i^Y \mu_i^y) d \log q_i + \sum_{i \in N} \frac{\lambda_i^B}{\gamma_i} d \log \mu_i^Y d \log M_i \tag{46}$$

■

Proof of Proposition 7. To prove this, for each industry with heterogeneous firms, we construct an isomorphic industry with homogeneous firms which has the same price, quantity and mass of entrants. To do this, consider some industry with heterogeneous firms, where we drop the industry subscript to cut down on notation. The equations that determine the industry's mass of entrants, prices and quantity produced are

$$\begin{aligned}
Y &= \left(\sum_i M_i y_i \right)^{1/\gamma} \\
y_i &= b_i q_i^{1-\varepsilon} \\
p_i^y &= \frac{\mu_i^y}{b_i} \frac{p_i^q}{1-\varepsilon} q_i^{\varepsilon} \\
p_i^q &= \frac{\mu_i^q p^{inputs}}{A_i} \\
P^Y &= \mu^Y \gamma_i p_i^y Y^{1-\frac{1}{\gamma}} \\
M_i &= b_i M, \\
M &= \frac{1}{\gamma} \mu^Y \sum_i \left(1 - \frac{1-\varepsilon}{\mu_i^q \mu_i^y} \right) \lambda_i,
\end{aligned}$$

where b_i are the exogenous taste/productivity shifters for each firm. The comparison industry with homogeneous firms is

$$\begin{aligned}
Y_* &= (M_* y_*)^{\frac{1}{\gamma}} \\
y_* &= q_*^{1-\varepsilon} \\
p_*^y &= \mu_*^y \frac{p_*^q}{1-\varepsilon} q_*^{\varepsilon} \\
p_*^q &= \frac{\mu_*^q p_*^{inputs}}{A_*} \\
P_*^Y &= \mu_*^Y p_*^y Y^{1-1/\gamma} \\
M_* &= \frac{1}{\gamma} \mu_*^Y \left(1 - \frac{1-\varepsilon}{\mu_*^q \mu_*^y} \right) \lambda_*.
\end{aligned}$$

We want to have μ^q, μ^y, μ^Y, A, such that we match the quantity $Y = Y_*$ and the price $P = P^*$ in the two cases. We need also want the mass of entrants to be the same.

$$M_* = M \tag{47}$$

hence

$$\begin{aligned}\mu_*^Y\left(1-\frac{1-\varepsilon}{\mu_*^q\mu_*^y}\right) &= \mu^Y\sum_i\left(1-\frac{1-\varepsilon}{\mu_i^q\mu_i^y}\right)\frac{\lambda_i}{\lambda_*}\\ &= \mu^Y\left(1-(1-\varepsilon)\sum_i\frac{\delta_i}{\mu_i^q\mu_i^y}\right),\end{aligned}$$

where δ_i is firm i's sales shares in the industry. So, set

$$\mu_*^Y = \mu^Y \tag{48}$$

$$\mu_*^q\mu_*^y = \left(\sum_i\frac{\delta_i}{\mu_i^q\mu_i^y}\right)^{-1}. \tag{49}$$

To ensure that

$$P^Y = P_*^Y, \tag{50}$$

we need

$$\mu_*^y\mu_*^q\frac{1}{A_*}q_*^\varepsilon = \frac{\mu_i^y\mu_i^q}{b_i}\frac{1}{A_i}q_i^\varepsilon \tag{51}$$

and we know that

$$P^Y = \mu^Y\frac{\mu_i^y}{b_i}\frac{\mu_i^q p^{inputs}}{(1-\varepsilon)A_i}q_i^\varepsilon\gamma Y^{1-1/\gamma} \tag{52}$$

Hence

$$\left(\frac{b_iP^Y(1-\varepsilon)A_i}{Y^{1-1/\gamma}\gamma\mu^Y\mu_i^y\mu_i^q p^{inputs}}\right) = q_i^\varepsilon \tag{53}$$

Therefore,

$$\begin{aligned}\mu_*^y\mu_*^q\frac{1}{A_*}q_*^\varepsilon &= \mu_i^y\mu_i^q\frac{1}{A_i}\left(\frac{P^Y(1-\varepsilon)A_i}{Y^{1-1/\gamma}\gamma\mu^Y\mu_i^y\mu_i^q p^{inputs}}\right)\\ q_* &= \left(\left(\frac{P^Y(1-\varepsilon)}{Y_*^{1-1/\gamma}\gamma\mu^Y p^{inputs}}\right)\frac{A_*}{\mu_*^y\mu_*^q}\right)^{\frac{1}{\varepsilon}}\end{aligned}$$

But we also must have

$$Y = \left(\sum_i M_i q_i^{1-\varepsilon}\right) = \left(M_* q_*^{1-\varepsilon}\right) = Y_* \tag{54}$$

In other words

$$M_*\left(\frac{P^Y(1-\varepsilon)}{Y_*^{1-1/\gamma}\gamma\mu^Y p^{inputs}}\frac{A_*}{\mu_*^y\mu_*^q}\right)^{\frac{1-\varepsilon}{\varepsilon}} = \left(\sum_i b_i M\left(\frac{P^Y(1-\varepsilon)A_i}{Y^{1-1/\gamma}\gamma\mu^Y\mu_i^y\mu_i^q p^{inputs}}\right)^{\frac{1-\varepsilon}{\varepsilon}}\right) \tag{55}$$

$$\left(\frac{A_*}{\mu_*^y\mu_*^q}\right)^{\frac{1-\varepsilon}{\varepsilon}} = \left(\sum_i b_i\left(\frac{A_i}{\mu_i^y\mu_i^q}\right)^{\frac{1-\varepsilon}{\varepsilon}}\right) \tag{56}$$

or

$$\begin{aligned} A_* &= \mu_*^y\mu_*^q\left(\sum_i b_i\left(\frac{A_i}{\mu_i^y\mu_i^q}\right)^{\frac{1-\varepsilon}{\varepsilon}}\right)^{\frac{\varepsilon}{1-\varepsilon}}. \\ &= \left(\sum_i b_i\left(\frac{A_i}{\mu_i^y\mu_i^q/(\mu_*^y\mu_*^q)}\right)^{\frac{1-\varepsilon}{\varepsilon}}\right)^{\frac{\varepsilon}{1-\varepsilon}}. \end{aligned}$$

Let sectoral productivity be given by A_* and sectoral markups be given by μ_* where recall that $\mu_i^q\mu_i^y = \mu_i$. ■

Proof of Lemma 5. First, we solve out for A as a function of primitives. To that end, note that

$$\lambda_i = \frac{Mp_i^y y_i}{P^Y Y}.$$

Use the fact that

$$p_i^y = \frac{1}{1-\varepsilon}\mu_i^q\mu_i^y\left(\frac{q_i}{b_i}\right)^{\varepsilon} p^{inputs} = P^Y. \tag{57}$$

Hence

$$y_i = b_i^\varepsilon q_i^{1-\varepsilon} = b_i^\varepsilon\left(\frac{P^Y(1-\varepsilon)b_i^\varepsilon}{\mu_i^q\mu_i^y p^{inputs}}\right)^{\frac{1-\varepsilon}{\varepsilon}}. \tag{58}$$

Next note that, firm i's market share δ_i is given by

$$\delta_i = \frac{My_i}{Y} = \frac{Mb_i^\varepsilon\left(\frac{P^Y(1-\varepsilon)b_i^\varepsilon}{\mu_i^q\mu_i^y p^{inputs}}\right)^{\frac{1-\varepsilon}{\varepsilon}}}{\sum_j Mb_j^\varepsilon\left(\frac{P^Y(1-\varepsilon)b_j^\varepsilon}{\mu_j^q\mu_j^y p^{inputs}}\right)^{\frac{1-\varepsilon}{\varepsilon}}}$$

$$= \frac{b_i \mu_i^{\frac{\varepsilon-1}{\varepsilon}}}{\sum_j b_j \mu_j^{\frac{\varepsilon-1}{\varepsilon}}}.$$

Hence, substituting in, we have

$$\begin{aligned}\mu_* &= \left(\sum_i \frac{\delta_i}{\mu_i}\right)^{-1}, \\ &= \left(\sum_i \frac{b_i \mu_i^{-\frac{1}{\varepsilon}}}{\sum_j b_j \mu_j^{\frac{\varepsilon-1}{\varepsilon}}}\right)^{-1}\end{aligned}$$

which means we can write

$$A = \left(\sum_i \frac{b_i \mu_i^{-\frac{1}{\varepsilon}}}{\sum_j b_j \mu_j^{\frac{\varepsilon-1}{\varepsilon}}}\right)^{-1} \left(\sum_i b_i \left(\mu_i\right)^{\frac{\varepsilon-1}{\varepsilon}}\right)^{\frac{\varepsilon}{1-\varepsilon}}. \tag{59}$$

First consider the derivatives of the sectoral productivity shifter

$$\begin{aligned}\log A &= -\frac{1}{\left(\sum_i b_i \left(\mu_i^{-\frac{1}{\varepsilon}}\right)\right)} \left(\sum_i b_i \mu_i^{-\frac{1}{\varepsilon}} (-\frac{1}{\varepsilon}) d\log\mu_i\right) + \frac{\frac{1}{\varepsilon}}{\frac{1}{\varepsilon}-1} \frac{1-\frac{1}{\varepsilon}}{\left(\sum_i b_i \left(\mu_i\right)^{1-\frac{1}{\varepsilon}}\right)} \left(\sum_i b_i \mu_i^{1-\frac{1}{\varepsilon}} d\log\mu_i\right) \\ &= \frac{\frac{1}{\varepsilon}}{\left(\sum_i b_i \left(\mu_i^{-\frac{1}{\varepsilon}}\right)\right)} \left(\sum_i b_i \mu_i^{-\frac{1}{\varepsilon}} d\log\mu_i\right) - \frac{1}{\varepsilon} \frac{1}{\left(\sum_i b_i \left(\mu_i\right)^{1-\frac{1}{\varepsilon}}\right)} \left(\sum_i b_i \mu_i^{1-\frac{1}{\varepsilon}} d\log\mu_i\right) \\ d^2\log A &= -\frac{\frac{1}{\varepsilon}^2}{\left(\sum_i b_i \left(\mu_i^{-\frac{1}{\varepsilon}}\right)\right)} \left(\sum_i b_i \mu_i^{-\frac{1}{\varepsilon}} d\log\mu_i^2\right) + \frac{\frac{1}{\varepsilon}^2}{\left(\sum_i b_i \mu_i^{-\frac{1}{\varepsilon}}\right)} \left(\sum_i b_i \mu_i^{-\frac{1}{\varepsilon}} d\log\mu_i\right)^2 \\ &\quad + \frac{1}{\varepsilon} \frac{(1-\frac{1}{\varepsilon})}{\left(\sum_i b_i \left(\mu_i\right)^{1-\frac{1}{\varepsilon}}\right)} \left(\sum_i b_i \mu_i^{1-\frac{1}{\varepsilon}} d\log\mu_i\right)^2 - \frac{1}{\varepsilon} \frac{(1-\frac{1}{\varepsilon})}{\left(\sum_i b_i \left(\mu_i\right)^{1-\frac{1}{\varepsilon}}\right)} \left(\sum_i b_i \mu_i^{1-\frac{1}{\varepsilon}} d\log\mu_i^2\right) \\ &= -\frac{1}{\varepsilon} \left(\sum_i b_i d\log\mu_i^2\right) + \frac{1}{\varepsilon} \left(\sum_i b_i d\log\mu_i\right)^2.\end{aligned}$$

Obviously, at the efficient point $d\log A = 0$.

Now consider the log-derivative of the sectoral markup

$$\begin{aligned} d\log\mu_* &= -\frac{1}{\mu_*}\left(-\sum_i \frac{1}{\varepsilon}\frac{b_i\mu_i^{-\frac{1}{\varepsilon}}d\log\mu_i}{\sum_j b_j\mu_j^{\frac{\varepsilon-1}{\varepsilon}}} - \frac{\varepsilon-1}{\varepsilon}\frac{\sum_i b_i\mu_i^{-\frac{1}{\varepsilon}}}{\left(\sum_j b_j\mu_j^{\frac{\varepsilon-1}{\varepsilon}}\right)^2}\sum_j b_j d\log\mu_j\right) \\ &= -\left(-\frac{1}{\varepsilon}\sum b_i d\log\mu_i - \frac{\varepsilon-1}{\varepsilon}\sum_j b_j d\log\mu_j\right) \\ &= \sum_j b_j d\log\mu_j. \end{aligned}$$

■

Proof of Proposition 6. We assume there is one primary factor with no incumbents, no input-output in entry costs. For a model with entry in sectors, we can assume away within-industry heterogeneity momentarily. Therefore, we can assume entry is fully directed. We use the deadweight loss triangles formula, along with the fact that for each $i \in \mathcal{N}$

$$\mathrm{d}\log Y_i = \mathrm{d}\log\lambda_i^B - \mathrm{d}\log P_i.$$

So,

$$d\log\lambda_l^B = \sum_k\left(\delta_{lk} - \frac{\lambda_k^B}{\lambda_l^B}\Psi_{kl}^B\right)d\log\mu_k^q - \sum_j\frac{\lambda_j}{\lambda_l}(\theta_j-1)Cov_j(d\log P, \Psi_{(l)}^B), \tag{60}$$

where δ_{lk} is Kronecker's delta, and

$$d\log\lambda_{\pi_i} = d\log\lambda_i^B + \left(\frac{1}{\varepsilon_i}-1\right)d\log\mu_i, \tag{61}$$

$$\begin{aligned} d\log P &= \Psi^F\left(\frac{1-\varepsilon}{\gamma}d\log\mu\right) + \Psi^F\left(\varepsilon\left(d\log\lambda - d\log\hat{\lambda}_\pi\right)\right), \\ &= \Psi^F(1-\varepsilon)d\log\mu - \Psi^F(1-\varepsilon)\,d\log\mu = 0. \end{aligned}$$

Hence

$$d\log\lambda_l^B = \sum_k\left(\delta_{lk} - \frac{\lambda_k^B}{\lambda_l^B}\Psi_{kl}^B\right)d\log\mu_k^q, \tag{62}$$

Furthermore, letting Λ denote labor's share of income

$$d\log M_E = d\log\lambda_\pi - d\log\Lambda,$$

$$= d\log\lambda_i^B + \left(\frac{1}{\varepsilon_i} - 1\right) d\log\mu_i$$

$$\sum_l \lambda_l^B \left(d\log\lambda_l^B - d\log p_l\right) d\log\mu_l = \sum_l \sum_k \left(\lambda_l^B \delta_{lk} - \lambda_k^B \Psi_{kl}^B\right) d\log\mu_k d\log\mu_l$$

Next

$$\sum_{j\in E}\sum_{i\in N} \frac{\lambda_i^B \tilde{\zeta}_{ij}}{\mu_i^Y \gamma_i} d\log\left(\mu_i^y \mu_i^q\right) d\log M_{E,j} = \sum_i \lambda_i^B d\log\mu_i d\log\lambda_i^B + \sum_i \lambda_i^B \left(\frac{1}{\varepsilon_i} - 1\right) d\log\mu_i d\log\mu_i$$

Combining everything gives

$$\mathcal{L} = \sum_l \sum_k \left(\lambda_l^B \delta_{lk} - \lambda_k^B \Psi_{kl}^B\right) d\log\mu_k d\log\mu_l - \sum_i \lambda_i d\log\mu_i d\log\lambda_i^B - \sum_i \lambda_i^B \left(\frac{1}{\varepsilon_i} - 1\right) d\log\mu_i d\log\mu_i.$$

$$d\log\lambda_i^B = -\frac{1}{\lambda_i^B} \sum_j \left(\lambda_j^B \Psi_{ji}^B - \lambda_j^B \delta_{ij}\right) d\log\mu_j, \tag{63}$$

Or

$$\begin{aligned}
\mathcal{L} &= \sum_k \lambda_k^B \sum_l \left(\delta_{lk} - \Psi_{kl}^B\right) d\log\mu_k d\log\mu_l + \sum_j \lambda_j^B \sum_i \left(\Psi_{ji}^B - \delta_{ij}\right) d\log\mu_j d\log\mu_i \\
&\quad - \sum_i \lambda_i^B \left(\frac{1}{\varepsilon_i} - 1\right) d\log\mu_i d\log\mu_i, \\
&= -\sum_i \lambda_i^B \left(\frac{1}{\varepsilon_i} - 1\right) d\log\mu_i d\log\mu_i.
\end{aligned}$$

This is the loss function for a model with homogeneous sectors.

To extend this into a sectoral model with within-sector heterogeneity, consider the isomorphic sectoral model. We know that

$$d\log Y = \frac{d\log Y}{d\log A} d\log A + \frac{d\log Y}{d\log\mu} d\log\mu \tag{64}$$

$$\frac{1}{2} d^2\log Y = \frac{1}{2} d\log A' \frac{d^2\log Y}{d\log A^2} d\log A + \frac{d\log Y}{d\log A} d^2\log A + \frac{1}{2} d\log\mu' \frac{d^2\log Y}{d\log\mu^2} d\log\mu + \frac{d\log Y}{d\log\mu} d^2\log\mu \tag{65}$$

At the efficient point, $d\log A = 0$ and $d\log Y/d\log\mu = 0$,

$$-\mathcal{L} = \frac{1}{2}\frac{d\log Y}{d\log A}d^2\log A + \frac{1}{2}d\log\mu'\frac{d^2\log Y}{d\log\mu^2}d\log\mu$$

where, from the proof of the previous proposition, we know that

$$d^2\log A_k = -\frac{1}{2}\frac{1}{\varepsilon}Var_{\delta_k}\left(d\log\mu_{(k)}\right). \tag{66}$$

Finally, recall note that at the efficient point, from Hulten's theorem, $d\log Y/d\log A = \lambda^B(1-\varepsilon)$, so we get

$$d^2\log Y = -\frac{1}{2}\sum_I \lambda_I^B\left(\frac{1}{\varepsilon_I}-1\right)Var_{\delta_I}\left(d\log\mu_{(I)}\right) - \frac{1}{2}\sum_I \lambda_I^B\left(\frac{1}{\varepsilon_I}-1\right)E_{\delta_I}\left(d\log\mu_{(I)}\right)^2$$

■

Appendix B Mapping Model to Data

Our calibrated model is sectoral in the formal sense defined in the paper. Our calibration is very similar to Baqaee and Farhi (2019a), and we borrow much of the following discussion from the Appendix of that paper.

We have two principal datasources: (i) aggregate data from the BEA, including the input-output tables and the national income and product accounts; (ii) firm-level data from Compustat. Below we describe how we treat the input-output data, merge it with firm-level estimates of markups, and how we estimate markups at the firm-level.

B.1 Input-Output and Aggregate Data

Our input-output data comes from the BEA's annual input-output tables. We calibrate the data to the use tables from 1997-2015 before redefinitions. We also ignore the distinction between commodities and industries, assuming that each industry produces one commodity. For each year, this gives us the backward expenditure share matrix Ω^B at the industry level. We drop the government, scrap, and noncomparable imports sectors from our dataset, leaving us with 66 industries. We define the gross-operating surplus of each industry to be the residual from sales minus intermediate input costs and compensation of employees. The expenditures on capital, at the industry level, are equal to the gross operating surplus minus the share of profits (how we calculate the profit share is described

shortly). If this number is negative, we set it equal to zero. If any value in Ω^B is negative, we set it to zero.

In Appendix C, we supplement the markup estimates that are used in the main text with three other estimates of markups. For each markup series, we compute the profit share (amongst Compustat firms) for each industry and year, and then we use that profit share to separate payments to capital from gross operating surplus in the BEA data for that industry and year. Conditional on the harmonic average of markups in each industry-year, we can recover the forward matrix $\Omega^F = \mu\Omega$, also at the industry level. If for an industry and year we do not observe any Compustat firms, then we assume that the profit share (and the average markup) of that industry is equal to the aggregate profit share (and the industry-level markup is the same as the aggregate markup).

We assume that the economy has an sectoral structure along the lines of Section 8, so that all producers in each industry have the same production function up to a Hicks-neutral productivity shifter. This means that for each producer i and j in the same industry $\Omega^F_{ik} = \Omega^F_{jk}$. To populate each industry with individual firms, we divide the sales of each industry across the firms in Compustat according to the sales share of these firms in Compustat. In other words, if some firm i's markup is μ_i and share of industry sales in Compustat is x, then we assume that the mass of firms in that industry whose markups are equal to μ_i is also equal to x. These assumptions allow us to use the markup data and market share information from Compustat, and the industry-level IO matrix from the BEA, to construct the firm-level cost-based IO matrix.

B.2 Estimates of Markups

Now, we briefly describe how our firm-level markup data is constructed. Firm-level data is from Compustat, which includes all public firms in the U.S. The database covers 1950 to 2016, but we restrict ourselves to post-1997 data since that is the start of the annual BEA data. We exclude firm-year observations with assets less than 10 million, with negative book or market value, or with missing year, assets, or book liabilities. We exclude firms with BEA code 999 because there is no BEA depreciation available for them; and Financials (SIC codes 6000-6999 or NAICS3 codes 520-525). Firms are mapped to BEA industry segments using 'Level 3' NAICS codes, according to the correspondence tables provided by the BEA. When NAICS codes are not available, firms are mapped to the most common NAICS category among those firms that share the same SIC code and have NAICS codes available.

B.2.1 Production Function Estimation Approach

This is our benchmark method for estimating markups, and the results in the main body of the paper use this approach. For reference, we will call this the production function estimation (or PF) markups.

For the production function estimation approach markups, we follow the procedure PF1 described by De Loecker et al. (2019) with some minor differences. We estimate the production function using Olley and Pakes (1996) (OP) rather than Levinsohn and Petrin (2003). We use CAPX as the instrument and COGS as a variable input. We use the classification based on SIC numbers instead of NAICS numbers since they are available for a larger fraction of the sample. Finally, we exclude firms with COGS-to-sales and XSGA-to-sales ratios in the top and bottom 2.5% of the corresponding year-specific distributions. As with the other series, we use Compustat excluding all firms that did not report SIC or NAICS indicators, and all firms with missing sales or COGS. Sales and COGS are deflated using the gross output price indices from KLEMS sector-level data. CAPX and PPEGT – using the capital price indices from the same source. Industry classification used in the estimation is based on the 2-digit codes whenever possible, and 1-digit codes if there are fewer than 500 observations for each industry and year.

To compute the PF Markups, we need to estimate elasticity of output with respect to variable inputs. This is because once we know the output-elasticity with respect to a variable input (in this case, the cost of goods sold or COGS), then following **?**, the markup is

$$\mu_i = \frac{\partial \log F_i / \partial \log COGS_i}{\Omega_{i,COGS}},$$

where $\Omega_{i,COGS}$ is the firm's expenditures on COGS relative to its turnover.

The output-elasticities are estimated using Olley and Pakes (1996) methodology with the correction advocated by Ackerberg et al. (2015) (ACF). To implement Olley-Pakes in Stata, we use the *prodest* Stata package. OP estimation requires:

(i) outcome variable: log sales,

(ii) "free" variable (variable inputs): log COGS,

(iii) "state" variable: log capital stock, measured as log PPEGT in the Compustat data,

(iv) "proxy" variable, used as an instrument for productivity: log investment, measured as log CAPX in Compustat data.

(v) in addition, SIC 3-digit and SIC 4-digit firm sales shares were used to control for markups .

Given these data, we run the estimation procedure for every sector and every year. Since panel data are required, we use 3-year rolling windows so that the elasticity estimates based on data in years $t-1$, t and $t+1$ are assigned to year t. The estimation procedure has two stages: in the first stage, log sales are regressed on the 3-rd degree polynomial of state, free, proxy and control variables in order to remove the measurement error and unanticipated shocks; in the second stage, we estimate elasticities of output with respect to variable inputs and the state variable by fitting an AR(1) process for productivity to the data (via GMM). Just like in De Loecker et al. (2019), we control for markups using a linear function of firm sales shares (sales share at the 4-digit industry level).

In our benchmark estimates, we treat SG&A as a fixed cost. However, for robustness, following De Loecker et al. (2019), we also compute markups using an approach where SG&A is treated as a variable input in production. We call these the PF2 markups. The overall estimation is still done via the ACF-corrected OP method (with CAPX as a proxy).

Finally, before feeding these markup estimates into the structural model, we winsorize the markups at the 20th and 80th percentile to reduce the influence of outliers.

B.2.2 User Cost Approach

Our second approach to measuring markups is the user-cost approach (UC) markups. The idea here is to recover the profits of a firm by subtracting total costs from revenues. To compute total cost, we must measure the cost of capital. For this measure, we rely on the replication files from Gutiérrez and Philippon (2016) provided German Gutierrez. For more information see Gutiérrez and Philippon (2016). To recover markups, we assume that operating surplus of each firm is equal to payments to both capital as well as economic rents due to markups. We write

$$OS_{i,t} = r_{k_i,t} K_{i,t} + \left(1 - \frac{1}{\mu_i}\right) sales_{i,t},$$

where $OS_{i,t}$ is the operating income of the firm after depreciation and minus income taxes, $r_{k_i,t}$ is the user-cost of capital and $K_{i,t}$ is the quantity of capital used by firm i in industry j in period t. This equation uses the fact that each firm has constant-returns to scale. In other words,

$$\frac{OS_{i,t}}{K_{i,t}} = r_{k_i,t} + \left(1 - \frac{1}{\mu_i}\right) \frac{sales_{i,t}}{K_{i,t}}, \tag{67}$$

To solve for the markup, we need to account for both the user cost (rental rate) of capital as well as the quantity of capital. The user-cost of capital is given by

$$r_{k_i,t} = r_t^s + KRP_j - (1 - \delta_{k_i,t})E(\Pi_{t+1}^k),$$

where r_t^s is the risk-free real rate, KPR_j is the industry-level capital risk premium, δ_j is the industry-level BEA depreciation rate, and $E(\Pi_{t+1}^k)$ is the expected growth in the relative price of capital. We assume that expected quantities are equal to the realized ones. To calculate the user-cost, the risk-free real rate is the yield on 10-year TIPS starting in 2003. Prior to 2003, we use the average spread between nominal and TIPS bonds to deduce the real rate from nominal bonds prior to 2003. *KRP* is computed using industry-level equity risk premia following Claus and Thomas (2001) using analyst forecasts of earnings from IBES and using current book value and the average industry payout ratio to forecast future book value. The depreciation rate is taken from BEA's industry-level depreciation rates. The capital gains $E(\Pi_{t+1}^k)$ is equal to the growth in the relative price of capital computed from the industry-specific investment price index relative to the PCE deflator. Finally, we use net property, plant, and equipment as the measure of the capital stock. This allows us to solve equation (67) for a time-varying firm-level measure of the markup. We winsorize markups at the 5-95th percentile by year.

B.2.3 Accounting Profits Approach

The final approach to estimating markups is the accounting profits approach (AC). For the accounting-profit approach markups, we use operating income before depreciation, minus depreciation to arrive at accounting profits. Our measure of depreciation is the industry-level depreciation rate from the BEA's investment series. The BEA depreciation rates are better than the Compustat depreciation measures since accounting rules and tax incentives incentivize firms to depreciate assets too quickly. We use the expression

$$profits_i = \left(1 - \frac{1}{\mu_i}\right) sales_i,$$

to back out the markups for each firm in each year. We winsorize markups and changes in markups at the 5-95th percentile by year. Intuitively, this is equivalent to assuming that the cost of capital is simply the depreciation rate (equivalently, the risk-adjusted rate of return on capital is zero). The advantage of this approach is its simplicity.

Appendix C Additional Quantitative Results

IRS, $1-\varepsilon=0.875$	No Entry	Entry uses Factors	Entry uses Goods and Factors
PF2 Markups	12%	39%	47%
UC Markups	3.0%	23%	34%
AC Markups	4.5%	54%	75%
IRS, $1-\varepsilon=0.75$	No Entry	Entry uses Factors	Entry uses Goods and Factors
PF2 Markups	24%	32%	31%
UC Markups	7.2%	15%	17%
AC Markups	11%	14%	14%
DRS, $1-\varepsilon=0.875$			
PF2 Markups	19%	25%	25%
UC Markups	6.0%	11%	11%
AC Markups	8.2%	13%	12%
DRS, $1-\varepsilon=0.75$			
PF2 Markups	9.0%	28%	29%
UC Markups	4.8%	40%	43%
AC Markups	2.6%	18%	19%

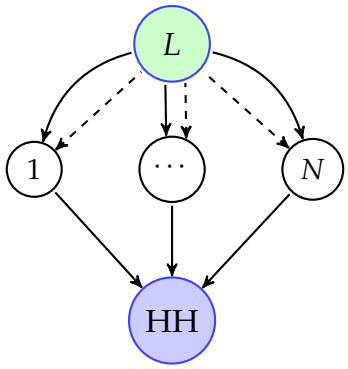
Table 3: The gains from moving to the efficient allocation. The IRS specification sets $\gamma_{\mathcal{I}} = 1-\varepsilon_{\mathcal{I}} = 0.875$ and uses an imperfect-substitutes interpretation. The DRS specification sets $\gamma_{\mathcal{I}} = 1$, $1-\varepsilon_{\mathcal{I}} = 0.875$ and uses a perfect-substitutes interpretation.

Appendix D Additional Examples

This section contains a detailed analysis of the "horizontal" and "vertical" economy.

D.1 Horizontal Economy

Next, to see the importance of directed versus undirected entry, consider the economy depicted in Figure 4. For simplicity, suppose entry costs are paid in units of labor. Since this economy is more complex, for brevity, we consider only productivity shocks to the producers 1 through to N (we do not shock labor). For the economy in example 4, it is easy to verify using the definition that the forward and backward Domar weight coincide $\lambda_i^B = \lambda_i^F = \lambda_i^Y$ (even though the economy may be inefficient). Accordingly, for this example, we drop the superscripts and write λ_i.



$$Y_i = (M_i y_i)^{\frac{1}{\gamma_i}}, \qquad y_i = l_i^{1-\varepsilon_i},$$
$$M_{E,j} = \sum_{i=1}^{N} \zeta_{ji} \frac{\lambda_{\pi_i}}{w_L}, \quad L = \sum_{i=1}^{N} l_i + \sum_{j=1}^{E} M_{E,j},$$
$$M_i = \sum_{j=1}^{N} \zeta_{ji} M_{E,j}, \quad Y = \left(\sum_{i=1}^{N} \bar{\lambda}_i Y_i^{\frac{\theta_0 - 1}{\theta_0}} \right)^{\frac{\theta_0}{\theta_0 - 1}}.$$

Figure 4: Horizontal Economy. The solid and dashed arrows represent the flow of resources for production and for entry. The sole factor for this economy is indexed by L. The equations assume free-entry, for the no-entry case, treat M_i as exogenous.

DRS with Directed Entry. In this case $E = N$, ζ is the identity matrix, and $1 - \varepsilon_i < \gamma_i = 1$. Theorem 3 takes a very simple form

$$\mathrm{d} \log Y = \mathbb{E}_\lambda (\mathrm{d} \log A),$$

where $\mathbb{E}_\lambda$ is the expectation operator with respect to the sales shares λ, using the fact that $\lambda_i^F = \lambda_i^B = \lambda_i^Y$. In this economy, a productivity shock to i could increase or decrease the sales of industry i (depending on the elasticity of substitution across industries). The change in the size of industry i will change the pattern of entry, as entrants enter into the industries that expand and leave the industries that shrink. In equilibrium, no individual producer changes their scale of operation, and all the adjustment in industry size comes through the extensive margin. Therefore, in equilibrium, the only reason prices changes is because of the changes in productivity (i.e. the prices of producer-specific quasi-fixed factors do not change).

Superficially, it looks like this example satisfies Hulten's theorem, since the elasticity of output with respect to each productivity shock is given by the sales share. Surprisingly, the initial value of markups μ are not relevant in this case! Although the allocation matrix clearly does change in this example, the reallocation happens purely on the entry margin, and resources along the intensive margin are not reallocated. Proposition 2 shows that the reallocations caused by technology shocks in this example are neutral (much like they would be if the initial allocation had been efficient). Using the notation of Section 5.2, while reallocations do happen in this economy $\mathrm{d}\mathcal{X} / \mathrm{d} \log A \neq 0$, these reallocations do not affect allocative efficiency $\mathrm{d} \log Y / \mathrm{d}\mathcal{X} \, \mathrm{d}\mathcal{X} = 0$. In fact, as example 3 shows, this property always holds as long as (1) entry is directed and (2) there is only one incumbent industry (i.e. one primary factor).

DRS with Undirected Entry. Since entry is undirected, ζ is $1 \times |\mathcal{N}|$ and entrants are randomly assigned to different products according to the existing market share of those products $\zeta_{1i} = \lambda_i$. In this case, in response to productivity shocks $\mathrm{d}\log A$, we have

$$\mathrm{d}\log Y = \mathbb{E}_\lambda(\mathrm{d}\log A) - \sum_j \varepsilon_j \lambda_j \left(\mathrm{d}\log \lambda_{\pi_j} - \mathrm{d}\log \hat{\lambda}_{\pi_j}\right),$$

where we again use $\lambda_i^F = \lambda_i^B = \lambda_i^Y$. The first term is the direct effect of the productivity shock on consumer prices, holding fixed the price of fixed factors, and the second term is the change in the price of those factors. Equivalently, the first term is the direct effect of the productivity shock, holding fixed the allocation matrix, and the second term is the change in the allocation matrix.

After some algebra, we can break the reallocation effect into two terms

$$\begin{aligned}\mathrm{d}\log Y - \mathbb{E}_\lambda(\mathrm{d}\log A) = &\left[\mathbb{E}_\lambda(\varepsilon) - \mathbb{E}_\lambda\left(1 - \frac{1-\varepsilon}{\mu}\right)\right] \frac{Cov_\lambda\left(1 - \frac{1-\varepsilon}{\mu}, d\log\lambda\right)}{\mathbb{E}_\lambda\left(1 - \frac{1-\varepsilon}{\mu}\right)} \\ &+ \left[Cov_\lambda\left(1 - \frac{1-\varepsilon}{\mu}, d\log\lambda\right) - Cov_\lambda(\varepsilon, d\log\lambda)\right]. \end{aligned} \tag{68}$$

Consider the first term and note that

$$\mathrm{d}\log M = d\log\hat{\lambda}_\pi = \frac{Cov_\lambda\left(1 - \frac{1-\varepsilon}{\mu}, d\log\lambda\right)}{\mathbb{E}_\lambda\left(1 - \frac{1-\varepsilon}{\mu}\right)}.$$

In this example, producers are enjoying both Ricardian and monopolistic rents, and the rent share of sales is $1 - (1-\varepsilon_i)/\mu_i$. Therefore, when $Cov_\lambda\left(1 - \frac{1-\varepsilon}{\mu}, d\log\lambda\right) > 0$, this means relatively more profitable industries are expanding, and this stimulates entry. If we were to fix the fraction of production labor working in each i constant, more entry would be beneficial if, and only if, the total rent share is lower than the Ricardian rent share:

$$\mathbb{E}_\lambda(\varepsilon) > \mathbb{E}_\lambda\left(1 - \frac{1-\varepsilon}{\mu}\right).$$

This is because the optimal amount of entry occurs whenever total rents equal Ricardian rents (i.e. there are no monopolistic rents in aggregate). If the condition above holds, then we have too little entry in equilibrium, and so an increase in entry, holding all else equal, is beneficial. Putting these two observations together, the first line of (68) is then the effect of the change in entry, holding fixed the share of production labor working in each i.

The second line of (68) accounts for all else not being equal. Because entry is untargeted, the share of workers across firms changes. This redistribution is beneficial when expansion covaries more with total (Ricardian and monopolistic) rents than with Ricardian profits, since in this case firms with relatively higher markups are scaling up. This is beneficial because a relatively higher markup implies those firms had relatively too few workers allocated to them.

Using the forward and backward propagation equations, we can solve this out in terms of primitives

$$\mathrm{d}\log Y = \mathbb{E}_\lambda\left(d\log A\right) + \mathbb{E}_\lambda\left(\varepsilon\right)\left[\frac{Cov_\lambda\left(1-\frac{1-\varepsilon}{\mu}, \frac{1}{\varepsilon}d\log A\right)}{\mathbb{E}_\lambda\left(1-\frac{1-\varepsilon}{\mu}\right)} - \frac{Cov_\lambda(\varepsilon, \frac{1}{\varepsilon}d\log A)}{\mathbb{E}_\lambda\left(\varepsilon\right)}\right].$$

In the case where every firm has the same returns to scale $\varepsilon_i = \varepsilon$ for every $i \in \mathcal{N}$, the expression simplifies further to just

$$\mathrm{d}\log Y = \mathbb{E}_\lambda(\mathrm{d}\log A) + Cov_\lambda\left(\frac{1-\frac{1-\varepsilon}{\mu}}{1-\frac{1-\varepsilon}{\bar{\mu}}}, \mathrm{d}\log A\right),$$

where $\bar{\mu} = \mathbb{E}_\lambda(\mu^{-1})^{-1}$ is the harmonic average markup. In this case, since all producers have the same returns to scale, a change in sales shares $\mathrm{d}\log\lambda$ can only stimulate entry if the high markup firms are expanding, and since entry is undirected, this means that the high markup firms are expanding their scale, which is beneficial.

IRS with Directed Entry. Now, suppose that $1-\varepsilon_i = \gamma_i \in (0,1)$, and adopt the product differentiation interpretation. In other words, each i is produced using a CES aggregator with elasticity of substitution $1/(1-\gamma_i)$. Assume that love-of-variety within each industry is weaker than love-of-variety across industries, or $1/(1-\gamma_i) \geq \theta_0$ for every i; otherwise, output may be non-differentiable.[33]

Applying Theorem 3, we get

$$d\log Y = \mathbb{E}_\lambda(\mathrm{d}\log A) + Cov_\lambda\left(\frac{1}{\gamma}, d\log\lambda\right).$$

The first term is the usual change in prices along the intensive margin, for a fixed mass of entrants. The second term is how the change in the market share of different product types affects profits and hence, entry. If i expands at the expense of other industries, then

[33]See Ciccone and Matsuyama (1996) for an explanation of why output becomes non-differentiable in this case.

this stimulates further entry into i, and this can increase or reduce output depending on whether i has greater or less external economies than the rest of the economy.[34] Unlike the case with DRS, in this case, the elasticities of substitution are very important in determining the response of output.

Define the vector $\theta = 1/(1-\gamma)$, where θ can be interpreted as the within industry elasticity of substitution (across different producers in each $i \in \mathcal{N}$) and θ_0 is the cross-industry elasticity of substitution. Using the backward and forward propagation equations we can write the response of output in terms of primitives

$$\mathrm{d}\log Y = \mathbb{E}_\lambda\left(d\log A\right) + \frac{Cov_\lambda\left(\frac{\theta-1}{\theta-\theta_0}, d\log A\right)}{\mathbb{E}_\lambda\left(\frac{\theta-1}{\theta-\theta_0}\right)}.$$

So, if the cross-industry elasticity of substitution $\theta_0 > 1$, and the shocks negatively covary with θ_i, then we have beneficial reallocation. Intuitively, when the shocks negatively covary with θ_i, then sectors with stronger scale effects are receiving more positive shocks, this causes more entry in those sectors as long as different sectors are substitutes, which allows the effects of entry to reinforce itself. If sectors are complements, then these forces operate in reverse.

Interestingly, the initial value of markups μ are not relevant to the comparative statics. As with the DRS directed entry example, this stems from the fact that the reallocation occurs purely on the entry margin, and resources along the intensive margin are not reallocated. This means we do not have to compare the marginal benefit of reallocating resources across different i (which would necessitate comparing relative markups).

IRS with Undirected Entry. Use the same elasticities of before, but now suppose that ζ is $1 \times |\mathcal{N}|$ with $\zeta_{1i} = \lambda_i$. Theorem 3 simplifies to give

$$\mathrm{d}\log Y = \mathbb{E}_\lambda\left(d\log A\right) + \mathbb{E}_\lambda\left(\frac{1}{\theta-1}\right) Cov_\lambda\left(\frac{1-\frac{1}{\mu}}{1-\frac{1}{\tilde{\mu}}}, d\log\lambda\right)$$

The first term is just the usual change in prices along the intensive margin, for a fixed mass of entrants. The second term considers how the change in the market share of

[34] Whether or not i expands in equilibrium, in turn, depends on the strength of returns to scale in i, since as more entry occurs the price effects of the initial entry are reinforced. Using Proposition 4,

$$d\log\lambda_i = (1-\theta_0)\frac{\theta_i - 1}{\theta - \theta_0}\left(-d\log A_i + d\log Y\right),$$

where $\theta_i = 1/(1-\gamma_i)$.

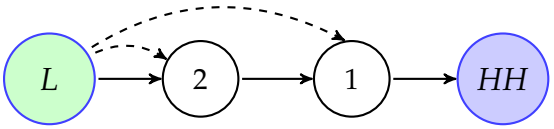
different product types affects profits and hence, entry. If i has relatively higher margins than the rest of the economy, then an increase in the sales of i will stimulate additional entry, but because entry is undirected, the benefits of this entry depend on the average of external economies across all products.[35] For brevity, suppose that θ_i is the same for all i, then we can write

$$\mathrm{d}\log Y = \mathbb{E}_\lambda\left(d\log A\right) + \mathbb{E}_\lambda\left(\frac{\theta_0 - 1}{\theta - 1}\right) Cov_\lambda\left(\frac{1-\frac{1}{\mu}}{1-\frac{1}{\bar{\mu}}}, \mathrm{d}\log A\right).$$

In words, high-markup sectors expand if productivity shocks covary positively with the Lerner index $1 - 1/\mu$, and if the elasticity of substitution across industries $\theta_0 > 1$, so that sectors receiving positive shocks expand. Compared to the previous example, reallocations do occur on the intensive margin of production, and so the level of markups is relevant for determining the welfare consequences of reallocation.

D.2 Vertical Economy

The final example we consider is the supply-chain in Figure 5 where entry costs are paid in units of labor but there are input-output linkages in production. Once again, consider the DRS model $\gamma = 1$ (with the perfect substitutes interpretation) and the IRS model $\gamma = 1 - \varepsilon$ (with the imperfect substitutes interpretation).



$$Y_i = (M_i y_i)^{\frac{1}{\gamma_i}}, \qquad y_2 = l_2^{1-\varepsilon_2},$$
$$y_1 = (Y_2/M_1)^{1-\varepsilon_1}, \qquad M_i = \frac{\lambda_{\pi_i}}{w_L},$$
$$L = M_2 l_2 + M_1 + M_2, \qquad Y = Y_1.$$

Figure 5: Vertical Economy. The solid and dashed arrows represent the flow of resources for production and for entry, respectively. The sole factor for this economy is indexed by L. The equations assume free-entry, for the no-entry case, M_i are exogenous.

[35] Whether or not i expands in equilibrium, in turn, depends on the strength of returns to scale in i relative to the other sectors; as more entry occurs, all product types experience more entry, and relative prices diverge due to differences in external economies of scale. Consider a univariate shock to i. Using Proposition 4, whether this shock results in substitution towards or away from i depends on the elasticity of substitution across products θ_0:

$$d\log\lambda_i = (1-\theta_0)\left(-\frac{1}{\theta_i - 1}\frac{1}{1-\bar{\mu}} Cov_\lambda(1-\frac{1}{\mu}, d\log\lambda) - d\log A_i + d\log Y\right).$$

No Entry. Without entry, this model's behavior is trivial. Since there is only one feasible allocation of resources, the equilibrium is efficient regardless of the value of μ^Y and μ. In other words, the allocatin matrix does not change $\mathrm{d}\mathcal{X} = 0$. Hence, Corollary 2 implies that

$$\mathrm{d}\log Y = \lambda_1^F \,\mathrm{d}\log A_1 + \lambda_2^F \,\mathrm{d}\log A_2 + 0\,\mathrm{d}\log\mu_1 + 0\,\mathrm{d}\log\mu_2.$$

The model's behavior only becomes interesting once we all for the possibility of free entry.

DRS with Directed Entry. In this case, Theorem 3 implies that

$$\begin{aligned} d\log Y = d\log A_1 + \mu_1\lambda_2 d\log A_2 - \frac{(1-\varepsilon_1)/\mu_1}{1-(1-\varepsilon_1)/\mu_1}(\mu_1 - 1)\, d\log\mu_1 \\ - \mu_1\lambda_2\frac{(1-\varepsilon_2)/\mu_2}{1-(1-\varepsilon_2)/\mu_2}(\mu_2 - 1)\, d\log\mu_2. \end{aligned}$$

Whether an increase in markups raises or reduces output depends on whether or not firms in each product market were above or below their efficient scale (i.e. μ_i is greater than or less than 1).

IRS with Directed Entry. In this case, Theorem 3 becomes

$$\begin{aligned} d\log Y = d\log A_1 + \mu_1\lambda_2 d\log A_2 \\ + \left(\frac{1}{(\varepsilon_1 - 1)(\mu_1 - 1)} - \frac{1}{(\varepsilon_2 - 1)} - 1\right) d\log\mu_1 + \mu_1\lambda_2\left(\frac{1}{(\varepsilon_2 - 1)(\mu_2 - 1)} - 1\right) d\log\mu_2. \end{aligned}$$

Unlike the DRS case, the scale elasticities interact with one another. First, consider the upstream markups $\mathrm{d}\log\mu_2$. When $\mu_2 = 1/(1-\varepsilon_2)$, a change in upstream markups have no effect on welfare. Intuitively, at this level of markups, $\mathrm{d}\log\mu_2$ does not affect the price of industry 2 — an increase in markups raises the individual price but, by stimulating entry, increases product variety. When $\mu_2 = 1/(1-\varepsilon_2)$, these two effects exactly cancel (this is exactly the markup associated with a Dixit-Stiglitz market structure).

Furthermore, regardless of the value of μ_2, a change in μ_2 has no effect on production and entry downstream, since downstream profitability is not a function of upstream prices. This is because profits generated downstream are $\lambda_{\pi_1} = (1 - 1/\mu_1)$ and do not depend on μ_2. So, when $\mu_2 = 1/(1-\varepsilon_2)$, a change in μ_2 has no effect on upstream or downstream prices.

Markups downstream μ_1 are a very different story. If $\mu_1 = 1/(1-\varepsilon_1)$, then an increase $\mathrm{d}\log\mu_1$ reduces output by $-1/(\varepsilon_2 - 1)$. Intuitively, this is because a higher μ_1 reduces

profitability upstream. The stronger are external economies upstream (the closer is ε_2 to one), the more costly are increases in markups downstream. An increase in downstream markups $\mathrm{d}\log\mu_1$ increases entry downstream at the expense of entry upstream. Contrast this with the DRS example where the scale elasticities upstream were irrelevant.

This example can rationalize why upstream or "linkage" industries, like semiconductors, which appear in supply chains of many other industries, are commonly touted as prime candidates for industrial policy. Upstream industries are likely to be those industries which are double-marginalized most intensely, and therefore, even if all industries have the same external economies of scale, double-marginalization implies that industries that are relatively upstream experience too little entry in the decentralized equilibrium. Therefore, subsidizing their entry can improve welfare. These questions, about the conduct of and gains from industrial policy lead naturally to the next section of the paper.

D.2.1

Baqaee and Farhi (2019a) show that in one-factor models without entry, the loss function is a linear combination of elasticities of substitution — each elasticity is weighted by some sufficient statistic that depends on forward and backward input-output matrix. This is no longer true once we allow free-entry.

For example, consider the simple supply chain depicted in Figure 5. In this economy, without free entry, the losses from markups are equal to zero since there is only one feasible allocation of resources, and wedges do not distort any decisions.

However, if we allow for free entry into each industry, the loss function is given by

$$\mathcal{L} = \frac{1}{2}\frac{\theta_1^2\theta_2}{\theta_1\theta_2 - 1}\left(d\log\mu_1\right)^2 + \frac{1}{2}\theta_2\left(d\log\mu_2\right)^2.$$

The losses from an increase in downstream markups now depend on the elasticity of substitution upstream. Intuitively, an increase in markups downstream deprives upstream producers from sales — this lowers entry upstream. In other words, the markup downstream distorts the entry margin upstream. In the case where upstream entry is irrelevant $\theta_2 \to \infty$, the losses in downstream markups simplify to the usual $1/2\theta_1(d\log\mu_1)^2$, which is what we had in the one-sector example.

Appendix E Additional Examples for Theorem 3

To build more intuition, consider the CRS, DRS, and IRS special cases of Theorem 3, limiting attention to univariate perturbations $\mathrm{d}\log A_i$ and $\mathrm{d}\log \mu_i$. Assume $\mu_i^Y = 1/\gamma_i$ throughout, consistent with both the perfect-substitutes and imperfect-substitutes interpretation.[36]

Example 1 (CRS without Entry)**.** No entry means that $\zeta = 0$, and since there is no entry, constant-returns can be achieved by $\gamma_i = (1 - \varepsilon_i)$. Hence, Theorem 3 reduces to

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \mathrm{d}\log \mu\right) - \sum_{f \in \mathcal{F}} \lambda_f^F \, \mathrm{d}\log \lambda_f^B,$$

where $\mathcal{F}$ is the set of primary factors. Recall that a primary factor is simply an incumbent with zero returns to scale $\varepsilon_i = 1$, produced by incumbents. This recovers the main result in Baqaee and Farhi (2019a). For comparison to the next example, note that when there is only one primary factor, say labor indexed by L, we get

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \mathrm{d}\log \mu_i\right) - \mathrm{d}\log \lambda_L^B,$$

where λ_L^B is labor's share of income (or equivalently, the backward Domar weight of labor).

Example 2 (CRS with Entry)**.** Next, modify the example above to allow for entry. To make the model CRS, assume that $\gamma_i = 1 - \varepsilon_i = 1$ for all non-factor goods with directed entry. In this case, Theorem 3 reduces to

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \mathrm{d}\log \mu_i\right) - \sum_{f \in \mathcal{F}} \lambda_f^F \, \mathrm{d}\log \lambda_f^B.$$

Superficially, this looks identical to the previous example, but it is not. The reason is that with entry, the factor shares behave differently, even though the same equation holds. The clearest example is when we restrict the model to have one primary factor, where the comparative static becomes[37]

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \mathrm{d}\log \mu_i\right) - \mathrm{d}\log \lambda_L^B = \lambda_i^F \left(\mathrm{d}\log A_i - \mathrm{d}\log \mu_i\right). \tag{69}$$

The change in the labor share $\mathrm{d}\log \lambda_L^B$ disappears because it is always equal to zero. This is because in the model with entry, labor's share in aggregate income must always equal

[36] Setting $\mu_i^Y = 1/\gamma_i$ ensures that the aggregator $Y_i = (M_i y_i)^{1/\gamma_i}$ generates neither profits nor losses. This is implicitly the assumption one makes whenever one uses a CES aggregator.

[37] Equation (69) is reminiscent of Liu (2017).

one. Whereas, the model without entry necessitated knowing changes in the labor share, an object that depends on details like the elasticities of substitution and the shape of the production network, the one-factor model with entry does not require this information because the labor share never changes.[38]

Example 3 (DRS with Entry)**.** Consider again a model with only one factor of production, decreasing returns $\varepsilon_i \in (0, 1]$, and no external economies $\gamma_i = 1$. To build even more intuition, suppose entry is fully directed so that Theorem 3 simplifies to

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \frac{1-\varepsilon_i}{1-\frac{1-\varepsilon_i}{\mu_i}} \left(1 - \frac{1}{\mu_i}\right) \mathrm{d}\log \mu_i \right).$$

The first term corresponds to the technology effect working its way through the forward linkages.

The second term, which depends on $\mathrm{d}\log \mu_i$ is zero around the efficient point $\mu_i = 1$. At the efficient point, an increase in markups induces entry but shrinks each individual producer's scale of operation. At the efficient point, these effects cancel out to a first order.

A surprising property of this model is that the elasticities of substitution in consumption and in production across $i \in \mathcal{N}$ are irrelevant to a first-order. In this model, relative prices are pinned down independently of demand despite the fact that the model is inefficient and has decreasing returns to scale technology. Intuitively, this is because a shock to i only reallocates resources across the intensive and extensive margin in i and there is no substitution across i's. [39,40]

[38] In this version of the model, since $1 - \varepsilon = \gamma = 1$, entry is socially wasteful. This means the model has an unusual property at odds with a basic finding of neoclassical economics (exposited most famously by Harberger, 1964). In 'standard' models, introducing a distorting wedge has no effect on output to a first order starting at the efficient point (for a formal proof see e.g. Baqaee and Farhi, 2019a). In this example, at the efficient equilibrium, a marginal increase in markups does reduce output: $\mathrm{d}\log Y / \mathrm{d}\log \mu_i = -\lambda_i^F$. Why does the classic intuition fail in this model? The reason is that in this model, the marginal increase in markups around the efficient point induces entry. However, because producers have constant returns to scale, entry is socially wasteful, so the marginal social benefits between entry and non-entry are not equated, and therefore, reallocating resources towards entry reduces output to a first-order. In this version of the model, an increase in markups acts exactly like a negative productivity shock — destroying, rather than reallocating, resources. Of course, even in this version of the model, a wedge shock (i.e. $\mathrm{d}\log \mu_i^Y$) does not behave this way. For shocks to μ^Y, the labor share changes, which means that we would then need to know something about the input-output network and the elasticities of substitution (information required to solve for the change in the labor share).

[39] Unlike Example 2, this example does not have the exotic property described before. At the efficient allocation $\mathrm{d}\log Y / \mathrm{d}\log \mu_i = 0$, even when entry is not fully directed. This is because entry, even when its not fully directed, is not socially wasteful. Entry helps to overcome diminishing returns to scale in the individual production functions, and so around the efficient point, marginal benefits of all activities are being equated. Therefore, introducing a markup at the efficient point does not reduce output.

[40] This result is reminiscent of the no-substitution theorem (Georgescu-Roegen, 1951; Samuelson, 1951).

If we drop the assumption that entry is fully directed, Theorem 3 becomes

$$\mathrm{d}\log Y = \lambda_i^F \left(\mathrm{d}\log A_i - \frac{1-\varepsilon_i}{1-\frac{1-\varepsilon_i}{\mu_i}} \left(1 - \frac{1}{\mu_i}\right) \mathrm{d}\log \mu_i \right) - \lambda^F \varepsilon \cdot \left(d\log \lambda_\pi - \widehat{\mathrm{d}\log \lambda}_\pi \right).$$

Since entry is not directed, we must account for the residual rents. If for some product $k \in \mathcal{N}$, there is a large increase in profit shares $\mathrm{d}\log \lambda_{\pi_k}$ without an accompanying increase in the projection $\widehat{\mathrm{d}\log \lambda}_{\pi_k}$, then this means that entry is not able to respond to the increase in profitability. In this case, the individual producers in k are earning higher rents, their producer-specific factors are becoming more expensive, and this negatively affects output by a degree dependent on how valuable those producer-specific fixed factors are to production ε_k.

Example 4 (IRS with Entry)**.** Finally, consider the IRS case with $\gamma_i = 1 - \varepsilon_i < 1$ for every $i \in \mathcal{N}$ (except the factors). This coincides with using a CES aggregator with love-of-variety, where the product-level elasticity of substitution is $1/\varepsilon_i = 1/(1-\gamma_i)$. Since products are aggregated using a CES aggregator, adopt the imperfect substitutes interpretation in Proposition 1 $\mu_i^Y = 1/\gamma_i$ and $\mu_i^y = (1-\varepsilon_i)$. Since $\mu_i^Y > 1$, this implies the model is no longer efficient. For simplicity, continue to assume there is only one primary factor and no incumbents. Then Theorem 3 becomes

$$\mathrm{d}\log Y = \lambda_i^F \left(d\log A_i - \mathrm{d}\log \mu_i \right) + \sum_{k \in \mathcal{N}} \lambda_k^F \left(\frac{1}{\gamma_k} - 1 \right) \widehat{\mathrm{d}\log \lambda}_{\pi_k}.$$

In the previous DRS example, the key object were the residuals. In this case, the key objects is the projection in the regression. This is because now, a large projected value $\widehat{\mathrm{d}\log \lambda}_{\pi_k}$ means that increased rents are being captured by new entrants, and increased entry into product type k boosts output because k is produced with increasing returns on the margin. The closer is γ to one, the weaker is the love-of-variety effect, and the less important are the increasing returns to scale and changes in quasi-rents.

Appendix F Extensions

In this section, we consider some extensions to the basic model. First, we consider relaxing Assumption 1. Next, we consider an extension where factor supply is endogenous.

However, it holds under different assumptions: in particular, one does not need to assume constant returns to scale, nor perfect competition. The classic no-substitution theorem requires both assumptions, and will fail if either postulate is vioalted.

F.1 Relaxing Assumption 1

In this section, we relax Assumption 1 by considering how the model changes if (i) entry happens via a non-iso-elastic Kimball (1995) aggregator, and (ii) if the extent of decreasing returns to scale is variable.

F.1.1 Relaxing IRS

For the IRS benchmark, we can relax the assumption that entry happens via a CES aggregator by using the Kimball demand system instead. In other words, index firms in market i by some parameter θ, and suppose the production function is given by

$$y_i(\theta) = A_i(\theta)\left[f_i(x_{ij}(\theta))\right]^{1-\varepsilon_i}, \tag{70}$$

where f_i has constant returns to scale. Next, suppose that the inputs into the production function are defined implictly via the equation:

$$1 = \int \Upsilon_{ij}\left(\frac{x_{ij}(\theta,\theta')}{x_{ij}(\theta)}\right) M_j(\theta')d\theta', \tag{71}$$

where Υ_{ij} is an increasing concave function and $M_j(\theta)$ is the mass of type θ firms in $j \in \mathcal{N}$. The resource constraint for the output of this firm is then

$$y_i(\theta) = \sum_j \int x_{ji}(\theta',\theta)M_j(\theta')d\theta' + c_i(\theta'). \tag{72}$$

Let $P(i,j)$ be the marginal cost of input $x_{ij}(\theta)$. Because of homotheticity, we can consider the marginal cost of $x_{ij}(\theta)$ as depending only on $\{p_j(\theta'), M_j(\theta')\}_{\theta'}$. Define for each $(i,j) \in \mathcal{N}^2$, the linear operator $s(i,j) : L_2(\mathbb{R}) \to \mathbb{R}$

$$s(i,j)\cdot z = \int \left(\frac{p_i(\theta')x_{ij}(\theta,\theta')M_i(\theta')}{P(i,j)x_{ij}(\theta)}\right) z(\theta')d\theta. \tag{73}$$

Then we can write the change in the marginal cost of x_{ij}

$$d\log P(i,j) = s(i,j)\cdot d\log p_j - s(i,j)\cdot\left[(\delta_{ij}-1)d\log M_j\right], \tag{74}$$

where

$$\delta_{ij}(\theta) = \left(\int \Upsilon'\left(\frac{x_{ij}(\theta,\theta')}{x_{ij}(\theta)}\right)\frac{x_{ij}(\theta,\theta')}{x_{ij}(\theta)}M_j(\theta')d\theta'\right)^{-1}. \tag{75}$$

By homotheticity, $\delta_{ij}(\theta)$ is not a function of θ. The variable $\delta_{ij} > 1$ measures the love-of-variety effect in this model.

By Shephard's lemma

$$\begin{aligned}
d\log p_i(\theta) &= -d\log A_i(\theta) + d\log \mu_i(\theta) + \sum_j \Omega^F_{ij} d\log P(i,j) + \frac{\varepsilon_i}{1-\varepsilon_i} d\log y_i(\theta) \\
&= -d\log A_i(\theta) + d\log \mu_i(\theta) + \sum_j \Omega^F_{ij} d\log P(i,j) + \frac{\varepsilon_i}{1-\varepsilon_i} d\log \lambda_i(\theta) \\
&- \frac{\varepsilon_i}{1-\varepsilon_i} d\log M_i(\theta) - \frac{\varepsilon_i}{1-\varepsilon_i} d\log p_i(\theta) \\
&= -(1-\varepsilon_i) d\log A_i(\theta) + (1-\varepsilon_i) d\log \mu_i(\theta) + (1-\varepsilon_i) \sum_j \Omega^F_{ij} d\log P(i,j) \\
&+ \varepsilon_i \left(d\log \lambda_i(\theta) - d\log M_i(\theta)\right)
\end{aligned}$$

Therefore

$$\begin{aligned}
d\log P(i,j) &= s(i,j) \cdot \left(-(1-\varepsilon_i) d\log A_i(\theta) + (1-\varepsilon_i) d\log \mu_i(\theta) + (1-\varepsilon_i) \sum_j \Omega^F_{ij} d\log P(i,j) \right) \\
&+ s(i,j) \cdot \left(\varepsilon_i \left(d\log \lambda_i(\theta) - d\log M_i(\theta)\right)\right) - s(i,j) \cdot \left[(\delta_{ij} - 1) d\log M_j \right] \\
&= \left(s(i,j) \cdot (1-\varepsilon_i) d\log \frac{\mu_i(\theta)}{A_i(\theta)} + (1-\varepsilon_i) \sum_j \Omega^F_{ij} d\log P(i,j) \right) \\
&+ \varepsilon_i s(i,j) \cdot \left(d\log \lambda_i(\theta) - d\log M_i(\theta)\right) - s(i,j) \cdot \left[(\delta_{ij} - 1) d\log M_j \right]
\end{aligned}$$

We also have that

$$\lambda_{\pi_i}(\theta) = \left(1 - \frac{1-\varepsilon_i}{\mu_i(\theta)}\right) \lambda_i(\theta). \tag{76}$$

Define the function $\zeta_j(i,\theta)$ to be the mass of entrant j mapped to (i,θ). Zero-profit condition for type j entrant is

$$E_{\zeta_j}\left(\lambda_{\pi_i}(\theta)\right) = P_E, \tag{77}$$

where the expectation is with respect to ζ_j. We also have

$$M_i(\theta) = \int_E \zeta_j(i,\theta) M_{E,j} dj. \tag{78}$$

So we can write

$$d\log\lambda_{\pi_i}(\theta) = d\log\lambda_i(\theta) + \frac{\frac{1-\varepsilon_i}{\mu_i(\theta)}}{\left(1-\frac{1-\varepsilon_i}{\mu_i(\theta)}\right)} d\log\mu_i(\theta) \tag{79}$$

$$\sum_i \int \zeta_j(i,\theta)\frac{\lambda_{\pi_i}(\theta)}{M_i(\theta)}\left(d\log\lambda_{\pi_i}(\theta) - d\log M_i(\theta)\right)d\theta = P_E d\log P_E \tag{80}$$

$$d\log M_i(\theta) = \frac{\int_E \zeta_j(i,\theta) M_{E,j} d\log M_{E,j} dj}{\int_E \zeta_j(i,\theta) M_{E,j} dj}. \tag{81}$$

Let $\tilde{\zeta} : E \to \mathbb{R}^{\mathcal{N}}$and $\lambda_\pi : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$be linear operators. Then we can write

$$d\log M_i(\theta) = \tilde{\zeta}\cdot d\log M_E \tag{82}$$

$$\tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log\lambda_\pi - \tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log M = P_E d\log P_E \tag{83}$$

$$\begin{aligned}
\tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log\lambda_\pi - \tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\cdot d\log M_E &= P_E d\log P_E \\
\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\right)^{-1}\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log\lambda_\pi - P_E d\log P_E\right) &= d\log M_E \\
\tilde{\zeta}\cdot\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\right)^{-1}\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log\lambda_\pi - P_E d\log P_E\right) &= d\log M,
\end{aligned}$$

where $\tilde{\zeta}^*$is the adjoint operator. Define

$$d\log\hat{\lambda}_\pi = \tilde{\zeta}\cdot\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\right)^{-1}\tilde{\zeta}^*\cdot\lambda_\pi\cdot d\log\lambda_\pi$$

Hence, the forward equation becomes

$$d\log P(i,j) = \left(s(i,j)\cdot(1-\varepsilon_i)d\log\frac{\mu_i(\theta)}{A_i(\theta)} + (1-\varepsilon_i)\sum_j \Omega_{ij} d\log P(i,j)\right) \tag{84}$$

$$+ \left(\varepsilon_i s(i,j)\cdot\left(d\log\lambda_i(\theta) - d\log M_i(\theta)\right)\right) - s(i,j)\cdot\left[(\delta_{ij}-1)d\log M_j\right] \tag{85}$$

$$= \left(s(i,j)\cdot(1-\varepsilon_i)d\log\frac{\mu_i(\theta)}{A_i(\theta)} + (1-\varepsilon_i)\sum_j \Omega_{ij} d\log P(i,j)\right) \tag{86}$$

$$+ \left(\varepsilon_i s(i,j)\cdot\left(d\log\lambda_i(\theta) - d\log\lambda_\pi(\theta) + \tilde{\zeta}\cdot\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\right)^{-1}\tilde{\zeta}^*\cdot\lambda_\pi\cdot\lambda_E\cdot\Omega^E d\log P(i,j)\right)\right)$$

$$- s(i,j)\cdot(\delta_{ij}-1)d\log\lambda_\pi(\theta) - \left[s(i,j)\cdot(\delta_{ij}-1)\right]\tilde{\zeta}\cdot\left(\tilde{\zeta}^*\cdot\lambda_\pi\cdot\tilde{\zeta}\right)^{-1}\tilde{\zeta}^*\cdot\lambda_\pi\cdot\lambda_E\cdot\Omega^E d\log P(i,j) \tag{87}$$

This is a linear system in $d\log P(i,j)$. Group (i,j) together and write this linear system as a $\mathcal{N}^2\times 1$ vector, with an appropriately defined Ψ^F,then we have

$$\begin{aligned} d\log P(l,m) = &\sum_{ij}\Psi^F(lm,ij)\left(s(i,j)\cdot(1-\varepsilon_i)d\log\frac{\mu_i(\theta)}{A_i(\theta)} + \varepsilon_i s(i,j)\cdot\left(d\log\lambda_i(\theta) - d\log\hat{\lambda}_\pi(\theta)\right)\right) \\ &- \sum_{ij}\Psi^F(lm,ij)\left(s(i,j)\cdot(\delta_{ij}-1)d\log\hat{\lambda}_\pi(\theta)\right). \end{aligned}$$

This is the generalization to Theorem 3 and Proposition 3, showing that those results survive generalization.

Next, to pin down $d\log\lambda$, we need an analogue to the backward equations.

$$y_i(\theta) = \sum_j\int x_{ji}(\theta',\theta)M_j(\theta')M_j(\theta')d\theta' + c_i(\theta'). \tag{88}$$

$$\begin{aligned} \lambda_i(\theta) &= M_i(\theta)p_i(\theta)y_i(\theta) \\ &= M_i(\theta)p_i(\theta)\sum_j\int x_{ji}(\theta',\theta)M_j(\theta')d\theta'. \end{aligned}$$

Define

$$\sigma_{ji} = -\frac{\Upsilon'\left(\frac{y_{ji}(\theta)}{y_{ji}}\right)}{-\frac{y_{ji}(\theta)}{y_{ji}}\Upsilon''\left(\frac{y_{ji}(\theta)}{y_{ji}}\right)}, \tag{89}$$

where $y_{ji}(\theta) = \int x_{ji}(\theta',\theta)M_j(\theta')d\theta'$ and y_{ji} is defined implicitly via $1 = \int\Upsilon_{ji}\left(\frac{y_{ji}(\theta)}{y_{ji}}\right)M_j(\theta)d\theta$. Intuitively, because of homotheticity, we can assume that an intermediary purchases y_{ji} and sells it at marginal cost to all θ types in industry j. The quantity purchased by the intermediary from firm θ' in industry i is $y_{ji}(\theta')$ and the total output of the intermediary is y_{ji}.

The variable σ_{ji} is the price-elasticity of residual demand.

$$-d\log\delta_{ij} = \frac{\int\left(\Upsilon'\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)\frac{y_{ij}(\theta)}{y_{ij}}M_j(\theta)\right)\left[\frac{\frac{y_{ij}(\theta)}{y_{ij}}\Upsilon''\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)}{\Upsilon'\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)}d\log\left(\frac{y_{ij}(\theta)}{y_{ij}}\right) + d\log\frac{y_{ij}(\theta)}{y_{ij}} + d\log M_j(\theta)\right]d\theta}{\int\Upsilon'\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)\frac{y_{ij}(\theta)}{y_{ij}}M_j(\theta)d\theta.}$$

$$= \frac{\int \left(\frac{p_j(\theta)}{\delta_{ij}P(i,j)}\frac{y_{ij}(\theta)}{y_{ij}}M_j(\theta)\right)\left[\left(1-\frac{1}{\sigma_{ji}(\theta)}\right)d\log\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)+d\log M_j(\theta)\right]d\theta}{\int \frac{p_j(\theta)}{\delta_{ij}P(i,j)}\frac{y_{ij}(\theta)}{y_{ij}}M_j(\theta)d\theta.}$$

$$= s(i,j)\cdot\left[\left(1-\frac{1}{\sigma_{ji}(\theta)}\right)d\log\left(\frac{y_{ij}(\theta)}{y_{ij}}\right)+d\log M_j(\theta)\right] \tag{90}$$

$$d\log\left(\frac{y_{ij}(\theta)}{y_{ij}}\right) = d\log(\Upsilon')^{-1}_{ji}\left(\frac{p_i(\theta)}{\delta_{ji}P(j,i)}\right) \tag{91}$$

Hence

$$d\log\left(\frac{y_{ij}(\theta)}{y_{ij}}\right) = \sigma_{ij}(\theta)\left(d\log p_i(\theta) - d\log\delta_{ji} - d\log P(j,i)\right). \tag{92}$$

Use this in

$$\begin{aligned}\lambda_i(\theta) &= M_i(\theta)p_i(\theta)y_i(\theta)\\ &= M_i(\theta)p_i(\theta)\sum_j y_{ji}(\Upsilon')^{-1}_{ij}\left(\frac{p_i(\theta)}{\delta_{ji}P(j,i)}\right),\end{aligned}$$

where the final line follows from homotheticity. Hence

$$d\log\lambda_i(\theta) = d\log M_i(\theta) + d\log p_i(\theta) + \sum_j \frac{y_{ji}(\theta)}{y_{ji}}\left(\sigma_{ji}(\theta)\left(d\log p_i(\theta) - d\log\delta_{ji} - d\log P(j,i)\right)\right)$$
$$+\sum_j \frac{y_{ji}(\theta)}{y_{ji}}d\log y_{ji}. \tag{93}$$

Next, use

$$y_{ji} = \frac{1-\varepsilon_j}{\bar{\mu}_j}\frac{\Omega^F_{ji}\lambda_j}{P(j,i)} \tag{94}$$

coupled with

$$\Omega^F_{ji}d\log\Omega^F_{ji} = (1-\theta_j)Cov_j(d\log P(j,m), I_{(i)}) \tag{95}$$

to get

$$d\log y_{ji} = -d\log\bar{\mu}_j + (1-\theta_j)Cov_j(d\log P(j,m), I_{(i)}) + d\log\lambda_j - d\log P(j,i), \tag{96}$$

where

$$\bar{\mu}_j = \left(\int \frac{\lambda_i(\theta)M_i(\theta)}{\int\lambda_i(\theta)M_i(\theta)d\theta}\frac{1}{\mu_i(\theta)}d\theta\right)^{-1}, \tag{97}$$

so

$$-d\log\bar{\mu}_i = \int \frac{\lambda_i(\theta)M_i(\theta)}{\int \lambda_i(\theta)M_i(\theta)d\theta}\frac{1}{\mu_i(\theta)}\left[-d\log\mu_i(\theta) + d\log\lambda_i(\theta) + d\log M_i(\theta) - d\log\lambda_i\right]d\theta \tag{98}$$

Finally, use the fact that

$$\lambda_i = \sum_j \frac{1-\varepsilon_j}{\bar{\mu}_j}\Omega^F_{ji}\lambda_j + \sum_j \Omega^E_{ji}\lambda_{E,j} \tag{99}$$

to get

$$d\lambda_i = \sum_j d\lambda_j \frac{1-\varepsilon_j}{\bar{\mu}_j}\Omega^F_{ji} - \sum_j \frac{\lambda_j}{\bar{\mu}_j}(1-\varepsilon_j)\Omega^F_{ji}d\log\bar{\mu}_j + \sum_j \frac{\lambda_j}{\bar{\mu}_j}(1-\varepsilon_j)\Omega^F_{ji}d\log\Omega^F_{ji} + \sum_j d\lambda_{E,j}\Omega^E_{ji}. \tag{100}$$

Equations (87), (90), (93), (95), (96), (98), (100) jointly complete the characterization.

F.1.2 Relaxing DRS

Suppose that

$$Y = M_i A_i f_i\left(\left\{x_{ij}\right\}_j\right), \tag{101}$$

where we do not impose homotheticity on f_i. Define

$$\begin{aligned}
\lambda^Y &= PY = \mu^Y\lambda^y \\
\lambda^y &= pyM \\
\lambda_\pi &= \frac{1}{\mu^Y}\left(1 - \frac{1-\varepsilon}{\mu}\right)\lambda^Y
\end{aligned}$$

which implies that

$$\begin{aligned}
d\log\lambda_\pi &= -d\log\mu^Y - \frac{\frac{1-\varepsilon}{\mu}}{\left(1-\frac{1-\varepsilon}{\mu}\right)}\left[d\log(1-\varepsilon) - d\log\mu\right] + d\log\lambda \\
d\log M &= d\log\hat{\lambda}_\pi - d\log\hat{P}
\end{aligned}$$

$$\begin{aligned}
P &= \mu^Y\frac{dC}{dY} = \mu^Y C/Y \\
d\log P &= d\log\mu^Y + d\log p
\end{aligned}$$

$$
\begin{aligned}
d\log p &= d\log\mu - d\log(1-\varepsilon) - d\log y_i + \Omega^F d\log P + \frac{\partial \log C_i}{\partial \log y_i}\left(d\log y_i - d\log A_i\right) \\
&= d\log\mu - d\log(1-\varepsilon) - d\log y_i + \Omega^F d\log P + \frac{1}{1-\varepsilon}\left(d\log y_i - d\log A_i\right) \\
&= d\log\mu - d\log(1-\varepsilon) + \Omega^F d\log P + \frac{\varepsilon}{1-\varepsilon}\left(d\log\lambda - d\log p - d\log M - \frac{1}{\varepsilon}d\log A_i\right) \\
d\log p &= (1-\varepsilon)d\log\mu - (1-\varepsilon)d\log(1-\varepsilon) + (1-\varepsilon)\Omega^F d\log P + \varepsilon\left(d\log\lambda - d\log M\right) - d\log A \\
d\log P &= d\log\mu^Y + d\log p \\
&= d\log\mu^Y + \left[(1-\varepsilon)d\log\mu + d\varepsilon + (1-\varepsilon)\Omega^F d\log P + \varepsilon\left(d\log\lambda - d\log M\right) - d\log A\right] \\
&= d\log\mu^Y + (1-\varepsilon)d\log\mu + d\varepsilon + (1-\varepsilon)\Omega^F d\log P \\
&+ \varepsilon\left(d\log\lambda - d\log\hat{\lambda}_\pi + d\log\hat{P}\right) - d\log A \\
(I-\Omega^F)d\log P &= d\log\mu^Y + (1-\varepsilon)d\log\mu + \varepsilon\left(d\log\lambda + d\log\varepsilon - d\log\hat{\lambda}_\pi\right) - d\log A
\end{aligned}
$$

This last equation generalizes the forward equations in Proposition 3.

To get the backward equation, assuming some separability, we can write

$$
f_i\left(\left\{x_{ij}\right\}_j\right) = f_i(q_i), \tag{102}
$$

where q_i is CRS function of inputs. We can write

$$
\begin{aligned}
\Omega_{ij} &= \frac{M_i P_j x_{ij}}{P_i Y_i} = \frac{P_j x_{ij}}{\mu^Y p_i y_i} = \frac{P_j x_{ij}}{\mu^Y \mu_i^y (1-\varepsilon_i)\mu_i^q m c_i q_i} = \frac{1}{\mu_i \mu_i^Y}\frac{1}{1-\varepsilon_i}\frac{p_j x_{ij}}{m c_i q_i} \\
d\log\Omega_{ij} &= -d\log\mu_i\mu_i^Y + d\log\gamma_i - d\log(1-\varepsilon_i) + d\log\left(\frac{p_j x_{ij}}{m c_i q_i}\right)
\end{aligned}
$$

Denote the super-elasticity by $\frac{\partial^2 \log f_i}{\partial \log q_i^2} = \kappa_i$. Then we can write

$$
d(1-\varepsilon_i) = \kappa_i\left(d\log\lambda_i^q - d\log p_i^q\right) = d\log\lambda_i^q - d\log\lambda_i^y. \tag{103}
$$

Hence,

$$
d\varepsilon = d\lambda_i^y - d\lambda_i^q \tag{104}
$$

and

$$
d\log\lambda_i^q = \frac{1}{\kappa_i - 1}\left(\kappa_i d\log p_i^q - d\log\lambda_i^y\right). \tag{105}
$$

Hence

$$
d\log\Omega_{ij} = -d\log\mu_i\mu_i^Y - \kappa_i\left(d\log\lambda_i^q - d\log p_i^q\right) + d\log\left(\frac{p_j x_{ij}}{m c_i q_i}\right)
$$

$$
\begin{aligned}
&= -d\log\mu_i\mu_i^Y - \frac{\kappa_i}{1-\varepsilon_i}\left(\frac{1}{\kappa_i-1}\left(\kappa_i d\log p_i^q - d\log\lambda_i^y\right) - d\log p_i^q\right) + d\log\left(\frac{p_j x_{ij}}{mc_i q_i}\right)\\
&= -d\log\mu_i\mu_i^Y - \frac{\kappa_i}{1-\varepsilon_i}\left(\frac{1}{\kappa_i-1}\left(\kappa_i d\log p_i^q - d\log\lambda_i^y\right) - d\log p_i^q\right) + (1-\theta_i)Cov_i(d\log P, I_{(i)})\\
&= -d\log\mu_i\mu_i^Y - \frac{\kappa_i}{1-\varepsilon_i}\frac{1}{\kappa_i-1}\left(d\log p_i^q - d\log\lambda_i^y\right) + (1-\theta_i)Cov_i(d\log P, I_{(i)})\\
&= -d\log\mu_i\mu_i^Y - \frac{\kappa_i}{1-\varepsilon_i}\frac{1}{\kappa_i-1}\left(\sum_j \Omega_{ij}P_j - d\log\lambda_i^y\right) + (1-\theta_i)Cov_i(d\log P, I_{(i)})
\end{aligned}
$$

Finally, combine this with

$$
d\lambda^{y'} = d\lambda^{y'}\Omega + \lambda^{y'}d\Omega + d\lambda_E\Omega^E \tag{106}
$$

to pin down the backward equations, which is the equivalent of Proposition 4.

F.2 Variable Factor Supply

Suppose the supply of each factor $f \in \mathcal{F}$ is given by

$$
L_f = G_f(w_f/P_0, Y) = G_f(w_f Y, Y). \tag{107}
$$

and let $\zeta_f = \partial\log G_f/\partial\log w_f$ be the Marshallian price elasticity of supply and $\partial\log G_f/\partial\log Y = \gamma_f$. Hence, $\zeta_f - \gamma_f$ is the income elasticity of supply. Here, to make the notation more familiar, we refer to the quantity of each factor f by L_f and the price of the factor by w_f. We consider perturbations to $\mathrm{d}\log\mu$ and $\mathrm{d}\log A$. So

$$
\begin{aligned}
d\log P =& \frac{1-\varepsilon}{\gamma}d\log\mu - d\log A + \left(1-\frac{1-\varepsilon}{\gamma}\right)d\log\lambda^Y + \frac{(1-\varepsilon)}{\gamma}\mu^q\Omega^V d\log P\\
&+ \frac{\varepsilon}{\gamma}\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\left(\lambda_E\Omega^E d\log P\right) - \frac{\varepsilon}{\gamma}d\log\hat{\lambda}_\pi\\
&+ \frac{(1-\varepsilon)}{\gamma}\mu^q\Omega^V_{\mathcal{F}}d\log w,
\end{aligned}
$$

Now use the fact that

$$
\begin{aligned}
d\log w_f &= d\log\lambda_f - \zeta_f d\log w_f - \gamma_f d\log Y,\\
d\log w &= \frac{1}{1+\zeta_f}d\log\lambda_f - \frac{\gamma_f}{1+\zeta_f}d\log Y.
\end{aligned}
$$

$$d\log\lambda_\pi + \left(1 - \frac{1-\varepsilon_i}{\mu_i^q \mu_i^y}\right)^{-1} d\log\mu_i^Y - \left(1 - \frac{1-\varepsilon_i}{\mu_i^q \mu_i^y}\right)^{1} \frac{1-\varepsilon_i}{\mu_i^q \mu_i^y} (d\log\mu) = d\log\lambda_i \tag{108}$$

Hence

$$\begin{aligned}
d\log P &= \frac{1-\varepsilon}{\gamma} d\log\mu - d\log A + \left(1 - \frac{1-\varepsilon}{\gamma}\right) d\log\lambda^Y + \frac{(1-\varepsilon)}{\gamma} \mu^q \Omega^V d\log P \\
&\quad + \frac{\varepsilon}{\gamma} \tilde{\zeta}' (\tilde{\zeta}\lambda_\pi \tilde{\zeta}')^{-1} \left(\lambda_E \Omega^E d\log P\right) - \frac{\varepsilon}{\gamma} d\log\hat{\lambda}_\pi \\
&\quad + \frac{(1-\varepsilon)}{\gamma} \mu^q \Omega^V_{\mathcal{F}} \left[\frac{1}{1+\zeta_f} d\log\lambda_{\mathcal{F}} - \frac{\gamma_f}{1+\zeta_f} d\log Y\right], \\
d\log P &= \Psi^F \left[\frac{1-\varepsilon}{\gamma} d\log\mu - d\log A + \left(1 - \frac{1-\varepsilon}{\gamma}\right) d\log\lambda^Y - \frac{\varepsilon}{\gamma} d\log\hat{\lambda}_\pi\right] \\
&\quad + \Psi^F \left[\frac{(1-\varepsilon)}{\gamma} \mu^q \Omega^V_{\mathcal{F}} \left[\frac{1}{1+\zeta_f} d\log\lambda_{\mathcal{F}} - \frac{\gamma_f}{1+\zeta_f} d\log Y\right]\right] \\
&= \Psi^F \left[\frac{1-\varepsilon}{\gamma} d\log\mu - d\log A + \left(1 - \frac{1-\varepsilon}{\gamma}\right)\left(d\log\lambda^Y - d\log\hat{\lambda}_\pi\right)\right] \\
&\quad + \Psi^F \left[\left(1 - \frac{1}{\gamma}\right) d\log\hat{\lambda}_\pi\right] \\
&\quad + \Psi^F \left[\frac{(1-\varepsilon)}{\gamma} \mu^q \Omega^V_{\mathcal{F}} \left[\frac{1}{1+\zeta_f} d\log\lambda_{\mathcal{F}} - \frac{\gamma_f}{1+\zeta_f} d\log Y\right]\right].
\end{aligned}$$

This is the counterpart to the forward equations of Proposition 3 when factor supply is elastic.

Defining $\varepsilon = 1$ and $\gamma = 1$ for factors, and $\gamma = 0$ and ζ=0 for non-factors. Hence we can combine to write the aggregation equation of Theorem 3 as

$$\begin{aligned}
d\log Y &= -\lambda^F \left[\frac{1-\varepsilon}{\gamma} d\log\mu - d\log A\right] \\
&\quad - \lambda^F \left[\left(1 - \frac{1-\varepsilon}{\gamma}\right)\left(\frac{1}{1+\zeta} d\log\lambda_\pi - \frac{\gamma}{1+\zeta} d\log Y - d\log\hat{\lambda}_\pi\right) - \left(1 - \frac{1}{\gamma}\right) d\log\hat{\lambda}_\pi\right].
\end{aligned}$$

The backward equations are unchanged relative to ones in Proposition 4.

Appendix G Intuition for Markup/Wedge Shocks

We briefly discuss markup/wedge shocks in Theorem 3. We consider a univariate markup shock $\mathrm{d}\log\mu_i$ or a univariate wedge shock $\mathrm{d}\log\mu_i^Y$. For simplicity, we only treat the case where entry into i is directed, so that entrants in market i are from a single type and cannot

enter in any other market. It is then more convenient to rewrite Theorem 3 as

$$\begin{aligned}\mathrm{d}\log Y = -\lambda_i^{\mathrm{F}}\frac{1-\varepsilon_i}{\gamma_i}\frac{\mu_i-1}{\mu_i-(1-\varepsilon_i)}d\log\mu_i - \lambda_i^{\mathrm{F}}\frac{1}{\gamma}d\log\mu_i^{Y} \\ -\sum_{j\in\mathcal{N}}\lambda_j^{\mathrm{F}}\left(1-\frac{1-\varepsilon_j}{\gamma_j}\right)\left(\mathrm{d}\log\lambda_j^{\mathrm{B}}-\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}\right)+\sum_{j\in\mathcal{N}}\lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j}-1\right)\widehat{\mathrm{d}\log\lambda_j^{\mathrm{B}}}.\end{aligned}$$

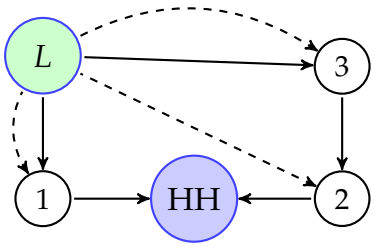
We have broken down changes in rents into changes in profit margins and changes in sales. This is because markups/wedge shocks directly affect the former but not the latter. The second line is the indirect effect of markup/wedge shocks operating through changes in market sizes. It takes the same form as in the case of productivity shocks where profit margins are constant, and has the same intuition.

The first line is the direct effect of markup/wedge shocks, holding market sizes constant. Consider first a univariate markup increase $\mathrm{d}\log\mu_i > 0$ with $\mathrm{d}\log\mu_i^{Y} = 0$. Intuitively, there are countervailing effects on final-demand prices. On the one hand the markup increase raises the price of market good i by increasing the gap between price and marginal cost of individual producers . On the other hand, it also raises the profitability of i, encourages entry, and hence reduces the price of market good i through love for variety. The overall direct effect of the markup increase depends on the balance of these two effects, which ultimately depends on whether there is too much or too little entry to begin with. When $\mu_i < 1$ ($\mu_i > 1$) there is too little (too much) entry to begin with, and so the direct effect of the markup increase is positive (negative). When $\mu_i = 1$, entry is efficient, and there is no direct effect of the markup increase on entry. Next consider a univariate increase in the output wedge $\mathrm{d}\log\mu_i^{Y} > 0$ with $\mathrm{d}\log\mu_i = 0$. The direct effect of the wedge shock is an increase in final-demand prices, which, in this case, is not counterbalanced by an increase in entry.

Appendix H Intervening in a Cobb-Douglas Example

To make this intuition even more concrete, we consider the example in Figure 6. For this economy, consumption goods can either be produced directly using labor or they can be produced via a two-step supply chain. In addition, we have $\lambda_L^{\mathrm{B}} = 1$ and $\lambda_L^{\mathrm{F}} = \lambda_1^{\mathrm{B}}/\gamma_1 + \lambda_2^{\mathrm{B}}(1/\gamma_2 + 1/\gamma_3 - 1)$.

An entry subsidy extracts labor from the rest of the economy and funnels it into entry in market i. This adjustment comes about via an increase in the price of labor. The effect of a dollar spent subsidizing entry in 1 is $-(1/\lambda_{E,1}^{\mathrm{B}})\,\mathrm{d}\log Y/\,\mathrm{d}\log\mu_{E,1}^{Y} = (1/\gamma_1 - \lambda_L^{\mathrm{F}})$, where



$$Y_i = (M_i y_i)^{\frac{1}{\gamma_i}}, \qquad y_i = l_i^{1/\gamma_i} \ (i \in \{1,3\}),$$
$$y_2 = (Y_3/M_2)^{1/\gamma_2}, \qquad M_i = \frac{\lambda_{\pi,i}}{w_L},$$
$$L = \sum_{i\in\{1,3\}} M_i l_i + \sum_{i=1}^{3} M_i, \quad Y = Y_1^{\overline{\lambda}_1^B} Y_2^{\overline{\lambda}_2^B}.$$

Figure 6: The solid and dashed arrows represent the flow of resources for production and for entry. The sole factor for this economy is indexed by L.

the first term is the benefits from increased entry into 1 and the second term is the cost of having less resources for the rest of the economy. Similarly, the effect of a dollar spent subsidizing entry in 2 is $-(1/\lambda^{\mathrm{B}}_{E,2})\,\mathrm{d}\log Y/\,\mathrm{d}\log\mu^{Y}_{E,2} = 1/\gamma_2 - \lambda^{\mathrm{F}}_{L}$. Finally, the effect of a dollar spent subsidizing entry in 3 is $-(1/\lambda^{\mathrm{B}}_{E,3})\,\mathrm{d}\log Y/\,\mathrm{d}\log\mu^{Y}_{E,3} = 1/(\gamma_2\gamma_3) - \lambda^{\mathrm{F}}_{L}$. This means that subsidizing 3 dominates subsidizing 2, and subsidizing the supply chain is beneficial as long as cumulated increasing returns are stronger in the longer supply chain $1/\gamma_2 + 1/\gamma_3 - 1 > 1/\gamma_1$. It is then optimal to subsidize entry upstream of complex supply chains.

Starting at the monopolistically competitive equilibrium, a reduction in the markup of 1 has no effect on aggregate output since it leads to offsetting effects on the price of 1 from the reduction in the prices of individual producers and the increase in entry. The same goes for a reduction in the markup of 3. By contrast, a reduction in the markup of 2 is beneficial, since it allows 3 to expand $-(1/\lambda^{\mathrm{B}}_{2})\,\mathrm{d}\log Y/\,\mathrm{d}\log\mu^{q}_{2} = (1/\gamma_2)(1/\gamma_3 - 1)$. Therefore, as long as there is increasing returns to scale $1/\gamma_3 > 1$ in market 3, markup reductions should be targeted downstream to market 2. It is then optimal to reduce markup downstream of complex supply chains.

Appendix I Sectoral Models

For any sectoral model with heterogeneous firms in each sector, there is an isomorphic *companion* sectoral model with homogenous firms in each sector. The companion model assumes that all firms in a given sector $\mathcal{I}$ are identical with productivity shifter $A_{\mathcal{I}}$ and markup $\mu_{\mathcal{I}}$ defined by

$$A_{\mathcal{I}} = \frac{\mu_{\mathcal{I}}}{\overline{\mu}_{\mathcal{I}}}\left(\sum_{i\in\mathcal{I}} \overline{\lambda}_i^{\mathcal{I},\mathrm{B}} \left(\frac{A_i/\overline{A}_i}{\mu_i/\overline{\mu}_i}\right)^{\frac{1-\varepsilon_{\mathcal{I}}}{\varepsilon_{\mathcal{I}}}}\right)^{\frac{\varepsilon_{\mathcal{I}}}{1-\varepsilon_{\mathcal{I}}}} \qquad \text{and} \qquad \mu_{\mathcal{I}} = \frac{1}{\sum_{i\in\mathcal{I}} \lambda_i^{\mathcal{I},\mathrm{B}} \frac{1}{\mu_i}},$$

where for each $i \in \mathcal{I}$, we define $\lambda_{\mathcal{I}}^{\mathrm{B}} = \sum_{j \in \mathcal{I}} \lambda_j^{\mathrm{B}}$ and $\lambda_i^{\mathcal{I},\mathrm{B}} = \lambda_i^{\mathrm{B}} / \lambda_{\mathcal{I}}^{\mathrm{B}}$. Here we remind the reader that we use overlines to signal initial values when there is an ambiguity but we drop them when there is none. We denote by $\check{Y}$ the aggregate output in the companion model without heterogeneity within sectors.

If Y denotes aggregate output in a sectoral model with heterogeneity, we denote by $\check{Y}$ denote aggregate output in the companion model without heterogeneity.

Proposition 7 (Sectoral Aggregation)**.** *For any sectoral model with within-sector heterogeneity, the nonlinear response $\Delta \log Y$ of aggregate output to shocks to productivities and markups is equal to the nonlinear response $\Delta \log \check{Y}$ of aggregate output to shocks to sectoral productivities and markups in the companion model with no within-sector heterogeneity.*

The outer-elasticity $\gamma_{\mathcal{I}}$, which distinguishes models with IRS from those with DRS, is not relevant to how we aggregate firms within the sector since neither $A_{\mathcal{I}}$ nor $\mu_{\mathcal{I}}$ depend on $\gamma_{\mathcal{I}}$.

To break this problem into a within-sector and cross-sector problem, in vector notation, write

$$\mathrm{d} \log Y = \mathrm{d} \log \check{Y} = \sum_{\mathcal{I}} \frac{\mathrm{d} \log \check{Y}}{\mathrm{d} \log A_{\mathcal{I}}} \mathrm{d} \log A_{\mathcal{I}} + \sum_{\mathcal{I}} \frac{\mathrm{d} \log \check{Y}}{\mathrm{d} \log (\mu_{\mathcal{I}}, \mu_{\mathcal{I}}^{Y})} \mathrm{d} \log (\mu_{\mathcal{I}}, \mu_{\mathcal{I}}^{Y}).$$

We now differentiate a second time and evaluate the second derivative at the efficient marginal-pricing equilibrium. We use the fact that at the efficient point $\mathrm{d} \log A_{\mathcal{I}} = 0$ and, from the envelope theorem, that $\mathrm{d} \log \check{Y} / \mathrm{d} \log \mu_{\mathcal{I}} = 0$, we get a simpler expression for the loss function $\mathcal{L} = -(1/2) d^2 \log \check{Y}$ using

$$d^2 \log \check{Y} = \sum_{\mathcal{I}} \frac{\mathrm{d} \log \check{Y}}{\mathrm{d} \log A_{\mathcal{I}}} d^2 \log A_{\mathcal{I}} + \sum_{\mathcal{I},\mathcal{J}} \mathrm{d} \log (\mu_{\mathcal{I}}, \mu_{\mathcal{I}}^{Y})' \frac{d^2 \log \check{Y}}{\mathrm{d} \log (\mu_{\mathcal{I}}, \mu_{\mathcal{I}}^{Y}) \, \mathrm{d} \log (\mu_{\mathcal{J}}, \mu_{\mathcal{J}}^{Y})} \mathrm{d} \log (\mu_{\mathcal{J}}, \mu_{\mathcal{J}}^{Y}), \quad (109)$$

where $\mathrm{d} \log \check{Y} / \mathrm{d} \log A_{\mathcal{I}} = \lambda_{\mathcal{I}}^{B} (1 - \varepsilon_{\mathcal{I}})$ by Theorem 2. This expression can then be combined with the following lemma.

Using Lemma 5 below, it becomes apparent that: the first term in the loss function captures misallocation arising from distortions in relative producer sizes driven by the *dispersion* of markups/wedges within sectors; the second term captures misallocation arising from distortions in entry within sectors and relatives sizes across sectors arising driven by the *levels* of markups. The losses increase with the returns to scale and go to

infinity in the constant-returns limit where $\varepsilon_{\mathcal{I}}$ goes to zero.

Lemma 5. *At the efficient marginal-cost pricing equilibrium, changes in sectoral markups and productivities in the companion model are related to changes in markups/wedges in the original model according to*

$$\mathrm{d}\log\mu_{\mathcal{I}} = \mathbb{E}_{\lambda^{\mathcal{I},B}}(\mathrm{d}\log\mu), \quad \mathrm{d}\log A_{\mathcal{I}} = 0, \quad \textit{and} \quad d^2\log A_{\mathcal{I}} = \frac{1}{\varepsilon_{\mathcal{I}}} Var_{\lambda^{\mathcal{I},B}}(\mathrm{d}\log\mu),$$

where these expressions denote within-sector weighted expectations and variances of the changes in markups/wedges $\mathrm{d}\log\mu_i$ *in the original model, with weights given by the within-sectoral sales share distribution* $\lambda_i^{\mathcal{I},B}$.