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, Revised November 2020

Emmanuel Farhi tragically passed away in July, 2020. Hewas a once in a lifetime friend and collaborator. We thank Pol Antras, Natalie Bau, Elhanan Helpman, Hugo Hopenhayn, Oleg Itskhoki, Xavier Gabaix, Marc Melitz, and especially Jonathan Vogel for comments. Veronica De Falco provided exceptional research assistance. 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, Revised November 2020
JEL No. E0,E1,E3,O0,O11,O21,O25,O3,O4,O41

ABSTRACT

A tension between entry and rents lies at the core of a general theory of aggregation with scale effects. This paper characterizes 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 entry across markets. In addition, we characterize the social costs of distortions. We prove that while first best industrial policy is network-independent, second-best policy does depend on the network, and boosts upstream industries that intensively supply downstream industries with high returns to scale. As an application, we quantify the impact of misallocation from markups on aggregate efficiency in the US. In our preferred specification, losses are 40% of GDP whereas if we abstract from endogenous entry they are only 20%. Our baseline is sensitive not only to modeling entry, but also to the specifics of how entry is modeled, in ways that our social-costs-of-distortions formulas clarify.

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

Emmanuel Farhi
and CEPR

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

1 Introduction

We analyze how microeconomic shocks translate into aggregate productivity and output in the presence of non-constant returns to scale and product entry and exit. We use this analysis to characterize the efficiency losses from misallocation and derive optimal first- and second-best industrial policies.

Our analysis is relatively general, and applies to many of the models typically used to analyze scale effects, for example, Dixit and Stiglitz (1977), Krugman (1979), Romer (1987), Aghion and Howitt (1992), 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 (TFP) 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 exogenous 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 divided by GDP. The intuition for this term is exactly as in Hulten (1978)'s theorem.

If the equilibrium is efficient, the envelope theorem implies that reallocation effects can be ignored to a first order, $\Delta \log TFP^{\text{alloc}} = 0$. Once we stray from efficiency, we show that changes in allocative efficiency play a dominant role in determining the aggregate consequences of disturbances. These reallocation effects depend on which markets expand and shrink, and on whether these adjustments in market sizes occur through 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.

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 (variable profit). 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 paid by entrants into that market.

Our treatment of entry is novel due to its generality and nests the cases of directed technical change, in which entrants choose their technology from a menu, and undirected innovation, in which they do not. We show that entry by 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 the change in quasi-rents associated with each market, denoted by

$$\Delta \text{ Quasi-rents} = \widehat{\Delta \log \lambda}_\pi,$$

is the projection of changes in rents, $\Delta \log \lambda_\pi$, on a vector space representing the entry technology. This projection determines the amount of entry into the different markets. Therefore, in a least-squares sense, new entry minimizes new rents claimed by existing producers. The residual from this projection, denoted by

$$\Delta \log \lambda_\pi - \widehat{\Delta \log \lambda}_\pi,$$

measures the inability of entry to respond to changes in rents. If new entrants can perfectly direct their entry decisions, then this residual is always zero.

We show that this projection and its residuals 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

[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 sum is over all markets i.[3]

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, raising the costs of production due to diminishing returns. If the 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 (love of variety). With increasing external returns to scale ($\mathcal{E}_i^{\text{ext}} > 0$), allocative efficiency further improves if reallocations increase entry in markets with relatively high external economies. The sum of these terms is always zero in efficient economies, but not in inefficient economies.[4]

There is folk wisdom 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 intuition is generally invalid 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 responses to shocks.

We complete this new perspective by providing propagation equations that show how changes in rents $d \log \lambda_\pi$ are determined in equilibrium as functions 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 respond 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 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 depend on network structure. In particular, for economies with external increasing returns, we rationalize and revise Hirschman (1958)'s influential argument that policy should encourage expansion in 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 the sales of sectors that intensively supply other industries with strong scale economies.

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

[3]We derive similar formulas for the response to changes in markups and other distortions.

[4]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 the suboptimal level of 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%.[5] Accounting for entry can double these losses. Furthermore, the specific number one attaches to these losses is sensitive to assumptions on 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. While our results show that these features are critical theoretically and quantitatively, little is know about them in practice, and more empirical work is needed to bridge theory and measurement.

The structure of the paper is as follows. In Section 2, we set up the general model and define the equilibrium. 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 7, we analyze the normative properties of the economy, including first- and second-best optimal policy and the social costs of distortions. Section 8 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.

[5]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.

Related Literature. Our results apply to a broad range of influential models in the macro, trade, and growth literatures. For instance, our framework encompasses and generalizes 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), Jones (2013), Bigio and La'O (2016), 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.

This paper also relates to the literature 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 economies with entry losses can be approximated using Harberger triangles to non-neoclassical environments, 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 (perhaps randomly) assigned to be producers. 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.

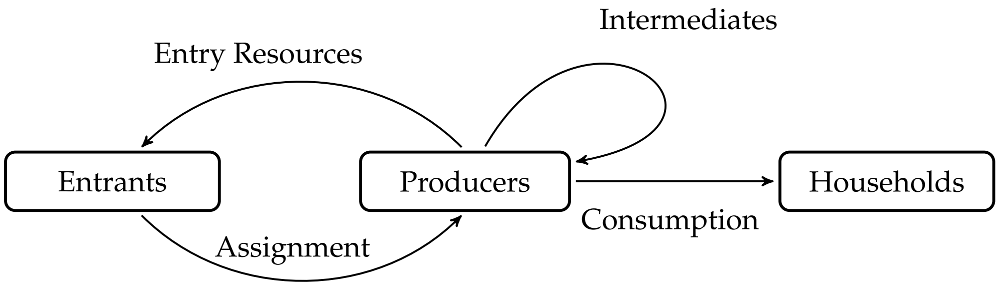


Figure 1: Circular flow schematic of the economy showing the flow of resources.

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

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

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) used by i. 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. This distinction matters because revenues generated by μ_i incentivize entry, whereas revenues generated by μ_i^Y do not. 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 denote a bundle of inputs and ignore productivity by setting $A_i = 1$ as an argument.

Assume that $f_i(x) = x^{1-\gamma_i}$ for some $\gamma_i \in [0,1]$, and $F_i(x) = x^{\frac{1}{\gamma_i}}$. Then $Y_i = (M_i x^{\gamma_i})^{\frac{1}{\gamma_i}}$ captures a CES aggregator with an elasticity of substitution $1/(1-\gamma_i)$ between differentiated varieties produced under constant returns to scale. Suppose instead that $F_i(x) = x$. Then $Y_i = M_i x^{\gamma_i}$ 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 factor 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 type-j entrants 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).

Entry Technology. The entry matrix ζ is an $|E| \times |\mathcal{N} - \mathcal{F}|$ positive-valued matrix. Type-j entrants who pay the fixed 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.[6] 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.

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

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 to operate in market j without having to enter.

We refer to markets where entry is not possible as *uncontested markets* and denote their collection by $\mathcal{N}^u$. 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 $\mathcal{N}^c$. Note that since we are flexible in the way we define and combine markets, we can capture a situation where incumbents and entrants compete by having them operate in different markets that are highly-substitutable with one another.

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 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.[7]

Zero-Profit Conditions. The zero-profit condition for type-j entrants equates expected profits post-entry with the costs of entry

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

where

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

[7]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. 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 B for more details and for a more general formalization which sidesteps these issues.

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 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

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

in words, the total supply of good i is equal to demand by households, producers (as intermediate inputs), and entrants (as fixed costs). The mass of producers in a contested market $i \in \mathcal{N}^c$ is the sum of the share of entrants $j \in E$ that obtained technology i:

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

The mass of producers M_i in uncontested markets $i \in \mathcal{N}^u$ is exogenous.

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. Since these are reduced-form wedges, 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.

Of course, many wedges, like variable markups or nominal rigidities, are themselves endogenous. In this paper, we do not commit to a specific theory of wedge determination and instead ask how changing wedges affect output. Some questions, like the economy's distance from the Pareto efficient frontier, can be answered without endogenizing the wedges. Endogenizing wedges requires additional assumptions and results in additional equations for changes in wedges. Those equations can then be combined with our comparative statics, using the chain rule, to generate comparative statics (for example, along the lines of Rubbo, 2020).

Going forward, the primary object of interest is the response of aggregate output $d\log Y$ (or welfare) to productivity shocks and shocks to wedges.[8] Since the supply of primary factors is fixed, changes in aggregate output coincide with changes in aggregate productivity $d\log TFP$.[9]

[8]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 and shock their productivity. Finally, using the Arrow-Debreu trick of indexing commodities by dates and states of the world, we can capture dynamic stochastic models.

[9]We abstract away from the well-understood issues related to the treatment of new goods in the measurement of aggregate output. Therefore, in this paper, real GDP coincides with welfare.

2.6 Noteworthy Special Cases

At this level of abstraction, the model nests many 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 version of Melitz (2003), models with product variety like Dixit and Stiglitz (1977), Krugman (1979), (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 Acemoglu et al. (2012), Baqaee and Farhi (2019a), and Bigio and La'O (2016) 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 $i \in \mathcal{N}$.

Theorem 1, below, shows that the marginal cost-pricing benchmark is efficient. Therefore, it generalizes the first welfare theorem to an environment with fixed and sunk costs of operation and entry. From a normative perspective, it clarifies how the optimal allocation can be implemented using linear taxes, and we use this implementation in Section 7 when we approximate the decentralized economy's distance from the Pareto-efficient frontier.

Theorem 1 (First Welfare Theorem)**.** *The marginal-cost pricing equilibrium is Pareto-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 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.[10,11] In the next section, we derive comparative statics for shocks when the economy is inefficient.

4 Framework for Inefficient Equilibrium

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 since reallocation effects can no longer be ignored. In this section, we impose some additional assumptions on the model, and introduce input-output notation, to analyze inefficient equilibria.

4.1 Additional Assumptions

To emphasize our mechanisms of interest, we specialize the general framework. Assume that each good is *either* produced as differentiated varieties using a CES aggregator, as in Dixit and Stiglitz (1977) or Melitz (2003), or as perfect substitutes with decreasing returns, as in Hopenhayn (1992).[12,13]

Assumption 1. For each $i \in \mathcal{N} - \mathcal{F}$, either i is a CES aggregate of imperfectly substitutable varieties

$$Y_i = \left(M_i y_i^{\gamma_i}\right)^{\frac{1}{\gamma_i}}, \qquad y_i = A_i f_i\left(\left\{x_{ij}\right\}_{j \in \mathcal{N}}\right), \tag{2}$$

or i is a linear aggregate of perfectly substitutable varieties

$$Y_i = \left(M_i y_i\right), \qquad y_i = A_i f_i\left(\left\{x_{ij}\right\}_{j \in \mathcal{N}}\right)_i^{\varepsilon_i}, \tag{3}$$

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

[11]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.

[12]Appendix A discusses the relationship between these two assumptions.

[13]Appendix E relaxes Assumption 1 and extends our results to the case where internal diseconomies are non-isoelastic (allowing for variable returns to scale at the producer level) and external economies are non-isoelastic (along the lines of Kimball, 1995).

where are $\gamma_i, \varepsilon_i \in [0,1]$ and f_i has constant returns to scale.

The parameters ε_i and γ_i control internal and external returns to scale (on the margin) respectively.[14] Since goods that are produced according to (2) have increasing external returns (due to love-of-variety), we refer to them as IRS goods. On the other hand, since goods according to (3) have decreasing internal returns, we refer to them as DRS goods. Denote the set of IRS goods by $\mathcal{N}^{IRS}$ and the set of DRS goods by $\mathcal{N}^{DRS}$.[15]

The next assumption rules out corners in M_i by ensuring that markups are not so low that producer i always makes negative profits. The rent (or variable profit) of market i is

$$\lambda_{\pi,i} = \frac{P_i Y_i}{GDP}\pi_i, \quad \text{with} \quad \pi_i = \left(1 - \frac{\varepsilon_i}{\mu_i}\right)\mathbf{1}_{i\in\mathcal{N}^{DRS}} + \left(1 - \frac{1}{\mu_i}\right)\mathbf{1}_{i\in\mathcal{N}^{IRS}}. \tag{4}$$

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 and the rents due to diminishing returns.

Assumption 2. If i is a DRS good, then $\mu_i > \varepsilon_i$, and if i is an IRS good, then $\mu_i > 1$.

Our final assumption, Assumption 3, is made without loss of generality, and we make it to simplify notation. We say entry is non-overlapping when multiple entrant types do not enter into the same market.

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: $\zeta(j,i) \neq 0$.

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 $Y = F(My)$ and $M = M_{E,1} + M_{E,2}$. To turn this into a model with non-overlapping entry, create two fictitious markets $Y_i = F(M_i y_i)$ with non-overlapping entry $M_i = M_{E,i}$ for $i \in \{1,2\}$. Now create a third fictitious market, with no entry, where Y_3 aggregates Y_1 and Y_1 in the same way as F. Since $Y = Y_3$, we have recast a model with overlapping entry into an equivalent model with non-overlapping entry. We impose this assumption throughout.

[14] We normalize A_i to ensure that Y_i is unit-elastic in the productivity shock. For comparison, note that we did not impose this unit-elasticity normalization in Section 2.

[15] To map IRS goods into the framework in Section 2, note that a CES aggregate is isomorphic to an industry aggregator with increasing returns $F_i(x) = x^{1/\gamma_i}$ and a production function with decreasing returns $f_i(x) = x^{\gamma_i}$, where x is the bundle of inputs. Since the aggregators do not generate any revenue (but have curvature), they must set price equal to average costs. This implies that, for a CES model, the outer aggregator is charging an implicit markup $\mu_i^Y = 1/\gamma_i$, and the inner aggregator is charging an implicit markdown $\mu_i^y = \gamma_i$. We come back to this in Section 7, in particulare footnote 32, when we discuss optimal policy and misallocation.

4.2 Notation and Other Preliminaries

We normalize nominal GDP to one throughout, so all prices are quoted in the nominal GDP numeraire and 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 Normalized Entry Matrix

Define the $|E| \times |\mathcal{N}|$ 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 (if i is contested, then the ith column of $\tilde{\zeta}$ sums to one).

IO Matrices

We introduce the *forward* and *backward* Input-Output (IO) matrices Ω^{F} and Ω^{B} and their accompanying Leontief inverses Ψ^{F} and Ψ^{B}. Intuitively, the backward matrix encodes the 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.[16]

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

$$\Omega^V_{ij} \equiv \frac{M_i P^Y_j x_{ij}}{P^Y_i Y_i}.$$

[16] "Backward" and "forward" are due to Hirschman (1958), though he did not define them mathematically.

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

$$\Omega^E_{ij} \equiv \frac{P^Y_j x_{E,ij}}{\sum_{k \in \mathcal{N}} P^Y_k x_{E,ik}}.$$

Finally, let π be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix of profit shares defined in (4).

The *backward* IO matrix combines variable and fixed expenditures

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

Its ijth element Ω^{B}_{ji} is the fraction of the revenues of j directly paid out to i for variable 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.[17]

Forward IO Matrix. Let $\mathcal{E}$ be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix whose ith diagonal element is equal to $\mathcal{E}_{ii} = (1 - \varepsilon_i)$ if $i \in \mathcal{N}^{DRS}$ or $\mathcal{E}_{ii} = (1/\gamma_i - 1)$ if $i \in \mathcal{N}^{IRS}$. Intuitively, $\mathcal{E}_{ii}$ measures how an increase in entry in i affects the price of good i.

The forward IO matrix is defined by

$$\Omega^{\mathrm{F}} = \mu \Omega^V + \mathcal{E} \tilde{\zeta}' \Omega^E,$$

where μ is a diagonal matrix of markups. The 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

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

[17]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 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 holds Ω^B and other sales λ^B 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 holds Ω^B constant but accounts for changes in sales λ^B. As we shall see, this is equivalent to holding relative prices constant, since when relative prices are constant, Ω^{B} is also held constant.

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.[18]

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

$$\lambda^{\mathrm{B}}_i = \frac{P^Y_i Y_i}{GDP} = P^Y_i Y_i,$$

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

$$\lambda^{\mathrm{B}}_i = \sum_j \Omega^{\mathrm{B}}_{0j} \Psi^{\mathrm{B}}_{ji} = \Psi^{\mathrm{B}}_{0i}.$$

This 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

$$\lambda^{\mathrm{F}}_i = \Psi^{\mathrm{F}}_{0i} = \sum_j \Omega^{\mathrm{F}}_{0j} \Psi^{\mathrm{F}}_{ji}.$$

This 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 $\lambda^{\mathrm{B}}_i = \lambda^{\mathrm{F}}_i$, 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 $\lambda^{\mathrm{B}}_i \neq \lambda^{\mathrm{F}}_i$ and their ratio $\lambda^{\mathrm{F}}_i / \lambda^{\mathrm{B}}_i$ 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.

[18] By Shepard's lemma, the ijth 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.

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

5 Aggregation

We now generalize Theorem 2 to inefficient economies. We provide our comparative statics in two steps. First, in this section , we provide an aggregation equation that 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 propagation 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 and markups/wedges which we write in vector form as $(\mathrm{d}\log A, \mathrm{d}\log \mu)$.[20]

5.1 The Aggregation Equation

Let λ_π be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix of rents, and let $\mathrm{d}\log\lambda_\pi$ be the $|\mathcal{N}| \times 1$ vector of changes in rents.[21] Define the (rent-weighted) projection of $\mathrm{d}\log\lambda_\pi$ on the entry matrix $\tilde{\zeta}$ by

$$\widehat{\mathrm{d}\log\lambda}_\pi = \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\tilde{\zeta}\lambda_\pi\,\mathrm{d}\log\lambda_\pi. \tag{5}$$

Denote ith component of this projection by $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$. Intuitively, $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$ is the change in the *quasi-rent* associated with market i, that is, changes in rents that are dissipated by entry costs (we discuss the reason in Section 5.2).

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

$$\begin{aligned} d\log Y = &\sum_{i\in\mathcal{N}} \lambda_i^F d\log A_i - \sum_{i\in\mathcal{N}^{IRS}} \lambda_i^F d\log\mu_i - \sum_{i\in\mathcal{N}^{DRS}} \lambda_i^F \varepsilon_i d\log\mu_i \\ &- \sum_{i\in\mathcal{N}^{DRS}} \lambda_i^F (1-\varepsilon_i)\left(d\log\lambda_i^B - \widehat{\mathrm{d}\log\lambda}_{\pi,i}\right) + \sum_{i\in\mathcal{N}^{IRS}} \lambda_i^F\left(\frac{1}{\gamma_i}-1\right)\widehat{\mathrm{d}\log\lambda}_{\pi,i}. \end{aligned} \tag{6}$$

Theorem 3 is a key result of the paper, and we spend the rest of this section unpacking its intuition and working through some examples. Since nominal GDP is normalized to one, changes in real GDP are the negative of changes in the household price index

[20] An output wedge on i not rebated back to the proprietor, in our notation μ_i^Y, can be captured by adding a fictitious incumbent middleman who buys i's output and sells to the rest of the economy. A markup on this fictitious middleman is isomorphic to an output wedge on i. Therefore, comparative statics in μ encompass both output wedges and markups. In Appendix B, which contains the proofs, we explicitly distinguish between markups, μ_i, and output wedges μ_i^Y.

[21] 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 paper, in order to lighten the notation, we often use the same symbol to denote vectors and their counterparts as diagonalized matrices.

$\mathrm{d}\log Y = -\mathrm{d}\log P_0^Y$.[22]

Focus on the first line, which captures changes in consumer prices when sales and quasi-rents are held constant. The first term captures the direct effect of productivity shocks, which are weighted by their forward Domar weights. The second and third terms capture the effect of an increase in markups on consumer prices, which are weighted by the forward Domar weight of the bundle of inputs (λ_i^F for IRS goods and $\lambda_i^F \varepsilon_i$ for DRS goods).

The second line accounts for changes in sales and quasi-rents. Intuitively, the first term on the second line captures how for each DRS 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 IRS 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.

5.2 The Role of Entry

The following lemma helps clarify the role played by $\widehat{\mathrm{d}\log\lambda}_\pi$, and its relation to entry.

Lemma 1. *In equilibrium, the change in the mass of producers in each market is given by*

$$\mathrm{d}\log M = \widehat{\mathrm{d}\log\lambda}_\pi - \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\,\mathrm{d}\log P_E, \tag{7}$$

where $\mathrm{d}\log M$ *is the* $|\mathcal{N}|\times 1$ *vector of changes in masses of producers,* λ_E *is the* $|E|\times|E|$ *diagonal matrix of expenditures on entry, and* $\mathrm{d}\log P_E$ *is the* $|E|\times 1$ *vector of changes in entry prices.*[23]

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 minimize the new 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

[22]Recall that $\mathrm{d}\log\lambda_\pi = \mathrm{d}\log\lambda^B + \mathrm{d}\log\pi$, hence, $\widehat{\mathrm{d}\log\lambda}_\pi = \widehat{\mathrm{d}\log\lambda^B} + \widehat{\mathrm{d}\log\pi}$. In words, changes in quasi-rents are either due to changes in sales or due to changes in profit margin. Since $\mathrm{d}\log\pi_i = (1-\pi_i)/\pi_i\,\mathrm{d}\log\mu_i$, changes in profit margin are zero unless markup μ_i is perturbed.

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

by those entrants who go on to become producers of type i.[24] Changes in quasi-rents in market i can depend on the changes in rents $\mathrm{d}\log\lambda_{\pi,j}$ in all markets $j \in \mathcal{N}$.

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.

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 $\widehat{\mathrm{d}\log\lambda}_{\pi,i} = 0$.

5.3 The Role of Reallocation

Theorem 3 provides an interpretable decomposition of changes in output into changes in technical and allocative efficiency along the lines of Baqaee and Farhi (2019a). To see this, 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 $i \in \mathcal{N} + E$. Together with the vector of productivity shifters A, the allocation matrix pins down the whole allocation, and hence aggregate output $Y(A, \mathcal{X})$.

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

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

where the first term is the direct effect of changes in technology, holding the allocation of

[24]To see this, note that (7) can be rearranged to $\tilde{\zeta}\lambda_\pi\widehat{\mathrm{d}\log\lambda}_\pi = \tilde{\zeta}\lambda_\pi\,\mathrm{d}\log M + \tilde{\zeta}\lambda_\pi\tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\,\mathrm{d}\log P_E = \lambda_E d\log\lambda_E = d\lambda_E$. Since λ_E is the vector of quasi-rents, $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$ is the change in quasi-rents that can be attributed to market i.

resources constant, and the second term is the indirect effect of equilibrium reallocations

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

Proposition 1 breaks Theorem 3 into two components.

Proposition 1 (Decomposition with Inefficiencies)**.** *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log \mu)$*, changes into aggregate output can be decomposed in changes in technical efficiency*

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

and changes in allocative efficiency

$$\begin{aligned}\frac{\partial \log Y}{\partial \mathcal{X}}\,\mathrm{d}\,\mathcal{X} = &-\sum_{i\in\mathcal{N}^{IRS}} \lambda_i^F d\log\mu_i - \sum_{i\in\mathcal{N}^{DRS}} \lambda_i^F \varepsilon_i d\log\mu_i \\ &-\sum_{i\in\mathcal{N}^{DRS}} \lambda_i^F (1-\varepsilon_i)\left(d\log\lambda_i^B - \widehat{\mathrm{d}\log\lambda}_{\pi,i}\right) + \sum_{i\in\mathcal{N}^{IRS}} \lambda_i^F\left(\frac{1}{\gamma_i}-1\right)\widehat{\mathrm{d}\log\lambda}_{\pi,i}.\end{aligned} \tag{8}$$

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, and the household's exposure in prices Ψ_{0i}^F to each good i is given by λ_i^F not λ_i^B.

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 (8) is zero. This leaves two terms on the second line.

The first term depends on decreasing internal returns to scale $1-\varepsilon_i$. When $\mathrm{d}\log\lambda_i^{\mathrm{B}} - \widehat{\mathrm{d}\log\lambda}_{\pi,i} > 0$, 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 $\lambda_i^F(1-\varepsilon_i)$ of these specific fixed factors.[25,26] When

[25]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.

[26]Recall that primary factors $f\in\mathcal{F}\subset\mathcal{N}$ are captured as producer-specific fixed factors in factor markets

$\mathrm{d}\log\lambda_i^B - \widehat{\mathrm{d}\log\lambda}_{\pi,i} = 0$, decreasing returns to scale do not matter since adjustments in market size are taking place along the extensive margin (individual producers are not change their scale).

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.

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 $\varepsilon_i, \gamma_i < 1$ for all $i \in \mathcal{N}$, changes in technical and allocative efficiency are given by*[27]

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

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.

5.4 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).[28] Consider the following cases: (CRS) when all goods are produced with constant-returns to scale so that for every $j \in \mathcal{N} - \mathcal{F}$ either $\varepsilon_j = 1$ or $\gamma_j = 1$;

with zero-returns-to-scale individual producers ($1-\varepsilon_f = 1$) aggregated linearly with no entry ($\widehat{\mathrm{d}\log\lambda}_{\pi,f} = 0$).

[27]The assumption that $\varepsilon_i, \gamma_i < 1$ 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 J for more details.

[28]The intuition for a shock to markups/wedges is similar, but for brevity, we relegate this discussion to Appendix K.

(DRS) where all goods are produced with decreasing returns $\mathcal{N} - \mathcal{F} = \mathcal{N}^{DRS}$; and (IRS) where all goods are produced with increasing returns $\mathcal{N} - \mathcal{F} = \mathcal{N}^{IRS}$.

Productivity shocks with CRS. When $\varepsilon_j = 1$ or $\gamma_j = 1$ 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 that primary factor markets $f \in \mathcal{F}$ are produced with zero returns to scale ($\varepsilon_f = 0$), and are uncontested (no entry) ($\mathrm{d}\log \lambda_f^{\mathrm{B}} - \widehat{\mathrm{d}\log \lambda}_{\pi,f} = \mathrm{d}\log \lambda_f^{\mathrm{B}}$).

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. This happens if reallocations, by making better use of resources, make the primary factors relatively cheaper (less scarce).[29]

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

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

There are decreasing internal economies but no increasing external economies. 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 $\lambda_j^F (1 - \varepsilon_j)$ 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,

[29] An easy special case 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.

when there are no markups, $\lambda_{\pi,j} = \lambda_j^F(1-\varepsilon_j)$ and so

$$\sum_{j\in\mathcal{N}} \lambda_j^F(1-\varepsilon_j)(d\log\lambda_{\pi,j} - \widehat{\mathrm{d}\log\lambda}_{\pi,j}) = \sum_{j\in\mathcal{N}} \lambda_{\pi,j}(d\log\lambda_{\pi,j} - \widehat{\mathrm{d}\log\lambda}_{\pi,j}) = 0,$$

because the weighted sum of residuals must be zero.

When there is directed entry ($|E| = |N|$), this expression simplifies further to just

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

If there is only one primary factor, then $\mathrm{d}\log\lambda_f^{\mathrm{B}} = 0$, and $d\log Y = \lambda_i^F d\log A_i$, so that there are 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, even though the equilibrium may be inefficient, reallocations happen entirely on the extensive margin of entry and exit and offset one another.

Productivity shocks with IRS. When all non-primary factor markets are IRS, Theorem 3 becomes

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

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).

6 Propagation

Theorem 3 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. 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 and quantities.

6.1 Forward Propagation

We start by describing the response of prices to shocks.

Proposition 2 (Forward Propagation). *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log\mu)$*, changes in prices are given by*

$$d\log P_i^Y = -\sum_{j\in\mathcal{N}} \Psi_{ij}^F d\log A_j + \sum_{j\in\mathcal{N}^{IRS}} \Psi_{ij}^F \,\mathrm{d}\log\mu_j + \sum_{j\in\mathcal{N}^{DRS}} \Psi_{ij}^F \varepsilon_j \,\mathrm{d}\log\mu_j$$
$$+ \sum_{j\in\mathcal{N}^{DRS}} \Psi_{ij}^F \left(1-\varepsilon_j\right)\left(d\log\lambda_j^B - \widehat{\mathrm{d}\log\lambda}_{\pi,j}\right) - \sum_{j\in\mathcal{N}^{IRS}} \Psi_{ij}^F \left(\frac{1}{\gamma_j}-1\right)\widehat{\mathrm{d}\log\lambda}_{\pi,j}$$

Proposition 2 is similar to Theorem 3. In fact, since nominal GDP is normalized to one, changes in real output are just the negative of the changes in the consumer price index $\mathrm{d}\log Y = -\mathrm{d}\log P_0$. Therefore, Proposition 2 can be specialized to yield Theorem 3 by setting i to be the price of the final consumption good 0. Therefore, the intuition for Proposition 2 is similar to the one for Theorem 3.[30]

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* $(\mathrm{d}\log A, \mathrm{d}\log\mu)$*, changes in sales, rents, and profit margins are related through*

$$\mathrm{d}\log\lambda_{\pi,i} = \mathrm{d}\log\lambda_i^B + \mathrm{d}\log\pi_i, \quad \textit{where} \quad \mathrm{d}\log\pi_i = \frac{1-\pi_i}{\pi_i}\,\mathrm{d}\log\mu_i.$$

[30] An interesting special case of Proposition 2 is when every good is DRS, there is only one primary factor, and entry is fully-directed. In this case, the change in prices simplifies to

$$d\log P_i^Y = -\sum_{j\in\mathcal{N}} \Psi_{ij}^{\mathrm{F}} d\log A_j + \sum_{j\in\mathcal{N}^{DRS}} \Psi_{ij}^{\mathrm{F}} \left(1 - \frac{(1-\varepsilon_j)}{\pi_j}\right)\mathrm{d}\log\mu_j.$$

In other words, the change in relative prices does not depend on final demand or on the elasticities of substitution in production. This is reminiscent of the no-substitution theorem (Georgescu-Roegen,1951; Samuelson, 1951). However, it holds under different assumptions: in particular, unlike the classic no-substitution theorem, one does not need to assume constant returns to scale nor perfect competition.

Hence, changes in rents in each sector are driven, either by changes in sales $d\log\lambda^B$ or changes in profit margins $d\log\pi$. Therefore, it is enough to characterize changes in sales $\mathrm{d}\log\lambda_i^{\mathrm{B}}$ since we can then immediately obtain changes in rents $\mathrm{d}\log\lambda_{\pi,i}$ and quasi-rents $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$ from Lemma 2 and equation (5).

For simplicity, we assume that all production and entry functions in the economy f_i and g_i are CES production functions. Without loss of generality (by relabelling the input-output network), we can assume that each CES production function i has a single elasticity of substitution θ_i associated with it.[31] We make this assumption for clarity, not tractability, and Appendix C generalizes our results to non-CES production functions. We also assume that there are perfectly competitive incumbents that produce the goods required for paying the entry cost, and these perfectly competitive incumbents are added to the input-output network as additional "producers" whose outputs are only used by entrants.

To state our results, we use the *input-output covariance operator*:

$$Cov_m(X,\Psi^{\mathrm{B}}_{(i)}) = \sum_{k\in\mathcal{N}}(1-\pi_m)^{-1}\Omega^V_{mk}X_k\Psi^{\mathrm{B}}_{ki} - \left(\sum_{k\in\mathcal{N}}(1-\pi_m)^{-1}\Omega^V_{mk}{\Psi^{\mathrm{B}}}_{ki}\right)\left(\sum_{k\in\mathcal{N}}(1-\pi_m)^{-1}\Omega^V_{mk}X_k\right),$$

where ${\Psi^{\mathrm{B}}}_{(i)}$ is the ith column of the backward Leontief inverse Ψ^{B}. This is the covariance between the vector X and the ith column of Ψ^{B}, using the mth row of $(1-\pi)^{-1}\Omega^V$ as the probability distribution. We call it a covariance since $\sum_{k\in\mathcal{N}}(1-\pi_m)^{-1}\Omega^V_{mk} = 1$ for $m\in\mathcal{N}-\mathcal{F}$.

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

$$d\lambda_i^B = -\sum_{m\in\mathcal{N}}\lambda_m^B\sum_{k\in\mathcal{N}}\left[\Omega^V_{mk}-(1-\pi_m)\sum_{j\in E}\tilde{\zeta}_{jm}\Omega^E_{jk}\right]\Psi^B_{ki}d\log\mu_m - \sum_{m\in\mathcal{N}}\lambda^B_m(1-\pi_m)(\theta_m-1)Cov_m\left(d\log P^Y,\Psi^B_{(i)}\right).$$

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

The first term is the effect of changes in markups/wedges on the demand for i holding fixed relative prices. The term in square brackets is how an increase in m's markup $\mathrm{d}\log\mu_m > 0$ affects spending on some input k. On the one hand, a higher markup reduces m's variable spending on input k by $\lambda_m^B\Omega^V_{mk}$. On the other hand, a higher markup increases

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

entry, and this increases spending on k by entrant j by $\lambda^B_m(1-\pi_m)\tilde{\zeta}_{jm}\Omega^E_{jk}$. These two effects in turn change spending on i in proportion to the exposure Ψ^B_{ki} of k to i.

The term on the second line captures the effect of substitutions on the intensive margin. Changes in relative prices $\mathrm{d}\log P^Y$ caused by the shocks lead individual producers in every market $m \in \mathcal{N}$ to shift their expenditures on their inputs. If $\theta_m > 1$, then m's inputs are gross substitutes. Hence, m substitutes its expenditures towards those inputs that have become relatively cheaper. If those inputs intensively rely on i, then $(\theta_m - 1)Cov_m(\mathrm{d}\log P, \Psi^{\mathrm{B}}_{(i)})$ is negative. Hence, substitution by m changes i's sales in proportion to $-\lambda^{\mathrm{B}}_m(1-\pi_m)(\theta_m - 1)Cov_m(\mathrm{d}\log P, \Psi^{\mathrm{B}}_{(i)})$.

Combining Forward and Backward Propagation. Proposition 2 can be plugged into Proposition 3 to give a linear system in $\mathrm{d}\log\lambda$. The solution pins down changes in sales shares in every market, and these can then be plugged back into Theorem 3 for welfare changes. In Appendix G, we apply these Propositions to some specific examples, and show how the framework nests many popular models in the literature.

7 Optimal Policy and Misallocation

In this section, we turn to optimal policy. We describe first-best policies when instruments are unrestricted. Then we characterize the gains from implementing optimal policy (i.e. the distance to the efficient frontier). Finally, we consider second-best policies when only limited instruments are available.

7.1 First-Best Policy

Theorem 1 implies that the first best is attained when $\mu^Y_i = \mu_i = 1$ for all $i \in \mathcal{N}$. Consider a planner with access to unrestricted linear taxes. Then applying Theorem 1 to the class of economies that satisfy Assumption 1 results in the following.

Corollary 2. *The decentralized equilibrium is efficient if in each market $i \in \mathcal{N}$, the planner introduces output taxes τ^Y_i and τ^y_i on industry-level and firm-level output satisfying*

$$\tau^Y_i = \mathbf{1}_{\{i\in\mathcal{N}^{DRS}\}} + \gamma_i \mathbf{1}_{\{i\in\mathcal{N}^{IRS}\}}, \qquad \tau^y_i = \frac{1}{\mu_i}\mathbf{1}_{\{i\in\mathcal{N}^{DRS}\}} + \frac{1}{\gamma_i\mu_i}\mathbf{1}_{\{i\in\mathcal{N}^{IRS}\}},$$

where revenues collected by the industry-level tax are paid to the household, and revenues collected by the firm-level tax are paid to producers of the good.

Corollary 2 shows 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 in that market. In particular, for DRS markets, optimal policy ensures marginal cost pricing. For IRS markets, optimal policy sets the markup charged by each producer equal to $1/\gamma_i$, which is the Dixit and Stiglitz (1977) markup, to incentivize entry and subsidises output by γ_i to offset markup.[32]

7.2 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 the importance of accounting for entry.

For any equilibrium variable X, we denote by $\mathrm{d}\log X$ the log-deviation of X from its value at the efficient allocation, which can also be thought of as the change in X caused by the deviations of $\mathrm{d}\log\tau_i$ and $\mathrm{d}\log\tau_i^Y$ of the firm-level and industry-level output wedges from their efficient values in Corollary 2. We provide a second-order approximation in these deviations $(\mathrm{d}\log\tau, \mathrm{d}\log\tau^Y)$ of the associated aggregate efficiency loss $\mathcal{L} = -(1/2)\,\mathrm{d}^2\log Y$.[33]

Proposition 4 (Deadweight-Loss). *As long as $\varepsilon_i, \gamma_i < 1$, the efficiency loss can be approximated, up to second-order approximation, as*

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

Hence, the social cost of distortions is, up to a second-order approximation, a Domar-weighted sum of Harberger triangles associated with variable production and entry. In

[32] At first glance, it may seem that Corollary 2 contradicts the marginal cost pricing result in Theorem 1, since for IRS markets, the firms do not seem to be marginal cost pricing. To see that Corollary 2 is a consequence of Theorem 1, one needs to recognize that an IRS industry with production function $Y_i = \left(M_i y_i^{\gamma_i}\right)^{1/\gamma_i}$ is equivalent to $Y_i = F_i(M_i f_i(y_i))$, where $F_i(x) = x^{1/\gamma_i}$ and $f_i(x) = x_i^{\gamma}$ are as in Section 3. Since aggregators F_i and f_i have curvature but generate no income, this means that they implicitly charge price equal to average cost (as opposed to marginal cost). Therefore, to restore efficiency, the planner must introduce taxes that undo these implicit markups and markdowns.

[33] By Corollary 1, around the efficient point, the first-order loss is zero as long as $\varepsilon_i, \gamma_i < 1$. If either $\varepsilon_i = 1$ or $\gamma_i = 1$, and there is entry into i, then the losses are first-order, and we must instead use Theorem 3.

conjunction with the forward and backward propagation equations in Propositions 2 and 3, we can rewrite these loss functions in terms of microeconomic primitives (the input-output matrix, the elasticities of substitution, and returns to scale).[34] We relegate this general formula to Appendix B, and focus on a few prominent examples obtained by considering a special class of models with a sectoral structure.

7.2.1 Sectoral Models

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

1. every non-primary-factor-market $i \in \mathcal{N} - \mathcal{F}$ is assigned to a unique sector $\mathcal{I}$, with common returns to scale so that its output matters only through sectoral output. Sectoral output is

$$Y_{\mathcal{I}} = \sum_{i \in \mathcal{I}} y_i^{\varepsilon_{\mathcal{I}}}, \qquad \text{or} \qquad Y_{\mathcal{I}} = \left(\sum_{i \in \mathcal{I}} y_i^{\gamma_{\mathcal{I}}} \right)^{\frac{1}{\gamma_{\mathcal{I}}}},$$

   depending on whether $\mathcal{I}$ is DRS or IRS;[35]

2. individual producers i in sector $\mathcal{I}$ have the same production function $y_i = A_i f_{\mathcal{I}}(\{x_{i\mathcal{J}}\})$, 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 $i \in \mathcal{I}$ according to some fixed distribution;

4. individual producers i in sector $\mathcal{I}$ charge different markups τ_i^y but share common industry-level wedges $\tau_i^Y = \tau_{\mathcal{I}}^Y$.

Sectoral models, common in the literature, are worth singling out because their within-sector heterogeneity can be aggregated. For sectoral models, we can break the problem of computing the distance to the frontier into two blocks, within and across sectors. See Appendix D 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 $\lambda_i^{\mathcal{I},\mathrm{B}} = \lambda_i^{\mathrm{B}} / \lambda_{\mathcal{I}}^{\mathrm{B}} \mathbf{1}_{\{i \in \mathcal{I}\}}$. We will denote by

[34] To do this, note that $d\log Y_i = d\log \lambda_i^B - d\log P_i$, where Proposition 2 gives $d\log P_i$ and Proposition 3 gives $d\log \lambda_i^B$. Next, observe that $d\log Y_i = d\log y_i + 1/\gamma_i \, d\log M_i$ if i is IRS and $d\log Y_i = d\log y_i + d\log M_i$ if i is DRS. Finally, note that $d\log M$ is given by Lemma 1. Putting this altogether will allow us to write Proposition 4 in terms of primitives.

[35] Formally, to capture this, we assume there is an incumbent producer indexed by $\mathcal{I}$ that aggregates the inputs of different markets i and prices at average cost (thereby, making no profits).

$\mathbb{E}_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\tau^y)$ and $Var_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\tau^y)$ the within-sector weighted expectations and variances of changes in markups/wedges $\mathrm{d}\log\tau_i^y$ of producers $i\in\mathcal{I}$ with weights $\lambda_i^{\mathcal{I},\mathrm{B}}$.

7.2.2 Sectoral DRS Example

For sectoral models, we can provide a straightforward characterization of the loss function with DRS. 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 $\mathrm{d}\log\tau_{\mathcal{I}}^Y = 0$.

Proposition 5. *Consider a sectoral model where every sector is DRS, there is only one primary factor, entry is paid in units of the factor, and there are no deviations of output wedges from their efficient benchmarks* $\mathrm{d}\log\tau_{\mathcal{I}}^Y = 0$*. To a second order, the loss function is given by*

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

Because there are no output wedges, we know from Proposition 4 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 5 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. Intuitively, this is because the wedges do not cause cross-sectoral misallocation. 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.

7.2.3 Sectoral IRS Examples

Consider a sectoral model with IRS sectors. We denote by $\theta_{\mathcal{I}} = 1/(1-\gamma_{\mathcal{I}})$ the elasticity of substitution associated with every sector $\mathcal{I}$. In other words, we can think of $\theta_{\mathcal{I}}$ as the elasticity of substitution across different varieties i in 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 are 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 a one-sector model heterogenous-firm economy with IRS and free entry. Aggregate output is given by

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

Each good k is produced from labor with constant returns and productivity A_k. The aggregate efficiency loss is, to a second-order, given by

$$\mathcal{L} = \frac{1}{2}\theta Var_{\lambda^{\mathrm{B}}}\left(d\log\tau^{y}\right) + \frac{1}{2}\theta\mathbb{E}_{\lambda^{\mathrm{B}}}\left(d\log\tau^{y}\right)^{2}.$$

The first term captures misallocation on the intensive margin and comes from the fact that high-markup firms are too small and low-markup firms are too big. This term depends on the elasticity of substitution and on the dispersion of markups, and is standard in the literature (see e.g. Hsieh and Klenow, 2009; Baqaee and Farhi, 2019a). The second term captures misallocation on the extensive margin and comes from the fact that there is too much or too little entry. This term depends on the elasticity of substitution and on the level of wedges, and is new to the literature.[36]

[36] If instead of inefficient markups τ_i^y, we considered inefficient output wedges τ_i^Y instead, then the extensive margin would be unaffected and we would only have misallocation on the intensive margin, leading to $\mathcal{L} = \frac{1}{2}\theta Var_{\lambda^{\mathrm{B}}}\left(d\log\tau^{Y}\right)$. This 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. Now, consider a multi-sector economy where output is an aggregate of different sectors Y_k indexed by k with elasticity of substitution θ_0

$$Y^{IRS} = \left(\sum_{k=1}^{N} Y_k^{\frac{\theta_0 - 1}{\theta_0}} \right)^{\frac{\theta_0}{\theta_0 - 1}},$$

where each industry's output is itself a CES aggregate of product varieties in industry k

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

Here, M_k is the total mass of varieties of k. We assume that the within-sector elasticity of substitution θ_k is greater than the between-sector one θ_0. Each good in each industry is produced linearly using labor $y_k = A_k l_k$. Labor is the only primary factor, there is free entry into each sector, and entry costs are paid in units of labor.

Applying Proposition 4, the aggregate efficiency loss is

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

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, Proposition 4 implies the aggregate efficiency loss function is

$$\mathcal{L} = \frac{1}{2} \left[\theta_0 Var_{\lambda^{\mathrm{B}}} \left(\mathbb{E}_{\lambda^{\mathcal{I},\mathrm{B}}} (d \log \tau^y) \right) + \sum_{\mathcal{I}} \theta_{\mathcal{I}} Var_{\lambda^{\mathcal{I},\mathrm{B}}} \left(d \log \tau^y \right) \right],$$

where the first variance is the weighted variance of sectoral markups with weights given by sectoral sales shares $\lambda_{\mathcal{I}}^{\mathrm{B}}$, and the last set of variances are within-sector variances weighted by within-sector sales shares $\lambda_i^{\mathcal{I},\mathrm{B}}$. Whereas $\theta_{\mathcal{I}}$ controls losses from within-sector dispersion, the cross-sectoral elasticity of substitution θ_0 controls the extent to which sectoral misallocation reduces output. Since there is no entry, the level of wedges is no longer relevant, only their dispersion.

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

Roundabout Economy. We finish with an economy that has intermediate inputs. Consider a roundabout-entry economy with one sector populated by homogenous firms and free entry with entry costs requiring the use of both labor and goods. Aggregate output

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

is the output of sector 1 not used for entry. Gross output is given by

$$Y_1 = \left(M_1 y_1^{\frac{\theta_1 - 1}{\theta_1}} \right)^{\frac{\theta_1}{\theta_1 - 1}},$$

where the representative producer has a production function $y_1 = A_1 l_1$ that transforms labor into goods linearly.

Suppose there is free entry and potential entrants pay a fixed entry cost that relies on both goods and factors. In particular, the entry good is produced from labor and products and sold at marginal cost

$$Y_2 = (l_2)^{\Omega^V_{2L}} (x_{21})^{1-\Omega^V_{2L}}.$$

When entry only uses labor $\Omega^V_{2L} = 1$, applying Proposition 4, the aggregate efficiency loss from markups is

$$\mathcal{L} = \frac{1}{2}\theta_1 (d \log \tau^y_1)^2.$$

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 zero as θ_1 goes to infinity. In this limit, entry is socially wasteful and the losses from any amount of entry are first-order (which is why the second-order approximation explodes).

Next, suppose that entry uses only products $\Omega^V_{2L} = 1$. Applying Proposition 4, the aggregate efficiency loss from markups is

$$\mathcal{L} = \frac{1}{2} \frac{(\theta_1 - 1)^3}{(\theta_1 - 2)^2} (d \log \tau^y_1)^2.$$

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.[37]

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

7.3 Second-Best Policy

Whereas the first-best policy is network-independent, second-best policies do depend on the details of the network. This section provides bang-for-buck formulas to compare the merits of different marginal interventions. These formulas revive 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, the shape of the production network, and the strength of scale economies.

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 $\mathrm{d}\log\mu_i < 0$ of the producers of market i. This can be achieved by placing a subsidy on i and taxing owners of i to fund the subsidy. Applying Theorem 3, the response of aggregate output, normalized by the revenues $-\lambda_i^{\mathrm{B}}\,\mathrm{d}\log\mu_i > 0$ transfered away from the producers by the associated implicit subsidy, is

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

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

[37] In Appendix G.5.1, we show that this U-shaped pattern also arises with input-output linkages in variable production rather than entry.

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 is lower or higher than the infra-marginal surplus created by new varieties $1/\gamma_i$.

Under Dixit and Stiglitz (1977) monopolistic competition, the direct effects exactly cancel $\mu_i = 1/\gamma_i$, leaving us the second term

$$-\frac{1}{\lambda_i^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_i} = \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j}-1\right)\left(-\frac{1}{\lambda_i^{B}}\frac{\mathrm{d}\widehat{\log \lambda_j^{\mathrm{B}}}}{\mathrm{d}\log \mu_i}\right). \tag{9}$$

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 interacted with increasing returns to scale $1/\gamma_j - 1$.[38]

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 $\mu_{E,i} = 1$. 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 quasi-rents of type i entrants, or by the zero profit condition, 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 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

$$-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_{E,i}^{Y}} = \left(\frac{\lambda_{E,i}^{\mathrm{F}}}{\lambda_{E,i}^{\mathrm{B}}} - \frac{\lambda_{L}^{\mathrm{F}}}{\lambda_{L}^{\mathrm{B}}}\right) + \sum_{j\in\mathcal{N}-\mathcal{F}} \lambda_j^{\mathrm{F}}\left(\frac{1}{\gamma_j}-1\right)\left(-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\mathrm{d}\widehat{\log \lambda_j^{\mathrm{B}}}}{\mathrm{d}\log \mu_{E,i}^{Y}}\right), \tag{10}$$

where, at the no-intervention equilibrium, the sales share of labor $\lambda_L^{\mathrm{B}} = 1$ since all markets

[38] Since the sum of backward linkages $\sum_i \lambda_i^{B}$ is a natural measure of intermediate input use. There is a sense in which the best marginal intervention acts to encourage intermediate input usage.

are contested (there are no pure profits), 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 (10). The second term is exactly the same as that for measuring the bang-for-buck impact of markup regulations in equation (9) 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 entry $\lambda_{E,i}^{\mathrm{F}}/\lambda_{E,i}^{\mathrm{B}}$ where the intervention takes place, and that of labor $\lambda_L^{\mathrm{F}}/\lambda_L^{\mathrm{B}}$.[39] The ratio of forward to backward Domar weights i measures the reduction in the size i caused by cumulated wedges 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 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 raises final-demand prices.

In Appendix L, we apply these formulas to a Cobb-Douglas economy, showing that ceteris paribus, markup policies should lower markups in industries that are downstream and entry subsidies should subsidize entry into industries that are upstream.

8 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 H.

8.1 Description of Quantitative Model

The 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 with 66 industries, 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.

[39] This first term is also related to Liu (2017), who studied marginal interventions around the decentralized equilibrium of a production network economy with constant returns and without entry.

We adopt the baseline estimates of De Loecker et al. (2019) to obtain firm-level markups using a production function estimation approach. In Appendix I, we perform robustness checks by recomputing our results using three alternative methods for estimating markups: an alternative implementation of the production function estimation approach with different categories of costs (including SG&A in variable costs, as in Traina, 2018), and alternative approaches that compute markups by netting out the cost of capital from gross surplus. Although the numbers depend on the specific approach, the qualitative message that accounting for entry and returns to scale is very important 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 $\varepsilon_{\mathcal{I}}$ or $\gamma_{\mathcal{I}}$ depending on whether we assume the industry is 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 $\gamma_{\mathcal{I}} = \gamma$ and consider two scenarios: (1) every sector is assumed to be IRS with scale elasticity γ; (2) every sector is assumed to be DRS with scale elasticity ε. In either case, we consider two different scale elasticities, in the DRS case, we set $\varepsilon = 0.875$ or $\varepsilon = 0.75$. In the IRS case, we set $\gamma = 0.875$ or $\gamma = 0.75$, which corresponds to a within-industry elasticity of substitution of 8 or 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.

8.2 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 benchmark calibration shows that the losses from inefficiency are higher (roughly double) when we allow entry than when we do not, refuting the notion that endogenizing entry necessarily reduces the social cost of markups.

Decomposing the Results. For each calibration, Table 1 decomposes the sources of the distance to the frontier. 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 rescales the level of markups in the data so that their harmonic average within each sector is equal to the efficient level (so sectoral markups are equal to the Dixit and Stiglitz (1977) markups when we adopt the IRS benchmark and equal to one when we adopt the DRS benchmark) but keeps their dispersion constant.

IRS, $\gamma = 0.875$	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, $\gamma = 0.75$			
Level only	4.6%	17%	20%
Dispersion only	22%	23%	20%
Benchmark	19%	32%	37%
DRS, $\varepsilon = 0.875$			
Level only	1.5%	7.8%	7.6%
Dispersion only	23%	23%	23%
Benchmark	26%	35%	32%
DRS, $\varepsilon = 0.75$			
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 $\gamma = 0.875$ under IRS corresponds to elasticity of substitution across firms within sectors equal to 8, whereas $\gamma = 0.75$ corresponds to elasticity of substitution equal to 4.

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 of substitution 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 and γ are lower), the gains tend to be higher when entry also uses intermediates.

The efficiency losses are different in the IRS benchmark than in the DRS benchmark. This is because quantities are less elastic in the DRS economy and entry distortions are less costly. To understand the latter point, it is useful to think about the limit where ε and γ go to zero, 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 because love of variety becomes extreme and so do the distortions in entry, as can be seen in Proposition 4. By contrast, under DRS, the efficiency losses go to zero as made clear by Propositions 4 and 5.

Comparison to Simplified Models. 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 by Jones (2011) or 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 intermediate 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

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 $\varepsilon = 0.875$ under DRS, and $\gamma = 0.875$ under IRS. For IRS, this corresponds to an elasticity of substitution across firms within industries of 8.

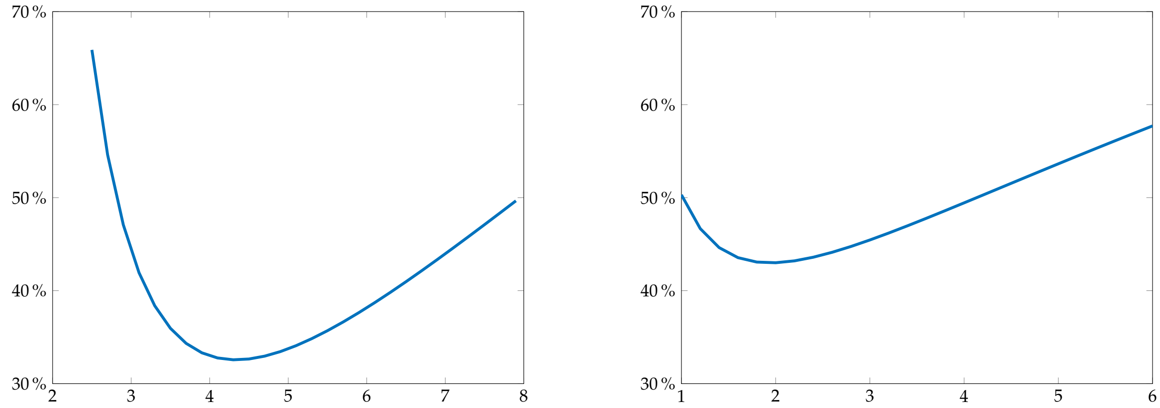
producers by more, since firms in two different industries are treated as if they are highly substitutable.

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 quantitative importance of modelling and measuring the details as best we can.

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 7.2, this intuition fails when there is entry and markets are of the IRS type.

Figure 2a shows that for the IRS benchmark, the distance to the frontier is U-shaped as a function of the within-sector elasticity of substitution $1/(1-\gamma)$. For instance, the losses are 50% when $1/(1-\gamma) = 8$. This number falls to 32% when the elasticity is lowered to 4, before rising to close to 65% when the elasticity is lowered further to 2.5. This is consistent with the theoretical discussion in the last example of Section 7.2. 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



(a) Elasticity of substitution

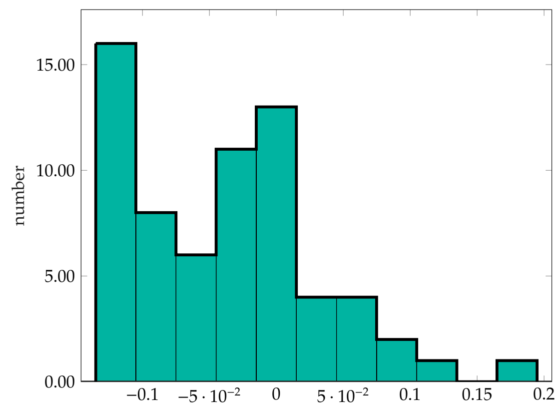
(b) Entry wedges

Figure 2: Efficiency losses for the benchmark IRS model when entry uses factors as a function of the within-industry elasticity of substitution and entry wedges for the benchmark IRS model.

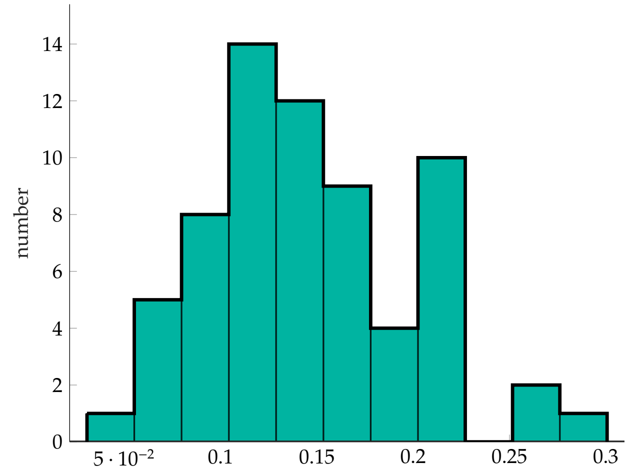
where the elasticity goes to one (γ goes to zero), misallocation along the extensive margin becomes infinitely costly.

Role of Barriers to Entry. In our benchmark specifications with entry, we assume that no friction interferes with 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.

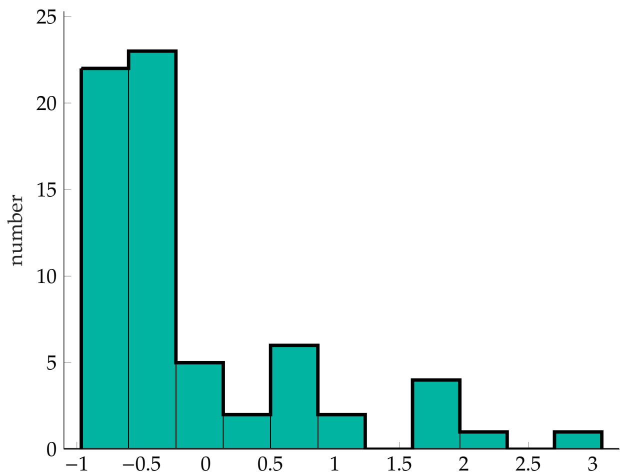
Figure 2b displays the estimated distance to the frontier as a function of the view that one takes on the size of entry barriers in the data, where the size of entry barriers are measured by the size of the implicit entry tax/wedge (a value of one means that there are no barriers to entry). Perhaps surprisingly, 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.



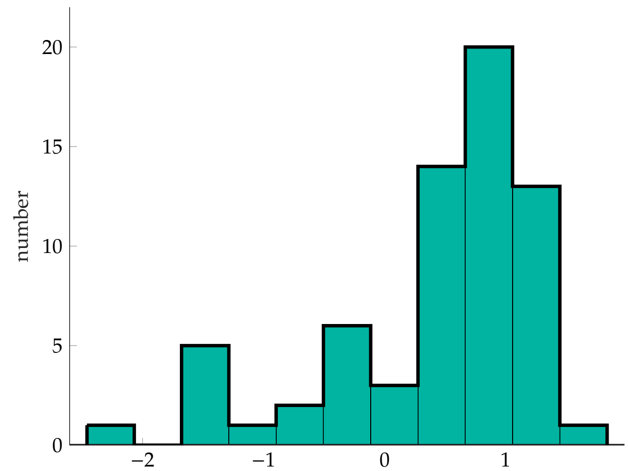
(a) Entry subsidies for CES markups



(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.

8.3 Bang for Buck of Marginal Policy Interventions

We end this section by considering 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 7.3, to make the magnitudes comparable.

For this exercise, we focus on the IRS case where $\gamma = 0.875$. 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 estimated values. We begin by discussing the case where all markups are set equal to their CES Dixit and Stiglitz (1977) values. Then we discuss the case where markups are equal to their estimated values in the data.

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, 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 comes from reducing markups for those sectors 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, on the other hand, 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.[40]

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, note that the elasticities in the top row are an order of magnitude smaller than the elasticities in the bottom row of Figure 3.

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, some of them are poisoned because 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.

9 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

[40]For more intuition about this, in Appendix L, we work through a Cobb-Douglas example.

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 — that is, productivity shocks. In the data, these exogenous changes, or total factor productivity shocks, are responsible for a large fraction of both the cycle and the trend in aggregate output.

For inefficient economies, this first-order approach is untenable. In a disaggregated economy, where many different margins can be misallocated, total factor productivity is endogenous and affected, to a first order, by reallocation effects. Furthermore, unlike exogenous productivity shocks, which are likely gradual and positive, reallocation effects can be abrupt and have either sign. This paper shows that these reallocation effects can be very potent in the presence of non-convexities and entry, and provides a framework for studying them.

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.

Bigio, S. and J. La'O (2016). Financial frictions in production networks. Technical report.

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

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.

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. *Econo-*

metrica, 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.

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

Jones, C. I. (2013). Input-Output economics. In *Advances in Economics and Econometrics: Tenth World Congress*, Volume 2, pp. 419. Cambridge University Press.

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.

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.

Rubbo, E. (2020). Networks, phillips curves and monetary policy. Technical report, mimeo, Harvard University.

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.

Traina, J. (2018). Is aggregate market power increasing? production trends using financial statements. Technical report.

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.