NBER WORKING PAPER SERIES

ENTRY VS. RENTS:
AGGREGATION WITH ECONOMIES OF SCALE

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 December 2021

Emmanuel Farhi tragically passed away in July, 2020. He was a once-in-a-lifetime friend and collabora-tor. We thank Pol Antras, Natalie Bau, Ariel Burstein, Swati Dhingra, Elhanan Helpman, Hugo Hopenhayn, Oleg Itskhoki, Xavier Gabaix, Marc Melitz, Alireza Tahbaz-Salehi, Felix Tintelnot, and especially Jonathan Vogel for comments. Veronica De Falco provided exceptional research assistance. We acknowledge research financial support from the Ferrante fund at Harvard University and NSF grant #1947611. 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: Aggregation with Economies of Scale
David Baqaee and Emmanuel Farhi
NBER Working Paper No. 27140
May 2020, Revised December 2021
JEL No. E0,E1,E3,O0,O11,O21,O25,O3,O4,O41

ABSTRACT

We characterize the response of aggregate output to micro shocks in disaggregated economies with entry, non-constant returns to scale, input-output linkages, and distortions. We decompose output changes into technical and allocative efficiency components, and show that the latter depends on changes in rents and quasi-rents across markets. We use this to characterize the social costs of distortions and show the importance of accounting for entry both qualitatively and quantitatively. As an example, we show that the efficiency losses caused by markups in the US rise from around 20% of GDP to around 40% once we account for the entry margin. Our baseline is sensitive not only to the presence of entry, but also to the specifics of how entry is modeled, in ways that our social-costs-of-distortions formulas clarify. Entry can substantively alter the economy's response to shocks even if variable profits and fixed costs are small as a share of GDP.

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
Harvard University
*NA user is deceased

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

1 Introduction

A major challenge of macroeconomics is the aggregation problem: the problem of translating microeconomic disturbances into macroeconomic consequences. Aggregation results like those of Hulten (1978), or more recently Baqaee and Farhi (2019a), provide a framework for mapping microeconomic primitives to macroeconomic outcomes. In this paper, we generalize these results to environments with non-constant returns to scale and product entry and exit.

Our analysis is relatively general, and allows for scale effects due to both demand-side forces, as in Dixit and Stiglitz (1977), or supply-side forces, as in Lucas (1978). We also allow for an arbitrary pattern of distorting wedges and technological heterogeneity within and across industries, as well as unrestricted input-output linkages in both variable and fixed costs.

We characterize how aggregate output and aggregate productivity respond to changes in technology and changes in wedges.[1] We decompose changes in output into changes in technical and allocative efficiency. 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.[2]

We show that changes in technical efficiency are equal to a weighted sum of microeconomic technology shocks with weights that can sum to a number greater than one. The weight on each technology shock depends on expenditure shares and can be thought of as a cost-based or distortion-adjusted analogue to sales shares (i.e. Domar weights). The intuition for this cost-based Domar weight is similar to the logic in Hulten (1978), and captures the mechanical benefits of the technology shock on the production of the final good taking into account direct and indirect linkages but holding the allocation of resources fixed. This is the entirety of the effect if the equilibrium is efficient: when the equilibrium is efficient, reallocation effects can be ignored to a first-order approximation. Hence, even in models with non-convexities, fixed costs, and product creation and destruction (e.g. as in Hopenhayn, 1992), the logic of Hulten (1978) continues to apply as long as the equilibrium is efficient.

However, in economies with non-convexities and product entry and exit, efficiency is rarely attained. Once we stray from efficiency, we show that changes in allocative efficiency can play a theoretically and quantitatively important role in determining the

[1] In the body of the paper, we treat wedges as primitives but we discuss the implications of our results for endogenous wedges and give a specific illustration, using endogenous markups, in Appendix A.

[2] 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.

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 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*. Here, rent is variable profit due to either decreasing returns or markups, and quasi-rent is the part of this rent that is dissipated by the fixed cost of entry. We show that changes in rents and quasi-rents capture reallocation effects in equilibrium.[3]

Our treatment of entry is general. First, we allow for the possibility that entry costs be in terms of goods or factors. Second, the mapping from entry decisions to entry outcomes is flexible and nests two common extremes as special cases. At one extreme, entry is fully directed, and entrants can choose what production function they have post-entry. At the other extreme, entry is completely undirected, and entrants have no control over what production function they have post-entry. The first extreme is typically used in multi-sector models with homogeneous firms in each sector, whereas the second extreme is used in single-sector models with heterogeneous firms in each sector. Our results apply equally to both scenarios and to anything in between.

We use our comparative static results to study the social cost of distortions and the gains from industrial policy. We generalize the influential insights of Harberger (1954) to economies with non-convexities and entry and exit. In particular, we show that the social cost of inefficiencies is, up to a second-order approximation, equal to the sales-weighted sum of a series of Harberger triangles. Some of these triangles are associated with production and some are associated with entry.

We characterize these Harberger triangles in terms of microeconomic primitives — elasticities of substitution, expenditure shares, and returns-to-scale parameters. In doing so, we overturn a common intuition, valid in CES models without entry, that the social cost of misallocation is monotone in the elasticities of substitution. While a high elasticity of substitution increases the size of Harberger triangles associated with variable production, a low elasticity of substitution increases the size of Harberger triangles associated with entry. This results in non-monotonicity of losses with respect to elasticities of substitution.

We provide an application by quantifying the social costs of markups using a calibrated firm-level model for the U.S. We decompose the losses into losses arising from misallo-

[3] This generalizes the intuition in Baqaee and Farhi (2019a) that reallocation effects are summarized by changes in factor income shares. From the lens of this paper, income earned by factors are rents and in that paper quasi-rents are always equal to zero.

cation of resources in variable production and misallocation of resources in the amount entry versus variable production. 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% (this is similar to the results in Baqaee and Farhi, 2019a).[4] Accounting for entry can double these losses. The extent of misallocation along the entry margin depends on whether there is excessive or insufficient entry relative to first-best. In the baseline calibration, where entry costs are exactly offset by quasi-rents (there are no entry barriers), there is excessive entry in equilibrium. Somewhat surprisingly, this implies that the distance to the efficient frontier is smaller, conditional on estimates of markups, if one also believes that there are barriers to entry. Furthermore, the distance from the efficient frontier also depends on whether entry costs are paid in units of labor or goods, and whether the value of entry arises from consumer surplus or producer surplus.

We also consider how a marginal entry or production subsidy affects output starting at the distorted equilibrium. Unlike first-best policies, which are independent of network structure and simply ensure efficiency market-by-market, the effect of second-best policies are network-dependent. In particular, for economies with increasing returns to scale, 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 concepts. We show that the optimal marginal intervention aims to boost the sales of sectors that are upstream and have strong scale economies themselves and downstream from themselves.

Of course, there is a great deal of uncertainty about the specific number one attaches to these exercises. For us, the goal is to provide a sense of the order of magnitudes and to shed light on which microeconomic primitives determine these numbers (at least locally). In particular, our aim is to show how assumptions on the production structure, including the strength of scale economies, the extent to which entry is targeted, the type of resources used for entry, and the view one takes on the presence of entry restrictions affect the gains from policy and the losses from misallocation. While our results show that these features are critical theoretically and quantitatively, little is known about them empirically, and more empirical work is needed to bridge theory and measurement.

Despite their generality, our theoretical results also have some limitations. First, we focus on first- or second-order approximations, and extending this analysis to account for higher-order nonlinearities is an important extension we leave for future work.

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

Second, we model markups and other distortions as exogenous wedges. The advantage is that we can characterize the response of the equilibrium to changes in the wedges without committing to any specific theory of wedge determination (e.g. monopolistic competition, financial frictions, nominal rigidities, etc.) For some questions, like the economy's distance from the efficient frontier, the thought experiment specifies how wedges change, and so the absence of a theory of wedge determination is unimportant. However, our results cannot be directly used for counterfactuals where wedges change endogenously in unknown ways. Nevertheless, in these cases, our results are still relevant as part of a larger analysis that accounts for the endogenous response of wedges. Specifically, this paper characterizes how a change in wedges affect equilibrium outcomes. A theory of wedge-determination would pin down how wedges change in response to changes in equilibrium outcomes. The two parts could then be combined to conduct a full-blown counterfactual analysis. We provide an explicit example, using variable markups, in Appendix A.

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 provide 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 in terms of changes in rents and quasi-rents. Section 6 contains backward and forward propagation equations that determine how rents and quasi-rents respond to shocks as a function of primitives. In Section 7, we apply these results to analyze the social costs of distortions. Section 8 considers the bang-for-buck from competition and industrial policy at the decentralized equilibrium. Section 9 contains a quantitative application that computes and dissect the social costs of markups using firm-level data on markups.

Related Literature. Our results apply to a broad range of influential models. 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.[5]

There is a folk wisdom in the literature that modelling entry via diminishing returns,

[5] See Carvalho and Tahbaz-Salehi (2018) for a review of this literature.

in the spirit of Lucas (1978), Hopenhayn (1992), or Restuccia and Rogerson (2008), is in some sense isomorphic to modelling entry via diminishing marginal utility, in the spirit of Dixit and Stiglitz (1977), Melitz (2003), or Hsieh and Klenow (2009). By allowing for both possibilities, we show that this intuition is very fragile and fails outside of very simple single-sector models. In particular, for the first class, the relevant sufficient statistic for reallocation effects are changes in pure rents. On the other hand, for second class, the relevant sufficient statistic for reallocation effects are changes in quasi-rents.

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 a tightly-parameterized class of production networks with increasing returns, entry, and distortions. This paper dispenses with the parametric restrictions, allows for a more sophisticated handling of entry, 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 also relates to the literature on cross-sectional misallocation and policy interventions, with or without externalities, like Restuccia and Rogerson (2008), Hsieh and Klenow (2009), Epifani and Gancia (2011), Liu (2017), Osotimehin and Popov (2017), Behrens et al. (2016), Bartelme et al. (2019), Boehm and Oberfield (2020), Rubbo (2020), and La'O and Tahbaz-Salehi (2020). Our analysis of the economy's distance to the frontier is also related to Edmond et al. (2018), who analyze the social cost of markups. Our paper is also closely related to Bilbiie et al. (2012) and Bilbiie et al. (2019) who study the positive and normative implications of entry and exit in a dynamic context. Our paper also contributes to this literature by focusing on how the entry margin interacts with the input-output network to affect the costs of distortions. By showing that even in non-neoclassical economies with entry social losses can be approximated using Harberger triangles, the paper also extends the insights of Harberger (1954) and Harberger (1964).

Another strand of the literature which this paper relates to is the literature studying link-formation in production networks. In contrast to the approach in this paper, this literature takes discreteness of decisions seriously and is often studied with a non-Walrasian equilibrating mechanism. Some examples are Oberfield (2017), Lim (2017), Acemoglu and Azar (2020), Acemoglu and Tahbaz-Salehi (2020), Taschereau-Dumouchel (2020), Kikkawa et al. (2018), Dhyne et al. (2021), and Elliott et al. (2020). We abstract from these issues in our analysis, assuming that individual firms are infinitesimal and that the mass of entrants and number of links adjusts smoothly in response to perturbations of primitives. In exchange for these simplifications, we can provide a fairly general local characterization of the equilibrium.

2 Framework

This section describes the environment and equilibrium. There are three categories of agents: a representative household, a set of producers, and a set of entrants. The circular flow diagram of the economy is depicted in Figure 1. Each rectangle represents a type of agent. Entrants buy resources from producers to enter. Upon paying the start-up costs, entrants are (perhaps randomly) assigned to be producers. Producers produce using intermediate materials that they purchase from other producers. The representative household owns all resources in the economy and purchases final goods using aggregate income. We describe the problem each agent faces and then define the equilibrium.

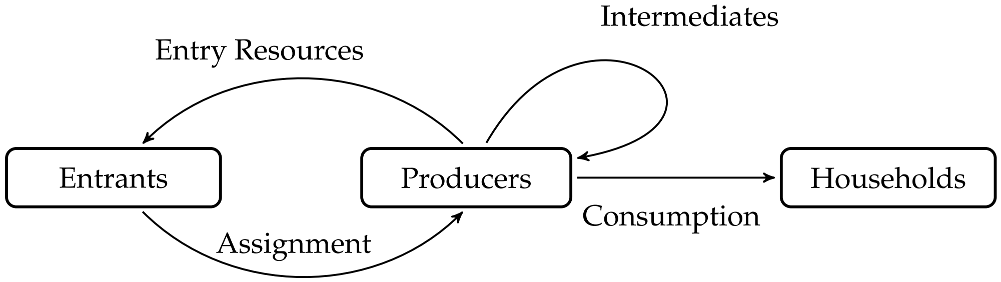


Figure 1: Circular flow schematic of the economy showing the flow of resources.

2.1 Producers

There is a set of producers indexed by their type $i \in \mathcal{N}$. Each producer of type i produces

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

units of good i, where f_i is a neoclassical production function, A_i is a productivity shifter, and x_{ij} is input j used by i. Each producer minimizes costs and sets its price p_i equal to its marginal cost times an exogenous markup/wedge μ_i.[6] The revenues generated by μ_i accrue to the owners of the firm.

[6] Appendix A endogenizes this markup wedge using monopolistic competition.

There is a potentially endogenous mass M_i of producers of each type i, and the overall output Y_i of producers of type i is defined by the homothetic aggregator

$$1 = \int_0^\infty F_i\left(\frac{y_i(\omega)}{Y_i}\right) d\omega = M_i F_i\left(\frac{y_i}{Y_i}\right), \tag{1}$$

where ω indexes varieties of type i, and the aggregator F_i is increasing, smooth, and weakly concave function with $F_i(0) = 0$. The last equality, which suppresses the ω index, follows from the fact that all varieties of type i are symmetric.[7] CES preferences are obtained when F_i in (1) is a power function.[8]

Agents who use good i buy the aggregated good Y_i. The price P_i of Y_i is equal to the marginal cost of producing Y_i times an exogenous wedge μ_i^Y.[9] Unlike the producer-level markup μ_i, revenues generated by the wedge μ_i^Y are *not* rebated to the owner of i and instead go directly to households. This distinction matters because revenues generated by μ_i incentivize entry, whereas revenues generated by μ_i^Y do not. Therefore, the wedge μ_i^Y acts like an output tax or subsidy on i.

Primary Factors. A subset of producers $\mathcal{F} \subset \mathcal{N}$ are the *primary factors* (e.g. labor, land, initial capital stock). For primary factors $f \in \mathcal{F}$, the mass M_f is exogenous, the production functions f_f have zero returns to scale (they are endowments), and the aggregator F_f is linear $Y_f = M_f y_f$. In addition, there are no markups/wedges $\mu_f = \mu_f^Y = 1$. In other words, there is no entry into the factor market (since M_f is fixed), each producer produces a fixed amount of output (since f_f has zero returns to scale), and outputs are aggregated linearly (since F_f is linear). This means that the total output of each factor is exogenous.

2.2 Entrants

Entrants have a choice of which entry opportunity they take up. Potential entry opportunities are indexed by $j \in E$, and entrants pay the corresponding fixed costs and draw a corresponding technology for variable production.

[7] Appendix F relaxes the assumption of symmetry of types in (1).

[8] The preferences in (1) are oftentimes called Kimball (1995) preferences. See Matsuyama and Ushchev (2017) for more details on this demand system which they call *Homothetic with Direct Additivity*. Matsuyama and Ushchev (2017) introduce two other generalizations of CES besides the one in (1), and our results are virtually unchanged if we use these alternatives.

[9] That is, the price of Y_i is given by $P_i = \mu_i^Y \min_{y_i}\{\int p_i(\omega) y_i(\omega) d\omega : Y_i = 1\}$.

Fixed Costs. To enter, type-j entrants must obtain a fixed bundle of inputs per entrant

$$x_j^E = g_j\left(\left\{x_{E,ji}\right\}_{i \in \mathcal{N}}\right), \tag{2}$$

where g_j has constant returns, and $x_{E,ji}$ is the input quantity of good i required for entry as a type j entrant.

Entry Technology. The entry matrix ζ is an $|E| \times |\mathcal{N} - \mathcal{F}|$ positive-valued matrix that gives the conditional probability that a type-j entrant becomes a type $i \in \mathcal{N} - \mathcal{F}$ producer:

$$\Pr(\text{ Producer i} \mid \text{Entrant } j) = \zeta(j, i).$$

Without loss of generality, assume that the rows of ζ are linearly independent.[10] We denote by $M_{E,j}$ the endogenous mass of type-j entrants who pay the entry cost for $j \in E$.

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 i without having to enter.

We refer to producers in markets where entry is not possible as *incumbents*. We refer to markets where entry is possible as *contested markets* and denote their collection by $\mathcal{N}^c$. Since we are flexible in the way we define and combine markets (via input-output linkages), 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 and 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 its technology 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.[11]

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

[11] 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 $g_j(\left\{x_{E,ji}\right\}_{i \in \mathcal{N}})$ have variable

Rents and Zero-Profit Conditions. The total *rent* or *variable profit* (we use the two terms interchangeably) earned by producers of type $i \in \mathcal{N}$ is equal to their revenues minus variable costs:

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

The zero-profit condition for type $j \in E$ entrants equates expected profits post-entry with the costs of entry

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

where $\lambda_{E,j}$ is expenditures on entry costs by entrants of type j. The left-hand side is the expected total profits 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* since they are dissipated by the costs of entry.

2.3 Households

There is a representative household with homothetic preferences

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

where Y is the money-metric measure of welfare. We also refer to Y as real GDP in this paper, abstracting from the well-understood issues related to the treatment of new goods in the measurement of aggregate output. To avoid corners, we impose Inada conditions on $\mathcal{D}$. The budget constraint of the representative household equates consumption expenditure to aggregate income, defined as revenues net of total costs,

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

Factor payments are rents earned by some zero-returns-to-scale incumbents in markets $\mathcal{F} \subset \mathcal{N}$.

(possibly increasing) returns to scale (for example, by making it a step function), but we do not pursue such an extension in this paper.

2.4 Resource Constraints and Equilibrium

The resource constraint for each 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:

$$M_i = \sum_{j \in E} \zeta(j, i) M_{E,j}. \tag{3}$$

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

The decentralized equilibrium is a collection of prices and quantities which clears markets and solves each agents' decision problem.

Definition 1. A *decentralized equilibrium* is a collection of prices $\{P_i, p_i\}$ and quantities $\left\{C_i, Y_i, y_i, x_{ij}, x_{E,ij}, M_{E,j}, M_i\right\}$, such that given technology $\{A_i\}$ 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) resource constraints are satisfied.

In this paper, we derive comparative statics with respect to changes in technologies A_i and wedges μ_i and μ_i^Y. These reduced-form wedges can be used to capture many distortions besides markups like taxes, financial frictions, or nominal rigidities.[12]

2.5 Some Examples

At this level of abstraction, the model nests many general equilibrium models with entry. The following examples highlight some important special cases.

Example 1 (Decreasing-Returns-to-Scale)**.** Let l denote labor input. Suppose that for each $i \in N$, individual production functions have decreasing returns to scale $f_i(x) = A_i l_i^{\epsilon}$ for some $\epsilon \in [0, 1]$ and suppose that each i is aggregated linearly $F_i(x) = x$. In addition, there is one extra producer, indexed by $N + 1$, with a linear production function

$$Y_{N+1} = y_{N+1} = \sum_{i \in N} Y_i = \sum_{i \in N} M_i A_i l_i^{\epsilon}.$$

[12] 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.

This captures demand systems used in firm-level models like Lucas (1978), Hopenhayn (1992), and Restuccia and Rogerson (2008), where goods are perfect substitutes but are produced with diminishing returns to scale.

Example 2 (Increasing-Returns-to-Scale)**.** Let l denote labor input. Suppose that for each $i \in N$, individual production functions have constant returns to scale $f_i(x) = A_i l_i$, but suppose that each i is aggregated non-linearly with $F_i(x) = x_i^{\frac{\theta-1}{\theta}}$ for some $\theta \in (1, \infty]$. In addition, there is one extra producer, indexed by $N+1$, with a CES production function with elasticity θ, then

$$Y_{N+1} = y_{N+1} = \left(\sum_{i \in N} Y_i^{\frac{\theta-1}{\theta}} \right)^{\frac{\theta}{\theta-1}} = \left(\sum_{i \in N} M_i \left(A_i l\right)_i^{\frac{\theta-1}{\theta}} \right)^{\frac{\theta}{\theta-1}} .$$

This captures demand systems used in firm-level models like Dixit and Stiglitz (1977), Melitz (2003), and Hsieh and Klenow (2009), where goods are imperfect substitutes but are produced with constant returns to scale.

Example 3 (Single-Sector with Heterogeneous Firms)**.** Suppose there is only one entrant type, then the zero-profit condition equates total profit to total entry costs

$$\sum_{i \in N} \lambda_{\pi,i} = \lambda_{E,1},$$

which pins down the mass of entrants $M_{E,1}$. The mass of producers of each type i is given by $M_i = \zeta(1, i) M_E$. This is similar to models where entrants pay the entry cost and are randomly assigned a production technology (e.g Hopenhayn, 1992 or Melitz, 2003).

Example 4 (Multi-Sector with Homogeneous Firms)**.** Suppose there is one entrant type for each type of producer. Then the zero-profit condition equates profits in each market to entry costs for that market

$$\lambda_{\pi,i} = \lambda_{E,i},$$

which pins down the mass of entrants $M_{E,i}$. The mass of producers of each type is then given by $M_i = M_{E,i}$. This is similar to a multi-sector model with entry, where each sector has its own entry condition.

3 The Value of Entry and an Efficient Benchmark

In this section, we analyze efficient equilibria as a preliminary step to understanding inefficient equilibria. To do so, we consider the value of product creation by using properties of demand and marginal cost curves.

The inverse demand curve for each variety of type i is

$$\frac{p_i}{P_i} = \mu_i^Y \gamma_i F_i'(\frac{y_i}{Y_i}), \tag{4}$$

where γ_i is a Lagrange multiplier. Equation (4) shows that the relative quantity of a given variety of i demanded is a decreasing function of the relative price of that variety. Figure 2 depicts the demand curve against the marginal cost curve for some variety of i.

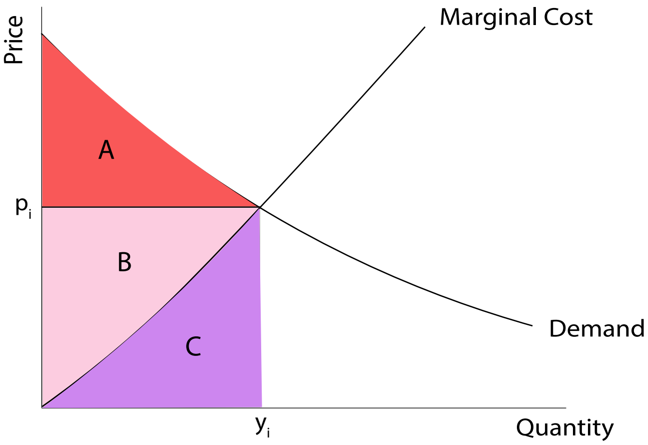


Figure 2: Residual demand and marginal cost curves for seller i as a function of quantity. The area under the demand curve is $A + B + C$, revenues are $B + C$, consumer surplus is A, producer surplus is B, and total variable costs are C.

Figure 2 supposes that i sets price equal to marginal cost. In this case, the area under the marginal cost curve, C, is equal to total variable cost. Sales are equal to the rectangle $B + C$. The total area under the demand curve is $A + B + C$. Hence, B is producer surplus (due to increasing marginal cost of y_i) and A is consumer surplus (due to diminishing marginal product/utility of y_i).

As pointed out by Baqaee et al. (2020), the Lagrange multiplier in (4) is equal to the area under the residual demand curve relative to sales:

$$\gamma_i = \frac{F(\frac{y_i}{Y_i})}{\frac{y_i}{Y_i}F_i'(\frac{y_i}{Y_i})} = \frac{\int_0^{y_i} p_i(y)dy}{p_i y_i} = \frac{A+B+C}{B+C} = \frac{A}{B+C} + 1 \geq 1. \tag{5}$$

In this model, firms generate value for society, in excess of total variable costs, both because of producer surplus (B) and consumer surplus (A). We refer to $\gamma_i - 1$ as the consumer surplus ratio. Below, we provide some sufficient conditions for efficiency of the decentralized equilibrium.

Theorem 1 (Conditions for Efficiency)**.** *There exists a decentralized equilibrium with markups $\mu_i = \gamma_i$ and output subsidies $\mu_i^Y = \gamma_i^{-1}$ for all $i \in \mathcal{N}$ that is Pareto-efficient.*

Theorem 1 implies that efficiency requires compensating entrants for the value their entry generates for society. The value of an additional firm for society is the sum of consumer and producer surplus ($A + B$ in Figure 2). Efficiency requires that entry takes place until entry costs are equal to the expected marginal value of entry. Since the entry cost paid is equal to expected profits, efficiency requires equating expected profits to the expected marginal value of entry ($A + B$). A firm that sets price equal to marginal cost has profits equal to B not $A + B$. Hence, if $A > 0$, which happens if, and only if, $\gamma_i > 1$, then the firm must be allowed to charge a markup commensurate with γ_i to incentivize the optimal amount of entry. However, this markup distorts input choices and so it must be offset by an output subsidy which restores marginal-cost-pricing conditional on entry.

If varieties do not generate consumer surplus, then the marginal cost pricing equilibrium is efficient. In this sense, Theorem 1 is a generalization of 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.

The following examples demonstrate the implications of Theorem 1 for some notable special cases.

Example 5 (Efficiency with only IRS)**.** If goods of type i are aggregated via CES aggregator $F_i(x) = x_i^{\frac{\theta-1}{\theta}}$, then, by the first equality in (5), the consumer surplus ratio is $\gamma_i - 1 = 1/(\theta - 1) \geq 0$. This is the so-called *love-of-variety* effect. In this case, efficiency is attained if each i charges a markup equal to $\theta/(\theta - 1)$ with an off-setting output subsidy equal to $(\theta - 1)/\theta$.

Example 6 (Efficiency with only DRS)**.** If goods of type i are aggregated via a linear aggregator $F_i(x) = x_i$, then $\gamma_i = 1$ and efficiency holds if we have marginal cost pricing $\mu_i = \mu_i^Y = 1$.

Theorem 1 is important from a normative perspective, but it also has important positive implications. In particular, it implies that, under the conditions of Theorem 1, the response of welfare or real GDP to technology shocks is straightforwardly tied to observables. Theorem 2 demonstrates.

Theorem 2 (Output Response with Efficiency)**.** *Under the conditions of Theorem 1, the response of aggregate output to a 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 type i sellers 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 generalizes Hulten (1978) to economies with fixed costs, increasing returns, and an extensive margin of product creation and destruction. It also extends Hulten (1978) to shocks to non-variable production, like the fixed costs of entry.

Theorem 2 shows that, in the efficient equilibrium, the response of welfare to microeconomic shocks is determined by simple and readily observable statistics and the details of the underlying production structure do not matter.[13] The rest of the paper studies inefficient equilibria.

4 Preliminaries for Studying Inefficient Equilibria

To emphasize our mechanisms of interest, we specialize the framework.

Assumption 1 (Iso-Elastic Cost Curves)**.** For each contested market $i \in \mathcal{N}^c$, either

$$1 = M_i F_i\left(\frac{y_i}{Y_i}\right) \quad \text{and} \quad y_i = A_i f_i\left(\left\{x_{ij}\right\}_{j\in\mathcal{N}}\right), \tag{6}$$

or

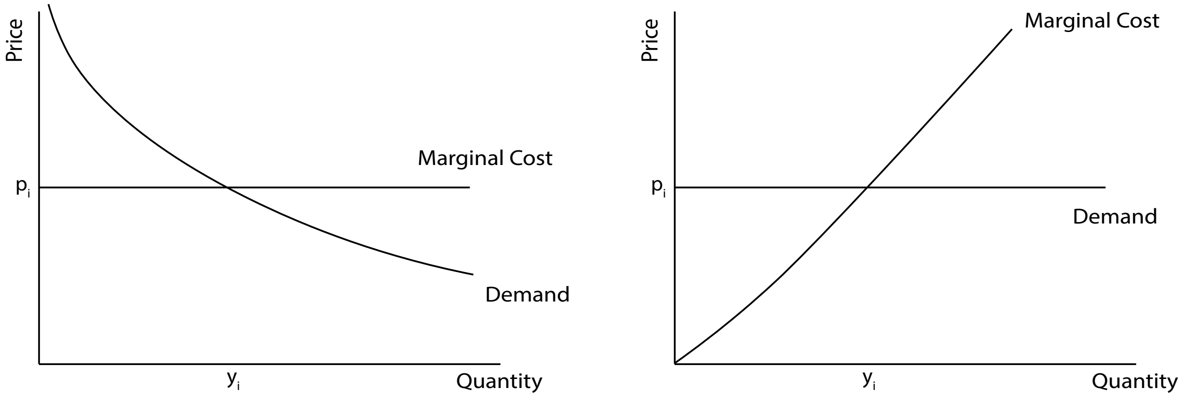
$$1 = M_i\left(\frac{y_i}{Y_i}\right) \quad \text{and} \quad y_i = A_i f_i\left(\left\{x_{ij}\right\}_{j\in\mathcal{N}}\right)_i^{\varepsilon_i}, \tag{7}$$

where $\varepsilon_i \in [0,1]$ and f_i has constant returns to scale. We refer to goods produced according to (6) as *IRS goods* and goods produced according to (7) as *DRS goods*. We denote each

[13] 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.

collection of goods by $\mathcal{N}^{IRS}$ and $\mathcal{N}^{DRS}$. For IRS goods, $\varepsilon_i = 1$ and $\gamma_i \geq 1$, whereas for DRS goods $\varepsilon_i \leq 1$ and $\gamma_i = 1$.

Figure 3 depicts the demand and marginal cost curves for IRS and DRS goods. We separate goods into $\mathcal{N}^{IRS}$ and $\mathcal{N}^{DRS}$ for exposition, but since the model allows for unrestricted input-output linkages, we can combine IRS and DRS goods to represent the production of goods where both forces are active at the same time.



(a) IRS goods

(b) DRS goods

Figure 3: Marginal cost and marginal product (demand) curves for $\mathcal{N}^{IRS}$ and $\mathcal{N}^{DRS}$.

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

Assumption 2 (Positive Profits)**.** If $i \in \mathcal{N}$ is contested, then $\mu_i > \varepsilon_i$.

Assumptions 1 and 2 are imposed throughout the rest of the paper.

We introduce notation that we rely on throughout the rest of the paper. All the objects introduced below are defined at the initial equilibrium (around which we provide first- or second-order approximations). We normalize the mass of entrants $M_{E,j}$ to one at the initial equilibrium, and treat nominal GDP as the numeraire.

The Normalized Entry Matrix. The $|E| \times |\mathcal{N} - \mathcal{F}|$ matrix $\tilde{\zeta}$ gives the conditional probability that a type i product is produced by a type j entrant. That is,

$$\tilde{\zeta}(j, i) = \Pr(\text{ Entrant } j \,|\, \text{Producer } i) = \frac{\zeta(j, i) M_{E,j}}{\sum_{k \in E} \zeta(k, i) M_{E,k}}$$

whenever market i is contested and zero otherwise. In other words, $\tilde{\zeta}$ is the reverse conditional probability compared to ζ. Each column i of $\tilde{\zeta}$ sums to one or zero depending on whether i is contested or not.

Rents and Quasi-Rents. The *rent* (or variable profit) of market $i \in \mathcal{N}$ is

$$\lambda_{\pi,i} = \frac{P_i Y_i}{GDP}\pi_i, \quad \text{with} \quad \pi_i = \frac{1}{\mu_i^Y}\left(1 - \frac{\varepsilon_i}{\mu_i}\right), \tag{8}$$

where π_i is the share of i's sales that are claimed as (gross) profits. The profit margin π_i consists of the rents due to market power and diminishing returns.

Quasi-rents are rents that are dissipated by entry costs. Some portion of rents, $\lambda_{\pi,i}$, earned by producers of product i are off-set by entry costs. Letting $\lambda_{E,j}$ be the entry cost paid by type-j entrants, define the change in quasi-rents of producers of type i to be

$$\mathbb{E}_{\tilde{\zeta}(:,i)}(d\log\lambda_E) = \sum_{j\in E}\tilde{\zeta}(j,i)d\log\lambda_{E,j}.$$

In words, the change in quasi-rent associated with product $i \in \mathcal{N}$ is the expected change in entry costs paid by entrants into i.

Denote the profit-weighted projection of changes in rents $\mathrm{d}\log\lambda_\pi$ on the entry matrix $\tilde{\zeta}$ by

$$\widehat{\mathrm{d}\log\lambda}_\pi = \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\tilde{\zeta}\lambda_\pi\,\mathrm{d}\log\lambda_\pi. \tag{9}$$

This rent-weighted projection is an important statistic with an intuitive interpretation. The following lemma shows why $\widehat{\mathrm{d}\log\lambda}_\pi$ is important.

Lemma 1. *The change in the quasi-rents associated with producer type $i \in \mathcal{N}$ is*[14]

$$\mathbb{E}_{\tilde{\zeta}(:,i)}(d\log\lambda_E) = \widehat{\mathrm{d}\log\lambda}_{\pi,i}.$$

Intuitively, an increase in rents ($d\log\lambda_{\pi,i} > 0$) earned by producer i increases quasi-rents, or expenditures on entry, associated with i by $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$. The examples below provide some intuition.

Example 7 (No Entry)**.** Consider an example where all producers are incumbents and the entry matrix $\tilde{\zeta}$ has rank zero. Applying Lemma 1, the expected log change in quasi-rents

[14] For this Lemma, we let λ_π be the $|\mathcal{N}|\times|\mathcal{N}|$ diagonal matrix of rents and $\mathrm{d}\log\lambda_\pi$ be the $|\mathcal{N}|\times 1$ vector of changes in rents. We define λ_π as an $|\mathcal{N}|\times|\mathcal{N}|$ diagonal matrix and $d\log\lambda_\pi$ be the $|\mathcal{N}|\times 1$ vector to streamline the matrix expressions for projections below. Throughout the paper, to lighten the notation, we often use the same symbol to denote vectors and their counterparts as diagonalized matrices.

associated with $i \in N$ is given by

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

Since entry is impossible, quasi-rents are equal to zero and never change regardless of changes in rents $d\log\lambda_\pi$.

Example 8 (Undirected Entry)**.** Consider again an example with N products, no incumbents, and only one entrant type ($E = 1$). Therefore, $\tilde{\zeta}$ is a $1 \times N$ vector of all ones. In this case, applying Lemma 1, the expected log change in quasi-rents for product i is

$$\widehat{\mathrm{d}\log\lambda}_{\pi,i} = d\log\left(\sum_{j=1}^{N}\lambda_{\pi,j}\right),$$

where the right-hand side does not depend on i. Hence, the change in quasi-rents associated with every product type i is the same, since there is only one entrant type. If type i producers become more profitable ($d\log\lambda_{\pi,i} > 0$), quasi-rents associated with i increase only in so far as profitability overall rises ($d\log\left(\sum_j \lambda_{\pi,j}\right) > 0$). This is because entrants respond only to average profitability and cannot directly target the rents earned by type i.

Example 9 (Directed Entry)**.** Consider again an example with N producer types, but suppose that there are also N entrant types ($|E| = |\mathcal{N} - \mathcal{F}|$). Therefore, the matrix $\tilde{\zeta}$ is an $N \times N$ matrix with rank N. Applying Lemma 1, the expected log change in quasi-rents for producer i is

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

In this case, the change in quasi-rents for every market i is the same as the changes in rents in that market, since each market has its own entrants. Hence, if type i sellers become more profitable ($d\log\lambda_{\pi,i} > 0$), quasi-rents associated with i increase by the same amount.

Some important extreme cases of entry are defined below.

Definition 2. We say that entry is *fully directed* if the entry matrix ζ has rank $|\mathcal{N} - \mathcal{F}|$, this happens if there are as many entrant types as there are contested markets. We say that entry is *fully undirected* if the entry matrix ζ has rank 1, this happens if there is only one entrant type. We say that there is *no entry* if the entry matrix ζ has rank zero.

We call the first situation fully-directed entry because in this case, changes in rents are captured entirely by new entrants as quasi-rents. 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 $\tilde{\zeta}$ has rank one, then entrants only choose whether or not they enter, and they do not choose what type of product they would like to produce. Finally, when entry is impossible, $\tilde{\zeta}$ has zero rank.

IO Matrices. We now define input-output matrices. Without loss of generality, we represent the household's final demand function as the output of some incumbent producer standing in for the household. To emphasize its unique role, we index the household by the number 0 and add the household to the set of producers $\mathcal{N}$.

The variable spending IO matrix, Ω^V, is 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 i's revenues

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

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

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

We introduce the *backward* and *forward* input-output (IO) matrices and their accompanying Leontief inverses. Intuitively, The backward matrix captures how a change in the sales of a customer is transmitted to the sales of its suppliers via backward linkages. The forward matrix captures how a change in the price of a supplier is transmitted to the price of its customers via forward linkages.

Backward IO Matrix. Let π be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix of profit shares defined in (8). 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.[15]

Forward IO Matrix. Let $\mathcal{E}$ be the $|\mathcal{N}| \times |\mathcal{N}|$ diagonal matrix whose ith diagonal element is equal to $\mathcal{E}_{ii} = (\gamma_i - \varepsilon_i)$. Intuitively, $\mathcal{E}_{ii}$ measures how a change 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

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

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.[16]

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 the fact that nominal GDP is the 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*

[15] The sales of j can be broken down into its sales to 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.

[16] 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.

weight defined as the *i*th element of the zeroth row of the backward Leontief inverse

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

This captures the household's exposure to i via backward linkages or i's centrality in demand/sales.

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

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

This captures the household's exposures to i via forward linkages or i's centrality in supply/prices.[17]

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

5 Aggregation

We now generalize Theorem 2 to inefficient economies. We provide our results in two steps. In this section, we write the response of aggregate output to shocks as a function of changes in rents and quasi-rents. In the next section, we derive changes in rents and quasi-rents, as a function of microeconomic primitives. Putting the two steps together yields a complete characterization. The shocks that we consider are shocks to technologies and markups/wedges written in vector form as $(\mathrm{d}\log A, \mathrm{d}\log \mu)$.[18]

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

[18] 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.

5.1 The Aggregation Equation

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

$$\mathrm{d}\log Y = \sum_i \lambda_i^F d\log A_i - \sum_{i\in\mathcal{N}} \lambda_i^F\left[1 - \frac{1-\varepsilon_i}{\pi_i}\right] d\log\mu_i \tag{10}$$

$$- \sum_{i\in\mathcal{N}^{DRS}} \lambda_i^F(1-\varepsilon_i)\left(d\log\lambda_{\pi,i} - \widehat{d\log\lambda}_{\pi,i}\right) + \sum_{i\in\mathcal{N}^{IRS}} \lambda_i^F\left(\gamma_i - 1\right)\widehat{d\log\lambda}_{\pi,i},$$

where the projection $\widehat{d\log\lambda}_{\pi}$ *captures changes in quasi-rents (profits dissipated by entry costs) and the residual* $d\log\lambda_\pi - \widehat{d\log\lambda}_\pi$ *captures changes in the difference between rents and quasi-rents (profits not dissipated by entry costs).*

Theorem 3 generalizes Theorem 2 to economies with distortions, and we spend the rest of this section unpacking its intuition and working through some examples. The first line consists of exogenous objects and the second line of endogenous ones. The terms $\sum_{i\in\mathcal{N}} \lambda_i^{\mathrm{F}} d\log A_i$ in (10) capture changes in *allocative efficiency* — that is, the "mechanical" change in real GDP caused by changes in technology holding fixed the allocation of resources (see Appendix C for a formal discussion). When the equilibrium is efficient, as in Theorem 2, these are the only terms that matter. The remaining terms in (10) are changes in output caused by reallocation effects. We refer to these terms collectively as changes in *allocative efficiency* caused by reallocations. If the initial equilibrium is efficient, changes in allocative efficiency will be zero (barring corners).[19]

Since nominal GDP is normalized to one, changes in real GDP are the negative of changes in the household price index $\mathrm{d}\log Y = -\mathrm{d}\log P_0$. Therefore, one way to understand (10) is to think through how shocks affect the price of the consumer price index. 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 term captures the effect of an increase in markups on consumer prices, which are weighted by the forward Domar weight of the bundle of inputs.

The second line accounts for changes in rents and quasi-rents. For DRS producers, the relevant statistic is the gap between changes in rents and quasi-rents, whereas for IRS

[19] Corner cases occur if $\varepsilon_i = \gamma_i = 1$ for some contested i. If $\gamma_i = \varepsilon_i = 1$, then entry into i is socially wasteful. If $i \in \mathrm{span}\{\zeta\}$, then in the efficient equilibrium, the mass of i is zero (i.e. at a corner). If a change in a wedge causes entry in i to become positive, then this will reduce aggregate output to a first-order because in the initial equilibrium the marginal social benefit of entry into i is not equated to the marginal social benefit of alternative uses for those resources.

producers, the relevant statistic is the change is quasi-rents. Overall, the second line captures changes in welfare caused by changes in producer and consumer surplus weighted by their forward (cost-based) Domar weight. As shown by Lemma 1, $\widehat{\mathrm{d}\log\lambda}_{\pi,i}$ measures the change in fixed-cost spending by entrants that become producers of i. 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 changing consumer surplus. 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 Three Special Cases

To build more intuition for Theorem 3, consider three special cases: for every $i \in \mathcal{N} - \mathcal{F}$ either (i) $\varepsilon_i = \gamma_i = 1$ (CRS); (ii) $\varepsilon_i < 1$ and $\gamma_i = 1$ (DRS); or (iii) (IRS) $\varepsilon_i = 1$ and $\gamma_i > 1$ (IRS). For simplicity, for all of these examples, assume there is only one primary factor and that all other markets are contested. Consider a univariate productivity shock $\mathrm{d}\log A_i$ (holding constant other productivities, wedges, and markups).[20]

CRS. When $\varepsilon_i = \gamma_i = 1$ for all $i \in \mathcal{N} - \mathcal{F}$, Theorem 3 reduces to

$$\mathrm{d}\log Y = \lambda_i^{\mathrm{F}}\,\mathrm{d}\log A_i, \tag{11}$$

so only the direct technology shock matters and reallocations are irrelevant. The reason is because of free-entry. When $\varepsilon_j = 1$ and $\gamma_j = 1$, entry is socially wasteful because there is no consumer or producer surplus associated with entry. Nevertheless, entry always adjusts to ensure zero profits. This means that entry absorbs or exudes resources in such a way that there are no changes in allocative efficiency, even though there are reallocations and the economy is inefficient. If there was no free-entry, then the behavior of output would be substantially more complicated, since we would then have to account for how technology shocks reallocate resources across producers (as in Baqaee and Farhi, 2019a).

Comparing (11) to the economies considered by Baqaee and Farhi (2019a) shows that free entry can dramatically alter the behavior of output, even if entry itself is socially wasteful. Furthermore, the economy with and without entry behave qualitatively differently, even if profits and entry costs are small as a share of GDP. Hence, unless the

[20] The intuition for a shock to markups/wedges is similar, but for brevity, we omit this discussion.

equilibrium is efficient, one cannot dismiss the importance of explicitly modelling entry by arguing that production has almost constant-returns or that entry costs are small as a share of GDP.

DRS. Suppose that $\varepsilon_i < \gamma_i = 1$ for all $i \in \mathcal{N} - \mathcal{F}$, Theorem 3 then becomes

$$\mathrm{d}\log Y = \lambda_i^{\mathrm{F}} \,\mathrm{d}\log A_i - \sum_{j \in \mathcal{N}-\mathcal{F}} \lambda_j^F (1-\varepsilon_j) \left(d\log \lambda_{\pi,j} - \widehat{\mathrm{d}\log \lambda}_{\pi,j} \right). \tag{12}$$

If rents outpace quasi-rents $\mathrm{d}\log \lambda_{\pi,j} - \widehat{\mathrm{d}\log \lambda}_{\pi,j} > 0$ for some market $j \in \mathcal{N}$, this implies that entry is not keeping up with sales. Therefore, individual producers in j are increasing their scale, using relatively more inputs, and running into diminishing returns. This raises their marginal cost and price. This reallocation contributes to reducing aggregate output in proportion to the forward Domar weight λ_j^F of these producers. The total effect of reallocations is obtained by summing over all markets. Reallocations lead to a more efficient use of resources when these changes in the scale of producers cause the final price index to fall.

Such improvements in allocative efficiency cannot occur when the economy is efficient. To see this, note that by Theorem 1, efficiency is attained if all markups μ are equal to one. In this case, the profits earned by each firm are just equal to producer surplus $\lambda_{\pi,j} = \lambda_j^F (1-\varepsilon_j)$. Hence, the reallocation terms become

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

where the final equality follows from the fact that the weighted sum of residuals of a linear projection must be zero.

When there is directed entry, (12) also simplifies to

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

even when the initial equilibrium is inefficient. This is similar to (11). Intuitively, in this case, changes in the prices of goods are determined independently from changes in their sales because changes in sales are accommodated entirely through changes in entry. This means that no individual producer changes their scale in response to a productivity shock, and hence marginal costs do not change in response to changes in industry-level quantities. In other words, even though the equilibrium may be inefficient, reallocations happen entirely on the extensive margin and are welfare-neutral.

IRS. Now suppose that $\gamma_i > \varepsilon_i = 1$ for all $i \in \mathcal{N} - \mathcal{F}$. In this case, Theorem 3 implies that

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

If in some market j, quasi-rents increase, $\widehat{\mathrm{d}\log\lambda}_{\pi,j} > 0$, then resources are reallocated towards entry into j and this boosts consumer surplus according to $\gamma_j - 1 > 0$. This increase in consumer surplus contributes to increasing aggregate output in proportion to the forward Domar weight λ_i^F of this market. The total effect of reallocations is obtained by summing over all markets.

6 Propagation

Theorem 3 in the previous section gives changes in aggregate output as a function of changes in rents and quasi-rents. In this section, we complete the theory by deriving equations for changes in sales and prices, and hence rents and quasi-rents. We do this in two steps: forward and backward propagation.

In Section 6.1, we characterize the propagation of shocks through forward linkages: how changes in prices feed forward from suppliers to consumers. In Section 6.2, we characterize the propagation of shocks through backward linkages: 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. We consider some worked-out examples in Section 6.4.

6.1 Propagation Through Forward Linkages

We start by describing the response of prices to shocks.

Proposition 1 (Forward Propagation)**.** *In response to shocks* $(\mathrm{d}\log A, \mathrm{d}\log\mu)$*, changes in prices are given by*

$$\begin{aligned} d\log P_i = &-\sum_{j\in\mathcal{N}} \Psi_{ij}^F d\log A_j + \sum_{j\in\mathcal{N}} \Psi_{ij}^F \left(1 - \frac{1-\varepsilon_i}{\pi_i}\right) \mathrm{d}\log\mu_j \\ &+ \sum_{j\in\mathcal{N}^{DRS}} \Psi_{ij}^F \left(1-\varepsilon_j\right)\left(d\log\lambda_{\pi,j} - \widehat{\mathrm{d}\log\lambda}_{\pi,j}\right) - \sum_{j\in\mathcal{N}^{IRS}} \Psi_{ij}^F \left(\gamma_j - 1\right) \widehat{\mathrm{d}\log\lambda}_{\pi,j} \end{aligned}$$

Proposition 1 is similar to Theorem 3. 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 1 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 1 is similar to the one for Theorem 3.[21]

6.2 Propagation Through Backward Linkages

Assume that all production and entry functions in the economy f_i and g_i are CES production functions. We make this assumption for clarity, not tractability, and Appendix D generalizes our results to non-CES production functions. Given the assumption that all production functions are nested-CES, 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.[22] For notational convenience, we also assume entry goods are assembled by perfectly competitive constant-returns-to-scale incumbents who are added to the input-output network as additional "producers."

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 the backward Leontief inverse Ψ^{B}, using the mth row of $(1-\pi)^{-1}\Omega^V$ as the probability distribution. This is a covariance since $\sum_{k\in\mathcal{N}}(1-\pi_m)^{-1}\Omega^V_{mk} = 1$ for $m \in \mathcal{N} - \mathcal{F}$.

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

$$d\lambda^B_i = -\sum_{m\in\mathcal{N}} \lambda^B_m \sum_{k\in\mathcal{N}} \left[\Omega^V_{mk} - (1-\pi_m)\sum_{j\in E} \tilde{\zeta}_{jm}\Omega^E_{jk}\right] \Psi^B_{ki} d\log\mu_m$$

[21] An interesting special case of Proposition 1 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

$$d\log P^Y_i = -\sum_{j\in\mathcal{N}} \Psi^{\mathrm{F}}_{ij} d\log A_j + \sum_{j\in\mathcal{N}} \Psi^{\mathrm{F}}_{ij}\left(1 - \frac{1-\varepsilon_j}{\pi_j}\right)\mathrm{d}\log\mu_j.$$

In other words, the change in relative prices does not depend on final demand or 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.

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

$$- \sum_{m \in \mathcal{N} - \mathcal{F}} \lambda^B_m (1 - \pi_m)(\theta_m - 1) Cov_m \left(d \log P, \Psi^B_{(:,i)} \right). \quad (13)$$

Proportional changes in sales are given by $\mathrm{d} \log \lambda^B_i = \mathrm{d} \lambda^B_i / \lambda^B_i$.

We discuss each line of (13) in turn. The first line is the effect of changes in markups on the sales of 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 $\lambda^B_m \Omega^V_{mk}$. On the other hand, a higher markup increases 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 second line captures the effect of expenditure-switching due to changes in relative prices. Changes in relative prices $\mathrm{d} \log P$ caused by the shocks lead individual producers in every market $m \in \mathcal{N} - \mathcal{F}$ 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 $-\lambda^{\mathrm{B}}_m (1 - \pi_m)(\theta_m - 1) Cov_m (\mathrm{d} \log P, \Psi^{\mathrm{B}}_{(:,i)})$. The overall effect of this expenditure-switching on i's sales are attained by summing over all m.

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

Lemma 2 (Changes in Rents). *In response to shocks* $(\mathrm{d} \log A, \mathrm{d} \log \mu)$*, changes in rents are given by*

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

and $d \log \lambda^B_i$ *is given by Proposition 2.*

Hence, changes in rents in each sector are driven, either by changes in sales $d \log \lambda^B$ or changes in profit margins $d \log \pi$. Given changes in sales $\mathrm{d} \log \lambda^{\mathrm{B}}_i$, from Proposition 2, it is easy to obtain changes in rents $\mathrm{d} \log \lambda_{\pi,i}$ from Lemma 2, and changes in quasi-rents by applying the linear projection formula (9).

6.3 Combining Forward and Backward Propagation

To recap, Proposition 1 pins down changes in prices in terms of changes in sales shares, and Proposition 2 pins down changes in sales shares in terms of changes in prices. Together, they pin down changes in both sales shares and prices in every market. This, in turn, determines changes in rents and quasi-rents, which can be plugged back into Theorem 3 for welfare changes.

Going beyond this, once in possession of changes in prices $d\log P$ and sales shares $d\log\lambda^B$, it is simple to solve for other equilibrium objects. For instance, changes in the aggregated output of i are $d\log Y_i = d\log\lambda_B^i - d\log P_i$ and changes in individual varieties of i are $d\log y_i = d\log Y_i - \gamma_i d\log M_i$. Finally, changes in the mass of entrants in each market are

$$\mathrm{d}\log M = \widehat{\mathrm{d}\log\lambda}_\pi - \tilde{\zeta}'(\tilde{\zeta}\lambda_\pi\tilde{\zeta}')^{-1}\lambda_E\,\mathrm{d}\log P_E, \tag{14}$$

where $\mathrm{d}\log M$ and $\widehat{d\log\lambda}_\pi$ are $\mathcal{N} - \mathcal{F} \times 1$ vectors, λ_E is the $E \times E$ diagonal matrix of expenditures on entry, and $d\log P_E$ is the $E \times 1$ vector of changes in the prices of the entry goods given by Proposition 1. Equation (14) shows that the mass of producers in each contested market is increasing in quasi-rents and decreasing in entry costs.

6.4 Illustrative Examples

In this section, we consider the simplest example and show how the response of output to shocks changes as we vary assumptions about returns to scale and the form of entry. This example has no intermediate inputs, a single factor (labor), and entry costs are assumed to be paid in units of labor. We compare constant, decreasing, and increasing returns and vary no entry, directed entry, and undirected entry.

We use the following notation throughout. Given three vectors u, v, and w with $\sum_k w_k = 1$, we write $\mathbb{E}_u(v) = \sum_k u_k v_k$ and $Cov_u(u, w) = \sum_k u_k(v_k w_k) - (\sum_k u_k v_k)(\sum_k u_k w_k)$.

No Entry. To start, consider an economy without entry. Aggregate output,

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

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

$$Y_k = \left(M_k y_k^{\gamma_k}\right)^{\frac{1}{\gamma_k}},$$

is itself a CES aggregate of some mass M_k of differentiated varieties with an elasticity of substitution $\theta_k = 1/(1 - \gamma_k) \geq \min\{\theta_0, 1\}$. Each variety in sector k is produced from labor with constant returns and productivity A_k by a single firm and sold at a markup $\mu_k > 1$

over marginal cost

$$y_k = A_k l_k, \quad p_k = \mu_k mc_k.$$

Consider a vector of k-level productivity shocks $\mathrm{d}\log A$. To apply Theorem 3, we use the fact that the backward and forward Domar weight are equal to each other $\lambda_k^{\mathrm{B}} = \lambda_k^{\mathrm{F}}$ in this example. Applying Theorem 3, the change in aggregate output is

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

where $d\log\lambda_L^{\mathrm{B}}$ is the change in labor's share of income. The first term is the change in technical efficiency, holding fixed the allocation of resources, and the second term is the change caused by reallocations. Applying Proposition 2, this can be written in terms of primitives as the second term is

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}\left(\mathrm{d}\log A\right) - (\theta_0 - 1)\frac{1}{\lambda_L^{\mathrm{B}}} Cov_{\lambda^{\mathrm{B}}}\left(\frac{1}{\mu}, \mathrm{d}\log A\right). \tag{15}$$

To understand the intuition, suppose that k's are substitutes ($\theta_0 > 1$) and that the shock disproportionately increases the productivity of high-markup k's. Since the shock disproportionately increase the productivity of high-markup firms ($Cov_{\lambda^B}(\mathrm{d}\log A, 1/\mu) < 0$), and since goods are substitutes ($\theta_0 > 1$), the shock reallocates labor towards high-markup firms and reduces the labor share (rents earned by labor). This reallocation improves allocative efficiency, because high-markup firms were too small to begin with from a social perspective, and boosts aggregate output.

IRS with Directed Entry. Consider the same model as above, but now suppose that there is directed entry into every k, with potential entrants choosing which k to enter into after paying a fixed cost in units of labor. From Theorem 3, changes in aggregate output are now given by

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

where the first term is the direct technology effect and the second term is the reallocation effect. Note that the reallocation effect is very different from what it was without entry. From Proposition 2, we can write

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) + \frac{1}{\mathbb{E}_{\lambda^{\mathrm{B}}}\left(\frac{\theta-1}{\theta-\theta_0}\right)} Cov_{\lambda^{\mathrm{B}}}\left(\frac{\theta-1}{\theta-\theta_0}, d\log A\right), \tag{16}$$

where $(\theta - 1)/(\theta - \theta_0)$ is a vector whose ith element is $(\theta_i - 1)/(\theta_i - \theta_0) > 0$. To understand the intuition, suppose that k's are substitutes ($\theta_0 > 1$) and that the shock disproportionately increases the productivity of k's with high consumer surplus $\gamma_i = (\theta_i - 1)/\theta_i$ (low elasticities of substitution). Then $Cov_{\lambda^B}((\theta - 1)/(\theta - \theta_0), d\log A) > 0$ and so the shock leads to improvements in allocative efficiency. Intuitively, the shock triggers beneficial reallocations of labor towards k's with strong scale economies which were too small to begin with from a social perspective. These forces operate in reverse when sectors are complements with $\theta_0 < 1$.

Comparing (15) to (16) reveals the importance of entry. The correlation between productivity shocks and markups, which was key in the economy without entry is now irrelevant. This is because now labor reallocations happen purely on the extensive margin via changes in entry in the different sectors, while the intensive margin remains unchanged as individual producers in the different sectors keep operating at the same scale. Instead, the key is now the correlation between productivity shocks and returns to scale.

DRS with Directed Entry. We now show that changes in aggregate output are very different under DRS. Consider the same example as above but assume that each k's output,

$$Y_k = M_k y_k,$$

is a linear aggregate of an endogenous mass M_k of undifferentiated varieties. Each variety in k is produced from labor with decreasing returns ε_k and sold at a markup μ_k over marginal cost

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

Changes in aggregate output are given by

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^B}(\mathrm{d}\log A). \tag{17}$$

Comparing (16) with (17) reveals the difference between IRS and DRS forms of entry. Changes in technical efficiency are captured by the same Hulten-like term as in the IRS case. By contrast, there are no longer any changes in allocative efficiency. This occurs even though there are equilibrium reallocations and the initial equilibrium is inefficient. The adjustment in the sizes of the different sectors happens entirely on the extensive margin via changes in entry. Individual producers in the different sectors keep operating at the same scale so that there is no change in the intensive margin. Since in addition there is no consumer surplus $\gamma_i = 1$, the price of each good at the sectoral level is the same as the price of the good for individual firms. Therefore, reallocations are therefore neutral on efficiency grounds. This example clarifies that the IRS and DRS models are, in general, very different.

IRS with Undirected Entry. Now consider again the IRS version of the example, but now suppose that there is only one entrant type. For simplicity, suppose that $\theta_k = \theta_0 > 1$ for this example. Now Theorem 3 implies that

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) + (\theta_0 - 1)^{-1}\,\mathrm{d}\log\left(\mathbb{E}_{\lambda^{B}}(\pi)\right), \tag{18}$$

where the second term captures the change in profits. Hence, allocative efficiency increases if profits increase as a share of GDP. Using Proposition 2, we can rewrite this in terms of primitives as

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) - \frac{1}{\lambda_E} Cov_{\lambda^{\mathrm{B}}}\left(\frac{1}{\mu}, \mathrm{d}\log A\right), \tag{19}$$

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

Comparing (19) to (16) reveals the importance of directed entry. Whereas for directed entry, it is a comparison of θ_k's that determines allocative efficiency, for undirected entry, it is a comparison of markups μ_k that does so. In that sense, the model with undirected entry is more similar to the one without entry (15) but there are differences here too. Although the sign of the change in allocative efficiency is the same in both (19) and (15), their magnitude is different. In both cases, improvements in efficiency, brought about by reallocation of labor to high-markup firms, economize on labor. However, when there is no entry, the labor saved is used towards variable production by incumbents, but in the model with entry it is used for the entry and variable production of new firms.

DRS with Undirected Entry. Finally, we consider the DRS model with undirected entry. For simplicity, we assume that $\varepsilon_k = \varepsilon$ for every k. Applying Theorem 3 gives

$$d\log Y = \mathbb{E}_{\lambda^{B}}(d\log A) + (1 - \varepsilon) d\log\left(\mathbb{E}_{\lambda^{B}}(\pi)\right),$$

which is very similar to (18). Indeed, using Proposition 2 we can rewrite this in terms of primitives as

$$\mathrm{d}\log Y = \mathbb{E}_{\lambda^{\mathrm{B}}}(\mathrm{d}\log A) - \frac{\epsilon}{\lambda_E} Cov_{\lambda^{\mathrm{B}}}\left(\frac{1}{\mu}, \mathrm{d}\log A\right). \tag{20}$$

Therefore, in this example with a single sector and undirected entry, the IRS and DRS models behave similarly as a comparison of (19) and (20) shows. This is in keeping with folk intuition that Hopenhayn (1992) and the closed-economy version of Melitz (2003) are isomorphic to one another. However, these examples show that this intuition is highly fragile, and a simple change like the extent to which entry is directed, can break the equivalence.

Overall, these examples underscore the importance of modeling the extent to which entry is directed as well as whether or not firms provide value because of consumer surplus (IRS) or producer surplus (DRS).

7 First-Best Policy and Misallocation

In this section, we use the results in Section 5 and 6 to 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 diminishing marginal cost, the distance to the frontier can be approximated via a Domar-weighted sum of Harberger triangles associated with variable production and entry. We specialize this 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 Theorem 1. We provide a second-order approximation in these deviations $(\mathrm{d}\log \tau, \mathrm{d}\log \tau^Y)$ of the associated aggregate efficiency loss $\mathcal{L} = -(1/2)\,\mathrm{d}^2\log Y$.[23]

Proposition 3 (Deadweight-Loss). *As long as either $\varepsilon_i < 1$ or $\gamma_i > 1$ for each $i \in \mathcal{N}^c$, the*

[23] Around the efficient point, the first-order loss is zero as long as $\min\{\varepsilon_i, \gamma_i\} < 1$ (see Corollary 1 in Appendix C). If $\varepsilon_i = \gamma_i = 1$, and $i \in \mathrm{span}\{\zeta\}$, then the losses from inefficiencies are first-order, and we must use Theorem 3 instead.

efficiency loss can be approximated, up to second-order approximation, as

$$\mathcal{L} \approx \underbrace{\frac{1}{2}\sum_{i\in\mathcal{N}} \lambda_i^B \,\mathrm{d}\log y_i \,\mathrm{d}\log\left(\tau_i\tau_i^Y\right)}_{\text{Harberger Triangles for Variable Production}} + \underbrace{\frac{1}{2}\sum_{i\in\mathcal{N}} \lambda_i^B \gamma_i \,\mathrm{d}\log M_i \,\mathrm{d}\log \tau_i^Y}_{\text{Harberger Triangles for Entry}}.$$

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 conjunction with the forward and backward propagation equations in Propositions 1 and 2, we can rewrite these loss functions in terms of microeconomic primitives (the input-output matrix, the elasticities of substitution, and returns to scale).[24] 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.

These examples help demonstrate how the social cost of distortions changes in models where entry occurs with *IRS* versus *DRS*.

7.1 Sectoral Models

To generate examples, we use *sectoral* models. Sectoral models are common in the literature, and they are worth singling out because for this class of economies we can break the problem of computing the distance to the frontier into two blocks: within and across sectors. A sectoral model satisfies the following conditions (see Appendix E for detailed derivations):

1. every producer type $i \in \mathcal{N} - \mathcal{F}$ 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

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

   depending on whether $\mathcal{I}$ is DRS or IRS, where $f_{\mathcal{I}}\left(\{x_{i\mathcal{J}}\}\right)$ has constant-returns, and $x_{i\mathcal{J}}$ indicates that inputs are purchased from other sectoral aggregates $\mathcal{J}$;

2. 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;

[24] To do this, note that $\mathrm{d}\log Y_i = \mathrm{d}\log\lambda_i^B - \mathrm{d}\log P_i$, where Proposition 1 gives $\mathrm{d}\log P_i$ and Proposition 2 gives $\mathrm{d}\log\lambda_i^B$. Next, observe that $\mathrm{d}\log Y_i = \mathrm{d}\log y_i + 1/\gamma_i \,\mathrm{d}\log M_i$. Finally, note that $\mathrm{d}\log M$ is given by (14). Putting this all together will allow us to write Proposition 3 in terms of primitives.

3. individual producers i in sector $\mathcal{I}$ charge different markups τ_i^y but face a common industry-level wedge $\tau_i^Y = \tau_{\mathcal{I}}^Y$.

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 $\mathbb{E}_{\lambda^{\mathcal{I},\mathrm{B}}}(\mathrm{d}\log\tau^y)$ 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 $\lambda_i^{\mathcal{I},\mathrm{B}}$.

We now discuss the distance to the frontier for IRS and DRS sectoral models, starting with the simpler DRS case.

7.2 Misallocation in DRS Economies

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 is 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 4 (Deadweight-Loss in DRS Economy with Entry)**.** *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*

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

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. 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 4 is surprising if one is familiar with the misallocation literature. Normally, elasticities of substitution are key pieces of information but here they are irrelevant. This is because changes in sectoral markups do not change relative sectoral prices to a first-order, meaning that allocations are not distorted across sectors to a first-order. An increase in markups reduces the scale and hence the marginal cost of producers and starting at the efficient point, this reduction in marginal cost exactly offsets the increase in markups to a first-order. Since relative sectoral prices do not change to a first-order, the elasticity of substitution across sectors is not relevant for how sectoral quantities adjust to a first-order. Since Harberger triangles are products of first-order changes in quantities and first-order changes in the wedges (see Proposition 3), the cross-sectoral elasticity of substitution is irrelevant to a second-order.

7.3 Misallocation in IRS Economies

Whereas Proposition 4 provides a relatively general characterization of losses for DRS economies, the behavior of IRS economies is substantially more complicated. Rather than writing 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 Economy. We start with a one-sector model heterogenous-firm economy. Aggregate output is given by

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

Each good i is produced from labor with constant returns and productivity A_i. If there is no entry, applying Proposition 3, the aggregate efficiency loss from markups is, to a second-order,

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

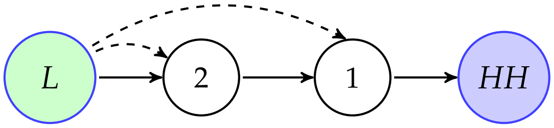
The loss is increasing in the elasticity of substitution θ and the dispersion of markups. Importantly, the level of markups do not matter, only their dispersion matters. This formula is standard in the misallocation literature (see e.g. Hsieh and Klenow, 2009; Baqaee and Farhi, 2019a).

Now consider the case where there is free-entry paid in units of labor. In this case, the equilibrium is efficient if markups are equal to $\theta/(\theta-1)$ for every producer. If markups deviate from this efficient benchmark, then the aggregate efficiency loss is, to a second-order, given by

$$\mathcal{L} = \frac{1}{2}\theta Var_{\lambda^{\mathrm{B}}}\left(d\log\tau^{y}\right) + \frac{1}{2}\theta\mathbb{E}_{\lambda^{\mathrm{B}}}\left(d\log\tau^{y}\right)^{2}. \qquad (21)$$

The first term is the same as before. 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 is also increasing in the elasticity of substitution, but it depends on the level of the wedges rather than their dispersion. Comparing this to Proposition 4 shows that, at least in this simple case, the loss function for IRS models is reminiscent of the one for DRS models. We shall see that these similarities disappear if allow for input-output linkages in either variable production or entry.

Intermediates in Variable Costs. Consider the supply chain depicted in Figure 4 where entry costs are paid in units of labor but there are input-output linkages in variable production. Proposition 4 shows that if the economy were of the DRS type, then the losses should simply be a linear combination of scale elasticities — each scale elasticity depends on the size and wedges in its own sector. Furthermore, losses are monotone in those scale elasticities. None of this is true in IRS models with input-output linkages.



$$Y_I = \left(\sum_{i\in I} M_i y_i^{\frac{\theta_i-1}{\theta_i}}\right)^{\frac{\theta_i}{\theta_i-1}}, \quad y_{i2} = A_{i2} l_{i2},$$

$$y_{i1} = A_{i1} x_{12}, \qquad L = \sum_{i\in 2} M_{i2} l_{i2} + M_1 + M_2.$$

Figure 4: The solid and dashed arrows represent the flow of resources for production and entry, respectively. The only primary factor is labor indexed by L. If variety i belongs to sector $I = \{1,2\}$, we write $i \in I$, and the output of i is denoted by y_{iI}. Variable intermediate and labor inputs by i in I are denoted by x_{iI} and l_{iI}.

Proposition 3 implies the loss function is given by

$$\mathcal{L} = \frac{1}{2}\sum_{I=1}^{2} \theta_i \left[Var_{\lambda^{I,B}}\left(d\log\tau^y\right) + \mathbb{E}_{\lambda^{i,B}}\left(d\log\tau^y\right)^2 \right] + \frac{1}{2}\frac{\theta_1}{\theta_1\theta_2 - 1} E_{\lambda^{1,B}}\left(d\log\tau^y\right)^2,$$

which is a nonlinear combination of the scale elasticities (in this case, θ_1 and θ_2). The first set of summands, which capture distortions along the intensive and extensive margin of each $I = \{1,2\}$ are similar to DRS goods (see Proposition 4 for comparison). However, unlike with DRS goods, it is not possible to separate sectoral distortions from one another. The last summand combines scale elasticities across the two sectors and makes the losses from an increase in downstream markups $d\log\tau_1^y$ depend on the elasticity of substitution upstream θ_2.

Furthermore, unlike in Proposition 4, losses can be U-shaped in the scale elasticities. This is because, on the one hand, a lower elasticity of substitution reduces the costs of misallocation by making variable quantities less sensitive to wedges. On the other hand, a lower elasticity of substitution makes misallocation along the extensive margin more costly since a lower elasticity of substitution. This non-monotonicity is also evident in our quantitative model in Section 9.

Intermediates in Fixed Costs. Non-monotonicity of losses in scale elasticities can also arise due to input-output linkages in fixed costs. To see this, consider a one-sector economy with a representative firm and free entry. Suppose that entry costs require the use of either labor or goods. The gross output of the sector 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 $y_1 = A_1 l_1$. Applying Proposition 3, the aggregate efficiency loss from a wedge τ_1^y when entry only uses labor is

$$\mathcal{L} = \frac{1}{2}\theta_1 (d \log \tau_1^y)^2.$$

This is just a special case (21) and the loss is increasing in the elasticity of substitution across products θ_1. Losses explode as θ_1 goes to infinity since in this limit, entry is socially wasteful and highly responsive to a change in the markup, so the losses from any amount of entry are first-order (this is why the second-order approximation explodes).

Next, suppose that entry uses only goods and labor is not required for entry (but it is required for variable production). Applying Proposition 3, the aggregate efficiency loss from markups is

$$\mathcal{L} = \frac{1}{2} \frac{(\theta_1 - 1)^3}{(\theta_1 - 2)^2} (d \log \tau_1^y)^2.$$

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. Once again, the long-standing intuition that efficiency losses are increasing in the elasticity of substitution is broken.

8 Second-Best Policy Interventions

Whereas Section 7 builds on Theorem 3 to analyze the gains from first-best policy, this section uses Theorem 3 to provide bang-for-buck formulas to compare the merits of small interventions starting at an inefficient equilibrium. 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 on the nature of the intervention, the shape of the production network, and the strength of scale economies.

In this section, we restrict attention to an analytically tractable and quantitatively relevant special case of Theorem 3. We focus on an economy where all goods are IRS, $\gamma_i > 1$ is constant for each i, there is one primary factor we call labor (indexed by L), and the production and entry functions f_i and g_i are Cobb-Douglas.[25] We focus on the no-intervention equilibrium with monopolistically competitive Dixit and Stiglitz (1977) markups. Finally, we assume that entry is directed. Given the fact that elasticities of substitution across producers i are equal to one and entry is directed, each i should be interpreted as an industry. We investigate markup regulation and entry subsidization, which can be thought of as capturing competition and industrial policy respectively.

Markup Regulation. To start with, consider a budget-neutral intervention reducing the markups $\mathrm{d}\log \mu_i < 0$ of industry 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 $-\lambda_i^{\mathrm{B}}\,\mathrm{d}\log \mu_i > 0$ transferred away from the producers by the associated implicit subsidy, is

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

Markup regulations are more effective the larger is the right-hand side. This happens when i is downstream from j's who have strong returns to scale and are themselves upstream of other industries with strong scale effects. Intuitively, Ψ_{ij}^{B} captures the reliance of i on inputs from j, whereas $(\gamma_j - 1)$ captures the strength of scale economies in j, and λ_j^F/λ_j^B captures the cumulation of markups downstream from j. If all markets have the same increasing returns to scale γ_j, then this formula favors markup reductions in industries that are relatively downstream.

Entry Subsidies. Next, consider entry subsidies to industry i at the no-intervention equilibrium. Denote a negative output tax on the fixed costs of i by $\mu_{E,i}$. At the no-intervention equilibrium, $\mu_{E,i} = 1$, the budgetary impact of this is just $-\lambda_{E,i}^{\mathrm{B}}\,\mathrm{d}\log \mu_{E,i}^{Y} > 0$. We normalize the response of aggregate output by its budgetary impact to allow bang-for-buck comparisons. When production functions are Cobb-Douglas, Theorem 3 implies that this is

$$-\frac{1}{\lambda_{E,i}^{\mathrm{B}}}\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_{E,i}^{Y}} = \frac{\lambda_i^{\mathrm{F}}}{\lambda_i^{\mathrm{B}}}\gamma_i - \lambda_L^{F}.$$

[25] The consumer surplus ratio γ_i is constant when the industry-level aggregator F_i is isoelastic.

Here, $\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}}$ is again a measure of the cumulated markups (and returns to scale) downstream from i and λ_L^F is the forward Domar weight of labor. The adjusted ratio $(\lambda_i^{\mathrm{F}}/\lambda_i^{\mathrm{B}})\gamma_i$ includes the gross increasing returns of i rather than only those of markets strictly downstream from i. Hence, the greatest improvements come from subsidizing entry into those markets that are upstream in supply chains with strong returns to scale. In other words, whereas markup regulations tend to target sectors that are downstream of long supply chains with strong scale economies, entry subsidies will tend to target sectors which are upstream of long supply chains with strong scale economies.[26] In either case, however, the goal is to boost the sales of, and entry into, sectors that are upstream of long supply chains with strong scale economies.

9 Quantitative Illustration

We end our analysis by illustrating the social cost of distortions, or equivalently the gains from optimal policy, using a quantitative model. We provide a brief account of how we calibrate the model to fit U.S. data in Section 9.1 and present the numerical results in Section 9.2. For more details on the data sources and calibration see Appendix G.

9.1 Description of Quantitative Model

The quantitative model has a sectoral structure with heterogenous firms within sectors and one primary factor capturing a composite of capital and labor. 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 (excluding government sectors), and assign each firm in our Compustat sample to a BEA industry. From the data, we have estimates of 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.

For firm-level markups, we adopt the benchmark procedure of De Loecker et al. (2019) using a production function estimation approach. In Appendix H, 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

[26] For a quantitative illustration of these bang-for-buck formulas, see Appendix I.1.

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 $1/\gamma = 0.875$ or $1/\gamma = 0.75$, 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.

9.2 Social Costs of Distortions

We solve the model nonlinearly and compute the efficiency loss from misallocation. We report the numbers as the percentage gain in welfare achieved by implementing optimal policy starting from the decentralized equilibrium outcome. The results are in Table 1 for different combinations of assumptions regarding entry and returns to scale. Across the board, the benchmark calibration shows that the losses from inefficiency are higher (roughly double) when we allow entry than when we do not because of the additional distortions along the entry margin.

Decomposing the Results. For each calibration, Table 1 breaks down 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 γ_i (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, $1/\gamma = 0.875$	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, $1/\gamma = 0.75$			
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 $1/\gamma = 0.875$ under IRS corresponds to elasticity of substitution across firms within sectors equal to 8, whereas $1/\gamma = 0.75$ corresponds to elasticity of substitution equal to 4.

The first column of Table 1 shows that when there is no entry, almost the entirety of the loss is 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 second and third columns), 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 high resulting in excessive entry.

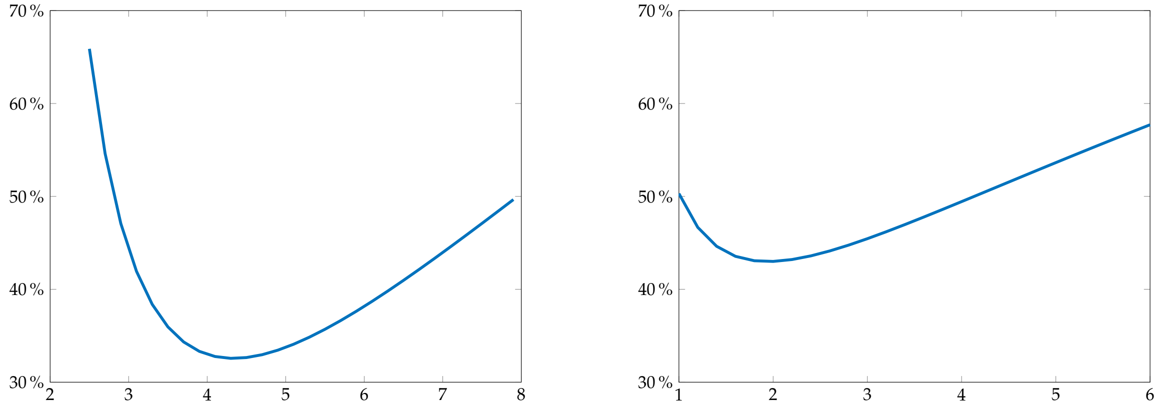
Whether entry only uses primary factors or also intermediates has ambiguous effects. Depending on the scale elasticities, the relative size of the gains can go either way. When the entry margin is more important (ε and $1/\gamma$ are lower), the gains tend to be higher when entry also uses intermediates.

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 first point, note that when firms have strong diminishing returns, the effect of markups on relative prices is off-set, to some extent, by a counteracting change in marginal costs. To understand the latter point, it is useful to think about the limit where ε and $1/\gamma$ 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 3. By contrast, under DRS, the efficiency losses go to zero as made clear by Proposition 4.

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 7.3, this intuition fails when there is entry, there is IRS, and input-output linkages.

Figure 5a shows that for the IRS benchmark, the distance to the frontier is U-shaped as a function of the within-sector elasticity of substitution $1/(1-1/\gamma)$. For instance, the losses are 50% when $1/(1-1/\gamma) = 8$. 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 two example of Section 7.3. Intuitively, a lower elasticity reduces the misallocation costs along the intensive margin but magnifies the misallocation costs along the extensive margin. With non-trivial input-output linkages, downstream markups shrink the scale, and therefore the mass of entrants, in upstream sectors. When scale elasticities in upstream sectors are large, this distortion in the mass of entrants upstream is very costly. In the limit where the elasticity goes to one (γ goes to infinity), this type of misallocation along the extensive margin becomes infinitely costly.

Role of Barriers to Entry. In our benchmark specifications with entry, we assume that, with the exception of rents earned by primary factors, all rents are quasi-rents rather than pure rents. That is, the zero-profit condition holds in every sector. 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



(a) Elasticity of substitution

(b) Entry wedges

Figure 5: 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.

entry in reduced form by introducing an entry tax/wedge.

Figure 5b displays the implied 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 if there is free-entry, then there is too much entry in the equilibrium. As a result, if one takes the view that there are entry barriers in the data (so that there is less entry than implied by profits), then one is lead to a lower estimate of the distance to the frontier up (up to some point, after which, entry becomes inefficiently too low).

10 Conclusion

Traditional theories of aggregation, by relying on aggregate envelope theorems, imply that the aggregate production function can be treated like a black-box whose contents are irrelevant to a first-order approximation. In this case, aggregate productivity changes are simply the sales-weighted averages of the exogenous microeconomic productivity shocks. Under this view, these exogenous changes in aggregate productivity are responsible for a large fraction of both the cycle and the trend in aggregate output.

For inefficient economies, this approach is untenable. In a disaggregated economy, where many different margins can be distorted, total factor productivity is endogenous and affected, to a first-order, by reallocation effects. Furthermore, unlike technical know-how, which likely grows gradually and always increases over time, reallocation effects can be abrupt, increase or decrease welfare, and plausibly explain a non-trivial fraction of both the cycle and the trend. This paper shows that these reallocation effects can be potent, and their sign and magnitude is intricately connected to assumptions about scale elasticities and the nature of entry.

References

Acemoglu, D. and P. D. Azar (2020). Endogenous production networks. *Econometrica 88*(1), 33–82.

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

Acemoglu, D. and A. Tahbaz-Salehi (2020). Firms, failures, and fluctuations: the macroeconomics of supply chain disruptions. Technical report, National Bureau of Economic Research.

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

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

Baqaee, D., E. Farhi, and K. Sangani (2020). The darwinian returns to scale. Technical report.

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.

Bilbiie, F., F. Ghironi, and M. Melitz (2012). Endogenous entry, product variety and business cycles. *Journal of Political Economy 120*(2), 304–345.

Bilbiie, F. O., F. Ghironi, and M. J. Melitz (2019). Monopoly power and endogenous product variety: Distortions and remedies. *American Economic Journal: Macroeconomics 11*(4), 140–74.

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

Boehm, J. and E. Oberfield (2020). Misallocation in the market for inputs: Enforcement and the organization of production. *The Quarterly Journal of Economics 135*(4), 2007–2058.

Carvalho, V. M. and A. Tahbaz-Salehi (2018). Production networks: A primer.

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.

Dhyne, E., A. K. Kikkawa, M. Mogstad, and F. Tintelnot (2021). Trade and domestic production networks. *The Review of Economic Studies 88*(2), 643–668.

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.

Elliott, M., B. Golub, and M. V. Leduc (2020). Supply network formation and fragility. *Available at SSRN 3525459*.

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.

Gupta, A. (2020). Firm heterogeneity, demand for quality and prices: Evidence from india. Technical report.

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

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

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

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

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

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

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

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

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.

Kikkawa, A. K., G. Magerman, E. Dhyne, et al. (2018). Imperfect competition in firm-to-firm trade. Technical report.

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

La'O, J. and A. Tahbaz-Salehi (2020). Optimal monetary policy in production networks. Technical report, National Bureau of Economic Research.

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

Lim, K. (2017). Firm-to-rm trade in sticky production networks.

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.

Lucas, R. E. (1978). On the size distribution of business firms. *The Bell Journal of Economics*, 508–523.

Matsuyama, K. and P. Ushchev (2017). Beyond ces: three alternative classes of flexible homothetic demand systems. *Global Poverty Research Lab Working Paper* (17-109).

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.

Oberfield, E. (2017). A theory of input-output architecture. Technical report.

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.

Taschereau-Dumouchel, M. (2020). Cascades and fluctuations in an economy with an endogenous production network. *Available at SSRN 3115854*.

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