Productivity and Misallocation in General Equilibrium

David Rezza Baqaee
LSE

Emmanuel Farhi*
Harvard

February 24, 2018

Abstract

We provide a general non-parametric formula for aggregating microeconomic shocks in general equilibrium economies with distortions such as taxes, markups, frictions to resource reallocation, and nominal rigidities. We show that the macroeconomic impact of a shock can be boiled down into two components: its "pure" technology effect; and its effect on allocative efficiency arising from the reallocation of resources (measured via changes in factor income shares). We derive a formula showing how these two components are determined by structural microeconomic parameters such as elasticities of substitution, returns to scale, factor mobility, and network linkages. Overall, our results generalize those of Solow (1957) and Hulten (1978) to economies with distortions. To demonstrate their empirical relevance, we pursue different applications, focusing on markup distortions. We operationalize our non-parametric results and show that improvements in allocative efficiency account for about 50% of measured TFP growth over the period 1997-2015. We also implement our structural results and conclude that eliminating markups would raise TFP by about 20%, increasing the economywide cost of monopoly distortions by two orders of magnitude compared to the famous 0.1% estimate ny Harberger (1954).

*Emails: d.r.baqaee@lse.ac.uk, efarhi@harvard.edu. We thank Philippe Aghion, Pol Antras, Susanto Basu, John Geanakoplos, Ben Golub, Gita Gopinath, Dale Jorgenson, Marc Melitz, Ben Moll, Matthew Shapiro, Dan Trefler, and Jaume Ventura for their valuable comments. We thank German Gutierrez, Thomas Philippon, Jan De Loecker, and Jan Eeckhout for sharing their data. We thank Thomas Brzustowski for excellent research assistance. We are especially grateful to Natalie Bau for detailed conversations.

1 Introduction

The foundations of macroeconomics rely on Domar aggregation: changes in a constant-returns-to-scale index are approximated by a sales-weighted average of the changes in its components.[1] Hulten (1978), building on the work of Solow (1957), provided a rationale for using Domar aggregation to interpret the Solow residual as a measure of aggregate TFP. He showed that in efficient economies

$$\frac{\Delta Y}{Y} - \sum_f \Lambda_f \frac{\Delta L_f}{L_f} \approx \sum_i \lambda_i \frac{\Delta TFP_i}{TFP_i},$$

where Y is real GDP, L_f is the supply of factor f, Λ_f is its of total income share in GDP, TFP_i is the TFP of producer i, λ_i is its sales as a share of GDP.

Although Hulten's theorem is most prominent for its use in growth accounting, where it is employed to measure movements in the economy's production possibility frontier, it is also *the* benchmark result in the resurgent literature on the macroeconomic impact of microeconomic shocks in mutisector models and models with production networks.[2]

The non-parametric power of Hulten's theorem comes from exploiting a macro-envelope condition resulting from the first welfare theorem. This requires perfect competition and Pareto-efficiency. Without these conditions, Hulten's theorem generally fails.[3]

Our paper generalizes Hulten's theorem beyond efficient economies, and provides an aggregation result for economies with arbitrary neoclassical production functions, input-output networks, and distortion wedges. Rather than relying on a macro-envelope condition like the first welfare theorem, our results are built on micro-envelope conditions: namely that all producers are cost minimizers. Our result suggests a new and structurally interpretable decomposition of changes in aggregate TFP into "pure" changes in technology and changes in allocative efficiency. It provides a unified framework for analyzing the effects of distortions and misallocation in general equilibrium economies, the study of which is the subject of a vibrant literature, recently reinvigorated by Restuccia and

[1] Although we refer to this idea as Domar aggregation, after Evesy Domar, the basic idea of using sales shares to weight changes in a price or quantity can be traced back at least to the early 18th century writer William Fleetwood. We refer to this idea as Domar aggregation, since Domar (1961) was the first to propose it in the context we are interested in: creating an index of aggregate technical change from measures of microeconomic technical change.

[2] See for example Gabaix (2011), Acemoglu et al. (2012), Carvalho and Gabaix (2013), Di Giovanni et al. (2014), Baqaee and Farhi (2017) amongst others.

[3] See for example the papers by Basu and Fernald (2002), Jones (2011), Jones (2013), Bigio and La'O (2016), Baqaee (2016), or Liu (2017) who explicitly link their inefficient models with the failure of Hulten's result. Some papers which study distorted networked economies (but place less of a focus on how their results compare to Hulten's), are Grassi (2017), Caliendo et al. (2017), Bartelme and Gorodnichenko (2015).

Rogerson (2008) and Hsieh and Klenow (2009).[4]

Loosely speaking, when a producer becomes more productive, the impact on aggregate TFP can be broken down into two components. First, given the initial distribution of resources, the producer increases its output, and this in turn increases the output of its direct and indirect customers; we call this the "pure" technology effect. Second, the distribution of resources across producers shifts in response to the shock, increasing some producers' output and reducing that of others; we call the impact of this reallocation of resources on aggregate TFP the change in allocative efficiency. In efficient economies, changes in allocative efficiency are zero to a first order, and so the overall effect characterized by Hulten (1978) boils down to the "pure" technology effect. In inefficient economies, changes in allocative efficiency are nonzero in general. Our theoretical contribution is to fully characterize the macroeconomic impact of microeconomic shocks as well as their decomposition into "pure" technology effects and changes in allocative efficiency in inefficient economies.

We present both ex-post and ex-ante results. The ex-post reduced-form results do not require any information about the microeconomic production functions besides input-output expenditure shares. The downside of these results is that they depend on the observation of factor income shares before and after the shock. The second set of results are ex-ante structural results. Although they do not necessitate ex-post information, they require information about microeconomic elasticities of substitution. Using this information in conjunction with input-output expenditure shares, we can deduce the implied changes in factor income shares. As a side benefit, our ex-ante results determine how factor income shares respond to shocks for a general neoclassical production structure, which is a question of independent interest in studies of inequality.[5]

To demonstrate the empirical relevance and the scope of applicability of our framework, we put it to use to answer three different questions about the role of markups on aggregate productivity. We focus on markups in light of the accumulating evidence that average markups have increased over the past decades in the US.[6]

1. How have changes in allocative efficiency contributed to measured TFP growth in

[4]Some other prominent examples are Hopenhayn and Rogerson (1993), Banerjee and Duflo (2005), Chari et al. (2007), Guner et al. (2008), Buera et al. (2011), Epifani and Gancia (2011), Fernald and Neiman (2011), Buera and Moll (2012), D'Erasmo and Moscoso Boedo (2012), Bartelsman et al. (2013), Caselli and Gennaioli (2013), Oberfield (2013), Peters (2013), Reis (2013), Asker et al. (2014), Hopenhayn (2014), Moll (2014), Midrigan and Xu (2014), Sandleris and Wright (2014), Edmond et al. (2015), Gopinath et al. (2017), and Sraer and Thesmar (2018).

[5]See, for example, Piketty (2014), Elsby et al. (2013), Barkai (2016), Rognlie (2016), Koh et al. (2016), and Gutierrez (2017).

[6]See Barkai (2016) and Caballero et al. (2017) for arguments using aggregate data, and Gutierrez (2017), and De Loecker and Eeckhout (2017) for evidence using firm-level data.

the US over the past 20 years?

We perform a non-parametric decomposition of measured TFP growth as captured by the Solow residual into a "pure" technology effect and an allocative efficiency effect.[7] We implement our Solow residual decomposition in the US over the period 1997-2014. Focusing on markups as a source of distortions, we find that the improvement in allocative efficiency accounts for about 50% of the cumulated Solow residual. This occurs despite the fact that average markups have been increasing.

A rough intuition for this surprising result is that average markups have been increasing primarily due to an across-firms composition effect, whereby firms with high markups have been getting larger, and not a within-firm increase in markups.[8] From a social perspective, these high-markup firms were too small to begin with, and so the reallocation of factors towards them has improved allocative efficiency and TFP.

2. What are the gains from reducing markups in the US, and how have these gains changed over time?

   Using our structural results, we find that in the US in 2014-2015, eliminating markups would raise aggregate TFP by about 20%. This increases the estimated cost of monopoly distortions by two orders of magnitude compared to the famous estimates of 0.1% of Harberger (1954).[9]

   Essentially, the reasons for this dramatic difference are that we use firm-level data, whereas Harberger only had access to sectoral data, and that the dispersion of markups is higher across firms within a sector than across sectors. Moreover, the relevant elasticity of substitution is higher in our exercise than in Harberger's since it applies across firms within a sector rather than across sectors. Finally, we properly take into account the general equilibrium input-output structure of the economy to aggregate the numbers in all industries whereas Harberger focused on manufacturing and ignored input-output linkages.

[7] There is also an additional effect, reminiscent of Hall (1990), due to the fact that the Solow residual does not weigh changes in factor shares correctly in the presence of distortions.

[8] This is consistent with Vincent and Kehrig (2017) and Autor et al. (2017) who argue that the labor share of income has decreased because more low labor share firms have become larger, and not because the labor share has declined within firms.

[9] Harberger's result had a profound impact on the economics discipline by providing an argument for de-emphasizing microeconomic inefficiencies in comparison to Keynesian macroeconomic inefficiencies. This impact is perhaps best illustrated by Tobin's famous quip that "it takes a heap of Harberger triangles to fill an Okun gap".

Like Harberger, we measure only the static gains from eliminating markups, holding fixed technology, abstracting away from the possibility that lower markups may reduce entry and innovation. In other words, even if markups play an important role in incentivizing entry and innovation, their presence also distorts the allocation of resources, and this latter effect is what we quantify.

Interestingly, we also find that the gains from reducing markups have increased since 1997. Roughly speaking, this occurs because the dispersion in markups has increased over time. This finding may appear to contradict our conclusion that allocative efficiency has made a positive contribution to measured TFP growth over the period. The resolution is that these results are conceptually different: one is about the contribution of changes in allocative efficiency to measured TFP growth along the observed equilibrium path of the economy, while the other one is about the comparison of the distance from the efficient production possibility frontier at the beginning and at the end of the sample. This distinction highlights the subtleties involved in defining and interpreting different notions of allocative efficiency.

3. How do markups affect the macroeconomic impact and diversification of microeconomic shocks?

   Our ex-ante structural results allow us to conclude that markups materially affect the impact of microeconomic productivity and markup shocks on output, both at the sector and at the firm level. They amplify some shocks and attenuate others. Unlike a perfectly competitive model, shocks to industries and firms have different effects on output, even controlling for size. Firm-level shocks trigger larger reallocations of resources across producers than industry-level shocks (since firms are more substitutable). On the whole, we find that output is more volatile than in a perfectly competitive model, especially with respect to firm-level shocks.

Finally, we compare our results to other decompositions of the effects of technology on aggregate TFP into "pure" technology effects and changes in allocative efficiency that have been proposed in the growth accounting literature. We focus in particular on the prominent contributions of Basu and Fernald (2002) and Petrin and Levinsohn (2012). These decompositions are closest in spirit to our ex-post non parametric results since they rely on ex-post data, but they are different. We contrast these decompositions and argue why we find ours preferable.

Despite their generality, our results have some important limitations. First, our basic framework accommodates neoclassical production with decreasing or constant returns to scale. It can also easily handle fixed costs, as long as production has constant or increasing

marginal cost. However, it is unable to deal with non-neoclassical production featuring increasing returns such as those studied by Baqaee (2016), where by *increasing returns*, we refer to a situation where marginal variable costs are decreasing in output. Towards the end of the paper, we sketch how our results can be extended to cover such cases, when we discuss entry and exit. Second, in this paper we focus on first-order approximations. We show that under some conditions, the nonlinear analysis of efficient economies in Baqaee and Farhi (2017) can be leveraged to characterize nonlinearities in the sort of inefficient economies studied in this paper. Finally, we model frictions using wedges, which we take as primitives. The advantage is that this means we can characterize the response of the decentralized equilibrium to a change in the wedges without committing to any particular theory of wedge determination. The downside is that this makes it hard to perform counterfactuals when wedges are endogenous. However, in these cases, our results are still relevant as part of a larger analysis that accounts for the endogenous response of wedges. As an example, in the appendix, we show how to use our results to analyze the effect of monetary policy and productivity shocks in a model with sticky prices.

The outline of the paper is as follows. In Section 2, we set up the general model, and we prove our non-parametric results. We also discuss how to interpret these results, and the data required to implement our formulas. In Section 3, we introduce a parametric version of the general model and present our structural results. We use a model with CES production and consumption functions, with an arbitrary number of nests, input-output patterns, returns to scale, and factors of production. In Section 4 we discuss some subtleties in implementing and interpreting our results, including how to deal with endogenous wedges. In Section 5, we apply our results to the data by performing non-parametric ex-post decompositions of the sources of growth in the US, as well as structural exercises measuring the TFP gains from markup reductions, aggregate volatility arising from micro shocks, and the macro impact of micro shocks, in a calibrated model. In Section 6, we describe extensions of our results to account for endogenous factor supply, fixed costs, entry, and nonlinearities.

2 General Framework and Non-Parametric Results

We set up our general framework, and characterize how shocks to wedges and productivity affect equilibrium output and TFP. We define our notion of change in allocative efficiency. We explain how it leads to a new decomposition of the Solow residual into changes in "pure" technology and changes in allocative efficiency. We end by discussing

the relationship between our results and the rest of the literature.

2.1 Set up

The model consists of N producers indexed by i and F inelastic factors indexed by f with supply L_f. The output of each producer is produced using intermediate inputs and factors, and is sold as an intermediate good to other producers and as a final good.

Final Demand

Final demand, or GDP, in the economy is represented as the maximization of a constant-returns aggregator of final demand for individual goods

$$Y = \max_{\{c_1,\ldots,c_N\}} \mathcal{D}(c_1, \ldots, c_N)$$

subject to the budget constraint

$$\sum_{i}^{N}(1 + \tau_{0i})p_i c_i = \sum_{f=1}^{F} w_f L_f + \sum_{i=1}^{N} \pi_i + \tau,$$

where p_i is the price of good i, w_f is the wage of factor f, τ_{0i} is the consumption wedge on good i, π_i is the profits of the producer of good i, and τ is a net lump-sum rebate.

Producers

Good i is produced by producer i according to a constant-returns technology described by the constant-returns cost function

$$\frac{1}{A_i}\mathbf{C}_i\left((1 + \tau_{i1})p_1, \ldots, (1 + \tau_{iN})p_N, (1 + \tau_{i1}^f)w_1, \ldots, (1 + \tau_{iF}^f)w_F\right) y_i,$$

where A_i is a Hicks-neutral productivity shifter, y_i is total output, τ_{ij} is the input-specific tax wedge on good j, and τ_{ig}^f is a factor-specific tax wedge on factor g. We assume that producer i sets a price $p_i = \mu_i \mathbf{C}_i/A_i$ equal to an exogenous markup μ_i over marginal cost $\mathbf{C}_i/A_i$.

General Equilibrium

Given productivities A_i, markups μ_i, wedges τ_{ij} and τ_{ij}^f, a competitive equilibrium is a set of prices p_i, factor wages w_f, intermediate input choices x_{ij}, factor input choices l_{if},

outputs y_i, and final demands c_i, such that: each producer minimizes its costs and charges the relevant markup on its marginal cost; final demand maximizes the final demand aggregator subject to the budget constraint, where profits and revenues from wedges are rebated lump sum; and the markets for all goods and factors clear.

Two Simplifications: Constant Returns to Scale and Markup-Wedge Equivalence

Without loss of generality, we exploit two simplifications. First, despite the specification of constant-returns cost functions, our setup can accommodate decreasing returns to scale. This is because decreasing returns to scale can be modeled with constant returns to scale and producer-specific fixed factors. Going forward, we proceed with our constant-returns setup with the understanding that it can be reinterpreted to capture decreasing returns provided that the original set of factor is expanded to include producer-specific fixed factors.

Second all the wedges τ_{ij} and τ_{ig}^{f} can be represented as a markups in a setup with additional producers. For example, the good-specific wedge τ_{ij} in the original setup can be modeled in a modified setup as a markup charged by a new producer which buys input j and sells it to producer i. Going forward, we take advantage of this equivalence and assume that all wedges take the form of markups.

2.2 Input-Output Definitions

To state our generalization of Hulten's theorem, we introduce some input-output notation and definitions. Our results are comparative statics describing how, starting from an initial decentralized equilibrium, the equilibrium level of output changes in responses to shocks to productivities A_k and markups μ_k. Without loss of generality, we normalize the initial productivity levels to one. We now define input-output objects such as input-output matrices, Leontief inverse matrices, and Domar weights. Each of these quantities has a revenue-based version and a cost-based version, and we present both. All these objects are defined at the initial competitive equilibrium.

Final Expenditure Shares

Let b be the $N \times 1$ vector whose ith element is equal to the share of good i in final expenditures

$$b_i = \frac{p_i c_i}{\sum_{j=1}^{N} p_j c_j},$$

where the sum of final expenditures $\sum_{j=1}^{N} p_j c_j$ is nominal GDP.

Input-Output Matrices

To streamline the exposition, we treat factors as special endowment producers which do not use any input to produce. We form an $(N+F)\times 1$ vector of producers, where the first N elements correspond to the original producers and the last F elements to the factors. For each factor, we interchangeably use the notation w_f or p_{N+f} to denote its wage, and the notation L_{if} or $x_{i(N+f)}$ to denote its use by producer i.

We define the revenue-based input-output matrix Ω to be the $(N+F)\times(N+F)$ matrix whose ijth element is equal to i's expenditures on inputs from j as a share of its total revenues

$$\Omega_{ij} \equiv \frac{p_j x_{ij}}{p_i y_i}.$$

The first N rows and columns of Ω correspond to goods, and the last F rows and columns correspond to the factors of production. Since factors require no inputs, the last F rows of Ω are identically zero.

Similarly, we define the cost-based input-output matrix $\tilde{\Omega}$ to be the $(N+F)\times(N+F)$ matrix whose ijth element is equal to the elasticity of i's marginal costs relative to the price of j

$$\tilde{\Omega}_{ij} \equiv \frac{\partial \log \mathbf{C}_i}{\partial \log p_j} = \frac{p_j x_{ij}}{\sum_{k=1}^{N+f} p_k x_{ik}}.$$

The second equality uses Shephard's lemma to equate the elasticity of the cost of i to the price of j to the expenditure share of i on j. Since factors require no inputs, the last F rows of $\tilde{\Omega}$ are identically zero.

Leontief Inverse Matrices

We define the revenue-based and cost-based Leontief inverse matrices as

$$\Psi \equiv (I-\Omega)^{-1} = I + \Omega + \Omega^2 + \ldots \quad \text{and} \quad \tilde{\Psi} \equiv (I-\tilde{\Omega})^{-1} = I + \tilde{\Omega} + \tilde{\Omega}^2 + \ldots.$$

While the input-output matrices Ω and $\tilde{\Omega}$ record the *direct* exposures of one producer to another, in revenues and in costs respectively, the Leontief inverse matrices Ψ and $\tilde{\Psi}$ record instead the *direct and indirect* exposures through the production network. This can be seen most clearly by noting that $(\Omega^n)_{ij}$ and $(\tilde{\Omega}^n)_{ij}$ measure the weighted sums of all paths of length n from producer i to producer j.

Domar Weights

We define the revenue-based Domar weight λ_i of producer i to be its sales share as a fraction of GDP

$$\lambda_i \equiv \frac{p_i y_i}{\sum_{j=1}^{N} p_j c_j}.$$

Note that $\sum_{i=1}^{N} \lambda_i > 1$ in general since some sales are not final sales but intermediate sales. The accounting identity

$$p_i y_i = p_i c_i + \sum_j p_i x_{ji} = b_i \left(\sum_{j=1}^{N} p_j c_j \right) + \sum_j \Omega_{ji} p_j y_j$$

relates Domar weights to the Leontief inverse via

$$\lambda' = b'\Psi = b'I + b'\Omega + b'\Omega^2 + \dots . \tag{1}$$

Similarly, we define the vector of cost-based Domar weights to be

$$\tilde{\lambda}' \equiv b'\tilde{\Psi} = b'I + b'\tilde{\Omega} + b'\tilde{\Omega}^2 + \dots .$$

We choose the name cost-based Domar weight for $\tilde{\lambda}$ to contrast it with the traditional revenue-based Domar weight λ. Intuitively, $\tilde{\lambda}_k$ measures the importance of k as a supplier in final demand, both directly and indirectly through the network.[10] This can be seen most clearly by noting that the i-th element of $b'\tilde{\Omega}^n$ measures the weighted sum of all paths of length n from producer i to final demand.

For expositional convenience, for a factor f we use Λ_f and $\tilde{\Lambda}_f$ instead of λ_f and $\tilde{\lambda}_f$. Note that revenue-based Domar weight Λ_f of factor f is simply its income share.

[10] The cost-based Domar weight only depends on k's role as a supplier rather than its role as a consumer. It is also sometimes referred to as the *influence* vector, since in a certain class of models (like Jones, 2013; Acemoglu et al., 2012), it maps micro productivity shocks to output. We avoid this language since influence is an ambiguous term, and while the cost-based Domar weights are often-times useful in characterizing equilibria, they do not generally map productivity shocks to output. In other words, they are not generally equivalent to "influence."

Cross-Entropy

Our final input-output definition is the cross-entropy, which loosely speaking, is a measure of difference between distributions.[11] The cross-entropy between $\tilde{\Lambda}$ and Λ is

$$H(\tilde{\Lambda}, \Lambda) \equiv -E_{\tilde{\Lambda}}(\log \Lambda) = -\sum_{f=1}^{F} \tilde{\Lambda}_f \log \Lambda_f.$$

Here $\tilde{\Lambda}$ and Λ are seen as measures on the set of factors. Since $\sum_f \tilde{\Lambda}_f = 1$, the total mass of $\tilde{\Lambda}$ is equal to one and hence it can be interpreted as a probability distribution. By contrast, Λ is typically not a probability distribution since $\sum_f \Lambda_f \neq 1$ in general. However, we can always supplement Λ with the pure profit share $\Lambda_{f^*} = 1 - \sum_f \Lambda_f$ accruing to an extra factor f^* for which the cost-based share is zero $\tilde{\Lambda}_{f^*} = 0$, and hence recover a probabilistic interpretation.

For a given change $\mathrm{d}\log\Lambda$ in the probability distribution for Λ resulting from a combination of productivity shocks $\mathrm{d}\log A$ and wedge shocks $\mathrm{d}\log\mu$. We denote by

$$\mathrm{d}\,H(\tilde{\Lambda}, \Lambda) \equiv H(\tilde{\Lambda}, \Lambda + \mathrm{d}\,\Lambda) - H(\tilde{\Lambda}, \Lambda) = -\sum_{f=1}^{F} \tilde{\Lambda}_f \,\mathrm{d}\log\Lambda_f,$$

the change in relative entropy of Λ with respect to the *fixed* initial distribution $\tilde{\Lambda}$.

2.3 Comparative-Static Results

In this section, we derive our comparative-static results. Take as given the factor supplies L_f, the cost functions $\mathbf{C}_i$, and final Demand $\mathcal{D}$. Specify the vector of productivities A and an $(N+F)\times(N+F)$ admissible allocation matrix $\mathcal{X}$, where $\mathcal{X}_{ij} = x_{ij}/y_j$ is the share of the physical output y_j of producer j used by producer i. Denote by $\mathcal{Y}(A, \mathcal{X})$ the output Y achieved by this allocation.[12,13] Finally, define $\mathcal{X}_{ij}(A, \mu)$ to be equal to $x_{ij}(A, \mu)/y_j(A, \mu)$

[11] Cross-entropy between two distributions is minimized when the two distributions are the same. This definition is due to Claude Shannon (1948). Note that the Kullback and Leibler (1951) divergence is cross-entropy plus a constant, so that $\mathrm{d}\,H(\tilde{\Lambda}, \Lambda) = \mathrm{d}\,D_{KL}(\tilde{\Lambda}\|\Lambda)$ in our context. Formally, this divergence is not a distance function, but a premetric.

[12] The allocation matrix is admissible if the following conditions are verified: $0 \leq \mathcal{X}_{ij} \leq 1$ for all i and j; $\mathcal{X}_{ij} = 0$ for all j and for $N+1 \leq i \leq N+F$; $\sum_{i=1}^{N+F} \mathcal{X}_{ij} \leq 1$ for all $1 \leq j \leq N$; $\sum_{i=1}^{N+F} \mathcal{X}_{ij} = 1$ for all $N+1 \leq j \leq N+F$; and there exists a unique resource-feasible allocation such that the share x_{ij}/y_j of the output y_j of producer i which is used by producer j is equal to $\mathcal{X}_{ij}$, so that $\mathcal{X}_{ij} = \frac{x_{ij}}{y_j}$.

[13] To see how to construct this allocation, consider the production functions F_i defined as the duals of the cost functions C_i in the usual way. Then the vector of outputs y_i solves the system of equations $y_i = F_i(\mathcal{X}_{1i}y_1, \ldots, \mathcal{X}_{(N+F)i}y_{N+F})$ for $1 \leq i \leq N$ and $y_{N+f} = L_f$ for $1 \leq f \leq F$. The corresponding level of final

at the decentralized equilibrium when the vector of productivities is A and the vector of wedges is μ. The level of output at this equilibrium is given by $\mathcal{Y}(A, \mathcal{X}(A, \mu))$.

Now consider how the general equilibrium level of output changes in response to shocks $\mathrm{d}\log A$ and $\mathrm{d}\log \mu$:

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

The change in output can be broken down into two components: the direct or "pure" effect of changes in technology $\mathrm{d}\log A$, holding the distribution of resources $\mathcal{X}$ constant; and the indirect effects arising from the equilibrium changes in the distribution of resources $\mathrm{d}\mathcal{X}$. Essentially, changes in allocative efficiency are the gap that opens up following a shock between the equilibrium level of output and a *passive* allocation that just scales the initial allocation proportionately without allowing any other form of reallocation through substitution. The passive allocation constitutes a benchmark without reallocation, and so it stands a useful yardstick against which to measure changes in allocative efficiency in the equilibrium allocation.[14]

Now, we extend Hulten (1978) to cover inefficient economies and provide an interpretation for the result. We also extend Hulten's theorem along another dimension by characterizing changes in output following not only changes productivities but also changes in wedges.

Theorem 1. *Consider some distribution of resources $\mathcal{X}$ corresponding to the general equilibrium allocation at the point (A, μ), then*

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

where

$$\frac{\partial \log \mathcal{Y}}{\partial \log A}\,\mathrm{d}\log A = \tilde{\lambda}'\,\mathrm{d}\log A$$

consumption of good i is $c_i = y_i(1 - \sum_{j=1}^{N+F} \mathcal{X}_{ji})$ and the level of output is $\mathcal{D}(c_1, \ldots, c_N)$.

[14]The changes in the passive allocation in response to productivity shocks $\mathrm{d}\log A$ are easily derived. Since the elasticity of a production function to an input is given by its cost share, we have $\partial \log \mathcal{Y}/\partial \log A_i = \sum_{j=1}^{N} b_j \partial \log y_j/\partial \log A_i$, where $\partial y_j/\partial A_i$ must solve the following system of equations: $\partial \log y_j/\partial \log A_i = \sum_{k=1}^{N} \tilde{\Omega}_{jk} \partial \log y_k/\partial \log A_i + \delta_{ji}$. This implies that $\partial \log y_j/\partial \log A_i = b_j \tilde{\Psi}_{ji}$ and $\partial \log \mathcal{Y}/\partial \log A = \tilde{\Psi}' b' = \tilde{\lambda}'$. Moreover, the passive allocation is invariant to wedge shocks so that $\partial \log y_j/\partial \log \mu_i = 0$.

and

$$\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}} \mathrm{d}\,\mathcal{X} = -\tilde{\lambda}' \,\mathrm{d} \log \mu + \mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = -\tilde{\lambda}' \,\mathrm{d} \log \mu - \tilde{\Lambda}' \,\mathrm{d} \log \Lambda.$$

This implies that

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_k} = \tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d} \log A_k} = \tilde{\lambda}_k - \sum_f \tilde{\Lambda}_f \frac{\mathrm{d} \log \Lambda_f}{\mathrm{d} \log A_k}, \tag{2}$$

and

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log \mu_k} = -\tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d} \log \mu_k} = -\tilde{\lambda}_k - \sum_f \tilde{\Lambda}_f \frac{\mathrm{d} \log \Lambda_f}{\mathrm{d} \log \mu_k}. \tag{3}$$

Theorem 1 not only provides a formula for the macroeconomic output impact of microeconomic productivity and wedges shocks, but it also provides an interpretable decomposition of the effect. Specifically, the first component $(\partial \log \mathcal{Y}/\partial \log A)\,\mathrm{d} \log A = \lambda' \,\mathrm{d} \log A$ is the "pure" technology effect: the change in output holding fixed the share of resources going to each user; the second component $(\partial \log \mathcal{Y}/\partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = -\tilde{\lambda}' \,\mathrm{d} \log \mu + \mathrm{d}\, H(\tilde{\Lambda}, \Lambda)$ is the change in output resulting from the reallocation of shares of resources across users.

In the proof of Theorem 1 in the appendix, we also provide an explicit characterization of $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)/\,\mathrm{d} \log A_k$ and $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)/\,\mathrm{d} \log \mu_k$ in terms of the microeconomic elasticities of substitutions of the production functions and final demand, the properties of the input-output network, and the wedges. We present this characterization in the main body of the paper for the more special parametric version of the model in Section 3.

We can obtain Hulten's theorem as a special case of Theorem 1 when there are no wedges. Even this special case is actually a slight generalization of Hulten's theorem since it only requires the initial equilibrium to be efficient, whereas Hulten's theorem applies only to the case where the equilibrium is efficient before and after the shock.

Corollary 1 (Hulten 1978)**.** *If the initial equilibrium is efficient so that there are no wedges so that* $\mu = 1$ *, then*

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_k} = \lambda_k \quad \textit{and} \quad \frac{\mathrm{d} \log Y}{\mathrm{d} \log \mu_k} = 0.$$

In efficient economies, the first-welfare theorem implies that the allocation matrix $\mathcal{X}(A, \mu)$ maximizes output given resource constraints over all admissible allocations. The envelope theorem then implies that $(\partial \log \mathcal{Y}/\partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = 0$ so that there are no changes in allocative efficiency. Furthermore, because of marginal cost pricing, the direct effect of

changes in technology are based on the vector of sales shares or revenue-based Domar weights λ and are given by $(\partial \log \mathcal{Y}/\partial \log A)\, \mathrm{d} \log A = \lambda' \, \mathrm{d} \log A$. Hence, Hulten's theorem is a macro-envelope theorem of sorts.

When the initial equilibrium is inefficient so that $\mu \neq 1$, this macro-envelope theorem fails. Intuitively, in equilibrium, from a social perspective, some shares are too large and some shares are too small. Equilibrium changes in shares $\mathrm{d}\mathcal{X}$ can therefore lead to changes in output. This is precisely what we call a change in allocative efficiency $(\partial \log \mathcal{Y}/\partial \mathcal{X})\, \mathrm{d}\mathcal{X} = -\tilde{\lambda}' \, \mathrm{d} \log \mu + \mathrm{d}\, H(\tilde{\Lambda}, \Lambda)$, which is nonzero in general. Furthermore, because of wedges between prices and marginal costs, the direct effect of changes in technology are now based on the vector of cost-based Domar weights $\tilde{\lambda}$ rather than on the vector of revenue-based Domar weights λ and are given by $(\partial \log \mathcal{Y}/\partial \log A)\, \mathrm{d} \log A = \tilde{\lambda}' \, \mathrm{d} \log A$.

In the case of productivity shocks, Theorem 1 implies that changes in allocative efficiency are given by a simple sufficient statistic: the weighted average of the change in factor income shares $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = -\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$. This remarkable property show that it is not necessary to track how the allocation of every single good is changing across its users. Instead, it suffices to track how factor income shares change.

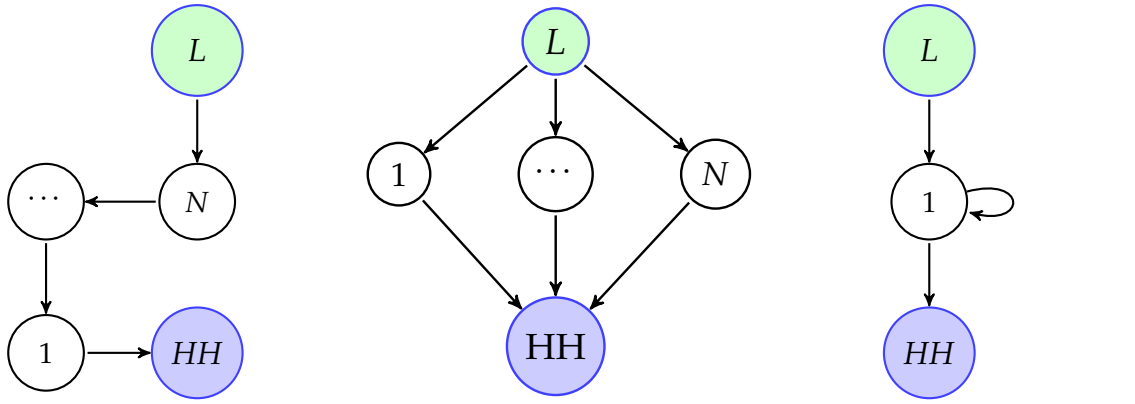
Given that $\Lambda = \tilde{\Lambda}$ in efficient equilibria, It may seem surprising that an improvement in allocative efficiency comes together with a movement of Λ away from $\tilde{\Lambda}$ so that $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) > 0$. However, the intuition is simple. This happens when $\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f < 0$ so that the weighted average of factor shares decreases. This means that the more monopolized or downwardly distorted parts of the economy are receiving more resources. This improves allocative efficiency, since from a social perspective, these monopolized or downwardly distorted parts of the economy receive too few resources to begin with.

Similarly, in the case of markup shocks, Theorem 1 implies that changes in allocative efficiency are given by a simple sufficient statistic: $-\tilde{\lambda}' \, \mathrm{d} \log \mu + \mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = -\tilde{\lambda}' \, \mathrm{d} \log \mu - \sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$. Now $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = -\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$ reflects both the direct effect $\tilde{\lambda}' \, \mathrm{d} \log \mu$ of the markup change on the profit share and the reallocation of workers towards or away from more distorted producers. To isolate the changes in allocative efficiency, which arise from the latter, we must net out the former.

2.4 Illustrative Examples

In this section, we introduce some bare-bones examples to illustrate the intuition of Theorem 1. In Section 3, we specialize Theorem 1 to the case of general nested constant-elasticity-of-substitution (CES) economies with arbitrary input-output linkages. The ex-

amples that we present here are simple special cases of these more general results.



(a) Vertical Economy (b) Horizontal Economy (c) Round-about Economy

Figure 1: The solid arrows represent the flow of goods. The flow of profits and wages from firms to households has been suppressed in the diagram. The sole factor for this economy is indexed by L.

Consider the three economies depicted in Figure 1. In all three economies, there is a single factor called labor. The only distortions in these examples are the markups charged by the producers.

The vertical economy in Figure 1a, drawn from Baqaee (2016), is a chain of producers. Producer N produces linearly using labor and downstream producers transform linearly the output of the producer immediately upstream from them. The household purchases the output of the most downstream producer. The horizontal economy in Figure 1b features downstream producers who produce linearly from labor. The household purchases the output of the downstream producers according to a CES aggregator with elasticity θ_0. The roundabout economy in Figure 1c features only one producer, who combines labor and its own products using a CES production function.

These different economies help illustrate the disappearance of two serendipities implied by the assumptions of Hulten's theorem: (1) the equality of revenue-based and cost-based Domar weights (used to weigh the direct or "pure" effects of technology); and (2) the absence of changes in allocative efficiency (reflecting the efficiency of the initial allocation). The vertical economy breaks (1) but not (2), the horizontal economy breaks (2) but not (1), and the round-about economy breaks (1) and (2).

Vertical Economy

First, consider the vertical economy in Figure 1a. In this economy, there is only one feasible allocation of resources, so the equilibrium allocation is efficient. An application

of Theorem 1 then confirms that

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \tilde{\lambda}_k = 1$$

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k + \tilde{\lambda}_k = 0.$$

Indeed, in this vertical economy, the verification of the theorem is trivial since $\tilde{\lambda}_k = 1$, $\tilde{\Lambda}_L = 1$, $\mathrm{d}\log \Lambda_L / \mathrm{d}\log A_k = 0$, and $\mathrm{d}\log \Lambda_L / \mathrm{d}\log \mu_k = -1$.

Hence, Hulten's theorem fails in the vertical economy, even though the equilibrium is efficient. Reassuringly, our decomposition detects no changes in allocative efficiency since $-\tilde{\lambda}' \,\mathrm{d}\log \mu - \tilde{\Lambda}_L \,\mathrm{d}\log \Lambda_L = 0$. The failure of Hulten's theorem is instead due to the gap between the revenue-based Domar weights $\lambda_k = \prod_{i=1}^{k-1} \mu_i^{-1}$ and the cost-based Domar weights $\tilde{\lambda}_k = 1$. Indeed, when markups are positive so that $\mu_i > 1$ for all i, we have $\tilde{\lambda}_k > \lambda_k$. This is a a consequence of downstream double-marginalization which divorces the revenues earned by a producer from that producer's share in the costs of faced by the household.

Horizontal Economy

Next we consider the horizontal economy represented in Figure 1b. The consumption of the household, or final demand, is given by

$$\frac{Y}{\overline{Y}} = \left(\sum_i \omega_{0i} \left(\frac{c_i}{\overline{c}_i} \right)^{\frac{\theta_0 - 1}{\theta_0}} \right)^{\frac{\theta_0}{\theta_0 - 1}},$$

where θ_0 is the elasticity of substitution in consumption, ω_{0i} are consumption weights, and variables with overlines in the denominator are normalizing constants measured in the same units as the numerator.

An application of Theorem 1 then yields

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \lambda_k - \lambda_k(\theta_0 - 1)\left(\frac{\mu_k^{-1}}{\sum_j \lambda_j \mu_j^{-1}} - 1 \right) \tag{4}$$

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = \lambda_k \theta_0 \left(\frac{\mu_k^{-1}}{\sum_i \lambda_i \mu_i^{-1}} - 1 \right). \tag{5}$$

In the horizontal economy, $\tilde{\lambda}_k = \lambda_k$ since there is no downstream double-marginalization. The direct or "pure" effects of a technology shock are still given by $\lambda' \,\mathrm{d}\log A$ exactly as in Hulten's theorem. However, technology shocks and markup shocks can now trigger nonzero changes in allocative efficiency $-\tilde{\lambda}' \,\mathrm{d}\log \mu - \tilde{\Lambda}_L \,\mathrm{d}\log \Lambda_L$.

Consider equation (4): the effects of a positive technology shock $\mathrm{d}\log A_k$ to producer k. Holding fixed the share of labor used by each producer, the productivity shock increases the output of producer k. However, the shock also reduces its price, which in turn increases the demand for its output via a substitution effect. Whether workers are reallocated towards or away from producer k depends on whether the increase in demand from the substitution effect is stronger than the increase in supply from the productivity shock. This in turn hinges on the whether θ_0 is greater than or less than 1, i.e. on the direction of the departure from Cobb Douglas. When $\theta_0 > 1$, workers are reallocated towards producer k. When $\theta_0 < 1$, workers are reallocated away from producer k. And when $\theta_0 = 1$, the allocation of workers is unchanged. Whether these reallocation of workers increase or decrease allocative efficiency and output in turn depends of the comparison of the markup of producer k to the (harmonic) average markup $(\sum_i \lambda_i \mu_i^{-1})^{-1}$.

When $\theta_0 > 1$, workers are reallocated towards producer k. If its markup is larger than the (harmonic) average markup $\mu_k > (\sum_i \lambda_i \mu_i^{-1})^{-1}$, then this producer is too small from a social perspective to begin with.[15] The reallocation of labor towards producer k therefore improves allocative efficiency and increases output.[16] The opposite occurs when the markup of producer k is smaller than the average markup. This effect works in the opposite direction when $\theta_0 < 1$, since in that case, the shock would reallocate workers away from producer k. Of course, in the Cobb-Douglas case when $\theta_0 = 1$, the allocation of labor remains unchanged, and hence there are no changes in allocative efficiency.[17]

All of this information is summarized by a simple sufficient statistic: the change in allocative efficiency is exactly the opposite of the change in the labor share $-\tilde{\Lambda}_L \,\mathrm{d}\log \Lambda_L =$

[15] Note that the average markup is simply the inverse of the labor share so that $(\sum_i \lambda_i \mu_i^{-1})^{-1} = 1/\Lambda_L$.

[16] When $\theta_0 > 1$ and producer k is significantly more competitive than the average producer $\mu_k < (\theta_0/(\theta_0 - 1) \sum_i \lambda_i \mu_i^{-1})^{-1}$, then the reduction in allocative efficiency can be so extreme that a positive productivity shock can actually reduce output.

[17] This last property is a more general property of Cobb-Douglas economies which we shall encounter in Section 3: productivity shocks do not lead to any change in allocative efficiency for Cobb-Douglas economies since their allocation matrix does not depend on the level of productivity.

$-\,\mathrm{d}\log\Lambda_L$ (since $\tilde{\Lambda}_L = 1$). The labor share of income decreases (increases), and allocative efficiency improves (worsens), when workers are reallocated to producers that were too small (large) from a social perspective to begin with because they were charging above-average (below-average) markups.

With productivity shocks, the benchmark with no changes in allocative efficiency is Cobb Douglas $\theta_0 = 1$. With markup shocks, the benchmark case with no changes in allocative efficiency is Leontief $\theta_0 = 0$ instead. For a markup shock $\mathrm{d}\log\mu_k$ to producer k, when $\theta_0 = 0$, complementarities are extreme and the household chooses to consume a fixed quantity of each good regardless of its price. As a result, the allocation of labor does not change in response to the shock, and there are therefore no associated changes in allocative efficiency. When $\theta_0 > 0$ instead, the price of producer k increases, the demand for its output decreases, and workers are reallocated away from it. Allocative efficiency and output decrease (increase) if its markup is larger (smaller) than the average markup.

All of this information is again summarized by a simple sufficient statistic $-\tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}_L\,\mathrm{d}\log\Lambda_L = -\lambda'\,\mathrm{d}\log\mu - \mathrm{d}\log\Lambda_L$. Now the opposite of the change in the labor share $-\,\mathrm{d}\log\Lambda_L$ reflects both the direct effect $\lambda\,\mathrm{d}\log\mu$ of the markup change on the profit share and the reallocation of workers towards or away from more distorted producers. To isolate the changes in allocative efficiency, which arise from the latter, we must net out the former.

Round-about Economy

Finally, we consider the round-about economy in Figure 1c. There is a single producer producing using labor and its own goods according to

$$\frac{y_1}{\overline{y}_1} = A_1\left(\omega_{11}\left(\frac{x_{11}}{\overline{x}_{11}}\right)^{\frac{\theta_0-1}{\theta_0}} + \omega_{1L}\left(\frac{L_1}{\overline{L}_1}\right)^{\frac{\theta_0-1}{\theta_0}}\right)^{\frac{\theta_0}{\theta_0-1}}.$$

An application of Theorem 1 then yields

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_1} = \tilde{\lambda}_1 - \tilde{\Lambda}_L\frac{\mathrm{d}\log\Lambda_L}{\mathrm{d}\log A_1} = \tilde{\lambda}_1 - (\theta_0 - 1)\lambda_1(\tilde{\lambda}_1 - 1)(\mu^{-1} - 1)$$

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_1} = -\tilde{\lambda}_1 - \tilde{\Lambda}_L\frac{\mathrm{d}\log\Lambda_L}{\mathrm{d}\log \mu_1} = \theta_0\lambda_1(\tilde{\lambda}_1 - 1)(\mu^{-1} - 1).$$

The round-about economy combines features of the vertical economy and of the horizontal economy. As in the vertical economy, revenue-based and cost-based Domar weights differ since $\tilde{\lambda}_1 = \mu_1/[\mu_1 - (1 - \lambda_1^{-1})] \neq \lambda_1$ as long as $\mu_1 \neq 1$. As in the horizontal economy, we have non-trivial changes in allocative efficiency in general so that $-\tilde{\lambda}' \,\mathrm{d}\log\mu - \tilde{\Lambda}_L \,\mathrm{d}\log\Lambda \neq 0$. The intuitions for these results combine those of the vertical economy and of the horizontal economy.

2.5 Growth Accounting

In this section, we discuss how to decompose time-series changes in aggregate TFP and the Solow residual into "pure" technology changes and allocation efficiency changes. For the purpose of this section, we introduce a small but simple modification to allow for changes in factor supplies. We denote the supply of factor f by L_f and by L the vector of factor supplies. The impact of a shock to the supply of a factor is given by $\mathrm{d}\log Y/\,\mathrm{d}\log L_f = \tilde{\Lambda}_f + \mathrm{d}\,H(\tilde{\Lambda},\Lambda)/\,\mathrm{d}\log L_f = \tilde{\Lambda}_f - \sum_g \tilde{\Lambda}_g \,\mathrm{d}\log\Lambda_g/\,\mathrm{d}\log L_f$.

Proposition 1 (TFP Decomposition)**.** *To the first order, we can decompose aggregate TFP as*

$$\underbrace{\Delta\log Y_t - \tilde{\Lambda}_{t-1}'\Delta\log L_t}_{\Delta\ \textit{Aggregate TFP}} \approx \underbrace{\tilde{\lambda}_{t-1}'\Delta\log A_t}_{\Delta\ \textit{Technology}} - \underbrace{\tilde{\lambda}_{t-1}'\Delta\log\mu_t - \tilde{\Lambda}_{t-1}'\Delta\log\Lambda_t}_{\Delta\ \textit{Allocative Efficiency}}. \tag{6}$$

The left-hand side of this expression, which we define to be aggregate TFP growth, differs from the Solow residual since it weighs the change in $L_{f,t}$ by the cost-based Domar weight $\tilde{\Lambda}_{f,t}$ rather than the revenue-based Domar weight $\Lambda_{f,t}$. The traditional Solow residual attributes all non-labor income to capital (and has no room for profit income). Therefore, with only labor (L) and capital (K) as factors, the traditional Solow residual would be

$$\Delta\log Y_t - \hat{\Lambda}_{t-1}'\Delta\log L_t \approx \tilde{\lambda}_{t-1}'\Delta\log A_t - \tilde{\lambda}_{t-1}'\Delta\log\mu_t - \tilde{\Lambda}_{t-1}'\Delta\log\Lambda_t + (\tilde{\Lambda}_{t-1} - \hat{\Lambda}_{t-1})'\Delta\log L_t, \tag{7}$$

where $\hat{\Lambda}_{L,t-1} = \Lambda_{L,t-1}$ for labor and $\hat{\Lambda}_{K,t-1} = 1 - \Lambda_{L,t-1}$ for capital. The key difference is that capital is weighed according to $1 - \Lambda_{L,t-1}$ and not to $\Lambda_{K,t-1}$.

Using $\tilde{\Lambda}$ to weigh factors in (6) is consistent with Hall (1990), who showed that for an aggregate production function, aggregate TFP should weigh changes in factor inputs by their share of total cost rather than their share of total revenue. In our context unlike in Hall's, the equilibrium can be distorted given factor supplies and there is no structural aggregate production function. We must weigh factors by their cost-based Domar weights. Proposition 1 therefore unifies the approach of Hulten (1978), who eschews aggregate

production functions, but maintains efficiency, with that of Hall (1990) who does not require efficiency but maintains aggregate production functions and therefore ignores the allocative efficiency issues that most concern us.

Turning to the right-hand side, in the case of an efficient economy, the envelope theorem implies that the reallocation terms are welfare-neutral (to a first order) and can be ignored. Furthermore, the appropriate weights on the technology shocks $\tilde{\lambda}_t$ coincide with the observable sales shares. In the presence of distortions, these serendipities disappear. However, given the input-output expenditure shares across producers, the level of wedges and their changes, and the changes in factor income shares, we can compute the right-hand side of equation (6) without having to make any parametric assumptions. This is an ex-post decomposition in the sense that it requires us to observe factor income shares and factor supplies at the beginning and at the end of the period.

It is important to note that Proposition 1 can be used in contexts where productivity or wedges are endogenous to some more primitive fundamental shocks, if these endogenous changes are actually observed in the data.

2.6 Comparison with Basu-Fernald and Petrin-Levinsohn

In their seminal work, Basu and Fernald (2002) provide an alternative decomposition of aggregate TFP changes into "pure" technology changes and changes in allocative efficiency for economies with markups. Their "pure" technology term, like ours, is a weighted average of technology changes $\Delta \log A_{kt}$ for each producer. The weight $\lambda_{kt}(1 - s_{Mkt})/(1 - \mu_{kt}s_{Mkt})$ attached to a given producer k is just its share in value added $\lambda_{kt}(1 - s_{Mkt})$ multiplied by a correction $1/(1 - \mu_{kt}s_{Mkt})$ involving its intermediate input share in revenues s_{Mkt} and its markup μ_{kt}. These weights therefore differ from the cost-based Domar weights $\tilde{\lambda}_{kt}$ prescribed by our decomposition. In fact, the information required to calculate their weights — the value added share, intermediate-input share, and markup of each producer — is not enough in general to calculate the cost-based Domar weights. Computing cost-based Domar weights requires the whole input-output matrix, and not just the intermediate input shares. As a result, their decomposition is different from ours.

Similar comments apply to Petrin and Levinsohn (2012), who build on the Basu and Fernald (2002) approach with extensions to cover non-neoclassical features of production like entry and fixed or sunk costs. The "pure" technology changes of Petrin and Levinsohn weigh the underlying technology changes $\Delta \log A_{kt}$ of all producers by the usual revenue-based Domar weights λ_{kt}, and define changes in allocative efficiency in terms of the gap between the Solow residual and the Domar-weighted growth in productivity. Although

we abstract away from fixed costs and entry in our benchmark model, we discuss how the results can be extended to incorporate these features of the data in Appendix C.

We demonstrate the difference between our approach and the Basu-Fernald and Petrin-Levinsohn decompositions by way of a revealing example. Consider an economy where the production network Ω is an acylic graph, as illustrated in Figure 2. The term acyclic here means that any two goods are connected to one another by exactly at most one undirected path, so that each factor and each good has a unique consumer. Such economies have a unique resource feasible allocation, simply because there is no option to allocate a given factor or good to different uses. This allocation is necessarily efficient. Markups and wedges have no effect on the allocation of resources, and as a result, there is no misallocation. In other words, misallocation requires cycles (undirected paths that connect a node back to itself) in the production network.

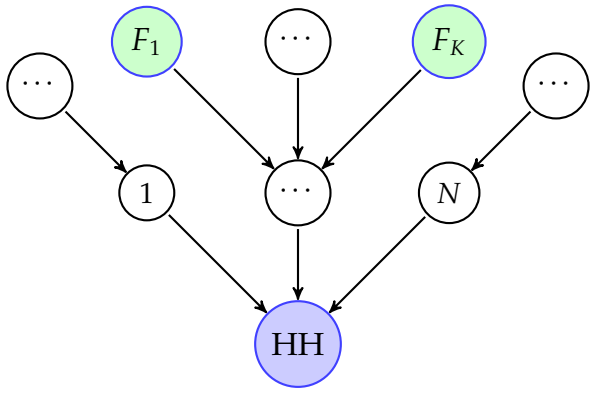


Figure 2: An acyclic economy, where the solid arrows represent the flow of goods. The factors are the green nodes. Each supplier (including factors) have at most one customer, whereas a single customer may have more than one supplier. Economies without cycles can be represented as directed trees with the household being the root.

Corollary 2 (Acyclic Economies)**.** *If the production network of the economy is an acyclic graph, then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k, \quad \frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = 0. \tag{8}$$

This follows immediately from the fact that by construction, acyclic economies hold fixed the share of resources across producers. An important consequence of this proposition is that for acyclic economies, where it is unambiguous a priori that there is no misallocation, our definition of changes in allocative efficiency indeed identifies that there are no changes in allocative efficiency in response to shocks. On the other hand, the Basu-Fernald and Petrin-Levinsohn decompositions do detect changes in allocative efficiency despite the fact that these economies are efficient. Note that although the equilibrium

allocation in this economy is efficient, Hulten's theorem still fails because the observed sales shares λ do not coincide with $\tilde{\lambda}$.[18,19]

2.7 Measuring Allocative Efficiency

Theorem 1 defines *changes* in allocative efficiency: the change in output owing to the reallocation of resources, relative to a benchmark that holds the initial distribution of resources fixed. This measure can be quantified without knowledge about the production functions over and above expenditure data. Theorem 1 does not offer a definition, nor a characterization, of the *level* of allocative efficiency.

To measure the level of allocative efficiency, along the lines of Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and others, one can compute how much output would increase if all wedges were eliminated. One can then follow Hsieh and Klenow (2009) by computing a different notion of change in allocative efficiency by tracking how the gains from removing all wedges change over time. Their measure gives the change over time in the distance from the unobserved efficient frontier of the economy.

Our notion of changes in allocative efficiency is an entirely different concept: it measures the contribution of reallocation to TFP along the equilibrium path, rather than the change in the distance of the allocation from the efficient production frontier. The latter concept requires strong parametric assumptions as one needs to specify the unobserved global efficient production frontier. Our concept, by contrast, is local, and can be measured without strong parametric assumptions.

In a given time-series, the two measures may easily move in different directions. To see how this could happen, consider a horizontal economy with two producers and $\theta_0 > 1$. Suppose $A_1/\mu_1 = A_2/\mu_2$, with $A_1 > A_2$. Then, at steady-state, the two producers split demand and workers evenly. Now, if the first producer receives a positive productivity shock, workers are reallocated from producer 2 to producer 1, allocative efficiency improves. However, the frontier can move by even more, and thereby increase the distance to the frontier.

These conceptual differences notwithstanding, our results can actually be used to compute the level of allocative efficiency understood as the gains from removing distortions, and hence also by implication its change over time. Call $Y(A, \mu)$ the level of output given

[18] See Appendix F for a fully worked-out example.

[19] Relative to Basu-Fernald and Petrin-Levinsohn, our approach also economizes on information by recognizing that the system of first-order conditions arising from cost-minimization by every producer gives rise to a system that can be solved. This is what allows us to summarize all the information into changes in the wedges, and changes in the primary factors, whereas the other decompositions require information on tracking how every input is used by every producer.

the productivity vector A and the markup vector μ. Consider a transformation of each markup $\hat{\mu}_i(t) = t\mu_i + (1-t)$. When $t = 1$, this transformation leaves markups as they are. On the other hand, $t = 0$ eliminates all distortions in the economy. The distance from the frontier is[20]

$$\begin{aligned}\log\left(\frac{Y(A,1)}{Y(A,\mu)}\right) = \log\left(\frac{Y(A,\hat{\mu}(0))}{Y(A,\hat{\mu}(1))}\right) &= -\int_0^1 \frac{\mathrm{d}\log Y(A,\hat{\mu}(t))}{\mathrm{d}\log\mu}\frac{\mathrm{d}\log\hat{\mu}(t)}{\mathrm{d}\,t}\,\mathrm{d}\,t \\ &= \frac{1}{2}\sum_i \frac{\mathrm{d}\log Y(A,\mu)}{\mathrm{d}\log\mu_i}\left(\frac{1-\mu_i}{\mu_i}\right) + O(\|\mu-1\|^3). \qquad (9)\end{aligned}$$

The first line links our comparative statics to the distance from the frontier by way of an integral. The second line shows that our comparative statics $\mathrm{d}\log Y(A,1)/\mathrm{d}\log\mu$ at the initial equilibrium with $t = 1$ can be used to approximate the distance from the frontier to a second order. When the output function $Y(A,\mu)$ is translog in μ, then this approximation holds exactly. The proof follows from an application of the trapezoidal rule (see for example Theil (1967)) and uses the fact that $\mathrm{d}\log Y(A,\hat{\mu}(0))/\mathrm{d}\log\mu = 0$ by an application of Corrolary 1 since the final equilibrium with $t = 0$ is efficient.

We can also combine this formula with the ex-ante structural results that we derive in Section 3 for the case of general nested constant-elasticity-of-substitution (CES) economies with arbitrary input-output linkages to compute $\mathrm{d}\log Y(A,\mu)/\mathrm{d}\log\mu_i$ as a function of the underlying structural microeconomic parameters of the model.

As an example, we can combine equation (5) with equation (9) to deduce, to a second-order, the gains from eliminating markups in the horizontal economy

$$\log\left(\frac{Y(A,1)}{Y(A,\mu)}\right) \approx \frac{1}{2}\theta_0\frac{Var_\lambda(\mu^{-1})}{E_\lambda(\mu^{-1})}.$$

The variance of a random variable divided by its mean is called the index of dispersion. Hence, the gain to output from shrinking all markups depends positively on the elasticity of substitution θ_0 and the expenditure-share weighted index of dispersion in $1/\mu_i$. Using the dispersion of wedges to measure misallocation is common in the literature following Hsieh and Klenow (2009). Our comparative statics, in these simple cases, delivers such results. Theorem 1 can therefore be leveraged to flexibly generalize such notions.[21] .

[20] In general, unless there is a representative consumer, the *level* of allocative efficiency, and of real GDP, is ambiguous to define. In practice, this means that the level of allocative efficiency, as defined in, say, Hsieh and Klenow (2009), is ill-defined. However, Theorem 1 holds even if there is no representative consumer as long as real output is defined using the Divisia GDP deflator. Hence, it can be used to define changes in allocative efficiency even in economies with heterogenous consumers.

[21] Although stated in terms of markups, Theorem 1 also characterizes the response of output to other

3 Parametric Model and Structural Results

Our results so far are non-parametric, but we can draw out some additional intuition by specializing them to the case of an arbitrary nested CES economy, with an arbitrary number of nests, weights, and elasticities. Working through this parametric class of models greatly helps build intuition and allows us to calibrate a structural model for quantifying the mechanisms that we identify.

We proceed in stages. After setting up the parametric model, we start with the one-factor case, which a fortiori, implies constant returns to scale, since by convention we model decreasing returns using fixed factors. The one-factor case is illustrative for showing how the production network, interacting with elasticities of substitution, can affect the degree of misallocation in the economy. Next, we extend our results to the case with multiple factors/decreasing returns.

3.1 Parametric Model Setup

Any CES economy with a representative consumer, an arbitrary numbers of nests, elasticities, and intermediate input use, can be re-written in what we call *standard form*, which is more convenient to study. Throughout this section, variables with over-lines are normalizing constants equal to the values in steady-state. Since we are interested in log changes, the normalizing constants are irrelevant.[22]

General Nested CES Economies in Standard Form

A CES economy in standard form is defined by a tuple $(\Omega, \boldsymbol{\theta}, \mu, F)$ where Ω is an $(N+F+1)\times(N+F+1)$ input-output matrix, the vector $\boldsymbol{\theta}$ is an $(N+1)\times 1$ vector of microeconomic elasticities of substitution, μ is the vector of markups/wedges, and F is the number of factors.

distortion shocks τ, as already explained above. For example, for a factor wedge shock to factor k for producer l, we have

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log(1+\tau^f_{lk})} = -\tilde{\lambda}_l\alpha_{lk} + \frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log(1+\tau^f_{lk})} = -\tilde{\lambda}_l\alpha_{lk} - \sum_g \tilde{\Lambda}_g \frac{\mathrm{d}\log\Lambda_g}{\mathrm{d}\log(1+\tau^f_{lk})}.$$

Similar formulas hold for shocks to other wedges. In Appendix E, we describe the relabelling in more detail. For reference, for an intermediate input wedge shock, we have $\mathrm{d}\log Y/\mathrm{d}\log(1+\tau_{lk}) = -\tilde{\lambda}_l\tilde{\omega}_{lk} + \mathrm{d}\,H(\tilde{\Lambda},\Lambda)/\mathrm{d}\log(1+\tau_{lk})$ and for a consumption wedge shock, we have $\mathrm{d}\log Y/\mathrm{d}\log(1+\tau^c_k) = -b_k + \mathrm{d}\,H(\tilde{\Lambda},\Lambda)/\mathrm{d}\log(1+\tau^c_k)$. For each formula, the first term corresponds to the impact on the GDP deflator holding fixed factor prices, and the second term measures the impact of the changing factor prices.

[22] We use normalized quantities since it simplifies calibration, and clarifies the fact that CES aggregators are not unit-less.

The F factors are modeled as non-reproducible goods, which we denote with uppercase letters, and the production function of these goods are endowments

$$\frac{Y_f}{\overline{Y}_f} = 1. \tag{10}$$

For each factor f, we set $\Omega_{fj} = 0$ for all j. The other $N + 1$ other goods are reproducible, and the production of a reproducible good k can be written as

$$\frac{y_k}{\overline{y}_k} = A_k \left(\sum_{l=1}^{N} \omega_{kl} \left(\frac{x_{lk}}{\overline{x}_{lk}} \right)^{\frac{\theta_k - 1}{\theta_k}} \right)^{\frac{\theta_k}{\theta_k - 1}},$$

where x_{lk} are intermediate inputs from l used by k. Each producer charges a markup over its marginal cost μ_k. Producer 0 represents the final-demand function of the household

$$\frac{Y}{\overline{Y}} = \frac{c_0}{\overline{c}_0}, \tag{11}$$

where c_0 is the final good. We set $\omega_{i0} = 0$ for all i. Through a relabelling, this structure can represent any CES economy with an arbitrary pattern of nests and wedges and elasticities. Intuitively, by relabelling each CES aggregator to be a new producer, we can have as many nests as we like.

The Input-Output Covariance Operator

In order to state our results, we introduce the following input-output covariance operator:

$$Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}, \Psi_{(f)}) = \sum_i \tilde{\Omega}_{ji} \tilde{\Psi}_{ik} \Psi_{if} - \left(\sum_i \tilde{\Omega}_{ji} \tilde{\Psi}_{ik} \right) \left(\sum_i \tilde{\Omega}_{ji} \Psi_{if} \right),$$

where $\tilde{\Omega}^{(j)}$ corresponds to the jth row of $\tilde{\Omega}$, $\tilde{\Psi}_{(k)}$ to kth column of $\tilde{\Psi}$, and $\Psi_{(f)}$ to the fth column of Ψ. In words, this is the covariance between the kth column of $\tilde{\Psi}$ and the fth column of Ψ using the jth row of $\tilde{\Omega}$ as the distribution.[23] Since the rows of $\tilde{\Omega}$ always sum to one for a reproducible (non-factor) good j, we can formally think of this as a covariance, and for a non-reproducible good, the operator just returns 0.

[23] This covariance operator is similar to a variance operator defined by Acemoglu et al. (2016) in the context of diffusion on graphs.

3.2 Single Factor

We begin by investigating the impact of productivity and markup/wedge shocks on output for the model with a single factor of production $F = 1$, which we index by L. We start with productivity shocks, since the intuition gained from these will be useful in understanding the impact of markup/wedges shocks as well.

3.2.1 Productivity Shocks

Proposition 2 (Productivity Shocks with One Factor). *Suppose there is only one factor, denoted by L. Then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k + \frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k},$$

where

$$\frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}\right), \tag{12}$$

and $\Psi_{(L)}$ is the column of the Leontief inverse Ψ corresponding to L.

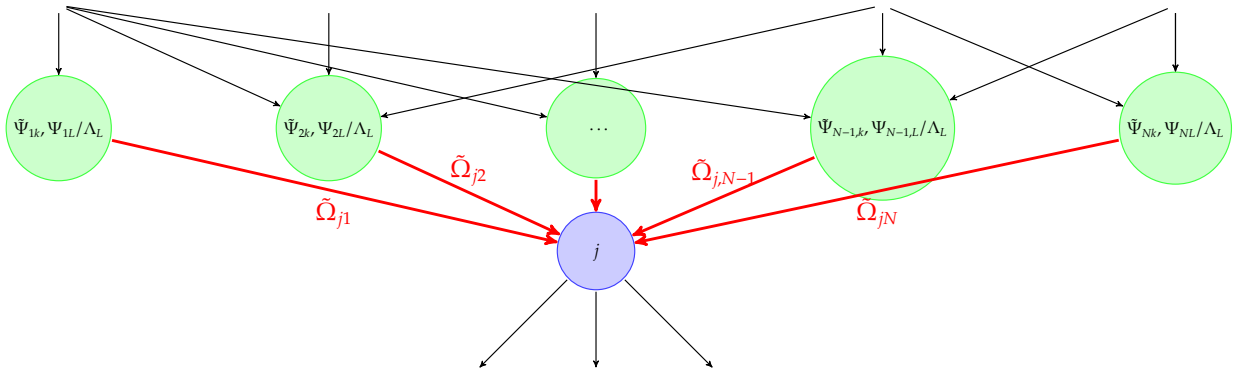


Figure 3: Illustration of the input-output covariance operator defined in equation (12).

Equation (12) characterizes the change in allocative efficiency as a function of the properties of the network and the elasticities of substitution. To get some intuition, consider Figure 3 which gives a graphical representation of the effects associated with a given term producer j in the sum on the right-hand side of equation 12 and its associated input-output covariance operator. The numbers Ψ_{iL} are the payments to labor as a share of the total revenue of producing a given good i used by producer j as an intermediate intput — what Baqaee (2015) calls the *network-adjusted labor share* of i — taking into account the entire supply chain of i. In an efficient economy with one factor, we have $\Psi_{iL} = \tilde{\Psi}_{iL} = 1$

and $\Lambda_L = \tilde{\Lambda}_L = 1$. By contrast, in an inefficient economy we still have $\tilde{\Psi}_{iL} = 1$, and $\tilde{\Lambda}_L = 1$ but we no longer necessarily have $\Psi_{iL} = 1$ or $\Lambda_L = 1$. For example, if all markups are positive, we have $\Psi_{iL} < 1$ and $\Lambda_L < 1$. A low value of Ψ_{iL} indicates that on average, markups are high along the supply chain of i, and a low value of Λ_L indicates that on average, markups are high in the economy along as a whole. The lower Ψ_{iL}/Λ_L, the more distorted is the supply chain of i relative to the economy as a whole. In other words, we can think of the Lth column of the Leontief inverse $\Psi_{(L)}$ as measuring the inverse degree of double-marginalization along the supply chain of each producer. The economy's labor content as a whole is given by labor's share of income Λ_L. Hence, producers with lower values of Ψ_{iL}/Λ_L have relatively too few workers in their supply chain.

With this interpretation, Proposition 2 becomes very intuitive. In response to a positive productivity shock to producer k, the relative prices of all producers i change according to their exposure to k, measured by $\tilde{\Psi}_{ik}$. If $\theta_j > 1$, the jth producer substitutes its cost share across its inputs towards the producers with higher exposure $\tilde{\Psi}_{ik}$ to k, since their relative prices decline by more. If those producers also happen to have lower Ψ_{iL}/Λ_L, then these producers are inefficiently too small in the initial pre-shock equilibrium. In this case, there is negative covariance between $\tilde{\Psi}_{(k)}$ and $\Psi_{(L)}$. This means that substitution, due to the productivity shock, lowers overall misallocation, sending more workers to produce goods which are inefficiently receiving too few workers. In this case, the changing allocation of workers boosts the impact of the productivity shock on output. Of course, j is not the only producer whose expenditure shares change and the same logic applies to all producers, so we sum over all j. If the elasticities are less than one, or the covariance is negative, the reallocation forces work against the positive impact of the technology shock.[24]

Note that the examples we started the paper with in Section 2.4 (the vertical, horizontal, and round-about economies), are simple applications of Proposition 2. Another example, which is immediate, is a Cobb-Douglas economy where $\theta_i = 1$ for all i. In that economy, $\mathrm{d}\log Y/\mathrm{d}\log A_k = \tilde{\lambda}_k$. Intuitively, for a Cobb-Douglas economy, the share of resources (in physical units) across users is invariant to productivity shocks, and hence the degree of misallocation in the economy is invariant to the value of A.

[24]Baqaee and Farhi (2017) show that for an economy like the one in Proposition 2, if the economy is efficient, then the output response to a shock to producer k depends *only* on k's role as a supplier. Proposition 2 shows that this fails if the equilibrium is inefficient. In particular, $\Psi_{(L)}$ — which captures information about how distorted the supply chain of each producer is (i.e. it depends on the producer's role as a consumer of inputs), also matters, since it affects the response of misallocation.

3.2.2 Markup/Wedge Shocks

Next, we consider shocks to wedges rather than productivity. The intuition we gained from the productivity shocks will prove useful here.

Proposition 3 (Makup/Wedge Shocks with One Factor)**.** *Suppose that there is only one factor, denoted by L. Then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k},$$

where

$$\frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = -\sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}\right) - \lambda_k \frac{\Psi_{kL}}{\Lambda_L}. \tag{13}$$

Proposition 3 implies that the effects of a positive markup shock are analogous to the effects of a negative productivity shock. In response to a change in markup/wedge, there is a direct effect $-\tilde{\lambda}_k$ on consumer prices from the higher markup/wedge, and an indirect effect from the changing factor income shares captured by $-\mathrm{d}\log \Lambda_L/\mathrm{d}\log \mu_k$.

Intuitively, the labor income share moves for two reasons. The first reason is exactly the same as it would be for a negative productivity shock: every producer j substitutes across its input branches in response to the change in the markup/wedge of k, and if this substitution pattern covaries positively with the measure of supply chain distortions $\Psi_{(L)}/\Lambda_L$, then this improves allocative efficiency. There is however a second set of upstream adjustments: compared to a negative productivity shock to k, an increase in the markup/wedge of k leads k to release some resources to the rest of the economy. These released resources can ultimately be expressed as released labor. The amount of labor released in proportion to total labor $\lambda_k \Psi_{kL}/\Lambda_L$ per unit of shock is given by k's sales share λ_k times the labor content of its revenue Ψ_{kL} divided by the economy's labor income share Λ_L.

Once again, the examples in Section 2.4 can be seen as simple applications of Proposition 3. Another useful example to consider is the Cobb-Douglas economy, which helps to isolate the importance of the new term in Proposition 3. For a Cobb-Douglas economy, the only source factor reallocation comes from the fact that the producer which increases its markup/wedge releases some labor. Let $\theta_j = 1$ for every j, which is the Cobb-Douglas special case. Now, applying Proposition 3, we get

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k + \lambda_k \frac{\Psi_{kL}}{\Lambda_L} = -\tilde{\lambda}_k \left(1 - \frac{\lambda_k}{\tilde{\lambda}_k}\frac{\Psi_{kL}}{\Lambda_L}\right).$$

As before, Ψ_{kL}/Λ_L is a measure of how distorted the supply chain of k is relative to the economy as a whole. If $\Psi_{kL}/\Lambda_L < 1$, then this means that for each dollar k earns, a smaller share reaches workers than it would if that dollar was spent by the household. In other words, producer k's supply chain has inefficiently too few workers. On the other hand, $\lambda_k/\tilde{\lambda}_k$ is a measure of how distorted the demand of chain of k is. If $\lambda_k/\tilde{\lambda}_k < 1$, this implies that k is facing double-marginalization. When the product of the downstream and upstream terms is less than one, this means that producer k is inefficiently starved of demand and workers. Hence, an increase in the markup/wedge of k reduces the allocative efficiency of the economy. On the other hand, when the product of these two terms is greater than one, the path connecting the household to labor via producer k is too large. Therefore, an increase in the markup/wedge of k reallocates resources to the rest of the economy where they are more needed and increases allocative efficiency.

3.3 Multiple Factors

So far, we have restricted ourselves to the case of a single factor of production/constant returns to scale. In this subsection, we extend our results to cover the case with multiple factors of production/decreasing returns to scale. We index all factors by f, but we also sometimes use L as a generic index for a factor.

3.3.1 Productivity Shocks

Proposition 4 (Productivity Shocks with Multiple Factors)**.** *In response to a productivity shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)} - \sum_g \tilde{\Psi}_{(g)}\frac{\mathrm{d}\log\Lambda_g}{\mathrm{d}\log A_k}, \frac{\Psi_{(f)}}{\Lambda_f}\right). \tag{14}$$

Given $\mathrm{d}\log\Lambda_f/\mathrm{d}\log A_k$*, we know, from Theorem 1 that*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k + \frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k}.$$

Proposition 2 can be seen immediately as a special case of Proposition 4. When there is only one factor, $\sum_f \tilde{\Psi}_{(f)}\,\mathrm{d}\log\Lambda_f/\mathrm{d}\log A_k$ is a constant vector and drops out of the covariance term, which allows us to recover the result for the case with a single factor.

As a bonus, Proposition 4 also determines how factor income shares for different factors move in response to productivity shocks in a distorted economy. This is a question of

independent interest for analyses of inequality and growth. We can rewrite equation (14) as the following linear system

$$\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} = \Gamma\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} + \delta^{(k)}, \tag{15}$$

with

$$\Gamma_{fg} = -\sum_j (\theta_j - 1)\lambda_j\mu_j^{-1} Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(g)}, \frac{\Psi_{(f)}}{\Lambda_f}\right),$$

and

$$\delta_f^{(k)} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(f)}}{\Lambda_f}\right).$$

When there is only factor, Γ is a zero 1×1 matrix, and we are left with only $\delta^{(k)}$, which is just equation (13). To see intuition for equation (15), imagine a negative shock $\mathrm{d}\log A_k < 0$ to producer k. For fixed factor prices, every producer i will substitute across its inputs in response to this shock. Suppose that $\theta_i < 1$, so that producer i substitutes its cost shares *towards* those inputs that are more reliant on producer k, captured by $\tilde{\Psi}_{ik}$. Now, if those inputs are also more reliant on factor f, captured by a high $Cov_{\Omega^{(i)}}\left(\tilde{\Psi}_{(k)}, \Psi_{(f)/\Lambda_f}\right)$, then substitution by i will increase demand for factor f and hence the income share of factor f and in turn the price of this factor.

If the economy has only a single factor denoted by L, then we simply need to consider the sum of $(\theta_i - 1)Cov_{\Omega^{(i)}}\left(\tilde{\Psi}_{(k)}, \Psi_{(L)/\Lambda_L}\right)$ weighted by the size λ_i and markups μ_i^{-1} of i for all i, and this is precisely what $\delta^{(k)}$ does.

However, when there are multiple factors, the change in demand for factors will affect relative factor prices, and the change in relative factor prices will set off additional rounds of substitution in the economy that we must account for, and this is the role Γ plays. For a given set of factor prices, the shock to k affects the demand for each factor, hence factor income shares and in turn factor prices, as measured by the $F\times1$ vector $\delta^{(k)}$. These changes in factor prices then cause further substitution through the network, leading to additional changes in factor demands and prices. The impact of the change in the relative price of factor g on the demand for factor f is measured by the fgth element of the $F \times F$ matrix Γ. The movements in factor shares are the fixed point of this process, i.e. the solution of equation (15). Crucially, the matrix Γ does not depend on which producer k has been shocked, since it encodes how changes in factor prices affect factor income shares.[25]

[25]Proposition 4 is tightly connected with the results in Baqaee and Farhi (2017), which characterize the change in sales shares (for goods and factors) in efficient economies. However, whereas for an efficient economy, changes in factor income shares determine the second-order impact of shocks on output, for a

To illustrate this intuition, consider the example depicted in Figure 4.

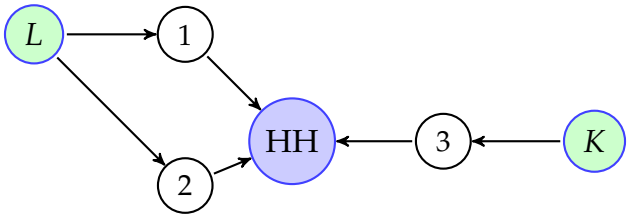


Figure 4: An economy with two factors of production L and K. The subgraph from L to the household contains a cycle, and hence can be subject to misallocation. On the other hand, there is only a unique path connecting K to the household, so there is no misallocation.

Example 3.1 (Figure 4)**.** We have

$$\Gamma = -(\theta_0 - 1)\begin{pmatrix} Cov_b(\tilde{\Psi}_{(L)}, \Psi_{(L)}) & Cov_b(\tilde{\Psi}_{(K)}, \Psi_{(L)}) \\ Cov_b(\tilde{\Psi}_{(L)}, \Psi_{(K)}) & Cov_b(\tilde{\Psi}_{(K)}, \Psi_{(K)}) \end{pmatrix}, \tag{16}$$

and

$$\delta^{(i)} = (\theta_0 - 1)\begin{pmatrix} Cov_b(\tilde{\Psi}_{(i)}, \Psi_{(L)}) \\ Cov_b(\tilde{\Psi}_{(i)}, \Psi_{(K)}) \end{pmatrix}.$$

Substituting in the values and solving the system of equations (15), using Proposition 4, and noting that $\lambda_i = \tilde{\lambda}_i$ for all i, we find that

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_i} = \lambda_i + \lambda_i(\theta_0 - 1)\left(1 - \frac{\mu_i^{-1}}{\frac{\lambda_1}{\lambda_1+\lambda_2}\mu_1^{-1} + \frac{\lambda_2}{\lambda_1+\lambda_2}\mu_2^{-1}}\right), \quad (i = 1, 2)$$

but

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_i} = \lambda_i, \quad (i = 3).$$

The details for this example are in Appendix B.[26] A lesson is that changes in allocative efficiency are only present for shocks to producers 1 and 2 which share a factor of production, but not for producer 3 which has its own factor of production. Moreover, the changes in allocative efficiency for shocks to producers 1 and 2 only depends on the markups of these two producers and not on the markup of producer 3.

We end this section but showing how our results can be extended to cover the impact of shocks to markups/wedges, in a manner similar to Proposition 3. The intuition for this result is the same as it was in the single factor case: a markup shock has the same effect as

distorted economy, this information is required even for the first-order impact of shocks.

[26]For this example $b_i = \lambda_i = \tilde{\lambda}_i$ for $i = 1, 2, 3$.

a negative productivity shock, with the additional fact that we must account for the fact that compared to a negative productivity shock, a markup shock leads the corresponding producer to release some resources to the rest of the economy. These released resources can eventually be translated into released factor uses.

3.3.2 Markup Shocks

Proposition 5 (Markup/Wedge Shocks with Multiple Factors)**.** *In response to a markup shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log\mu_k} = -\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}+\sum_g\tilde{\Psi}_{(g)}\frac{\mathrm{d}\log\Lambda_g}{\mathrm{d}\log\mu_k},\frac{\Psi_{(f)}}{\Lambda_f}\right)-\lambda_k\frac{\Psi_{kf}}{\Lambda_f}. \tag{17}$$

Given $\mathrm{d}\log\Lambda_f/\mathrm{d}\log\mu_k$*, we know, from Theorem 1 that*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_k} = -\tilde{\lambda}_k+\frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log\mu_k} = -\tilde{\lambda}_k-\sum_f\tilde{\Lambda}_f\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log\mu_k}.$$

To isolate the importance of this new term, we go back to the the Cobb-Douglas economy with a markup shock, where the only source factor reallocation comes from the fact that the producer which increases its markup releases some factors. Let $\theta_j = 1$ for every j, which is the Cobb-Douglas special case. Now, applying Proposition 5, we get

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_k} = -\tilde{\lambda}_k+\lambda_k\sum_f\tilde{\Lambda}_f\frac{\Psi_{kf}}{\Lambda_f} = -\tilde{\lambda}_k\left(1-\frac{\lambda_k}{\tilde{\lambda}_k}\sum_f\tilde{\Lambda}_f\frac{\Psi_{kf}}{\Lambda_f}\right).$$

This generalizes the intuitions discussed earlier for markup/wedge shocks in the Cobb-Douglas economy with a single factor to the case of multiple factors. In particular, the amount of factor f released by sector k as a fraction of total factor f per unit of shock is $\lambda_k\Psi_{kf}/\Lambda_f$ and the impact of that release on output per unit of shock is $\tilde{\Lambda}_f$. We also see again the roles of the index of downstream distortions $\lambda_k/\tilde{\lambda}_k$ and of the generalized index of upstream distortions $\sum_f\tilde{\Lambda}_f\Psi_{kf}/\Lambda_f$.

4 Applying and Interpreting the Results

Before moving on to applications, we pause to discuss some implementation and interpretation issues. Applying our formulas for inefficient economies requires more care than

when handling Hulten's theorem for efficient economies. Over and above the difficulties involved with reliably estimating wedges, there are two important issues that have to be confronted, namely the identification of the factors of production, and the level of aggregation of the data. We discuss these issues in turn after reviewing data requirements. We also discuss how to account for issues like biased technical change, demand shocks, and endogenous wedges in interpreting our results.

Observability

The vector of final expenditure shares b and the revenue-based input-output matrix Ω are directly observable from input-output data at the industry level and even sometimes even at the firm-level using value-added tax data. The vector of revenue-based Domar weights λ is observable even without input-output data. Unlike revenues, however, we do not typically directly observe costs, and so the cost-based input-output matrix $\tilde{\Omega}$ and the vector of cost-based Domar weights $\tilde{\lambda}$ are not readily observable from input-output data. Instead, these cost-based objects must be inferred from their observable revenue-based counterparts using the vector of wedges μ:

$$\Omega = (\text{diag}\,\mu)^{-1}\tilde{\Omega}, \quad \tilde{\Psi} = (I - \tilde{\Omega})^{-1}, \quad \text{and} \quad \tilde{\lambda}' = b'\tilde{\Psi}.$$

Of course, this requires knowledge of the vector of wedges. These wedges can sometimes be directly observed (taxes for example) or independently estimated (as in our markups applications for example). They could also arise from the structural estimation or calibration of a structural model super-imposed on our setup (incorporating a model of credit constraints generating capital wedges, or a model of industry competition generating markups for example). Alternatively, the cost-based objects can be inferred from elasticities of the production or cost function, if these objects can be directly estimated.

Identification of the Factors

One issue we have to confront when working with inefficient models is that we have to identify the factors of production. For an efficient economy, we do not need to worry about reallocation of resources, and hence we do not need to specifically identify and track the changes in factor income shares. For an inefficient economy, we must take a stance on this issue. The most challenging problem here is to identify "fixed" or quasi-fixed factors of production – namely, those factors whose presence gives rise to decreasing returns to scale for a producer, and whose factor payments need to be separated from pure profits.

In other words, when the equilibrium is inefficient, we need to take a stance about whether factors are "stuck" due to technological restrictions or market imperfections. In mapping the model to the data, we need to choose whether two factors that receive a different wage are being paid different wages due to frictions, or due to the fact that there are technological differences between the factors. These are issues that we do not have to confront when the equilibrium is efficient, since the consequences of reallocation are null to the first order.

Data Aggregation Level

The second issue is the aggregation of the data before it reaches the researcher. Up to a first-order approximation, efficient economies have a tremendously useful aggregation property: for a common productivity shock A to a collection of producers $S \subset \{1, \ldots, N\}$, the first order impact of the shock is given by $\mathrm{d}\, Y / \mathrm{d} \log A = \sum_{i \in S} p_i y_i$. In other words, the total sales of all producers in S will yield the impact of an aggregate shock to all producers in S.[27] So, we only need to observe sales data at the level of disaggregation at which a shock occurs.

This aggregation property does not hold for distorted economies, even in the Cobb-Douglas or acyclic cases where we do not need to account for changes in allocative efficiency. Unlike sales, cost-based Domar weights $\tilde{\lambda}$ are not directly observable, and instead need to be computed from input-output data *at* the level of disaggregation at which the markups and wedges appear. If wedges apply at the firm or establishment level, then firm or establishment-level input-output data is in general necessary, even if shocks are aggregate. See Appendix F for a worked-out example.[28]

Biased Technical Change and Demand Shocks

Although the model is written in terms of Hicks-neutral productivity shocks, this is done without loss of generality. We can always capture non-neutral productivity shocks, say factor-augmenting shocks, by relabelling the relevant factor of a given producer to be a separate producer. Then, Hicks-neutral productivity shocks to that industry would be

[27] Baqaee and Farhi (2017) present an important caveat to this observation: this first-order approximation can be highly unreliable in certain contexts.

[28] In Section 5, we apply our results in the case of markups using firm-level data. Firms are grouped into industries. We make the assumption that all firms within an industry have the same production function but have heterogenous markups and productivities. Given this assumption, we can recover, using the structure of the model, the input-output data at the firm level (which we do not observe) from the input-output data at the industrial level and the joint distribution of markups and size at the firm level within an industry (which we observe).

identical to factor-biased productivity shocks in the original model.

Demand shocks can also be modeled in this way. To capture demand shocks, we can use a mixture of consumer-specific productivity shocks: so for instance, an increase in demand by i for inputs from j can be modelled as a positive productivity shock when j sells to i and a series of negative productivity shocks when anyone besides j sells to i. In an efficient economy, Hulten's theorem implies that such changes in the composition of demand have no effect on aggregate TFP, since the positive demand shock cancels out the negative demand shock to the rest. However, in a model with distortions, the change in the composition of demand can affect TFP by changing allocative efficiency.

Frictions as Wedges

Throughout, we model distortions via monopoly markups and wedges. The wedges act like linear taxes, the revenues of which are rebated lump sum.[29,30] Beyond actual taxes, these wedges can also implicitly capture frictions preventing the reallocation of resources. For example, they can capture credit constraints and they must then be interpreted as the Lagrange multipliers on the constraints in the individual firms' cost minimization problem.

In mapping our results to the data, we assume that expenditures by i on inputs from j, and the revenues of i, are recorded *gross* of taxes and markups. In the case where these wedges are reduced-form representations of frictions like credit constraints, we adopt the convention of writing expenditures gross of these implicit wedges.[31] This is purely a convention which does not change anything to the economics of the problem. This convention might not coincide with the accounting convention for expenditures in the data. In that case, the data must be converted into the format required by our theory. This conversion is completely straightforward. For example, in the case of a credit constraint which increases the rental rate of capital perceived by a firm but not its true rental rate, the conversion requires inflating the firm's expenditure on capital measured in the data by a percentage equal to the equivalent implicit tax on capital given by the Lagrange multiplier on the constraint in the cost minimization problem of the firm.

[29] If the taxes were not rebated, then they would act as reductions in productivity since resources would actually be destroyed, and hence the first welfare theorem and Hulten's theorem would still apply.

[30] The question of how the distribution of lump sum rebates across the consumer and the different producers is merely an accounting convention, which is irrelevant to the economics of the problem, but which matters for mapping the model to the data. We expand on this issue below.

[31] See e.g. Bigio and La'O (2016).

Endogenous Productivities and Markups

Our framework treats productivity and wedges as exogenous primitives. However, our results can also be used to study situations in which these are endogenous to some more fundamental parameter.[32] For example, consider some parameter θ which gives rise to some endogenous vector of equilibrium productivities $A(\theta)$ and markups $\mu(\theta)$. Then a simple application of the chain rule will yield comparative statics of output in θ. The effect can then be decomposed in two: how A and μ respond to a change in θ, and how output responds to the change in A and μ. The advantage of our approach is that we can characterize the latter effect, in general, without committing to a specific theory of wedge determination or of innovation.

We provide a fully worked out example along this line in Appendix D when we consider a model with nominal rigidities. The model can be recast as a model with endogenous markups ensuring that the relevant prices stay constant. In this case, we explicitly solve these endogenous markups as a function of the underlying productivity and monetary policy shocks. We then apply the chain rule in conjunction with our results to characterize the effects of these shocks. One advantage of our approach is that we can cleanly separate the effects of monetary policy on output from the effects of monetary policy on aggregate TFP.

5 Applications

In this section, we pursue some quantitative applications of our results, focusing on markups as the source of inefficiency. First, we measure changes in allocative efficiency in the US over time, and decompose the Solow-residual into changes in pure-technology and changes in allocative efficiency. Next, we calibrate a simplified version of our parametric model to match firm-level markup and size data, as well as input-output data. We compute output elasticities with respect to firm-level and industry-level shocks to productivity and markups, and we compare these elasticities to those implied by the perfectly competitive and Cobb-Douglas models. We use these elasticities to estimate the amount of macroeconomic volatility arising from microeconomic shocks and the gains from eliminating wedges.

We work with the annual US input-output data from the BEA, dropping the government, noncomparable imports, and second-hand scrap industries. The dataset contains

[32] Of course, conditional on knowing the changes in productivity and the wedges, we can use our results without modification (for example, as we do in the growth accounting application in Section 2.5).

industrial output and inputs from 1997 to 2015 with 66 industries. We calibrate the expenditure share parameters to match the input-output table, and we use three alternative measures of markups estimated for Compustat firms. We only have markups and sales data at the microeconomic level for publicly listed firms in the US from Compustat. To extrapolate to the whole economy, we therefore make the assumption that Compustat firms are representative of the overall economy in the sense that the sales-weighted distributions of markups by industry and their transition matrices for the overall economy are the same as for Compustat. We then combine these data with input-output data at the industry level from the BEA to aggregate the economy.

The first markup series is estimated by Gutiérrez and Philippon (2016) and Gutierrez (2017), and relies on inferring markups from measured profits. These estimates are derived as residuals from gross operating surplus, after accounting for "normal" payments to capital. The "normal" payments to capital are computed via a user-cost of capital calculation, where the rental price takes into account the equity risk premium, following the framework of Caballero et al. (2017). We refer to these markups as GP markups. The second method for computing markups is to use the Lerner index, referred to as LI, which infers markups from average operating profit margin. The final set of estimates are from De Loecker and Eeckhout (2017), which we call DE markups, and rely on the production function estimation method laid out in De Loecker and Warzynski (2012). DE markups are given by the ratio of the elasticity of the production function to a variable input to the share of that input in revenues. All markup series are estimated for publicly listed firms in the US from Compustat.

The three markup series give different levels of markups: the GP markups are the smallest (and average around 5%), the LI markups are higher (averaging around 13%), and the DE markups are the largest (averaging around 30%). Whereas the GP and LI markups capture "average" markup margins (by stripping out expenses from revenues), the DE markups are designed to capture markups at the margin (gaps between the expenditure shares and output elasticities). For our empirical application, we maintain the assumption of constant returns, so there is no theoretical reason to prefer one set of markups over another.

Each markup series comes with its own pros and cons. The GP markups require measurements of the capital stock and industry-level estimates of the equity risk premium, both of which are notoriously difficult to measure. The DE markups on the other hand, rely on more parametric methods, and their changes over time are potentially biased in the presence of capital-biased technical change. We use the GP markups for our benchmark numbers, and we report numbers for the other two markup series in the tables and

in Appendix A.[33] Despite their differences, all three markup series show an increasing average markup over the sample.

5.1 Decomposing the Solow Residual

In this section, we implement our reduced-form results to decompose the sources of TFP growth as measured by the cumulated Solow residual in the US over the period 1997-2015, in the presence of these changing markups.

Conditional on markups and the input-output matrix at a given point in time t, we can approximate $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1}) = -\tilde{\Lambda}'_{t-1}\Delta \log \Lambda_{t-1}$ from $t-1$ to t using the change in observed factor income shares. Then we can decompose the Solow residual using equation (7). The results are plotted in Figure 5 using the GP markups. The sum of the red (allocative efficiency), yellow (factor under-counting), and purple ("pure" technology) lines add up to give the cumulative change in the Solow residual. Since we are interested in long-run trends, we assume that the only factors are labor and capital, and we abstract away from barriers to reallocation of factors like adjustment costs and variable capacity utilization.

We see that since the start of the sample, allocative efficiency has improved, and accounts for about 50% of TFP growth as measured by the cumulated Solow residual. The correction for the under-counting of factors in the Solow residual arising from the fact that $\Lambda_{t-1} \neq \tilde{\Lambda}_{t-1}$ is negative but small. Taken together, this implies that "pure" technology changes, which are computed as a residual, also account for about 50% of TFP growth as measured by the cumulated Solow residual. In other words, because allocative efficiency has improved considerably, "pure" technology has improved much less than would be implied by a naive interpretation of the Solow residual.

As documented by Gutierrez (2017), average markups have been increasing in all three of our markup series. Given the increase in the average markup, and the growing profit share in the economy, how then can we claim that allocative efficiency has increased over the same period? The key lies in realizing that markups on average have increased primarily because firms that charge large markups have gotten larger. While on average, markups are trending upwards, the average change in log markups has not been increasing. If on average, markups are not increasing, but firms with high markups are getting larger, then this implies that allocative efficiency in the economy *must* be increasing. Of course, to quantify and weigh the various changes correctly, we need to use the weights in equation (7).

[33] Note that our method allows for capital-biased technical change. In particular, we can measure changes in allocative efficiency independently of the nature of productivity shocks (Hicks neutral or factor biased). For more details see the discussion in Section 4.

In Figure 6 we plot the cumulative sum of $-\tilde{\lambda}'_{t-1}\Delta \log \mu_t$ and $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1})$ over the sample. Note that these are two components of changes in allocative efficiency. Both terms have contributed positively to allocative efficiency. The fact that the first term is positive means that (the appropriately weighted) average change in markups has been negative, even though the average markup has been increasing. The fact that the second term is positive means that there has been a reduction in the factor income shares, reflecting the fact that the average markup has been increasing. These two terms confirm the compositional origin of the increase in the average markup with high-markup firms expanding at the expense of low-markup firms, and the resulting improvement in allocative efficiency.

Overall, these patterns are also borne out when we use the LI and DE markups, although the magnitudes are different (see Appendix A). In particular, the contribution of allocative efficiency is similar at roughly 50% of the cumulated Solow residual, but the correction for factor under-counting is larger, simply because the markups are larger. As a result, the contribution of "pure" technology is also larger and is about equivalent to the cumulated growth of the Solow residual.

5.2 A Quantitative Structural Model

In this section, we use our structural results to explore quantitatively the importance of markup distortions. We calibrate a simplified version of the parametric model presented in Section 3.

To calibrate the model, we need estimates for industry-specific firm-level and industry-level structural elasticities of substitution. Unfortunately, disaggregated estimates of these elasticities do not exist. We consider a nested CES structure where each firm i in industry j produces using a CES aggregator of value-added VA and intermediate inputs X:

$$\frac{y_j(i)}{\overline{y}_j(i)} = A_i(j)\left(a_j\left(\frac{VA_j(i)}{\overline{VA}_j(i)}\right)^{\frac{\theta-1}{\theta}} + (1-a_j)\left(\frac{X_j(i)}{\overline{X}_j(i)}\right)^{\frac{\theta-1}{\theta}}\right)^{\frac{\theta}{\theta-1}}.$$

Value-added consists of labor and capital inputs

$$\frac{VA_j(i)}{\overline{VA}_j(i)} = \left(v_j\left(\frac{l_j(i)}{\overline{l}_j(i)}\right)^{\frac{\eta-1}{\eta}} + (1-v_j)\left(\frac{k_j(i)}{\overline{k}_j(i)}\right)^{\frac{\eta-1}{\eta}}\right)^{\frac{\eta}{\eta-1}},$$

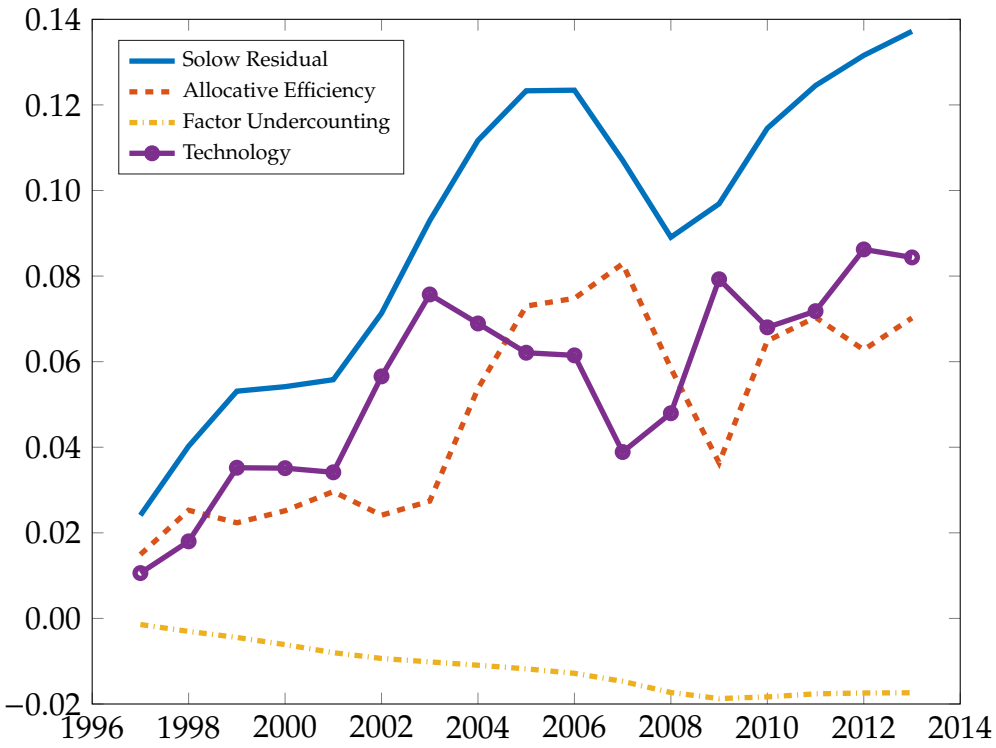


Figure 5: The decomposition in equation (7) using the Gutiérrez and Philippon (2016) markup data.

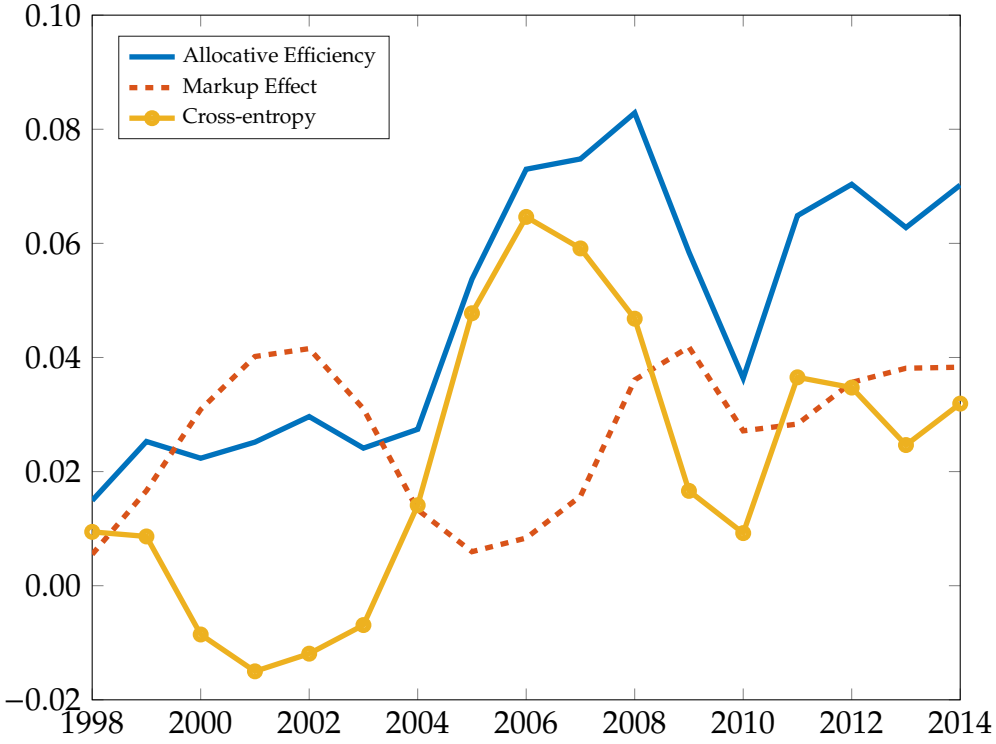


Figure 6: The cumulated contribution of (minus) changes in log markups $-\tilde{\lambda}'_{t-1}\Delta \log \mu_t$ and of changes in cross entropy $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1})$ using the Gutiérrez and Philippon (2016) markup data. The sum of the two components give the overall cumulated change in allocative efficiency.

and the intermediate input consists of inputs from other industries

$$\frac{X_j(i)}{\overline{X}_j(i)} = \left(\sum_{k=1}^{N} \omega_{jk} \left(\frac{x_{jk}(i)}{\overline{x}_{jk}(i)} \right)^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon-1}}.$$

Inputs purchased by firms in industry j from industry k are a CES aggregate of all varieties in that industry

$$\frac{x_{jk}(i)}{\overline{x}_{jk}(i)} = \left(\sum_{m=1}^{N_k} \delta_k(m) \left(\frac{x_{jk}(i,m)}{\overline{x}_{jk}(i,m)} \right)^{\frac{\xi-1}{\xi}} \right)^{\frac{\xi}{\xi-1}}.$$

Following our previous work (Baqaee and Farhi, 2017), and drawing on estimates from Atalay (2017) and Boehm et al. (2014), we set $\theta = 0.3$, $\theta_0 = 0.4$, and $\varepsilon \approx 0$. We set $\eta_i = 1$ which is a focal point in the literature about the micro-elasticity of substitution between labor and capital [34] Finally, we set $\xi = 8$, which is within the range of estimates of the variety-level elasticity of substitution from the industrial organization and international trade literatures. An elasticity of $\xi = 8$ is also consistent with our measure of average markups in the benchmark model, assuming a Dixit and Stiglitz (1977) market structure.

Gains from Reducing Markups

We use the calibrated model to approximate the gains to aggregate TFP from eliminating markups following the approximation result in equation (9) in Table 1. Using the benchmark GP markups and our second-order approximation, eliminating markups, holding fixed technology, increases aggregate TFP by around 20%.

Our second-order approximation uses the fact that, according to Corollary 1, *at* the efficient equilibrium the elasticity of aggregate TFP to a reduction in markups is zero. Hence, using the insights of Hotelling (1938) and Theil (1967), we construct a second-order estimate of the gains from reducing markups by averaging the first order gains at the initial and terminal (efficient) allocation. This corresponds to the equivalent of the area of the Harberger deadweight loss triangle in our general equilibrium model.[35]

The LI markups imply somewhat smaller gains, and the DE markups imply the largest gains. Indeed, with LI markups, the aggregate TFP gain is 17%, and with DE markups, it is 35%.

[34] There is quite a bit of disagreement in that literature, most estimates cluster a bit below 1, but there are also some estimates slightly above 1.

[35] In partial equilibrium, Harberger triangles are also second-order approximations, as shown by Hotelling.

Interestingly, we find that the gains from reducing markups have increased substantially since the start of the sample for all three series. For example, using our benchmark GP markups, we find that the gain from eliminating markups is 3% in 1997, much smaller that the 20% in 2014. As we described in Secion 2.7, this finding is logically consistent with our finding that allocative efficiency has improved since the start of the sample, since the counterfactual comparison in the two scenarios is different. Specifically, our calibrated model suggests that despite the fact that reallocations have improved the allocation, the frontier has moved more quickly, and so the gap has gotten larger.

In Table 2, we repeat the markup reduction exercise for some alternative specifications of the structural model. We consider the gains implied by a Cobb-Douglas-CES specification of the model which imposes that all elasticities apart from the elasticities of substitutions among firms within an industry are equal to 1, as well as the gains for a Cobb-Douglas-Cobb-Douglas specification of the model which imposes that all elasticities are equal to 1. We also compute the gains that would be implied by using value-added production functions which ignore the role of the production network. Value-added production functions are commonly used in the literature on misallocation, and our results suggest that relying on this simplification can substantively reduce the gains from eliminating frictions. In particular, we find that working with value-added productions can cut the estimated gains from reducing markups by more than half.

Our estimate that eliminating markups in the US economy in 2015 would increase TFP by about 20% raises the estimated cost of monopoly distortions by two orders of magnitude compared to the famous estimates of 0.1% of Harberger (1954). Essentially, the reasons for this dramatic difference is that we use firm-level data, whereas Harberger only had access to sectoral data, and that the dispersion of markups is higher across firms within a sector than across sectors. Moreover, the relevant elasticity of substitution is higher in our exercise than in Harberger's since it applies across firms within a sector rather than across sectors. Finally, we properly take into account the input-output structure of the economy to aggregate the numbers in all industries whereas Harberger only focused on manufacturing an essentially ignored input-output linkages.

Of course, both our estimate and Harberger's are static, taking as given the level of productivity in the economy. Markups may be playing an important role in incentivizing innovation and entry, so that exogenously eliminating markups may harm productivity. In Section 6, we discuss how one might try to account for these forces. Briefly, even if markups do play an important role in incentivizing innovation, they also distort the allocation of resources and our calculation is aimed at quantifying this latter effect.

	Gutiérrez-Philippon	Lerner Index	De Loecker-Eeckhout
End	20%	17%	35 %
Start	3%	5%	21%

Table 1: Gains from eliminating markups $\exp(1/2 \sum(\sum((\mathrm{d}\log Y/\mathrm{d}\log\mu_i(1-\mu_i)/\mu_i))))$ for the various markup series for the beginning (1997) and end of our sample (2014 for DE and 2015 for GP and LI).

	Benchmark	CD+CES	CD+CD	VA Benchmark	VA CD + CES	VA CD + CD
GP	20%	21%	4%	8%	8%	1%
LI	17%	18 %	4 %	7%	7%	1%
DE	35%	38%	7%	18 %	18%	3%

Table 2: Gains in aggregate TFP from eliminating markups towards for different markup series, and different structural models. CD+CES preserves the input-output structure, but sets all elasticities except ξ equal to one. CD + CD sets all elasticities to one. VA specifications eliminate the input-output matrix and use value-added production functions, a la Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). For VA, all elasticities except ξ equal to one, VA CES uses the same elasticities as the benchmark model, and VA CD sets all elasticities of substitution equal to one.

Volatility of Aggregate TFP

With comparative statics in both productivity and markup shocks at both the firm and industry level[36]

$$\log Y \approx \log \overline{Y} + \sum_i \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_i}\mathrm{d}\log A_i + \sum_i \frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_i}\mathrm{d}\log \mu_i,$$

we can also approximate the implied volatility of output in response to microeconomic shocks. Assuming productivity shocks and markup shocks are independent and identically distributed, we can approximate the volatility of output using

$$\begin{aligned} Var(\log Y) &\approx \sum_i \left(\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_i}\right)^2 Var(\mathrm{d}\log A_i) + \sum_i \left(\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_i}\right)^2 Var(\mathrm{d}\log \mu_i), \\ &= ||D_{\log A}\log Y||^2 Var(\mathrm{d}\log A) + ||D_{\log \mu}\log Y||^2 Var(\mathrm{d}\log \mu). \end{aligned}$$

[36]When we consider firm-level shocks, we assess only the contribution of shocks to Compustat firms. We focus on this exercise, for which we have the necessary data, because we do not have the data required to compute the contribution of shocks to all firms.

Hence, the Euclidean norm $\|D_{\log A} \log Y\|$ of the Jacobian of $\log Y$ with respect to $\log A$ gives the degree to which microeconomic productivity shocks are not "diverisified" away in the aggregate. Similarly, $\|D_{\log \mu} \log Y\|$ measures the diversification factor relative to markup shocks.[37]

Table 3 displays the diversification factor, for both markup shocks and productivity shocks at the firm level and at the industry level, for our benchmark model. We also compute the results for a Cobb-Douglas distorted economy where all elasticities are unitary, as well as for a perfectly competitive model without wedges. Across the board, the distorted model is more volatile than the competitive model, however the extent of this depends greatly on the type of shock and the level of aggregation. We discuss these different cases in turn.

First, consider the case of productivity shocks: as mentioned previously, the benchmark model is more volatile than the perfectly competitive model for both sets of shocks. However, the more interesting comparison is with respect to the distorted Cobb-Douglas economy. As explained in Section 3, the allocation of factors is invariant to productivity shocks in the Cobb-Douglas model. Hence, the Cobb-Douglas model lacks the reallocation channel, and hence can tell us in which direction the reallocation force is pushing. In the case of industry-level shocks, the benchmark model is slightly less volatile than the Cobb-Douglas model, whereas in the case of firm-level shocks, the benchmark model is significantly more volatile.

A partial intuition here relates to the elasticities of substitution: whereas industries are complements, firms within an industry are strong substitutes. Recall that loosely speaking, changes in allocative efficiency scale with the elasticity of substitution minus one. Firm-level shocks cause a considerable amount of changes in allocative efficiency whereas industry-level shocks cause much milder changes. At both levels of aggregation, these changes in allocative efficiency amplify some shocks and mitigate some others compared to the Cobb-Douglas model with no change in allocative effiency.[38] On the whole, at the firm level, the changes in allocative efficiency are so large that they dwarf the "pure" technology effects picked up by the Cobb-Douglas model and amplify the volatility of these shocks. By contrast, at the industry level, changes in allocative efficiency are more moderate and turn out to slightly mitigate the volatility of these shocks.

[37] Although Baqaee and Farhi (2017) suggest that log-linear approximations can be unreliable for modeling the mean, skewness, or kurtosis of output in the presence of microeconomic shocks, their results indicate the log-linear approximations of variance are less fragile (although still imperfect). In the final section of this paper, we discuss how our results can be extended to understanding the nonlinear impact of shocks.

[38] There is another difference: reallocation occurs towards the firm receiving a positive shock; but reallocation occurs away from the industry receiving a positive shock

	Benchmark	Competitive	Cobb-Douglas	Passive
Firm Productivity Shocks (GP)	0.0491	0.0376	0.0396	0.0396
Firm Markup Shocks (GP)	0.0296	0.0000	0.0077	0.0000
Industry Productivity Shocks (GP)	0.3162	0.3118	0.3259	0.3259
Industry Markup Shocks (GP)	0.0084	0.0000	0.0391	0.0000
Firm Productivity Shocks (LI)	0.0524	0.0376	0.0415	0.0415
Firm Markup Shocks (LI)	0.0368	0.0000	0.0085	0.0000
Industry Productivity Shocks (LI)	0.3188	0.3118	0.3375	0.3375
Industry Markup Shocks (LI)	0.0127	0.0000	0.0500	0.0000
Firm Productivity Shocks (DE)	0.0598	0.0376	0.0398	0.0398
Firm Markup Shocks (DE)	0.0321	0.0000	0.0112	0.0000
Industry Productivity Shocks (DE)	0.3299	0.3418	0.3618	0.3618
Industry Markup Shocks (DE)	0.0216	0.0000	0.0760	0.0000

Table 3: Diversification factor for different productivity and markup shocks at firm and industry level for different specifications of the model. A diversification factor of 1 means means that the variance of microeconomic shocks moves aggregate variance one-for-one. A diversification factor of 0 means that microeconomic shocks are completely diversified away at the aggregate level. GP corresponds to the Gutiérrez and Philippon (2016) markups, LI is markups according to the Lerner Index, and DE is using markup data from

These intuitions are confirmed in the first two columns of Figure 7, where we plot the output elasticity with respect to productivity shocks to specific firms or industries relative to their revenue-based Domar weight. This represents a comparison of our benchmark model to a competitive model where Hulten's theorem holds. We find considerable dispersion in the response of the model relative to both, but much more so at the firm level than at the industry level. We could plot the same graph but with the cost-based Domar weight as a reference point in order to represent the comparison of our benchmark model to the Cobb-Douglas model, and the results would be visually similar.

Next, consider the effects of markup shocks. In this case, the distorted Cobb-Douglas economy is not necessarily a very natural benchmark since even with Cobb-Douglas, shocks to markups will reallocate factors across producers. Nonetheless, it is still instructive to compare the benchmark model to the Cobb-Douglas one to find that a similar lesson applies as with productivity shocks. The volatility of firm-level shocks is amplified relative to Cobb-Douglas while the volatility of industry-level shocks is attenuated relative to Cobb-Douglas. This follows from the fact that industries are more complementary than firms, and hence, in line with the intuition from the horizontal economy, the effect of the shock are monotonically increasing in the degree of substitutability. The last two columns of Figure 7 plot the output elasticity with respect to markup shocks to specific firms or industries relative to their revenue-based Domar weight.

6 Robustness and Extensions

Our results can be extended to address some limitations of our analysis. In Appendix C we discuss, and in some cases fully characterize, how our basic framework could handle other complications: capital accumulation, adjustment costs, and variable capacity utilization, fixed costs, free entry, and nonlinearities. These issues introduce additional forces and mechanisms into the model, and we plan to squarely focus on these in future work. However, in the appendix, we show that the intuitions gleaned from the basic framework continue to be useful in analyzing these more complex scenarios. In this section, we show to generalize our analysis to allow for endogenous factor supplies.

To model elastic factor supplies, let $G_k(w_k/P_y, Y)$ be the supply of factor k, where w_k/P_y is the real price of the factor and Y is aggregate income. Let $\zeta_f = \partial \log G_f / \partial \log(w_k/P_y)$ be the elasticity of the supply of factor f to its real wage, and $\gamma_f = -\partial \log G_f / \partial \log Y$ be its

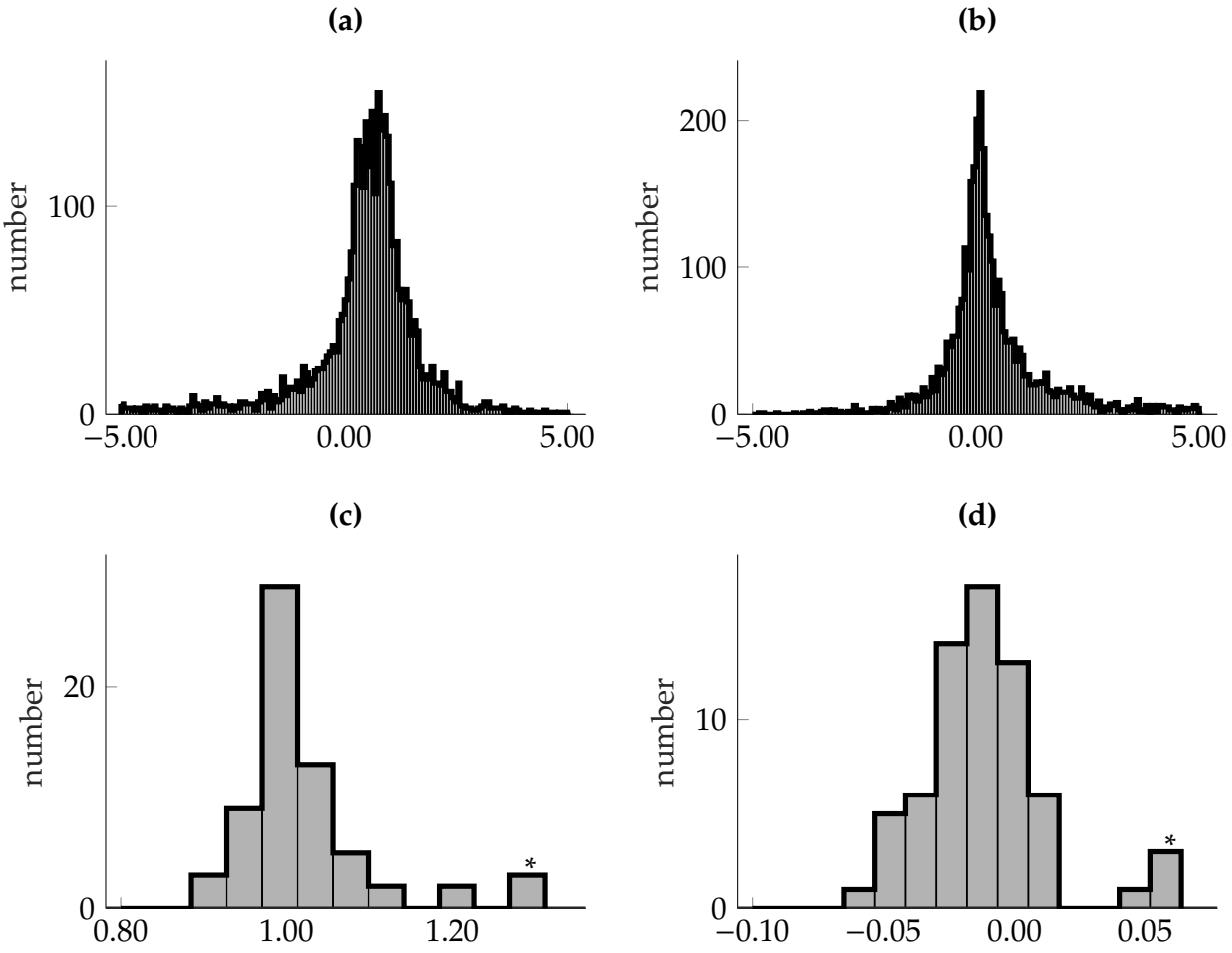


Figure 7: The left column contains histograms of $\mathrm{d}\log Y/\mathrm{d}\log A$ and the right $\mathrm{d}\log Y/\mathrm{d}\log \mu$ relative to λ for firm-level and industry-level shocks respectively. The bunching at the extremes, winsorizing at 4 standard deviations, marked with a star arise solely for displaying purposes. In all cases, the degree of dispersion around the response implied by the competitive model or the size of the producer is substantial.

income elasticity. We then have the following characterization:

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \varrho\left(\tilde{\lambda}_k - \frac{1}{1+\zeta}\frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log A_k}\right) = \varrho\left(\tilde{\lambda}_k - \sum_f \frac{1}{1+\zeta_f}\tilde{\Lambda}_f\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k}\right), \tag{18}$$

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = \varrho\left(-\tilde{\lambda}_k - \frac{1}{1+\zeta}\frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log \mu_k}\right) = \varrho\left(-\tilde{\lambda}_k - \sum_f \frac{1}{1+\zeta_f}\tilde{\Lambda}_f\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log \mu_k}\right), \tag{19}$$

where $\varrho = 1/(\sum_f \tilde{\Lambda}_f \frac{1+\gamma_f}{1+\zeta_f})$.

With inelastic factors, a decline in factor income shares, ceteris paribus, *increases* output since it represents a reduction in the misallocation of resources and an increase in aggregate TFP. With elastic factor supply, the output effect is dampened by the presence of $1/(1+\zeta_f) < 1$. This is due to the fact that a reduction in factor income shares, while increasing aggregate TFP, reduces factor payments and factor supplies, which in turn reduces output. Hence, when factors are elastic, increases in allocative efficiency from assigning more resources to more monopolistic producers are counteracted by reductions in factor supplies due to the associated suppression of factor demand.[39]

We can provide an explicit characterization of $\mathrm{d}\log\Lambda_f$ in terms of microeconomic elasticities of substitution by noting that changes in factor shares and output solve the following system of equations:

$$\begin{aligned} d\log\Lambda_f = &-\sum_k \lambda_k \frac{\Psi_{kf}}{\Lambda_f} d\log\mu_k + \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\left(\sum_k \tilde{\Psi}_{(k)} d\log A_k - \sum_k \tilde{\Psi}_{(k)} d\log\mu_k, \frac{\Psi_{(f)}}{\Lambda_f}\right) \\ &- \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\left(\sum_g \tilde{\Psi}_{(g)} \frac{1}{1+\zeta_g} d\log\Lambda_g - \sum_g \tilde{\Psi}_{(g)} \frac{\gamma_g + \zeta_g}{1+\zeta_g} d\log Y, \frac{\Psi_{(f)}}{\Lambda_f}\right), \end{aligned}$$

$$d\log Y = \frac{1}{\sum_f \tilde{\Lambda}_f \frac{1-\gamma_f}{1+\zeta_f}}\left[\sum_k \tilde{\lambda}_k d\log A_k - \sum_k \tilde{\lambda}_k d\log\mu_k - \sum_f \tilde{\Lambda}_f \frac{1}{1+\zeta_f} d\log\Lambda_f\right].$$

[39] In the limit where factor supplies become infinitely elastic, the influence of the allocative efficiency effects disappear from output, since more factors can always be marshaled on the margin at the same real price. To see this, consider the case with a single factor called labor, and factor supply function $G_L(w/P_y, Y) = \left(w/P_y Y\right)^{\nu}$, which can be derived from a standard labor-leisure choice model. In this case, $\gamma_L = \zeta_L = \nu$, and so equation (18) implies that $\mathrm{d}\log Y/\mathrm{d}\log A_k = \tilde{\lambda}_k + 1/(1+\nu)\,\mathrm{d}\,H(\tilde{\Lambda}_L, \Lambda_L)/\mathrm{d}\log A_k$. When labor supply becomes infinitely elastic $\nu \to \infty$, this simplifies to $\mathrm{d}\log Y/\mathrm{d}\log A_k = \tilde{\lambda}_k$, so that changes in allocative efficiency have no effect on output, even though they affect TFP.

Note that even without any frictions, Hulten's theorem cannot be used to predict how output will respond to microeconomic TFP shocks, due to endogenous responses of factors. So, equations (18) and (19) can also be applied to frictionless economies with endogenous factors.

7 Conclusion

We provide a non-parametric framework for analyzing and aggregating productivity and wedge shocks in a general equilibrium economy with arbitrary neoclassical production. Our results generalize the results of Solow (1957) and Hulten (1978) to economies with distortions. We show that, locally, the impact of a shock can be decomposed into a "pure" technology effect and an allocative efficiency effect. The latter can be measured non-parametrically using information about the wedges and the movements in factor income shares.

We apply our findings to the US, where our measure of wedges correspond to markups. We find that from 1997-2015, allocative efficiency in the US accounts for about half of aggregate TFP growth. We also find that the gains from reducing markups have increased since 1997, that eliminating markups will increase aggregate TFP by around 20% (up to a second order approximation). These numbers are substantially higher than classic estimates like those of Harberger (1954).

Although our results are comparative statics that take productivity and markups as exogenous, they can be used, in conjunction with the chain rule, to study models where productivity or markups are themselves endogenous. As an example, in Appendix D, we show how our framework could be applied to analyze the effects of productivity shocks and monetary policy shocks in models with sticky prices (where the effective markup of sticky firms is endogenous). In particular, we show how to characterize the oft-neglected potential effects of monetary policy shocks on TFP.

We end by speculating about some future areas for research, namely extending our analysis to allow for entry, increasing external economies, and nonlinearities. We view it as a very promising research direction to combine of our framework with more detailed industrial-organization models of market structure and imperfect competition, innovation, or more generally structural models of frictions in markets for credit, factors, and goods. We are pursuing these directions in ongoing work.

References

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., A. Ozdaglar, and A. Tahbaz-Salehi (2016). Networks, shocks, and systemic risk. In *The Oxford Handbook of the Economics of Networks*.

Asker, J., A. Collard-Wexler, and J. De Loecker (2014). Dynamic inputs and resource (mis) allocation. *Journal of Political Economy 122*(5), 1013–1063.

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

Atkeson, A. and A. Burstein (2008). Pricing-to-market, trade costs, and international relative prices. *The American Economic Review 98*(5), 1998–2031.

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

Banerjee, A. V. and E. Duflo (2005). Growth Theory through the Lens of Development Economics. In P. Aghion and S. Durlauf (Eds.), *Handbook of Economic Growth*, Volume 1 of *Handbook of Economic Growth*, Chapter 7, pp. 473–552. Elsevier.

Baqaee, D. R. (2015). Targeted fiscal policy.

Baqaee, D. R. (2016). Cascading failures in production networks.

Baqaee, D. R. and E. Farhi (2017). The macroeconomic impact of microeconomic shocks: Beyond Hulten's Theorem.

Barkai, S. (2016). Declining labor and capital shares.

Bartelme, D. and Y. Gorodnichenko (2015). Linkages and economic development. Technical report, National Bureau of Economic Research.

Bartelsman, E., J. Haltiwanger, and S. Scarpetta (2013, February). Cross-Country Differences in Productivity: The Role of Allocation and Selection. *American Economic Review 103*(1), 305–334.

Basu, S. (1995). Intermediate goods and business cycles: Implications for productivity and welfare. *The American Economic Review*, 512–531.

Basu, S. and J. G. Fernald (2002). Aggregate productivity and aggregate technology. *European Economic Review 46*(6), 963–991.

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

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

Bouakez, H., E. Cardia, and F. J. Ruge-Murcia (2009). The transmission of monetary policy in a multisector economy. *International Economic Review 50*(4), 1243–1266.

Buera, F. J., J. P. Kaboski, and Y. Shin (2011, August). Finance and Development: A Tale of Two Sectors. *American Economic Review 101*(5), 1964–2002.

Buera, F. J. and B. Moll (2012, January). Aggregate Implications of a Credit Crunch. NBER Working Papers 17775, National Bureau of Economic Research, Inc.

Caballero, R. J., E. Farhi, and P.-O. Gourinchas (2017, May). Rents, technical change, and risk premia accounting for secular trends in interest rates, returns on capital, earning yields, and factor shares. *American Economic Review 107*(5), 614–20.

Caliendo, L., F. Parro, and A. Tsyvinski (2017, April). Distortions and the structure of the world economy. Working Paper 23332, National Bureau of Economic Research.

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

Caselli, F. and N. Gennaioli (2013, 01). Dynastic Management. *Economic Inquiry 51*(1), 971–996.

Chari, V. V., P. J. Kehoe, and E. R. McGrattan (2007). Business cycle accounting. *Econometrica 75*(3), 781–836.

Chetty, R., A. Guren, D. Manoli, and A. Weber (2011). Are micro and macro labor supply elasticities consistent? a review of evidence on the intensive and extensive margins. *The American Economic Review 101*(3), 471–475.

De Loecker, J. and J. Eeckhout (2017). The rise of market power and the macroeconomic implications. Technical report, National Bureau of Economic Research.

De Loecker, J. and F. Warzynski (2012). Markups and firm-level export status. *The American Economic Review 102*(6), 2437–2471.

D'Erasmo, P. N. and H. J. Moscoso Boedo (2012). Financial structure, informality and development. *Journal of Monetary Economics 59*(3), 286–302.

Di Giovanni, J., A. A. Levchenko, and I. Méjean (2014). Firms, destinations, and aggregate fluctuations. *Econometrica 82*(4), 1303–1340.

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 (2015). Competition, markups, and the gains from international trade. *The American Economic Review 105*(10), 3183–3221.

Elsby, M. W., B. Hobijn, and A. Şahin (2013). The decline of the us labor share. *Brookings Papers on Economic Activity 2013*(2), 1–63.

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

Fernald, J. and B. Neiman (2011). Growth accounting with misallocation: Or, doing less with more in Singapore. *American Economic Journal: Macroeconomics 3*(2), 29–74.

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

Galí, J. (2008). Monetary policy, inflation, and the business cycle: An introduction to the New Keynesian Framework.

Gopinath, G. and O. Itskhoki (2010). Frequency of price adjustment and pass-through. *The Quarterly Journal of Economics 125*(2), 675–727.

Gopinath, G. and O. Itskhoki (2011). In search of real rigidities. *NBER Macroeconomics Annual 25*(1), 261–310.

Gopinath, G., Ş. Kalemli-Özcan, L. Karabarbounis, and C. Villegas-Sanchez (2017). Capital allocation and productivity in South Europe. *The Quarterly Journal of Economics*, qjx024.

Grassi, B. (2017). IO in I-O: Competition and volatility in input-output networks. Technical report.

Guner, N., G. Ventura, and X. Yi (2008, October). Macroeconomic Implications of Size-Dependent Policies. *Review of Economic Dynamics 11*(4), 721–744.

Gutierrez, G. (2017). Investigating global labor and profit shares.

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

Hall, R. E. (1990). *Growth/ Productivity/Unemployment: Essays to Celebrate Bob Solow's Birthday*, Chapter 5, pp. 71–112. MIT Press.

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

Hartman-Glaser, B., H. Lustig, and M. X. Zhang (2016, September). Capital Share Dynamics When Firms Insure Workers. NBER Working Papers 22651, National Bureau of Economic Research, Inc.

Hopenhayn, H. and R. Rogerson (1993, October). Job Turnover and Policy Evaluation: A General Equilibrium Analysis. *Journal of Political Economy 101*(5), 915–938.

Hopenhayn, H. A. (2014, August). On the Measure of Distortions. NBER Working Papers 20404, National Bureau of Economic Research, Inc.

Hotelling, H. (1938). The general welfare in relation to problems of taxation and of railway and utility rates. *Econometrica: Journal of the Econometric Society*, 242–269.

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

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

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

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

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

Koh, D., R. Santaeulàlia-Llopis, and Y. Zheng (2016). Labor share decline and intellectual property products capital.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics 22*(1), 79–86.

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

Midrigan, V. and D. Y. Xu (2014, February). Finance and Misallocation: Evidence from Plant-Level Data. *American Economic Review 104*(2), 422–458.

Moll, B. (2014, October). Productivity Losses from Financial Frictions: Can Self-Financing Undo Capital Misallocation? *American Economic Review 104*(10), 3186–3221.

Nakamura, E. and J. Steinsson (2010). Monetary non-neutrality in a multisector menu cost model. *The Quarterly journal of economics 125*(3), 961–1013.

Oberfield, E. (2013, January). Productivity and Misallocation During a Crisis: Evidence from the Chilean Crisis of 1982. *Review of Economic Dynamics 16*(1), 100–119.

Pasten, E., R. Schoenle, and M. Weber (2016). The propagation of monetary policy shocks in a heterogeneous production economy.

Pasten, E., R. Schoenle, and M. Weber (2017). Price rigidity and the granular origin of aggregate fluctuations.

Peters, M. (2013). Heterogeneous mark-ups, growth and endogenous misallocation.

Petrin, A. and J. Levinsohn (2012). Measuring aggregate productivity growth using plant-level data. *The RAND Journal of Economics 43*(4), 705–725.

Piketty, T. (2014). Capital in the 21st century.

Reis, R. (2013). The Portugese Slump and Crash and the Euro Crisis. *Brookings Papers on Economic Activity 44*(1 (Spring), 143–210.

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

Rognlie, M. (2016). Deciphering the fall and rise in the net capital share: accumulation or scarcity? *Brookings papers on economic activity 2015*(1), 1–69.

Sandleris, G. and M. L. J. Wright (2014, 01). The Costs of Financial Crises: Resource Misallocation, Productivity, and Welfare in the 2001 Argentine Crisis. *Scandinavian Journal of Economics 116*(1), 87–127.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal 27*(3), 379–423.

Solow, R. M. (1957). Technical change and the aggregate production function. *The review of Economics and Statistics*, 312–320.

Sraer, D. and D. Thesmar (2018). A sufficient statistics approach for aggregating firm-level experiments. Technical report, National Bureau of Economic Research.

Theil, H. (1967). *Economics and information theory*.

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

A Additional Figures

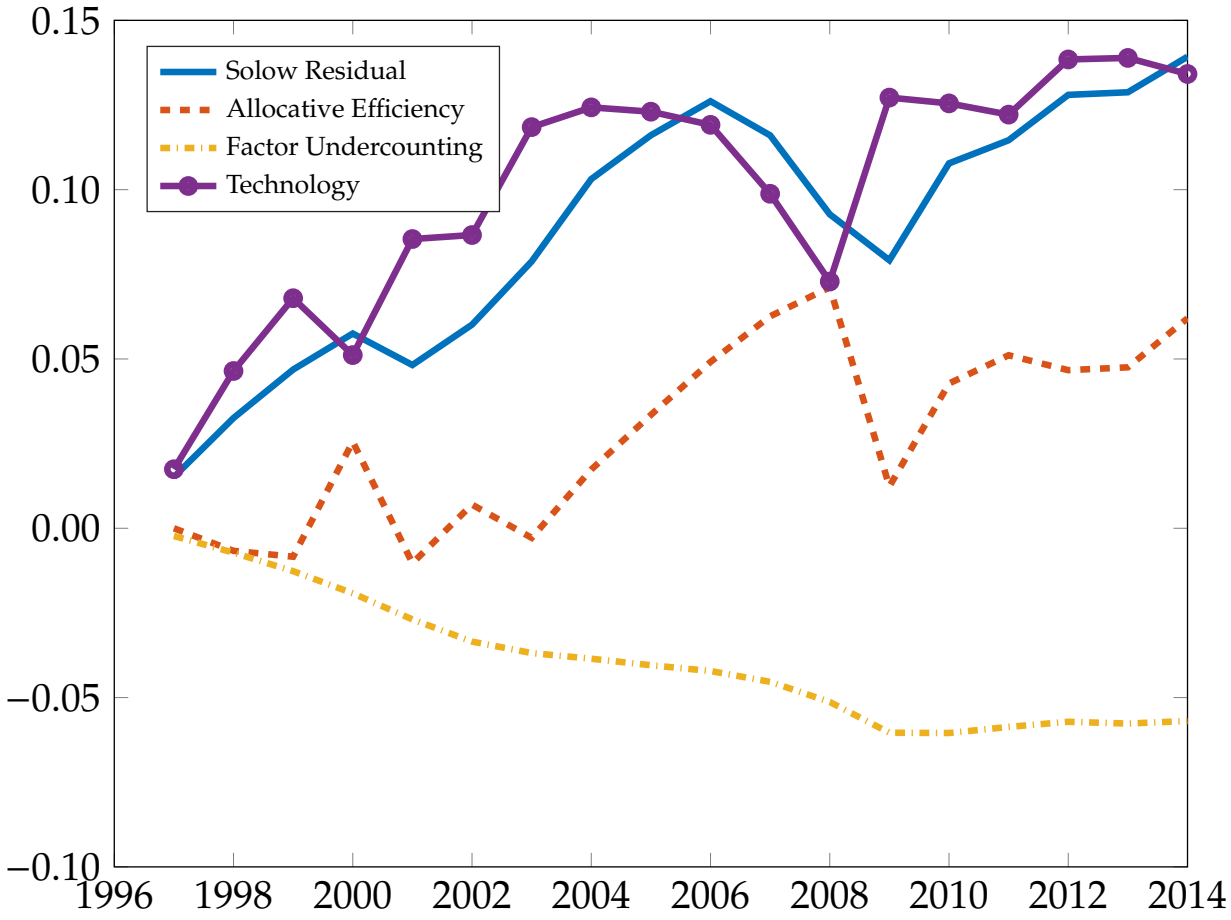


Figure 8: Decomposition of the Solow Residual using LI markups.

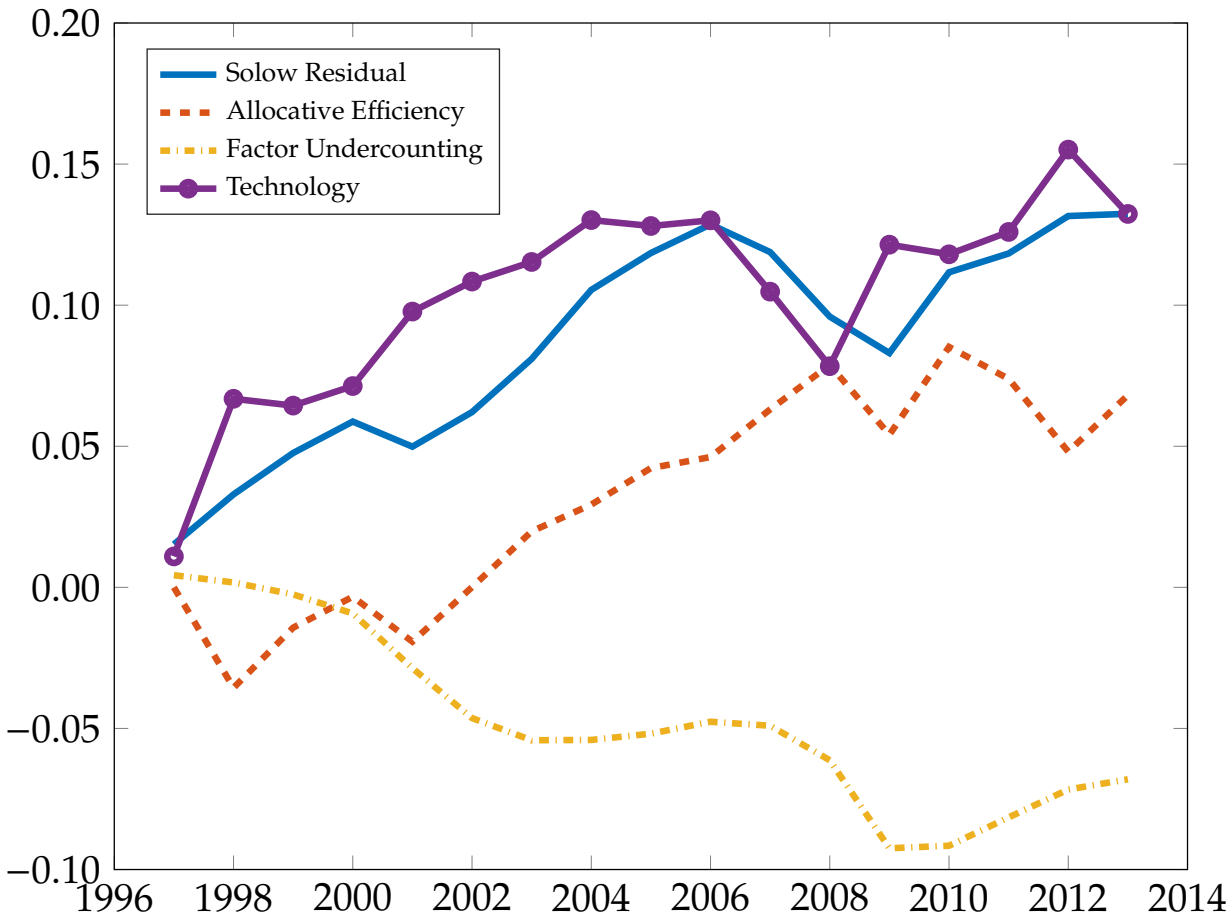


Figure 9: Decomposition of the Solow Residual using DE markups.

B Proofs

Proof of Theorem 1. By Shephard's lemma, we know that, to a first order, the productivity shock A_k will change the prices of any industry who purchases inputs, either directly or indirectly, from industry k

$$\frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} = -\mathbf{1}(i=k) + \sum_F \tilde{\Omega}_{if}\frac{\mathrm{d}\log w_F}{\mathrm{d}\log A_k} + \sum_j \tilde{\Omega}_{ij}\frac{\mathrm{d}\log p_j}{\mathrm{d}\log A_k}. \tag{20}$$

Denote the $N \times F$ matrix corresponding to $\tilde{\Omega}_{if}$ by $\tilde{\alpha}_{if}$. Then rewrite the equation above in matrix form to get

$$\frac{\mathrm{d}\log p}{\mathrm{d}\log A_k} = (I-\tilde{\Omega})^{-1}(\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} - e_k) = \tilde{\Psi}(\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} - e_k), \tag{21}$$

where e_k is the kth standard basis vector. Let the household's aggregate consumption good to be the numeraire, so that the household's ideal price index P_c is always equal to one. Then we know that

$$\frac{\mathrm{d}\log P_c}{\mathrm{d}\log A_k} = b'\frac{\mathrm{d}\log p}{\mathrm{d}\log A_k} = 0. \tag{22}$$

Combine this with the previous expression to get

$$-b'\tilde{\Psi}e_k + b'\tilde{\Psi}\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} = 0. \tag{23}$$

Note that, $b'\tilde{\Psi} = \tilde{\lambda}$ and $b'\tilde{\Psi}\tilde{\alpha} = \tilde{\Lambda}$. Hence,

$$-\tilde{\lambda}_k + \tilde{\Lambda}'\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} = 0. \tag{24}$$

Now, note that

$$\Lambda_f = \frac{w_f L_f}{P_c C}. \tag{25}$$

From this, we know that

$$\frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log A_k} = \frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log L_f}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k}. \tag{26}$$

Substitute this into the previous expression to get

$$-\tilde{\lambda}_k + \tilde{\Lambda}'\frac{\mathrm{d}\log \Lambda}{\mathrm{d}\log A_k} - \tilde{\Lambda}'\frac{\mathrm{d}\log L_f}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = 0, \tag{27}$$

where we use the fact that $\sum_f \tilde{\Lambda}_f = 1$. Rearrange this to get the desired result. To get an explicit characterization for $\mathrm{d}\log\Lambda_k/\mathrm{d}\log A_k$ in terms of structural parameters of the model, without loss of generality, assume that each good is produced from a distinct primary factor (this can be achieved by relabelling the input-output matrix). Now note that

$$\lambda_i = b_i + \sum_j \mu_j^{-1}\omega_{ji}\lambda_j, \tag{28}$$

$$w_i = \frac{\alpha_i p_i y_i}{\mu_i}, \tag{29}$$

$$\lambda_i = \frac{p_i y_i}{P_c C}, \tag{30}$$

$$p_i = \frac{1}{A_i} C_i(p_1, \ldots, p_N, w_i)\mu_i, \tag{31}$$

$$P_c = \frac{\sum_i p_i c_i}{C} = 1. \tag{32}$$

We can differentiate these to get to our answer. Denote the Moroshima elasticity of substitution between k and j for the total cost function of industry i by ρ^i_{kl}. Then we can write

$$\begin{aligned}\frac{\mathrm{d}\,\lambda_i}{\mathrm{d}\log A_k} = \frac{\mathrm{d}\,b_i}{\mathrm{d}\log A_k} + \sum_j \frac{\mathrm{d}\,\tilde{\Omega}_{ji}}{\mathrm{d}\log A_k}\mu_j^{-1}\lambda_j \\ + \sum_j \tilde{\Omega}_{ji}\mu_j^{-1}\frac{\mathrm{d}\,\lambda_j}{\mathrm{d}\log A_k} - \sum_j \tilde{\Omega}_{ji}\mu_j^{-1}\lambda_j\frac{\mathrm{d}\,\mu_j}{\mathrm{d}\log A_k},\end{aligned} \tag{33}$$

$$\frac{\mathrm{d}\,b_i}{\mathrm{d}\log A_k} = b_i \sum_{j\neq i} b_j\left(1 - \frac{1}{\rho^c_{ij}}\right)\left(\frac{\mathrm{d}\,p_i}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\,p_j}{\mathrm{d}\log A_k}\right), \tag{34}$$

$$\frac{\mathrm{d}\,\tilde{\Omega}_{ij}}{\mathrm{d}\log A_k} = \tilde{\Omega}_{ij}\sum_{k\neq j}\tilde{\Omega}_{ik}\left(1 - \frac{1}{\rho^i_{jk}}\right)\left(\frac{\mathrm{d}\,p_j}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\,p_k}{\mathrm{d}\log A_k}\right) + \tilde{\Omega}_{ij}\tilde{\alpha}_i\left(1 - \frac{1}{\rho^i_{jL}}\right)\left(\frac{\mathrm{d}\,p_j}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\,w_i}{\mathrm{d}\log A_k}\right), \tag{35}$$

$$\frac{\mathrm{d}\log w_i}{\mathrm{d}\log A_k} = \mu_i^{-1}\tilde{\alpha}_i\sum_j \tilde{\Omega}_{ij}\left(1 - \frac{1}{\rho^i_{jL}}\right)\left(\frac{\mathrm{d}\,w_i}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\,p_j}{\mathrm{d}\log A_k}\right) + \frac{\mathrm{d}\,p_i}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\,y_i}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\,\mu_i}{\mathrm{d}\log A_k}, \tag{36}$$

$$\frac{\mathrm{d}\log y_i}{\mathrm{d}\log A_k} = \frac{1}{\lambda_i}\frac{\mathrm{d}\,\lambda_i}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log P_c C}{\mathrm{d}\log A_k}, \tag{37}$$

$$\frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} = -\mathbf{1}(i=k) + \sum_j \tilde{\tilde{\Omega}}_{ij}\frac{\mathrm{d}\log p_j}{\mathrm{d}\log A_k} + (1 - \sum_j \tilde{\Omega}_{ij})\frac{\mathrm{d}\log w_i}{\mathrm{d}\log A_k}. \tag{38}$$

$$\frac{1}{\lambda_i}\frac{\mathrm{d}\,\lambda_i}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log P_c C}{\mathrm{d}\log A_k}, \tag{39}$$

The proof for the case with markups is very similar. ■

Proof of Proposition 2. This is a special case of Proposition 4. ■

Proof of Proposition 3. This is a special case of Proposition 5. ■

Proof of Proposition 4. Denote the $N \times F$ matrix corresponding to $\tilde{\Omega}_{if}$ by $\tilde{\alpha}_{if}$. By Shephard's lemma,

$$\frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} = -\mathbf{1}(i=k) + \sum_j \tilde{\Omega}_{ij}\frac{\mathrm{d}\log p_j}{\mathrm{d}\log A_k} + \sum_f \tilde{\alpha}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}. \tag{40}$$

Invert this system to get

$$\frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} = -\tilde{\Psi}e_k + \tilde{\Psi}_f\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k}, \tag{41}$$

where $\tilde{\Psi}_f = (I-\tilde{\Omega})^{-1}\tilde{\alpha}$ is $\tilde{\Lambda}$.

Now consider a factor L, we have

$$\begin{aligned}\frac{\mathrm{d}\,\Lambda_L}{\mathrm{d}\log A_k} =& \sum_i b_i(1-\theta_0)[-\tilde{\Psi}_{ik} + \sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}]\Psi_{iL},\\ &+ \sum_j (1-\theta_j)\lambda_j\mu_j^{-1}\sum_i \tilde{\Omega}_{ji}[-\tilde{\Psi}_{ik} + \sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k} + \tilde{\Psi}_{jk} - \sum_f \tilde{\Psi}_{jf}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}]\Psi_{iL}\\ &+ (\theta_k-1)\lambda_k\mu_k^{-1}\sum_i \tilde{\Omega}_{ki}\Psi_{iL}.\end{aligned}$$

Simplify this to

$$\begin{aligned}\mathrm{d}\,\Lambda_L =& (\theta_0-1)\left(\sum_i b_i\tilde{\Psi}_{ik}\Psi_{iL} - \sum_i b_i\Psi_{iL}\sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right),\\ &+ \sum_j (\theta_j-1)\lambda_j\mu_j^{-1}\left[\sum_i \tilde{\Omega}_{ji}\tilde{\Psi}_{ik}\Psi_{iL} - \sum_i \tilde{\Omega}_{ji}\tilde{\Psi}_{jk}\Psi_{iL}\right]\\ &+ \sum_j (1-\theta_j)\lambda_j\mu_j^{-1}\left[\sum_i \tilde{\Omega}_{ji}\sum_f \left(\tilde{\Psi}_{if} - \tilde{\Psi}_{jf}\right)\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\Psi_{iL}\right]\\ &+ (\theta_k-1)\lambda_k\mu_k^{-1}\sum_i \tilde{\Omega}_{ki}\Psi_{iL},\\ =& (\theta_0-1)\left(\sum_i b_i\tilde{\Psi}_{ik}\Psi_{iL} - \sum_i b_i\Psi_{iL}\sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right),\end{aligned}$$

$$
\begin{aligned}
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}\left[\sum_i\tilde{\Omega}_{ji}\tilde{\Psi}_{ik}\Psi_{iL}-\left(\sum_i\tilde{\Omega}_{ji}\Psi_{iL}\right)\left(\sum_i\tilde{\Omega}_{ji}\tilde{\Psi}_{ik}\right)\right]\\
&+\sum_j(1-\theta_j)\lambda_j\mu_j^{-1}\left[\sum_i\tilde{\Omega}_{ji}\sum_f\left(\tilde{\Psi}_{if}-\tilde{\Psi}_{jf}\right)\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\Psi_{iL}\right],\\
=&(\theta_0-1)\left(\sum_i b_i\tilde{\Psi}_{ik}\Psi_{iL}-\sum_i b_i\Psi_{iL}\sum_f\tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right),\\
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)},\Psi_{(L)})\\
&+\sum_j(1-\theta_j)\lambda_j\mu_j^{-1}\sum_f\left[\sum_i\tilde{\Omega}_{ji}\tilde{\Psi}_{if}\Psi_{iL}-\left(\sum_i\tilde{\Omega}_{ji}\Psi_{iL}\right)\left(\sum_i\tilde{\Omega}_{ji}\tilde{\Psi}_{if}\right)\right]\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\\
=&(\theta_0-1)\left(\sum_i b_i\tilde{\Psi}_{ik}\Psi_{iL}-\sum_i b_i\Psi_{iL}\sum_f\tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right),\\
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)},\Psi_{(L)})\\
&+\sum_j(1-\theta_j)\lambda_j\mu_j^{-1}\sum_f Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(f)},\Psi_{(L)})\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\\
=&(\theta_0-1)\left(\sum_i b_i\tilde{\Psi}_{ik}\Psi_{iL}-\sum_i b_i\Psi_{iL}\sum_f\tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right),\\
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)})\\
=&(\theta_0-1)\left(\sum_i b_i\Psi_{iL}\left(\tilde{\Psi}_{ik}-\sum_f\tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right)\right),\\
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)})\\
=&(\theta_0-1)Cov_b(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)})+(\theta_0-1)(\tilde{\lambda}_k-\sum_f\tilde{\Lambda}_f\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k})\lambda_L,\\
&+\sum_j(\theta_j-1)\lambda_j\mu_j^{-1}Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)}).
\end{aligned}
$$

Hence, for a productivity shock $\mathrm{d}\log A_k$, letting Λ_L be demand for factor L, and indexing

all factors by f, we have

$$\begin{aligned}\frac{\mathrm{d}\,\Lambda_L}{\mathrm{d}\log A_k} =&(\theta_0-1)Cov_b\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)}\right)\\&+\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}_j}\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k},\Psi_{(L)}\right)\\&+(\theta_0-1)\left(\tilde{\lambda}_k-\sum_f\tilde{\lambda}_f\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\right)\lambda_L.\end{aligned}\tag{42}$$

Combine this with the observation that

$$\Lambda_L(\mathrm{d}\log w_L+\mathrm{d}\log L_L-\mathrm{d}\log Y)=\mathrm{d}\,\Lambda_L.\tag{43}$$

and the fact that, from Theorem 1,

$$\mathrm{d}\log Y=\tilde{\lambda}_k+\sum_f\tilde{\lambda}_f\,\mathrm{d}\log\lambda_f.\tag{44}$$

Set $\mathrm{d}\,L_L=0$, since factors are inelastically supplied, and we have a linear system with $F+1$ equations and $F+1$ unknowns where F is the total number of factors. Substitute

$$\frac{1}{\Lambda_f}\frac{\mathrm{d}\,\Lambda_f}{\mathrm{d}\log A_k}+\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k}=\frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k}\tag{45}$$

back into network formula to get

$$\begin{aligned}\Lambda_L\,\mathrm{d}\log\Lambda_L =&(\theta_0-1)Cov_b\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k},\Psi_{(L)}\right)\\&+\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}_j}\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k},\Psi_{(L)}\right)\\&+(\theta_0-1)\left(\tilde{\lambda}_k-\sum_f\tilde{\lambda}_f\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k}-\mathrm{d}\log Y\right)\lambda_L.\end{aligned}\tag{46}$$

Use Theorem 1 to further simplify this

$$\Lambda_L\,\mathrm{d}\log\lambda_L =(\theta_0-1)Cov_b\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\,\mathrm{d}\log\lambda_f,\Psi_{(L)}\right)$$

$$+\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}-\sum_f\tilde{\Psi}_{(f)}\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k},\Psi_{(L)}\right). \tag{47}$$

Note that the final demand consumer needs to be labelled as producer 0 to complete the proof. ■

Proof of Example 3.1. We start with the matrices listed in (16) and (17), so that

$$\Gamma=\begin{pmatrix}(1-\theta_0)Cov_b(\widetilde{\Psi}_{(L)},\Psi_{(L)}) & (1-\theta_0)Cov_b(\widetilde{\Psi}_{(K)},\Psi_{(L)})\\ (1-\theta_0)Cov_b(\widetilde{\Psi}_{(K)},\Psi_{(L)}) & (1-\theta_0)Cov_b(\widetilde{\Psi}_{(K)},\Psi_{(K)})\end{pmatrix}. \tag{48}$$

$$\delta^{(i)}=\begin{pmatrix}(1-\theta_0)Cov_b(\widetilde{\Psi}_{(i)},\Psi_{(L)})\\ (1-\theta_0)Cov_b(\widetilde{\Psi}_{(i)},\Psi_{(K)})\end{pmatrix}. \tag{49}$$

From the structure of the problem, we can explicitly write the value of Γ as follows:

$$\Gamma=(1-\theta_0)\begin{pmatrix}b_3(b_1\mu_1^{-1}+b_2\mu_2^{-1}) & -b_3(b_1\mu_1^{-1}+b_2\mu_2^{-1})\\ -b_3(1-b_3)\mu_3^{-1} & b_3(1-b_3)\mu_3^{-1}\end{pmatrix}. \tag{50}$$

First we look at the cases for $i=1,2$ (the case where a factor of production is shared). We note that the case is symmetric for $i=1,2$ by the structure of the network and the problem. For these values, we have that

$$\delta^{(1)}=(1-\theta_0)\begin{pmatrix}b_1(\mu_1^{-1}-(b_1\mu_1^{-1}+b_2\mu_2^{-1}))\\ -b_1b_3\mu_3^{-1}\end{pmatrix}. \tag{51}$$

Generally (for all cases), in order to solve the system in equation (15), write

$$(\Lambda-\Gamma)=\begin{pmatrix}(b_1\mu_1^{-1}+b_2\mu_2^{-1})(1-(1-\theta_0)b_3) & (1-\theta_0)b_3(\mu_1^{-1}b_1+b_2\mu_2^{-1})\\ (1-\theta_0)b_3(1-b_3)\mu_3^{-1} & b_3\mu_3^{-1}(1-(1-\theta_0)(1-b_3))\end{pmatrix}. \tag{52}$$

Invert this to get

$$(\Lambda-\Gamma)^{-1}=\frac{1}{\det\Gamma}\begin{pmatrix}b_3\mu_3^{-1}(1-(1-\theta_0)(1-b_3)) & -(1-\theta_0)b_3(\mu_1^{-1}b_1+b_2\mu_2^{-1})\\ -(1-\theta_0)b_3(1-b_3)\mu_3^{-1} & (b_1\mu_1^{-1}+b_2\mu_2^{-1})(1-(1-\theta_0)b_3)\end{pmatrix}. \tag{53}$$

The determinant is

$$\begin{aligned}\det\Gamma=&(b_1\mu_1^{-1}+b_2\mu_2^{-1})(1-(1-\theta_0)b_3)b_3\mu_3^{-1}(1-(1-\theta_0)(1-b_3))-\\ &(1-\theta_0)^2b_3^2(1-b_3)\mu_3^{-1}(\mu_1^{-1}b_1+b_2\mu_2^{-1})),\end{aligned}$$

$$
\begin{aligned}
&= \Lambda_L\Lambda_K((1-(1-\theta_0)b_3)(1-(1-\theta_0)(1-b_3)) - (1-\theta_0)^2 b_3(1-b_3)) \\
&= \Lambda_L\Lambda_K(1-(1-\theta_0)b_3 - (1-\theta_0)(1-b_3) + (1-\theta_0)^2 b_3(1-b_3) - (1-\theta_0)^2 b_3(1-b_3)) \\
&= \Lambda_L\Lambda_K(1-(1-\theta_0)(b_3+1-b_3)) \\
&= \Lambda_L\Lambda_K\theta_0.
\end{aligned}
$$

Plug this back into (53) and simplify,

$$
(\Lambda-\Gamma)^{-1} = \frac{1}{\Lambda_L\Lambda_K\theta_0}\begin{pmatrix} \Lambda_K(1-(1-\theta_0)(1-b_3)) & -(1-\theta_0)b_3\Lambda_L \\ -\Lambda_K(1-\theta_0)(1-b_3) & \Lambda_L(1-(1-\theta_0)b_3) \end{pmatrix}. \tag{54}
$$

Returning to the specific case where $i = 1$,

$$
(\Lambda-\Gamma)^{-1}\delta^{(1)} = \begin{pmatrix} \frac{(1-(1-\theta_0)(1-b_3))}{\theta_0\Lambda_L} & \frac{(\theta_0-1)b_3}{\Lambda_K\theta_0} \\ -\frac{(1-\theta_0)(1-b_3)}{\theta_0\Lambda_L} & \frac{1-(1-\theta_0)b_3}{\Lambda_K\theta_0} \end{pmatrix}\begin{pmatrix} b_1(\theta_0-1)(\mu_1^{-1}-\Lambda_L) \\ -b_1(\theta_0-1)\Lambda_K \end{pmatrix}. \tag{55}
$$

Multiplying the values in (55), and using the identity in (15),

$$
\begin{aligned}
\mathrm{d}\log\Lambda =&(\Lambda-\Gamma)^{-1}\delta^{(1)}, \\
&= \frac{b_1(\theta_0-1)}{\theta_0}\begin{pmatrix} (\frac{\mu_1^{-1}}{\Lambda_L}-1)\theta_0 + b_3(1-\theta_0)\frac{\mu_1^{-1}}{\Lambda_L} + b_3(\theta_0-1) - b_3(\theta_0-1) \\ (\theta_0-1)(1-b_3)(\frac{\mu_1^{-1}}{\Lambda_L}-1) - (1-(1-\theta_0)b_3) \end{pmatrix}.
\end{aligned}
$$

Combine this with (2) gives

$$
\begin{aligned}
\mathrm{d}\log Y = b_1 &- (1-b_3)b_1\frac{\theta_0-1}{\theta_0}\Big(\Big[\frac{\mu_1^{-1}}{\Lambda_L}-1\Big]\theta_0 + b_3(1-\theta_0)\frac{\mu_1^{-1}}{\Lambda_L}\Big) \\
&- b_3b_1\frac{\theta_0-1}{\theta_0}\Big[(\theta_0-1)(1-b_3)\Big(\frac{\mu_1^{-1}}{\Lambda_L}-1\Big) - (1-(1-\theta_0)b_3),\Big].
\end{aligned}
$$

which further simplifies to

$$
\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_1} = b_1 + b_1(\theta_0-1)\Big[1-(1-b_3)\frac{\mu_1^{-1}}{\Lambda_L}\Big]. \tag{56}
$$

This gives the desired result for firm 1. Note, as mentioned before, a symmetric result holds for firm 2. In the case of firm 3, we have that

$$
\delta^{(3)} = (\theta_0-1)\begin{pmatrix} -b_3(b_1\mu_1^{-1} + b_2\mu_2^{-1}) \\ b_3\mu_3^{-1}(1-b_3) \end{pmatrix}. \tag{57}
$$

The results from (54) give the blueprint for solving for the value of $\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_3}$ as well. From this, we can conclude that

$$\frac{d\log Y}{d\log A_3} = b_3 - \begin{pmatrix} b_1+b_2 \\ b_3 \end{pmatrix}^T \begin{pmatrix} \frac{(1-(1-\theta_0)(1-b_3))}{\theta_0\Lambda_L} & \frac{(\theta_0-1)b_3}{\Lambda_K\theta_0} \\ -\frac{(1-\theta_0)(1-b_3)}{\theta_0\Lambda_L} & \frac{1-(1-\theta_0)b_3}{\Lambda_K\theta_0} \end{pmatrix} \begin{pmatrix} -b_3(b_1\mu_1^{-1}+b_2\mu_2^{-1}) \\ b_3\mu_3^{-1}(1-b_3) \end{pmatrix}, \tag{58}$$

which further simplifies to

$$= b_3 - \begin{pmatrix} 1-b_3 \\ b_3 \end{pmatrix}^T \begin{pmatrix} \frac{b_3(1-\theta_0)}{\theta_0}\Big[1-(1-\theta_0)(1-b_3)\Big] + \frac{(\theta_0-1)^2}{\theta_0}b_3(1-b_3) \\ -\frac{(\theta_0-1)^2}{\theta_0}b_3(1-b_3) + \Big[1-(1-\theta_0)(1-b_3)\Big]\frac{(1-b_3)(\theta_0-1)}{\theta_0} \end{pmatrix}. \tag{59}$$

This simplifies to give the required result

$$\frac{d\log Y}{d\log A_3} = b_3. \tag{60}$$

■

Proof of Proposition 5. From Theorem 1, we need only to characterize $\mathrm{d}\log\Lambda/\mathrm{d}\log\mu_k$. For a markup shock, we can differentiate demand for a quantity m to get

$$\frac{\mathrm{d}\,\lambda_m}{\mathrm{d}\log\mu_k} = \sum_i \left(b_i(1-\theta_0)(\mathrm{d}\log p_i) + \sum_j \tilde{\Omega}_{ji}\mu_j^{-1}(1-\theta_j)\lambda_j[\mathrm{d}\log p_i - \mathrm{d}\log p_j] - \mathbf{1}(j=k)\theta_k \right)\Psi_{im}. \tag{61}$$

By Shephard's lemma,

$$\mathrm{d}\log p_i = \mathbf{1}(i=k) + \sum_j \tilde{\Omega}_{ij}\,\mathrm{d}\log p_j + \sum_f \tilde{\alpha}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log\mu_k}. \tag{62}$$

Invert this system to get

$$\mathrm{d}\log p_i = \tilde{\Psi}e_k + \tilde{\Psi}_f\,\mathrm{d}\log w, \tag{63}$$

where $\tilde{\Psi}_f = (I-\tilde{\Omega})^{-1}\tilde{\alpha}$ is a $N\times K$ matrix of network-adjusted factor intensities by industry.

Substituting this back into (61) and set $m = L$ to get

$$\begin{aligned}\frac{\mathrm{d}\,\Lambda_L}{\mathrm{d}\log\mu_k} &= \sum_i b_i(1-\theta_0)[\tilde{\Psi}_{ik} + \sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log\mu_k}]\Psi_{iL}, \\ &\quad + \sum_j (1-\theta_j)\lambda_j\mu_j^{-1}\sum_i \tilde{\Omega}_{ji}[\tilde{\Psi}_{ik} + \sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log\mu_k} - \tilde{\Psi}_{jk} - \sum_f \tilde{\Psi}_{jf}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log\mu_k}]\Psi_{iL}\end{aligned}$$

$$- \theta_k \mu_k^{-1} \lambda_k \sum_i \tilde{\Omega}_{ki} \Psi_{iL}.$$

Simplify this to

$$\begin{aligned}
\mathrm{d}\,\Lambda_L =& (1-\theta_0)\left(\sum_i b_i \tilde{\Psi}_{ik}\Psi_{iL}\right) + (1-\theta_0)\left(\sum_i b_i \Psi_{iL} \sum_f \tilde{\Psi}_{if} \frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\right), \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left[\sum_i \tilde{\Omega}_{ji}\tilde{\Psi}_{ik}\Psi_{iL} - \sum_i \tilde{\Omega}_{ji}\Psi_{iL}\tilde{\Psi}_{jk}\right] \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left[\sum_i \tilde{\Omega}_{ji}\sum_f \tilde{\Psi}_{if}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\Psi_{iL} - \sum_i \tilde{\Omega}_{ji}\Psi_{iL}\sum_f \tilde{\Psi}_{jf}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\right] \\
&- \theta_k \lambda_k \mu_k^{-1}\Psi_{kL}, \\
=& (1-\theta_0)\left(Cov_b(\tilde{\Psi}_{(k)}, \Psi_{(L)}) + \tilde{\lambda}_k \Lambda_L\right) \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}, \Psi_{(L)}) - \mathbf{1}(j=k)\Psi_{jL}\right) \\
&+ (1-\theta_0)\left(Cov_b\left(\sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)}\right) + \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\Lambda_L\right) \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}\left(\sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)}\right) - \mathbf{1}(j=k)\Psi_{jL}\right) \\
&- \theta_k \lambda_k \mu_k^{-1}\Psi_{kL}, \\
=& (1-\theta_0)\left(Cov_b(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)})\right) \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)}) - \mathbf{1}(j=k)\Psi_{jL}\right) \\
&+ (1-\theta_0)\left(\tilde{\lambda}_k + \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\right)\Lambda_L - \lambda_k \mu_k^{-1}\Psi_{kL}, \\
=& (1-\theta_0)\left(Cov_b(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)})\right) \\
&+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}, \Psi_{(L)})\right) \\
&- \lambda_k \Psi_{kL}.
\end{aligned}$$

The final line follows from the fact that $\left(\tilde{\lambda}_k + \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log w_f}{\mathrm{d}\log \mu_k}\right) = b' \,\mathrm{d}\log p = \mathrm{d}\log P_c = 0$. Finally, substitute $\mathrm{d}\log w_f = \mathrm{d}\log \Lambda_f + \mathrm{d}\log Y$ into the expression above to get

$$\begin{aligned}\frac{\mathrm{d}\,\Lambda_L}{\mathrm{d}\log \mu_k} =& (1-\theta_0)\left(Cov_b(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}(\mathrm{d}\log\Lambda_f + \mathrm{d}\log Y), \Psi_{(L)})\right) \\ &+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}(\mathrm{d}\log\Lambda_f + \mathrm{d}\log Y), \Psi_{(L)})\right) \\ &- \lambda_k \Psi_{kL}.\end{aligned}$$

To complete the proof, note that

$$\mathrm{d}\log Y \sum_f \tilde{\Psi}_{(f)} = \mathrm{d}\log Y \mathbf{1}. \tag{64}$$

In other words, this is a vector of all ones multiplied by the scalar $\mathrm{d}\log Y$, and hence it drops out of the covariance operators, since the covariance of a vector of ones with any other vector under any probability distribution is always equal to zero. Hence,

$$\begin{aligned}\mathrm{d}\,\Lambda_L =& (1-\theta_0)\left(Cov_b(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\,\mathrm{d}\log\Lambda_f, \Psi_{(L)})\right) \\ &+ \sum_j (1-\theta_j)\lambda_j \mu_j^{-1}\left(Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} + \sum_f \tilde{\Psi}_{(f)}\,\mathrm{d}\log\Lambda_f, \Psi_{(L)})\right) \\ &- \lambda_k \Psi_{kL}.\end{aligned}$$

■

C Robustness and Extensions

In this section, we discuss some of the extensions mentioned in the body of the paper. Specifically, we address in more detail how are results extend to situations with elastic factors, capital accumulation/dynamics, fixed costs, entry, and nonlinearities. Proofs for the results are at the end of this section.

C.1 Elastic Factor Supplies

To be able to do counterfactuals, we provide the structural counterparts to these equations in terms of microeconomic elasticities of substitution. These are the formulas that we use to study the impact of monetary policy in models with nominal rigidities (where the endogenous supply of labor is important) in Appendix D, or the impact of changes in capital taxes (where the endogenous supply of capital is important).

Assume a nested CES structure similar to the one in Section 3. We have

$$d\log L_f = \zeta_f d\log(\frac{W_f}{P}) + \gamma_f d\log Y. \tag{65}$$

This implies that

$$d\log(\frac{W_f}{P}) = \frac{1}{1+\zeta_f} d\log\Lambda_f + \frac{1-\gamma_f}{1+\zeta_f} d\log Y, \tag{66}$$

$$d\log(\frac{W_f}{PY}) = \frac{1}{1+\zeta_f} d\log\Lambda_f - \frac{\gamma_f+\zeta_f}{1+\zeta_f} d\log Y, \tag{67}$$

and

$$d\log L_f = \frac{\zeta_f}{1+\zeta_f} d\log\Lambda_f + \frac{\gamma_f+\zeta_f}{1+\zeta_f} d\log Y. \tag{68}$$

Changes in factor shares and output solve the following system of equations:[40]

$$\begin{aligned} d\log\Lambda_f = &-\sum_k \lambda_k \frac{\Psi_{kf}}{\Lambda_f} d\log\mu_k + \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\Big(\sum_k \tilde{\Psi}_{(k)} d\log A_k - \sum_k \tilde{\Psi}_{(k)} d\log\mu_k, \frac{\Psi_{(f)}}{\Lambda_f}\Big) \\ &- \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\Big(\sum_g \tilde{\Psi}_{(g)} \frac{1}{1+\zeta_g} d\log\Lambda_g - \sum_g \tilde{\Psi}_{(g)} \frac{\gamma_g+\zeta_g}{1+\zeta_g} d\log Y, \frac{\Psi_{(f)}}{\Lambda_f}\Big), \end{aligned}$$

$$d\log Y = \frac{1}{\sum_f \tilde{\Lambda}_f \frac{1-\gamma_f}{1+\zeta_f}} \left[\sum_k \tilde{\lambda}_k d\log A_k - \sum_k \tilde{\lambda}_k d\log\mu_k - \sum_f \tilde{\Lambda}_f \frac{1}{1+\zeta_f} d\log\Lambda_f\right]. \tag{69}$$

[40] Note that the first equation can also be written as

$$\begin{aligned} d\log\Lambda_f = &-\sum_k \lambda_k \frac{\Psi_{kf}}{\Lambda_f} d\log\mu_k + \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\Big(\sum_k \tilde{\Psi}_{(k)} d\log A_k - \sum_k \tilde{\Psi}_{(k)} d\log\mu_k, \frac{\Psi_{(f)}}{\Lambda_f}\Big) \\ &- \sum_j (\theta_j - 1)\frac{\lambda_j}{\mu_j} Cov_{\tilde{\Omega}^{(j)}}\Big(\sum_g \tilde{\Psi}_{(g)} \frac{1}{1+\zeta_g} d\log\Lambda_g + \sum_g \tilde{\Psi}_{(g)} \frac{1-\gamma_g}{1+\zeta_g} d\log Y, \frac{\Psi_{(f)}}{\Lambda_f}\Big). \end{aligned}$$

C.2 Capital Accumulation, Adjustment Costs, and Capacity Utilization

In mapping this set-up to the data, there are two ways to interpret this model: either we could interpret final demand as a per-period part of a larger dynamic problem, or we could interpret final demand as an intertemporal consumption function where goods are also indexed by time à la Arrow-Debreu. When we interpret the model intertemporally, then output is the net present-value of consumption streams. When we interpret the model intratemporally, the output function encompasses demand for consumption goods and for investment goods, and treats the two as perfect substitutes. Our results actually generalize to the case where they are not, see Section 6 for more details. The existence of a constant-returns-to-scale aggregate final demand function allows us to unambiguously define real GDP using the corresponding ideal price index.[41]

In principle, we could also model factor accumulation in the usual Arrow-Debreu manner: treat goods in different time periods as different goods. Then, we could model the process of capital accumulation via intertemporal production functions that transform goods in one period into goods in other periods. This modeling choice would also be well-suited to handle technological frictions to the reallocation of factors such as adjustment costs and variable capacity utilization. Our formulas would apply to these economies without change, but of course, in such a world, the Domar weight of each producer would now be expressed in net-present value terms.

C.3 Fixed Costs

We start by adding fixed costs to the model, and then expand to allow for free entry. To add fixed overhead costs to the model, we need to separate variable cost from total cost. Under these conditions $\tilde{\Omega}$ becomes total-variable-cost based rather than total-cost based. It is the matrix whose ijth element is

$$\tilde{\Omega}_{ij} = \frac{\mathrm{d}\log \mathbf{C}_i}{\mathrm{d}\log p_j} = \frac{p_j x_{ij}}{VC_i}, \tag{70}$$

where VC_i is i's total variable cost rather than i's total costs. Then we have

$$\mathrm{d}\log Y = \tilde{\lambda}' \,\mathrm{d}\log A - \tilde{\lambda}\,\mathrm{d}\log \mu + \mathrm{d}\,H(\tilde{\Lambda}, \Lambda). \tag{71}$$

[41] We assume the existence of a representative consumer mostly for expositional convenience. Our ex-post reduced-form results could be generalized to cover the generic heterogenous consumers (as long as real GDP is defined using the Laspeyre index), and our ex-ante structural results could be generalized to cover heterogenous consumers with identical homothetic preferences.

where Λ is the share of income going to each factor including the payments to the infra-marginal fixed costs, but the cost-based Domar weights $\tilde{\lambda}$ and $\tilde{\Lambda}$ are constructed using the variable-cost based $\tilde{\Omega}$.[42]

If firms are earning positive economic profits, we might expect that this would induce entry and competition. In this section, we extend our basic results to cover the case where there is free entry. First, we establish that our results can easily be applied to cases where entry is "wasteful." Typically, entry can be economically meaningful due to several reasons: (1) it increases product variety; (2) it reduces markups; (3) it selects the most productive firms; (4) it counters decreasing returns to scale at the firm level. If we assume that firms have constant-returns-to-scale production functions, are ex-ante identical, markups/wedges are exogenous, and there is no returns to product variety, then entry is entirely socially wasteful.[43] Under these conditions, our results survive unchanged. Now we discuss how our results could be extended to cover (1), (2), (3), and (4).

C.4 Entry

First, we provide a simple case where we can micro-found our industry-level model with constant-returns industry cost functions using a model that has entry and decreasing returns to scale a the firm level. Here, channels (1), (2), and (3) are still shut off, but (4) is operating. Under these conditions, our results for the impact of productivity shocks survive unchanged. Finally, we sketch how a model which potentially allows for all four channels outlined above behaves. Here, our results connect to those of Baqaee (2016), who studies network economies with free entry and external economies of scale.

[42] With fixed costs, the interpretation of allocative efficiency as the *gap* between the passive allocation and the general equilibrium allocation still applies, but the passive allocation is no longer locally the same as general equilibrium when there are no frictions. The reason is that the passive allocation would send resources to the users of the fixed cost, even though the marginal benefit is zero. However, we can change the definition of the passive allocation so that it nets out fixed costs first (and hence becomes equivalent to general equilibrium when there are no frictions). Under this modification, the change in allocative efficiency is given by

$$\mathrm{d}\log Y^{p} - \mathrm{d}\log Y = (\hat{\lambda} - \tilde{\lambda})' \,\mathrm{d}\log A - \tilde{\lambda}' \,\mathrm{d}\log \mu + \mathrm{d}\, H(\tilde{\Lambda}, \Lambda), \tag{72}$$

where $\hat{\lambda} = b'(I - \varphi^{-1}\tilde{\Omega})^{-1}$ and φ is a diagonal matrix whose iith element is the share industry i's output that is not used for fixed costs in the initial allocation. If the equilibrium is efficient, then (72) is always zero. For reference, $\tilde{\lambda} = b'(I - \tilde{\Omega})^{-1}$ and $\lambda = b'(I - \Omega)^{-1}$.

[43] A constrained social planner would drive the mass of entrants in each industry to zero.

No External Economies

We start with the case where the entry margin has no effect on the marginal productivity of the industry so that mechanisms (1), (2), (3) and (4) are shut off.

Let industry k output be given by

$$y_k = A_k \left(M_k^{-1/\varepsilon_k} \int_{M_k} y(i,k)^{\frac{\varepsilon_k - 1}{\varepsilon_k}} \, \mathrm{d}\, i \right)^{\frac{\varepsilon_k}{\varepsilon_k - 1}},$$

and suppose the variable cost function of each product i in industry k is given by

$$\frac{1}{A_k z_k(i)} c(w,p) y(i,k),$$

so that firms have constant returns on the margin, and A_k is an industry level TFP. To enter, firms pay a fixed entry cost. After entry, each firm draws an idiosyncratic markup m_i and productivity z_i from some distribution $\Phi(z,m)$. There is also an industry level markup μ_k. The scaling term M_k^{-1/ε_k} is introduced to neutralize love for variety effects and thereby ensure that there are no external economies of scale.

Proposition 6 (No External Economies)**.** *Suppose that there is free entry subject to fixed costs, that there are no external economies of scale, and that firms have constant returns to scale. Then*

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_k} = \tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d} \log A_k},$$

and

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log \mu_k} = -\tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d} \log \mu_k}.$$

This set up allows us to span the basic framework in, for example, Autor et al. (2017), who argue that the decline in the labor share in the US in recent times is due to an increase in the size of low-labor-share firms.[44]. They consider a model with entry and where labor intensity falls with the scale of operation because of overhead labor costs. The mechanism Autor et al. (2017) emphasize is an increase in the industry-level elasticity of substitution, whereby more productive firms are able to capture more demand over time.[45] Our interpretation, consistent with their empirical evidence, is that low-labor share firms charge higher markups, and that these markups are part of the reason why they have a

[44]The same facts have also been documented by Vincent and Kehrig (2017) and Hartman-Glaser et al. (2016)

[45]Hartman-Glaser et al. (2016) offer a different explanation based on implicit contracts between firms and workers

low labor share. Our results in Section 5 indicate that these changes in the composition of firms within industries have increased aggregate productivity by improving allocative efficiency.

Constant External Economies

While the no-external-economies model described above allows us to accommodate entry, it does so within a very restrictive set up. Now, we sketch a version of the model where, due to decreasing returns to scale at the firm-level, free entry is not socially wasteful. We turn on mechanism (4) but keep (1), (2), and (3) off.

Although individual firms have decreasing returns to scale, industries have constant returns to scale. This simple model can microfound the use of constant-returns-to-scale industry level cost functions, and thereby allows our results to go through unchanged for productivity shocks, as long as we assume that each entrant in industry k has access to a homothetic production function, that all producers charge the same markup, and that overhead costs are paid in units of the industry good.

Proposition 7 (Constant External Economies)**.** *Suppose that there is free entry subject to fixed entry costs. In addition, suppose that entrants in each industry charge the same markup and have access to the same homothetic production function, and each entrant pays a fixed cost of entry in units of the industry's input. Then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log A_k}.$$

This extends our results for productivity shocks to an economy with entry. Unfortunately, the results for markup/wedge shocks do not apply any longer since a change in markups changes the scale of operations of firms, and these changes have associated efficiency changes that cannot be tracked in the same way.

Increasing External Economies

Finally, we sketch how our framework would relate to models with richer heterogeneity and entry properties by turning on mechanisms (1), (3), and (4).[46] In particular, we allow for the possibility that entry can induce increasing returns to scale at the industry level,

[46] We could also allow for mechanism (2) with endogenous markups in a model with Cournot competition, or with demand curves with non-constant elasticities. The formula would feature extra terms having to do with the elasticity of the markups to the shocks.

where the industry becomes more productive as more firms enter.[47]

To deal with the possibility of decreasing returns to scale at the firm-level, we introduce "fictitious" fixed factors, and assume that each entrant in an industry may use a fixed factor. Hence, firm i in industry k has a variable cost function that can be written as

$$\frac{1}{A_k z_k(i)} c_k(w, p, r_k(i)) y_{ik},$$

where A_k is industry TFP, $z(i)$ is individual TFP, and $r_k(i)$ is the wage paid to the fixed factor. We also allow for fixed overhead costs on top of the entry costs, which firms pay after observing the realization of their markup and productivity if they decide to be active (see below), thereby creating room for selection effects.

Index producers in industry k by i in such a way that their idiosyncratic productivity $z_k(i)$ is weakly increasing, and suppose that i is distributed according to the distribution $\phi_k(i)$. We assume that the price of the composite industry k good can be written as

$$p_k = \frac{\mu_k}{A_k} \left(\int_{\underline{i}_k}^{\infty} \left(\frac{\mu_k(z)}{z_k(i)} c_k(w, p, r_k(z)) \right)^{1-\varepsilon_k} \phi_k(i) \, \mathrm{d}\, i \right)^{\frac{1}{1-\varepsilon_k}},$$

where $\underline{i}_k$ denotes the cutoff below which firms that have paid the entry cost decide not to be active in order not to pay the fixed overhead cost. Movements in this cutoff drive the strength of selection effects and the external economies of scale arising from love for variety effects.

Proposition 8 (Increasing External Economies)**.** *Suppose that there is free entry, fixed entry costs, and fixed overhead costs paid after productivity and markup draws conditional on operating, and external economies of scale in the form of love for variety effects. Then*

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_k} = \tilde{\lambda}_k - \int_f \tilde{\Lambda}_f \frac{\mathrm{d} \log \Lambda_f}{\mathrm{d} \log A_k} - \sum_j \frac{\tilde{\lambda}_j}{\varepsilon_j - 1} \frac{p_j(\underline{i}_j)}{p_j} \phi_j(\underline{i}^j) \frac{\mathrm{d} \log \underline{i}_j}{\mathrm{d} \log A_k},$$

and

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log \mu_k} = -\tilde{\lambda}_k - \int_f \tilde{\Lambda}_f \frac{\mathrm{d} \log \Lambda_f}{\mathrm{d} \log \mu_k} - \sum_j \frac{\tilde{\lambda}_j}{\varepsilon_j - 1} \frac{p_j(\underline{i}_j)}{p_j} \phi_j(\underline{i}^j) \frac{\mathrm{d} \log \underline{i}_j}{\mathrm{d} \log \mu_k},$$

where f indexes the set of all factors including the "fictitious" fixed factors.

The model of Baqaee (2016) becomes a special case of this setup. Proposition 8 shows

[47] Baqaee (2016) shows that, even within the confines of a tightly parameterized model, these forces can significantly alter the propagation and amplification of shocks.

that by incorporating "fictitious" factors, and hence separating profits due to distortions from competitive rents from decreasing returns, we can extend our results to a much more general class of models going even beyond neoclassical production by allowing for increasing returns to scale at the macroeconomic level.

Importantly, allowing for increasing returns makes the "pure" technology effect of a shock endogenous, and this endogenous response of productivity is what the final (and new) term in Proposition 8 accounts for. There are two reasons why technology becomes endogenous: first, an increase in the mass of entrants improves the productivity of firms through its effect on product variety. Second, an increase in the mass of entrants can change the distribution of productivity in each industry, since the change in cutoff value at which firms enter the market changes. Interestingly, we see that these cut-off values for the productivity and their derivatives are sufficient statistics for understanding the first-order impact of the shocks. Analyzing this model in any more detail and expressing these cutoff values and their derivatives as a function of the structural microeconomic parameters of the model is well beyond the scope of this paper, but we pursue it in ongoing work.

C.5 Nonlinear Impact of Shocks

Another limitation of our results is that we neglect nonlinearities. As discussed by Baqaee and Farhi (2017), models with production networks can respond very nonlinearly to productivity shocks. We plan to extend these results to inefficient economies in full generality, but as a first step, we stipulate some conditions under which we can directly leverage these results to inefficient economies. In particular, we show that the amplification of negative shocks due to complementarities emphasized in Baqaee and Farhi (2017) can also work to amplify the negative effects of misallocation.

Consider the quantitative parametric model in Section 5. Let $\delta_k(i)$, $\mu_k(i)$, and $A_k(i)$ denote firm i in industry k's share of industry sales, markups, and productivity. Define

$$\mu_k = \left(\sum_i \frac{\delta_k(i)}{\mu_k(i)}\right)^{-1}$$

and

$$A_k = \mu_k \left(\sum_i \delta_k(i)(\mu_k(i)/A_k(i))^{1-\xi_k}\right)^{\frac{1}{\xi_k - 1}}.$$

Define the *efficiency* of each firm i in industry k to be $e_k(i) = 1/\mu_k(i)$. Consider a transformation $e_k(i) = t_k + (1 - t_k)\overline{e_k(i)}$ which shrinks dispersion in markups relative to its

steady-state value $\overline{\mu_k(i)} = 1/\overline{e_k(i)}$. This transformation keeps $\mu_k = 1/e_k$ constant. Define the revenue-based Domar weight of industry k by λ_k.

Proposition 9 (Competitive Isomorphism)**.** *Consider an economy where $\mu_k = 1$ for every k, then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log t_k} = \lambda_k \frac{\mathrm{d}\log A_k}{\mathrm{d}\log t_k} \tag{73}$$

and

$$\frac{\mathrm{d}^2\log Y}{\mathrm{d}\log t_k^2} = \lambda_k \frac{\mathrm{d}\log \lambda_k}{\mathrm{d}\log A_k}\left(\frac{\mathrm{d}\log A_k}{\mathrm{d}\log t_k}\right)^2 + \lambda_k \frac{\mathrm{d}^2\log A_k}{\mathrm{d}\log t_k^2}, \tag{74}$$

with $\mathrm{d}\log A_k/\mathrm{d}\log t_k \geq 0$.[48]

Hence, increases in the dispersion of markups, which keep the harmonic average of markups constant, are isomorphic to negative productivity shocks in a model which is efficient at the industry level. Hence, shocks which increase markup dispersion in an industry can have outsized nonlinear effects on output, if those industries are macro-complementary with other industries in the sense defined by Baqaee and Farhi (2017) so that $\mathrm{d}\log\lambda_k/\mathrm{d}\log A_k < 0$.

This helps flesh out the insight in Jones (2011) that complementarities can interact with distortions to generate large reductions in output, and that these can be quantitatively important enough to explain the large differences in cross-country incomes. Given the examples in Baqaee and Farhi (2017), it should be clear how misallocation in a key industry like energy production can significantly reduce output through macro-complementarities. Investigating these nonlinear forces more systematically is an interesting exercise that we leave for future work.

Proofs of this section

Proof of Proposition 6. The industry level price is then given by

$$\begin{aligned} p_k &= \left(\frac{1}{M_k}\int_{M_k}\left(c(w,p)\frac{\mu_k m_i}{A_k z_i}\right)^{1-\varepsilon_k}\Phi(z_i,m_i)\,\mathrm{d}\,m_i\,\mathrm{d}\,z_i\right)^{\frac{1}{1-\varepsilon_k}}, \\ &= c(w,p)\frac{\mu_k}{A_k}E\left(\left(\frac{m_i}{z_i}\right)^{1-\varepsilon_k}\right)^{\frac{1}{1-\varepsilon_k}}. \end{aligned}$$

[48] We typically also have $\mathrm{d}^2\log A_k/\mathrm{d}\log t_k^2 \geq 0$.

By Shephard's lemma, we know that,

$$\frac{\mathrm{d}\log p_i}{\mathrm{d}\log A_k} = -\mathbf{1}(i=k) + \sum_F \tilde{\tilde{\alpha}}_{iF}\frac{\mathrm{d}\log w_F}{\mathrm{d}\log A_k} + \sum_j \tilde{\Omega}_{ij}\frac{\mathrm{d}\log p_j}{\mathrm{d}\log A_k}, \tag{75}$$

where $\tilde{\alpha}_{iF}$ and $\tilde{\Omega}_{ij}$ are firm-level expenditures on factor F and industry j as a share of costs (excluding entry costs). Rewrite this in matrix form to get

$$\frac{\mathrm{d}\log p}{\mathrm{d}\log A_k} = (I-\tilde{\Omega})^{-1}(\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} - e_k) = \tilde{\Psi}(\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} - e_k), \tag{76}$$

where e_k is the kth standard basis vector. Let the household's aggregate consumption good to be the numeraire, so that the household's ideal price index P_c is always equal to one. Then we know that

$$\frac{\mathrm{d}\log P_c}{\mathrm{d}\log A_k} = b'\frac{\mathrm{d}\log p}{\mathrm{d}\log A_k} = 0. \tag{77}$$

Combine this with the previous expression to get

$$-b'\tilde{\Psi}e_k + b'\tilde{\Psi}\tilde{\alpha}\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} = 0. \tag{78}$$

Note that, $b'\tilde{\Psi} = \tilde{\lambda}$ and $b'\tilde{\Psi}\tilde{\alpha} = \tilde{\Lambda}$. Hence,

$$-\tilde{\lambda}_k + \tilde{\Lambda}'\frac{\mathrm{d}\log w}{\mathrm{d}\log A_k} = 0. \tag{79}$$

Now, note that

$$\Lambda_f = \frac{w_f L_F}{P_c C}. \tag{80}$$

From this, we know that

$$\frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log A_k} = \frac{\mathrm{d}\log w_f}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log L_f}{\mathrm{d}\log A_k} - \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k}. \tag{81}$$

Substitute this into the previous expression to get

$$-\tilde{\lambda}_k + \tilde{\Lambda}'\frac{\mathrm{d}\log \Lambda}{\mathrm{d}\log A_k} - \tilde{\Lambda}'\frac{\mathrm{d}\log L_f}{\mathrm{d}\log A_k} + \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = 0, \tag{82}$$

where we use the fact that $\sum_f \tilde{\Lambda}_f = 1$. ■

Proof of Proposition 7. Let the cost function of entrant i in industry k producing y_i units be

given by

$$c(w,p)h(y_i/A), \tag{83}$$

where A is industry level shock and k subscripts have been suppressed. Free entry then implies that

$$\mu c(w,p)h'(y/A)\frac{y}{A} - c(w,p)h(y/A) = \overline{f}c(w,p), \tag{84}$$

where $\overline{f}_k$ is the entry cost in units of the input good. We can simplify this to

$$\mu h'(y/A)\frac{y}{A} - h(y/A) = \overline{f}. \tag{85}$$

This equation pins down the efficient scale of operation $y = Ay^*(\overline{f},\mu)$.

The cost of producing q units of industry k output is then given by

$$nc(w,p)f(y^*(\mu,\overline{f})) + n\overline{f}c(w,p), \tag{86}$$

such that $ny = q$. Substitute the constraint into this to get

$$\mathfrak{C}(w,p)\frac{q}{A} = \frac{q}{y^*A}c(w,p)f(y^*(\mu,\overline{f})) + \frac{q}{Ay^*}\overline{f}c(w,p). \tag{87}$$

Hence the industry's cost function is linear q/A as needed.

Finally,

$$\frac{\partial\mathfrak{C}(w,p)q/A}{\partial w_j} = \frac{q}{y^*A}\frac{\partial c(w,p)}{\partial w_j}f(y^*(\mu,\overline{f})) + \frac{q}{Ay^*}\overline{f}\frac{\partial c(w,p)}{\partial w_j} = nl_k + n\overline{f}l_k^f. \tag{88}$$

Therefore, the industry level cost function obeys Shephard's Lemma and we can replicate the rest of the proof from Proposition 6 ■

Proof of Proposition 8.

$$\begin{aligned}\mathrm{d}\log p_k = {} & -\mathbf{1}(i=k)\,\mathrm{d}\log A_i + \sum_f\left(\int_{\overline{z}_k}^{\infty} s_k(z)\tilde{\alpha}_{kf}(z)\,\mathrm{d}F(z)\right)\mathrm{d}\log w_f \\ & + \sum_j\left(\int_{\overline{z}_k}^{\infty} s_k(z)\tilde{\omega}_{kj}(z)\,\mathrm{d}F(z)\right)\mathrm{d}\log p_j + \left(\int_{\overline{z}_k}^{\infty} s_k(z)\tilde{\beta}_k(z)\,\mathrm{d}\log r_k(z)\,\mathrm{d}F(z)\right) \\ & - \frac{1}{1-\varepsilon_k}s_k(\overline{z})f_k(\overline{z}_k)\,\mathrm{d}\log\overline{z}_k.\end{aligned}$$

where $s_k(z)$ is firm i's share of sales in industry k, and $\beta_k(z)$ is the cost share of firm i on its

fictitious fixed factor. We can solve this system to get

$$\mathrm{d}\log p = (I - \tilde{\Omega})^{-1}\left(-e_i\,\mathrm{d}\log A_i + \tilde{\alpha}\,\mathrm{d}\log w + \xi - \kappa\right), \tag{89}$$

where $\xi_k = \left(\int_{\overline{z}_k}^{\infty} s_k(z)\tilde{\beta}_k(z)\,\mathrm{d}\log r_k(z)\,\mathrm{d}F(z)\right)$ and $\kappa_k = \frac{1}{1-\varepsilon_k} s_k(\overline{z}) f_k(\overline{z}_k)\,\mathrm{d}\log \overline{z}_k$, and the elements of $\tilde{\Omega}$ and $\tilde{\alpha}$ are defined appropriately. Let output be the numeraire so that

$$b'\,\mathrm{d}\log p = 0. \tag{90}$$

Hence,

$$-\tilde{\lambda}_i + \tilde{\Lambda}'\,\mathrm{d}\log w + \sum_k \tilde{\lambda}_k \xi_k - \sum_k \tilde{\lambda}_k \kappa_k = 0. \tag{91}$$

For each factor, we know that

$$\mathrm{d}\log w = \mathrm{d}\log \Lambda + \mathrm{d}\log Y. \tag{92}$$

Substitute this in to get
$-\tilde{\lambda}_i + \tilde{\Lambda}'\,\mathrm{d}\log\Lambda + \tilde{\Lambda}'\,\mathrm{d}\log Y + \sum_k \int_{\overline{z}_k}^{\infty} \tilde{\lambda}_k s_k(z)\tilde{\beta}_k(z)\,\mathrm{d}\log\Lambda_{r_k(z)}\,\mathrm{d}F(z) + \sum_k \int_{\overline{z}_k}^{\infty} \tilde{\lambda}_k s_k(z)\tilde{\beta}_k(z)\,\mathrm{d}\log Y - \sum_k \tilde{\lambda}_k\kappa_k = 0.$
Finally, observe that $\sum_f \tilde{\Lambda}_f + \sum_k \int_{\overline{z}_k}^{\infty} \tilde{\lambda}_k s_k(z)\tilde{\beta}_k(z) = 1$.

The proof for markup shocks is very similar.

■

Proof of Proposition 9. On the other hand,

$$\frac{\mathrm{d}\log A}{\mathrm{d}\tau} = \sum_i s_i\left(\frac{1-\bar{e}_i}{\bar{e}_i}\right) = \sum_i s_i\left(\mu_i - 1\right) = \sum_i s_i\mu_i - 1 \geq 0. \tag{93}$$

On the other hand,

$$\frac{\mathrm{d}^2\log A}{\mathrm{d}\tau^2} = (\sigma-1)Var_s(\mu_i) - \sum_i s_i\left(\mu_i - 1\right)^2 = (\sigma-2)Var_s(\mu_i) - \left(\sum_i s_i\mu_i - 1\right)^2. \tag{94}$$

Consider an industry where: all firms i use the same upstream input bundle with cost C; firms transform this input into a firm-specific variety of output using constant return to returns to scale technology; each firm i has productivity a_i and charges a markup μ_i; the varieties are combined into a composite good by a competitive downstream industry according to a CES production function with elasticity σ on firm i.

We denote the quantity of composite good produced as

$$Q = \left[\sum b_i^{\frac{1}{\sigma}} p_i^{\frac{\sigma-1}{\sigma}}\right]^{\frac{\sigma}{\sigma-1}}. \tag{95}$$

Firm i charges a price

$$p_i = \frac{\mu_i}{a_i} C. \tag{96}$$

The resulting demand for firm i's variety is

$$q_i = (\frac{p_i}{P})^{-\sigma} b_i Q, \tag{97}$$

where the price index is given by

$$P = \left[\sum b_i p_i^{1-\sigma}\right]^{\frac{1}{1-\sigma}}. \tag{98}$$

Total profits are given by

$$\Pi = \sum_i (p_i - C)(\frac{p_i}{P})^{-\sigma} b_i Q. \tag{99}$$

We solve out the price index and profits explicitly and get

$$P = \left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} C, \tag{100}$$

$$\Pi = \sum_i \left(\frac{\mu_i}{a_i} - \frac{1}{a_i}\right) \left[\frac{\frac{\mu_i}{a_i}}{\left[\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}}\right]^{-\sigma} b_i C Q. \tag{101}$$

For completeness we can also solve for the sales of each firm as a fraction of the sales of the industry

$$\lambda_i = \frac{p_i q_i}{PQ} = \frac{b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}}{\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}}. \tag{102}$$

We want to understand how to aggregate this industry into homogenous industry with productivity A and markup μ. These variables must satisfy

$$P = \frac{\mu}{A} C, \tag{103}$$

$$\Pi = \left(\frac{\mu}{A} - \frac{1}{A}\right) CQ. \tag{104}$$

This implies that A and μ are the solutions of the following system of equations

$$\frac{\mu}{A} = \left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}, \tag{105}$$

$$\left(\frac{\mu}{A} - \frac{1}{A}\right) = \sum_i \left(\frac{\mu_i}{a_i} - \frac{1}{a_i}\right) \left[\frac{\frac{\mu_i}{a_i}}{\left[\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}}\right]^{-\sigma} b_i. \tag{106}$$

The solution is

$$A = \frac{1}{\left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} - \sum_i \left(1 - \frac{1}{\mu_i}\right) \left[\frac{\frac{\mu_i}{a_i}}{\left[\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}}\right]^{-\sigma} \frac{\mu_i}{a_i} b_i}, \tag{107}$$

$$\mu = \frac{\left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}}{\left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}} - \sum_i \left(1 - \frac{1}{\mu_i}\right) \left[\frac{\frac{\mu_i}{a_i}}{\left[\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}}\right]^{-\sigma} \frac{\mu_i}{a_i} b_i}. \tag{108}$$

We can also rewrite this in a useful way as

$$A = \frac{1}{\left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}} \frac{1}{\sum_i \frac{1}{\mu_i} \frac{\left(\frac{\mu_i}{a_i}\right)^{1-\sigma} b_i}{\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}}} = \frac{1}{\left[\sum_i b_i \left(\frac{\mu_i}{a_i}\right)^{1-\sigma}\right]^{\frac{1}{1-\sigma}}} \frac{1}{\sum_i \frac{1}{\mu_i} \lambda_i}, \tag{109}$$

$$\mu = \frac{1}{\sum_i \frac{1}{\mu_i} \frac{\left(\frac{\mu_i}{a_i}\right)^{1-\sigma} b_i}{\sum_j b_j \left(\frac{\mu_j}{a_j}\right)^{1-\sigma}}} = \frac{1}{\sum_i \frac{1}{\mu_i} \lambda_i}. \tag{110}$$

Define the efficiency of each firm i to be $e_i = \mu_i^{-1}$. Then, consider a steady state where $\sum_i s_i \mu_i^{-1} = \sum_i s_i \bar{e}_i = 1$. Consider a transformation $e_i = \tau + (1 - \tau)\bar{e}_i$. This transformation

keeps μ constant. On the other hand,

$$\frac{\mathrm{d}\log A}{\mathrm{d}\tau} = \sum_i s_i \left(\frac{1-\bar{e}_i}{\bar{e}_i}\right) = \sum_i s_i \left(\mu_i - 1\right) = \sum_i s_i \mu_i - 1 \geq 0. \tag{111}$$

On the other hand,

$$\frac{\mathrm{d}^2\log A}{\mathrm{d}\tau^2} = (\sigma-1)Var_s(\mu_i) - \sum_i s_i \left(\mu_i - 1\right)^2 = (\sigma-2)Var_s(\mu_i) - \left(\sum_i s_i \mu_i - 1\right)^2. \tag{112}$$

■

D Nominal Rigidities

In this Section, we apply our general framework to study the effects of sticky prices in economies with arbitrary production structures.[49] In general, sticky prices can be modeled via variable markups: markups which move to ensure that the relevant nominal prices stay constant. This is the point of connection with our framework. We show how to solve for these endogenous markups, and then to trace their impact on the economy.

In this section, we have two goals: first, we show how the existence of nominal rigidities changes the mapping from microeconomic productivity shocks to aggregate output or TFP in economies with distorted steady states; second, we show how monetary policy shocks can be analyzed using our results in economies with distorted steady states, leading to a clean separation the oft-neglected effects of monetary policy shocks on allocative efficiency from their traditional aggregate demand effects. In these applications, the steady-state distortions are the estimated markups discussed above in a given year. The endogenous response of markups to shocks is solved for to ensure that the relevant prices remain fixed.

These exercises are useful demonstrations of how to apply our results more generally in cases where markups are endogenous or variable. Typically, models with sticky prices are linearized around an efficient steady state, which ensures that reallocation terms disappear. We use our framework to study the model's behavior away from the efficient steady-state using empirically estimated steady-state markups, and with a realistic microeconomic production structure featuring input-output connections, complementarities in production, and substitutability among heterogenous firms within an industry.

[49] Starting with Basu (1995), a literature has grown to emphasize the importance of intermediate goods for understanding the business cycle properties of models with sticky prices. See for example Bouakez et al. (2009), Nakamura and Steinsson (2010), Pasten et al. (2016, 2017).

The size and direction of the TFP effects of monetary shocks depend crucially on the correlation pattern between price-stickiness and the level of markups. They can be large and positive if goods with higher markups have stickier prices. Our results draw attention to the fact that the typical loglinearization of New Keynesian models around undistorted steady states is potentially misleading since the allocative efficiency effects can be large away from the efficient steady state.

To model money demand we use the simplest formulation and assume that there is a cash in advance constraint

$$P_y Y = M,$$

where M is the instrument of monetary policy.

We index each individual producer by i, and write a firm-level input-output matrix in standard form. To model sticky prices, let s denote the set of producers with fixed prices, and let e_s be the $N \times |s|$ matrix given by $e_s = [e_i]_{\{i \in s\}}$, where e_i is the ith standard basis vector. Using the firm-level formulation, we can solve for the change in markups $\mathrm{d}\log\mu$ that would keep the price of sticky-firms constant, in response to the vector of productivity shocks $\mathrm{d}\log A$ the vector of changes in factor prices $\mathrm{d}\log w$:

$$\mathrm{d}\log\mu = (e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\,\mathrm{d}\log w). \tag{113}$$

To solve for how the change in factor prices, we use

$$\mathrm{d}\log w = \mathrm{d}\log\Lambda + \mathrm{d}\log M - \mathrm{d}\log L.$$

Combining these two equations characterizes how output responds to productivity or money shocks in general equilibrium

$$\mathrm{d}\log Y - \tilde{\Lambda}'\,\mathrm{d}\log L = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'e_s\,\mathrm{d}\log\mu + \mathrm{d}\,H(\tilde{\Lambda},\Lambda), \tag{114}$$

where $\mathrm{d}\log\mu$ is now determined according to (113).[50]

To finish our characterization, we need to make an assumption about labor supply, since equation (114) takes the change in factor supply as given. To fix ideas, we follow convention in the New Keynesian literature, and assume labor is the only factor of production and that the household utility function takes the form:

$$U(C,L) = \log(C) - \frac{L^{1+1/\nu}}{1+1/\nu},$$

[50]See Proposition 12 in the Appendix for a formal statement and proof of these results.

where ν is the Frisch elasticity of labor supply. Under these conditions, we can combine equations (114), (18), and (19) to explicitly solve out the labor supply decision.

Proposition 10 (Nominal Rigidities)**.** *Suppose that labor is the only factor of production. Then*

$$\mathrm{d}\log Y = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}' e_s\,\mathrm{d}\log\mu - \left(\frac{1}{1+\nu}\right)\tilde{\Lambda}\,\mathrm{d}\log\Lambda,$$

with

$$\mathrm{d}\log\mu = (e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\left(\mathrm{d}\log M - \left(\frac{1}{1+\nu}\right)\mathrm{d}\log\Lambda\right),$$

where Λ is the labor share of income. If the economy has a nested CES form, then

$$\mathrm{d}\log\Lambda = \sum_{k=1}^{N}\left(\sum_{j=0}^{N}(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\Omega^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}\right)\right)[\mathrm{d}\log A_k - \mathrm{d}\log\mu_k] - \sum_{k=1}^{N}\frac{\lambda_k\Psi_{kL}}{\Lambda_L}\,\mathrm{d}\log\mu_k.$$

Equation (114) gives the change in aggregate TFP in response to either monetary or technology shocks in a New Keynesian type environment and decomposes it into a "pure" change in technology component and change in allocative efficiency component. Proposition 10 characterizes the corresponding changes in output.

Typically, New Keynesian models are log-linearized around the efficient steady state, and in those cases, the changes in allocative efficiency are second-order and are therefore neglected.[51] In these cases, our results deliver the same conclusion.[52] Outside of these special cases, for inefficient steady states, changes in allocative efficiency are not zero, and our results then permit us to isolate these effects.

In their important study, Pasten et al. (2016) characterize the response of output to shocks in a model with Calvo frictions and production networks. They write the input-output matrix at the industry level, and suppose that some fraction δ_i of firms in industry i have flexible prices. Their sharpest analytical result is for the case with log utility in

[51] See, for example, Galí (2008).

[52] The case of the efficient steady-state is immediate in Proposition 10, since at the efficient steady-state $\Psi_{(L)}$ is a vector of all ones, and the covariances are all zero. Hence, at the efficient steady state, Proposition 10 implies that the change in allocative efficiency, to a first order, is

$$-\tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}\,\mathrm{d}\log\Lambda = -\sum_{k=1}^{N}\lambda_k\,\mathrm{d}\log\mu_k + \sum_{k=1}^{N}\frac{\lambda_k\Psi_{kL}}{\Lambda_L}\,\mathrm{d}\log\mu_k = -\sum_{k=1}^{N}\lambda_k\,\mathrm{d}\log\mu_k + \sum_{k}\lambda_k\,\mathrm{d}\log\mu_k = 0.$$

Of course, for a monopolistic economy, if the production network is irregular (or asymmetric), then the equilibrium is generically inefficient (due to the heterogeneity in markups implied by double marginalization) even if every producer charges the same markup.

consumption and an infinite Frisch elasticity of labor supply. In this special case, the response of output to shocks takes a very simple form — sticky prices act like shock absorbers to productivity shocks. Using Proposition 10, we can recover their result and shed light on why this happens.

Proposition 11. *[Pasten et al. 2016] Suppose that utility is log in consumption and the Frisch elasticity of labor supply is infinite. Then Proposition 10 implies*

$$\mathrm{d}\log Y = \tilde{\lambda}'\left(I - e_s(e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}\right)\mathrm{d}\log A + \tilde{\lambda}'e_s(e_s'\tilde{\Psi}e_s)^{-1}e_s'\mathbf{1}\,\mathrm{d}\log M.$$

In the special case where some fraction δ_i in industry i are flexible, then

$$\mathrm{d}\log Y = \mathrm{d}\log M - b'(I - \delta\tilde{\Omega})^{-1}\delta(\tilde{\alpha}\,\mathrm{d}\log M - \mathrm{d}\log A).$$

Proposition 11 is simple to interpret: for productivity shocks, the impact of a shock is the same as one in a model with flexible prices, but where productivity shocks affect *only* some fraction δ_i of each industry i's costs, or in other words, productivity shocks are attenuated by some weight δ_i at each industry. The impact of monetary policy shocks on output is given by $1 - b'(I - \delta\tilde{\Omega})^{-1}\delta\tilde{\alpha}$, or the total share of value-added which is sticky in the economy.

Crucially, information about elasticities of substitution and changes in allocative efficiency disappear from these calculations. This is due to assumption of infinitely elastic labor supply. In this model, labor supply moves exactly in such a way as to offset changes in allocative efficiency, so that output fluctuations boil down to only how the productivity shocks travel from suppliers into consumer prices. Hence, although output responses can easily be determined without information on elasticities of substitution, allocative efficiency *is* changing in this environment, and is given by

$$\mathrm{d}\log Y - \mathrm{d}\log L = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'e_s\,\mathrm{d}\log\mu - \mathrm{d}\log\Lambda,$$

which is nonzero.[53]

We now turn our attention to an application of our results. We calibrate a version of our quantitative model from Section 5, but augmented with a labor-leisure choice, and a Frisch elasticity of labor supply of $\nu = 1/2$, which is broadly consistent with the recommendation of Chetty et al. (2011). We create two copies of each firm in our sample,

[53]The fact that changes in allocative efficiency are not required to compute the changes in output is a generic property of infinitely elastic labor supply, and does not depend on parametric assumptions about the production functions, see footnote (39) for more details.

one copy has sticky prices while the other has flexible prices. We then use Proposition 10 to compute the impact of monetary policy shocks and firm-level productivity shocks.

Monetary Policy Shocks

In Table 4, we show the response of output and aggregate TFP for a shock to the money supply. We consider different specifications of the model with different elasticities of substitution. For each specification, we also consider some speculative scenarios where the flexible firms have slightly higher or lower markups than the sticky firms (keeping the average constant).

We find that the movements in aggregate TFP, which are purely caused by changes in allocative efficiency, can become very large if the elasticity of substitution across firms is high, and if average markups are not the same between the flexible and sticky firms. The size and sign of these movements depend crucially on the correlation between price rigidity and markups. These results suggest that empirical work on understanding the correlation between microeconomic price rigidities and the levels of markups could be of great importance.[54] They also suggest that the literature on the New Keynesian model, by assuming that the steady-state is efficient, and by assuming away correlations between price rigidities and levels of markups, could potentially be missing important first-order effects.[55]

Comparing the benchmark model to the one-sector model, we also recover the famous insight by Basu (1995) that intermediate goods can increase stickiness. If intermediate inputs are sticky, then flexible firms adjust their prices less in response to shocks. The degree of amplification caused by the intermediate-input share is hump-shaped in the fraction of firms δ that have sticky prices. In the limit, as all firms become sticky, the intermediate input share becomes irrelevant, and the same occurs when all firms become flexible. Here we show that these effects are stronger when the fraction of sticky prices is 20% (roughly corresponding to a horizon of 5 quarters) than if it is 50% (roughly

[54]The sign of this correlation is not ex-ante obvious. In models where the price elasticity of demand is not constant, the pass-through of costs to markups can depend on the level of the markup, so that the desired markups of firms with high markups are less sensitive to changes in costs. In the presence of price-adjustment costs, this means that high markup firms will have stickier prices (see Gopinath and Itskhoki, 2011, 2010; Atkeson and Burstein, 2008; Kimball, 1995). On the other hand, in a Calvo model where the markups are uncorrelated with stickiness on impact, in response to an expansion in the money supply, firms that do not adjust their markups for longer will over time have lower effective markups, inducing a negative correlation between stickiness and markups. Studying these sorts of effects requires a dynamic model however, and we leave this for future work.

[55]Typically, second-order effects on allocative efficiency are taken into account only in the computation of welfare, but not in the computation of the equilibrium allocation.

$\delta = 0.5$	Benchmark	CD + CES	CD + CD	One Sector
Uncorrelated	(0.154, 0.005)	(0.164, 0.017)	(0.164, 0.017)	(0.125, 0.000)
Sticky High Markup	(0.307, 0.204)	(0.319, 0.219)	(0.181, 0.039)	(0.250, 0.167)
Flex High Markup	(0.034,-0.153)	(0.041,-0.143)	(0.147,-0.006)	(0.023,-0.136)
$\delta = 0.2$	Benchmark	CD + CES	CD + CD	One Sector
Uncorrelated	(0.091, 0.003)	(0.0963, 0.012)	(0.096, 0.012)	(0.058, 0.000)
Sticky High Markup	(0.220, 0.182)	(0.223, 0.186)	(0.106, 0.024)	(0.142, 0.118)
Flex High Markup	(-0.015.,-0.143)	(-0.008,-0.133)	(0.086,-0.003)	(-0.017,-0.108)

Table 4: The elasticity of output and aggregate TFP with respect to monetary policy shocks ($\mathrm{d}\log Y/\mathrm{d}\log M, \mathrm{d}\log TFP/\mathrm{d}\log M$) for two different Calvo parameters δ. The parameter δ is the fraction of each industry with sticky prices. We show the results for the benchmark model, a Cobb-Douglas specification that keeps $\xi = 8$, but sets all other elasticities of substitution to one (CD + CES), a Cobb-Douglas specification that sets all elasticities equal to one (CD + CD), and a single-industry model with a value-added production function. The first row is where the sticky and flexible firms are identical. The second row is where the sticky firms have markups that are 5% larger than the industry average and the flexible firms have markups that are 5% smaller than the industry average. The third row reverses this, and has sticky firms with 5% lower and flexible firms with 5% higher markups than the industry average.

corresponding to a horizon of 2 quarters).[56]

Productivity Shocks

In general, we can decompose the effect on output, from a productivity shock, into three components: (1) the "pure" technology effect captured by the cost-based Domar weight, (2) changes in allocative efficiency, and (3) changes in quantity of factors supplied. In the context of a model with sticky prices, the second effect is subtle, since markups adjust in response to a productivity shock to ensure that sticky prices do not adjust. In Figure 10, we plot the allocative efficiency component of the response relative to the total output

[56]Relatedly, Nakamura and Steinsson (2010) and Pasten et al. (2016) have emphasized that heterogeneity in the frequency of price changes across industries is also quantitatively important. This is mostly because, even in the basic New Keynesian model with a trivial input-output structure, the mapping between the frequency of price changes and the degree of monetary non-neutrality is convex, and so for a given average frequency of price changes, increasing dispersion in the frequency of price changes increases monetary non-neutrality. This is an important dimension of heterogeneity that, for now, we abstract away from in our quantitative examples.

response for firm-level productivity shocks, as a histogram. We color the flexible and sticky priced firms differently. We see that changes in allocative efficiency, which are typically neglected by the positive literature on nominal rigidities, are sizable as a fraction of the total output response. Furthermore, we find that the sign on this term changes. For flexible firms, the "pure" technology impact of the productivity shock is amplified by improvements in allocative efficiency, while for sticky priced firms, the opposite is true.

In Figure 11, we plot the histogram of the output responses to firm-level productivity shocks relative to the sales shares of the affected firms. Since the sales share gives the output response in the competitive model, we can think of this as measure the degree of amplification or attenuation of productivity shocks relative to a model with perfectly flexible firms and perfect competition. We see that the impact of shocks is more attenuated when the firm is sticky than when it is flexible, but the effect is not necessarily zero, even when the firm has sticky prices. This is because of the upstream reallocation that happens in the case when a firm with sticky prices receives a productivity shock. On the whole, the model with nominal rigidities attenuates shocks relative to the competitive flexible model (most of the mass in the histogram is to the left of 1), but this is by no means universal, and there are some firms for whom shocks are amplified.

Finally, in Figure 12, we plot the response of the shock on output when the firm is sticky relative to when the firm is flexible. Since all the mass it to the left of 1, we can conclude that the flexible firm always affects output more than the sticky firm, but of course, the degree of attenuation is highly dispersed.

Proofs of this section

Proof of Proposition 12. Order the producers so that the first s producers are the ones with sticky prices. For a vector x, denote $x_{(s)} = e_s' x$. From the cash in advance constraint, we know that

$$\begin{aligned}
\mathrm{d}\log Y &= \mathrm{d}\log M - \mathrm{d}\log P_c, \\
&= \mathrm{d}\log M - b'\,\mathrm{d}\log p, \\
&= \mathrm{d}\log M - b'\tilde{\Psi}\left(\tilde{\alpha}\,\mathrm{d}\log w - \mathrm{d}\log A\right) + b'\tilde{\Psi}e_s\,\mathrm{d}\log\mu, \\
&= \mathrm{d}\log M - \tilde{\Lambda}'\,\mathrm{d}\log w + \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}e_s\,\mathrm{d}\log\mu, \\
&= \mathrm{d}\log M - \tilde{\Lambda}'\left(\mathrm{d}\log\Lambda + \mathrm{d}\log M - \mathrm{d}\log L\right) + \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}e_s\,\mathrm{d}\log\mu, \\
&= \mathrm{d}\log M - \tilde{\Lambda}'\left(\mathrm{d}\log\Lambda - \mathrm{d}\log L\right) - \mathrm{d}\log M + \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}e_s\,\mathrm{d}\log\mu,
\end{aligned}$$

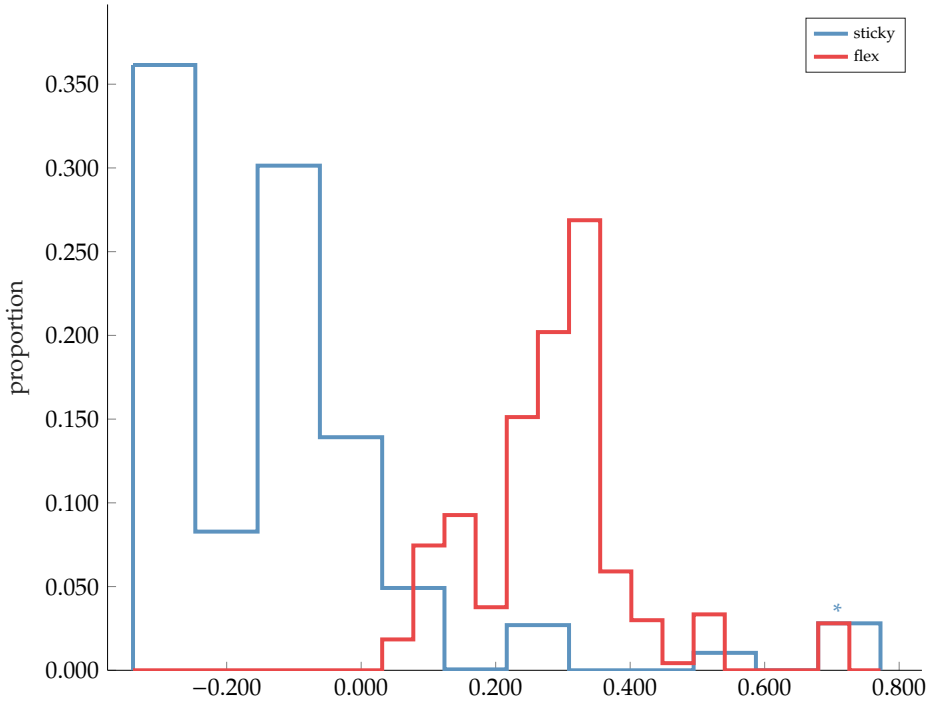


Figure 10: The change in allocative efficiency relative to the change in output for productivity shocks to the firms in Compustat with sticky and flexible prices. We set $\nu = 1/2$, and use GP markups.

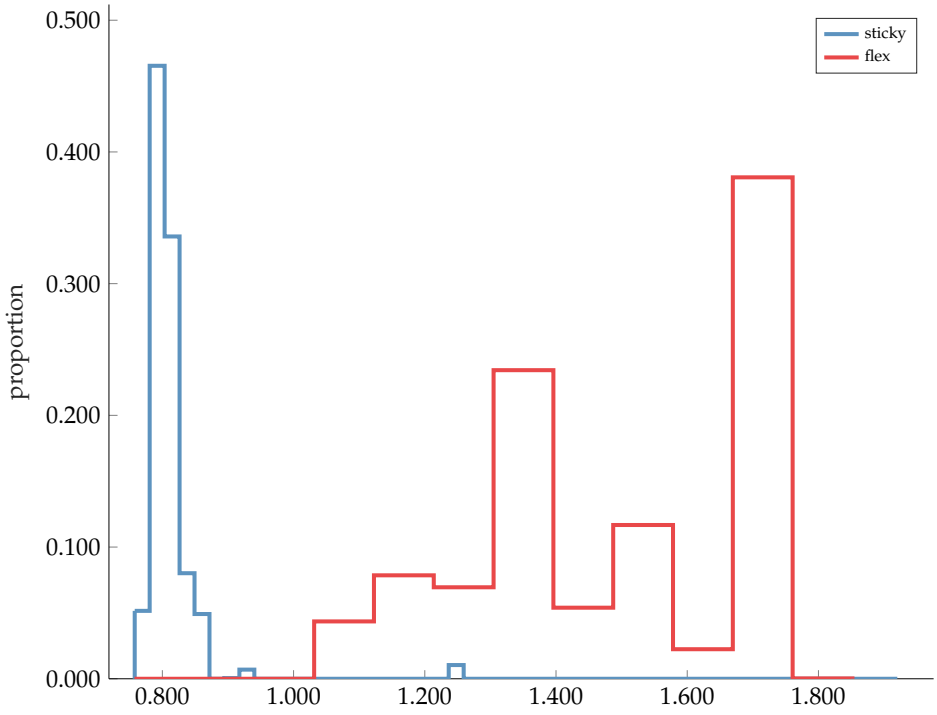


Figure 11: Response of output for productivity shocks to the sticky and flexible firms in Compustat relative to the response in a fully competitive model. We set $\nu = 1/2$, and use GP markups.

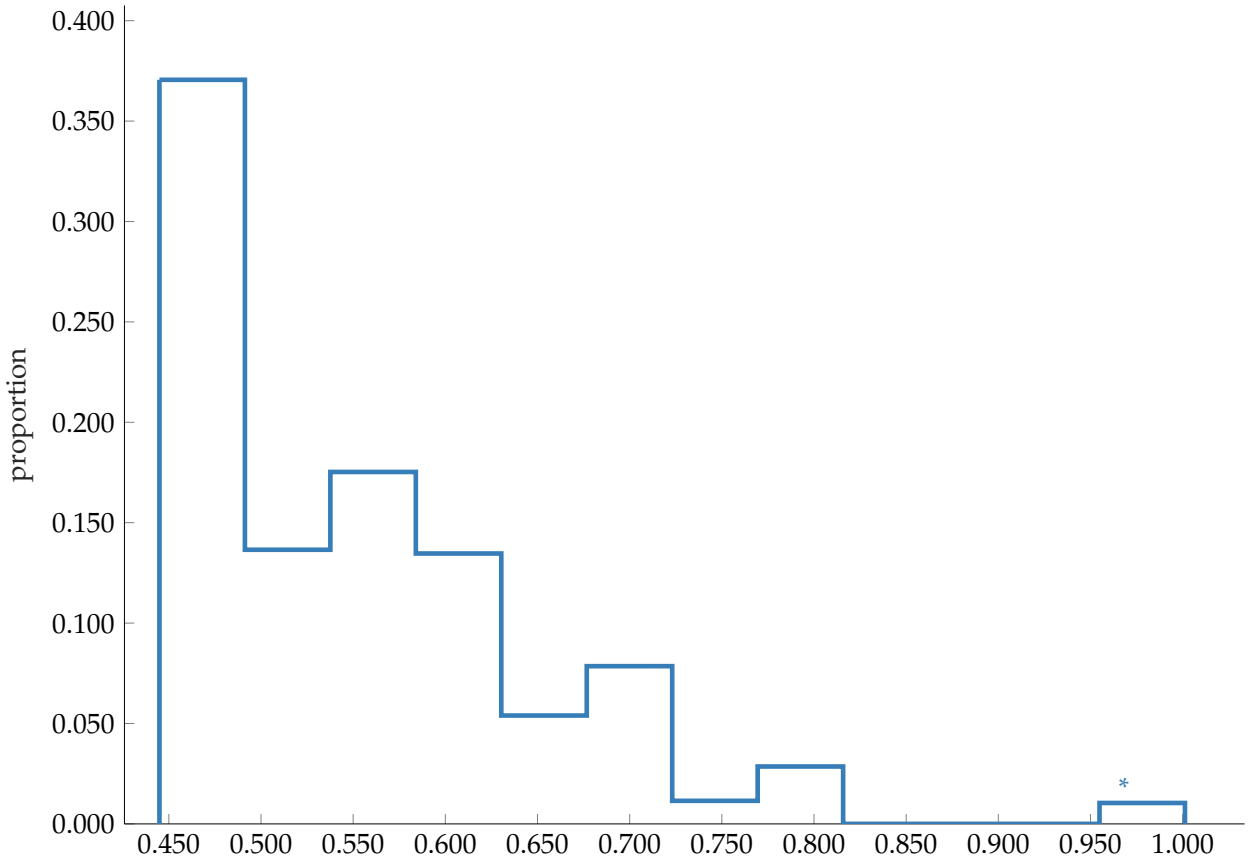


Figure 12: Response of output relative to a firm-level productivity shock for the sticky firm relative to the flexible firm. We set $\nu = 1/2$, and use GP markups.

which implies that

$$\mathrm{d}\log Y - \tilde{\Lambda}'\,\mathrm{d}\log L = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'_{(s)}\,\mathrm{d}\log\mu - \tilde{\Lambda}'\,\mathrm{d}\log\Lambda. \tag{115}$$

To get the markups necessary to keep $p_{(s)}$ sticky, we impose

$$\mathrm{d}\log p_{(s)} = \mathrm{d}\log\mu + e_s'\tilde{\Omega}\log p - \mathrm{d}\log A_{(s)} = 0. \tag{116}$$

This implies

$$\mathrm{d}\log\mu = -e_s'\tilde{\Omega}\,\mathrm{d}\log p - \tilde{\alpha}_{(s)}\,\mathrm{d}\log w + \mathrm{d}\log A_{(s)}. \tag{117}$$

On the other hand, we have

$$\mathrm{d}\log p = \tilde{\Psi}\left(\tilde{\alpha}\,\mathrm{d}\log w - \mathrm{d}\log A\right) + \tilde{\Psi}e_s\,\mathrm{d}\log\mu. \tag{118}$$

Substituting this back into the previous expression gives

$$\mathrm{d}\log\mu = -e_s'\tilde{\Omega}\tilde{\Psi}\tilde{\alpha}\,\mathrm{d}\log w - e_s'\tilde{\Omega}\tilde{\Psi}e_s\,\mathrm{d}\log\mu + e_s'\tilde{\Omega}\tilde{\Psi}\,\mathrm{d}\log A - e_s'\tilde{\alpha}\,\mathrm{d}\log w + e_s'\,\mathrm{d}\log A. \tag{119}$$

Solve this to get

$$\begin{aligned}\mathrm{d}\log\mu &= (e_s'(I+\tilde{\Omega}\tilde{\Psi})e_s)^{-1}e_s'(I+\tilde{\Omega}\tilde{\Psi})\left(\mathrm{d}\log A - \mathrm{d}\log w\right),\\ &= (e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}\left(\mathrm{d}\log A - \mathrm{d}\log w\right).\end{aligned}$$

■

To prove Proposition 10, we first prove the following:

Proposition 12. *Consider an economy with a cash-in-advance constraint, and nominal rigidities. Then,*

$$\mathrm{d}\log Y - \tilde{\Lambda}'\,\mathrm{d}\log L = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'e_s\,\mathrm{d}\log\mu + \mathrm{d}\,H(\tilde{\Lambda},\Lambda),$$

where

$$\mathrm{d}\log\mu = (e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\,\mathrm{d}\log w), \tag{120}$$

and

$$\mathrm{d}\log w = \mathrm{d}\log\Lambda + \mathrm{d}\log M - \mathrm{d}\log L.$$

In the special case where some fraction δ_i in industry i are flexible. Then,

$$\tilde{\lambda}'_{(s)}\,\mathrm{d}\log\mu = \left(b'(I-\tilde{\Omega})^{-1} - b'(I-\delta\tilde{\Omega})^{-1}\delta\right)\mathrm{d}\log A - (1 - b'(I-\delta\Omega)^{-1}\delta\alpha)\,\mathrm{d}\log w, \tag{121}$$

$$= \left(\tilde{\lambda} - b'(I - \delta\tilde{\Omega})^{-1}\delta\right) \mathrm{d}\log A - (1 - b'(I - \delta\Omega)^{-1}\delta\alpha)\,\mathrm{d}\log w.$$

Proof of Proposition 10. The labor-leisure condition and the cash-in-advance condition imply that

$$L^{1/(\nu)} = \left(\frac{w}{P_c C}\right) = \left(\frac{w}{M}\right). \tag{122}$$

Hence,

$$\frac{1}{\nu}\,\mathrm{d}\log L = \mathrm{d}\log w - \mathrm{d}\log M = \mathrm{d}\log \Lambda - \mathrm{d}\log L + \mathrm{d}\log M - \mathrm{d}\log M, \tag{123}$$

or

$$\mathrm{d}\log L = \frac{\nu}{1+\nu}\,\mathrm{d}\log \Lambda. \tag{124}$$

Therefore,

$$\mathrm{d}\log w = \frac{1}{\nu+1}\,\mathrm{d}\log \Lambda + \mathrm{d}\log M \tag{125}$$

To finish, apply Propositions 2 and 3. ■

Proof of Proposition 11. With log utility in consumption and infinite Frisch elasticity of labor supply we have that

$$\mathrm{d}\log w - \mathrm{d}\log P_c = \mathrm{d}\log Y, \tag{126}$$

or in other words, substituting in the cash in advance constraint

$$\mathrm{d}\log w = \mathrm{d}\log M. \tag{127}$$

Furthermore, the cash in advance constraint implies that

$$\begin{aligned}
\mathrm{d}\log Y &= \mathrm{d}\log M - \mathrm{d}\log P_c, \\
&= \mathrm{d}\log M - \tilde{\Lambda}'\,\mathrm{d}\log w + \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}' e_s\,\mathrm{d}\log \mu, \\
&= \mathrm{d}\log M - \mathrm{d}\log M + \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}' e_s\,\mathrm{d}\log \mu, \\
&= \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}' e_s\,\mathrm{d}\log \mu,
\end{aligned}$$

substituting equation (120) from Proposition 12 completes the proof

$$\begin{aligned}
&= \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}_{(s)}(e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\,\mathrm{d}\log w), \\
&= \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}_{(s)}(e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\,\mathrm{d}\log M).
\end{aligned}$$

In the special case where some fraction δ_i in industry i are flexible. Then,

$$\begin{aligned}\tilde{\lambda}'_{(s)}\,\mathrm{d}\log\mu &= \left(b'(I-\tilde{\Omega})^{-1} - b'(I-\delta\tilde{\Omega})^{-1}\delta\right)\mathrm{d}\log A - (1 - b'(I-\delta\Omega)^{-1}\delta\alpha)\,\mathrm{d}\log w, \qquad (128)\\ &= \left(\tilde{\lambda} - b'(I-\delta\tilde{\Omega})^{-1}\delta\right)\mathrm{d}\log A - (1 - b'(I-\delta\Omega)^{-1}\delta\alpha)\,\mathrm{d}\log w.\end{aligned}$$

At the industry level, equation (128) shows that the changes in markups can be interpreted *as if* some fraction of the firms in each industry change their markup in response to shocks. ■

E Relabelling

As mentioned previously, our results can be applied to any nested CES economy, with any arbitrary pattern of nested substitutabilities and complementarities among intermediate inputs and factors. For concreteness, we describe the relabeling for one specific example. Let the household's consumption aggregator be

$$\frac{C}{\overline{C}} = \left(\sum_k b_k \left(\frac{c_k}{\overline{c}_k}\right)^{\frac{\sigma-1}{\sigma}}\right)^{\frac{\sigma}{\sigma-1}}. \qquad (129)$$

Suppose each producer k produces using a CES aggregator of value-added VA and intermediate inputs X:

$$\frac{y_k}{\overline{y}_k} = A_k\left(\alpha_k\left(\frac{VA_k}{\overline{VA}_k}\right)^{\frac{\theta_k-1}{\theta_k}} + (1-\alpha_k)\left(\frac{X_k}{\overline{X}_k}\right)^{\frac{\theta_k-1}{\theta_k}}\right)^{\frac{\theta_k}{\theta_k-1}}.$$

Value-added consists of different factors

$$\frac{VA_k}{\overline{VA}_k} = \left(\sum_{j=1}^{F} \nu_{kj}\left(\frac{l_{kj}}{\overline{l}_{kj}}\right)^{\frac{\eta_k-1}{\eta_k}}\right)^{\frac{\eta_k}{\eta_k-1}},$$

where l_{kj} is factor of type j used by k. Intermediate input consists of inputs from other producers

$$\frac{X_k}{\overline{X}_k} = \left(\sum_{j=1}^{N} \omega_{kj}\left(\frac{x_{kj}}{\overline{x}_{jk}}\right)^{\frac{\varepsilon_k-1}{\varepsilon_k}}\right)^{\frac{\varepsilon_k}{\varepsilon_k-1}}.$$

Denote the matrix of ν_{kj} using $\mathcal{V}$ and the matrix of ω_{kj} as Ω.

We rewrite this economy in the standard form we require, and put hats on the new expenditure shares. The new input-output matrix $\hat{\Omega}$ has dimension $(1+3N+F)\times(1+3N+F)$. The first industry is the household's consumption aggregator, the next N industries are the original industries, the next N industries produce the value-added of the original industries, the next N industries produce the intermediate inputs of the original industries, and the final F industries correspond to the factors.

Under the relabeling, we have

$$\hat{b} = (1, 0, \ldots, 0)',$$

$$\hat{\Omega} = \begin{pmatrix} \boldsymbol{b} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & diag(a) & diag(1-a) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathcal{V} \\ \Omega & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{130}$$

with

$$\hat{\theta} = (\sigma, \boldsymbol{\theta}, \boldsymbol{\nu}, \boldsymbol{\varepsilon}, \mathbf{1}), \tag{131}$$

and

$$\hat{\mu} = (1, \boldsymbol{\mu}, 1, \ldots, 1). \tag{132}$$

F Aggregation of cost-based Domar weights

In this Appendix we show that recovering cost-based Domar weights from aggregated data is, in principle, not possible. We also show how the Basu-Fernald decomposition can detect changes in misallocation even in acyclic economies (where the equilibrium is efficient, and there is no possibility of reallocation of resources). The vertical economy in Figure 1a also shows the failure of the aggregation property implied by Hulten's theorem. The easiest way to see this is to consider aggregating the input-output table for the economy in Figure 1a. For simplicity, suppose that markups are the same everywhere so that $\mu_i = \mu$ for all i. Since there is no possibility of reallocation in this economy, and since markups are uniform, this is our best chance of deriving an aggregation result, but even in this simplest example, such a result does not exist. Suppose that we aggregate the whole economy $S = \{1, \ldots, N\}$. Then, in aggregate, the economy consists of a single industry that uses labor and inputs from itself to produce. In this case, the input-output

matrix is a scalar, and equal to the intermediate input share of the economy

$$\Omega_{SS} = \frac{1 - \frac{1}{\mu^{N-1}}}{1 - \frac{1}{\mu^N}}, \tag{133}$$

and the aggregate markup for the economy is given by μ. Therefore, $\tilde{\lambda}_S$ constructed using aggregate data is

$$\tilde{\lambda}_S = 1'(I - \mu\Omega)^{-1} = \frac{\mu^{N-1} - \frac{1}{\mu}}{1 - \frac{1}{\mu}}. \tag{134}$$

However, we know from the example that

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A} = \sum_{i \in S} \tilde{\lambda}_i = N \neq \tilde{\lambda}_S = \frac{\mu^{N-1} - \frac{1}{\mu}}{1 - \frac{1}{\mu}}, \tag{135}$$

except in the limiting case without distortions $\mu \to 1$. Therefore, even in this simplest case, with homogenous markups and no reallocation, aggregated input-output data cannot be used to compute the impact of an aggregated shock.

To compare our decomposition with that of Basu and Fernald (2002), consider the acyclic economy in Figure 13 with two factors L_1 and L_2. Now, we know that

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_2} = \tilde{\lambda}_2 = (1 - \alpha_1). \tag{136}$$

The Basu-Fernald decomposition for this economy gives

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_2} = \lambda_2 + R_M + \overline{\mu}^v R_M = \frac{1 - \alpha_1}{\mu_1} + \overline{\mu}^v R_M, \tag{137}$$

where the first term (in this case, the sales-share of 2) is the "pure" technology effect, and R_M is the reallocation of intermediate inputs, even though in this economy, there is no capacity for reallocating resources and the equilibrium is efficient.

G Extra Examples

Example G.1. Next, consider the minimal example with two elasticities of substitution, which demonstrates the principle that changes in misallocation are driven by how each node switches its demand across its supply chain in response to a shock. To this end, we apply Proposition 2 to the economy depicted in Figure 14.

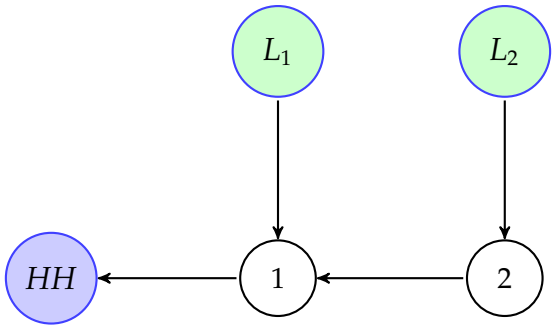


Figure 13: Acyclic economy where the solid arrows represent the flow of goods. The flow of profits and wages from firms to households has been suppressed in the diagram. The two factors in this economy are L_1 and L_2.

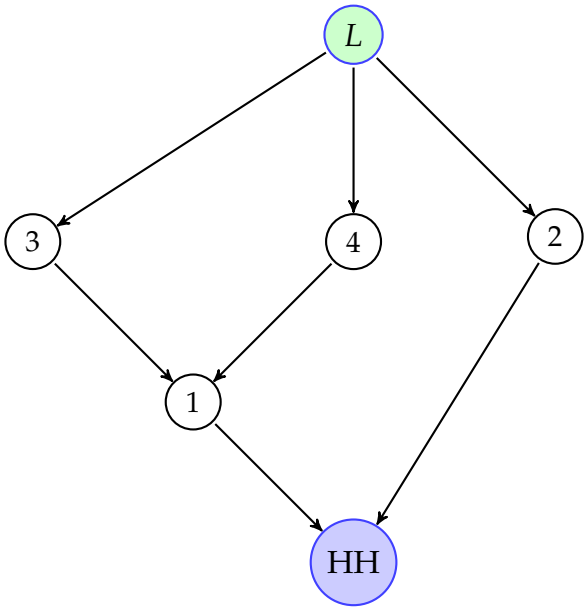


Figure 14: An economy with two elasticities of substitution.

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_3} = \tilde{\lambda}_3 - \frac{1}{\Lambda_L}\Big((\theta_0 - 1)\left[b_1(\omega_{13}\mu_1^{-1}(\omega_{13}\mu_3^{-1} + \omega_{14}\mu_4^{-1})) - \omega_{13}b_1\Lambda_L\right] + (\theta_1 - 1)\mu_1^{-1}\lambda_1\left[\omega_{13}\mu_3^{-1} - \omega_{13}\left(\omega_{13}\mu_3^{-1} + \omega_{14}\mu_4^{-1}\right)\right]\Big).$$

The term multiplying $(\theta_0 - 1)$ captures how the household will shift their demand across 1 and 2 in response to the productivity shock, and the relative degrees of misallocation in 1 and 2's supply chains. The term multiplying $(\theta_1 - 1)$ takes into account how 1 will shift its demand across 3 and 4 and the relative amount of misallocation of labor between 3 and 4. Not surprisingly, if instead we shock industry 1, then only the household's elasticity of

substitution matters, since industry 1 will not shift its demand across its inputs in response to the shock to industry 2:

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_1} = b_1 - \frac{1}{\Lambda_L}(\theta_0 - 1)\left[b_1\mu_1^{-1}(\omega_{13}\mu_3^{-1} + \omega_{14}\mu_4^{-1}) - b_1\Lambda_L\right].$$

This illustrates the general principle in Proposition 2 that an elasticity of substitution θ_j matters only if j is somewhere downstream from k.