Productivity and Misallocation in General Equilibrium

David Rezza Baqaee
LSE

Emmanuel Farhi*
Harvard

December 3, 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, which can be 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. As examples, we pursue some 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 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 by Harberger (1954).

*Emails: d.r.baqaee@lse.ac.uk, efarhi@harvard.edu. We thank Philippe Aghion, Pol Antras, Andrew Atkeson, Susanto Basu, John Geanakoplos, Ben Golub, Gita Gopinath, Dale Jorgenson, Marc Melitz, Ben Moll, Matthew Shapiro, Dan Trefler, Venky Venkateswaran 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 and Maria Voronina 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

[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, Carvalho, Ozdaglar, and Tahbaz-Salehi (2012), Carvalho and Gabaix (2013), Di Giovanni, Levchenko, and Méjean (2014), Baqaee and Farhi (2017a) 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, Parro, and Tsyvinski (2017), Bartelme and Gorodnichenko (2015).

of which is the subject of a vibrant literature, recently reinvigorated by Restuccia and 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]

Our ex-post reduced-form results provide a natural decomposition of aggregate TFP into "pure" technology effects and changes in allocative efficiency. We compare our decomposition to other decompositions in the growth accounting literature. We focus in particular on the prominent contributions of Basu and Fernald (2002) and Petrin and

[4]Some other prominent examples are Hopenhayn and Rogerson (1993), Banerjee and Duflo (2005), Chari, Kehoe, and McGrattan (2007), Guner, Ventura, and Yi (2008), Buera, Kaboski, and Shin (2011), Epifani and Gancia (2011), Fernald and Neiman (2011), Buera and Moll (2012), D'Erasmo and Moscoso Boedo (2012), Bartelsman, Haltiwanger, and Scarpetta (2013), Caselli and Gennaioli (2013), Oberfield (2013), Peters (2013), Reis (2013), Asker, Collard-Wexler, and De Loecker (2014), Hopenhayn (2014), Moll (2014), Midrigan and Xu (2014), Sandleris and Wright (2014), Edmond, Midrigan, and Xu (2015), Gopinath, Kalemli-Özcan, Karabarbounis, and Villegas-Sanchez (2017), and Sraer and Thesmar (2018).

[5]See, for example, Piketty (2014), Elsby, Hobijn, and Şahin (2013), Barkai (2016), Rognlie (2016), Koh, Santaeulàlia-Llopis, and Zheng (2016), and Gutierrez (2017).

Levinsohn (2012). We contrast these decompositions with ours and argue why we find ours preferable.

Although we view our main contribution as theoretical, we also demonstrate the empirical relevance and the scope of applicability of our framework via some examples. The examples are proof of concept illustrations of how our results can be used in practice. Specifically, we use our framework 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 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, a fact which, to the best of our knowledge, we were the first to document.[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]

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

[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, Dorn, Katz, Patterson, and Van Reenen (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

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.

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

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".

competitive model, especially with respect to firm-level shocks.

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. In the NBER working paper version of this paper (Baqaee and Farhi, 2017b) 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 (2017a) 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 we 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 NBER working paper version of this paper (Baqaee and Farhi, 2017b), 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 that 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 has 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 maximizer 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.[10]

Producers

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

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

[10]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. Our ex-post reduced-form results could be generalized to cover generic heterogenous consumers, as long as real GDP is defined using the Laspeyre index.

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^f_{ig} 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^f_{ij}, general 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 specifying 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^f_{ig} can be represented as 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 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.

The revenue-based and cost-based input-output matrices are related according to

$$\tilde{\Omega} = \text{diag}(\mu)\Omega$$

where μ is the vector of markups/wedges, and $\text{diag}(\mu)$ is the diagonal matrix with ith diagonal element given by μ_i.

Leontief Inverse Matrices

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

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

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.[11] 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.

Cross-Entropy

Our final input-output definition is the cross-entropy, which loosely speaking, is a measure of difference between distributions.[12] 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 $\tilde{\Lambda}$.

[11] 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."

[12] 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.

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 $\mathcal{D}$. Let $\mathcal{X}$ be an $(N+F)\times(N+F)$ admissible allocation matrix, where $\mathcal{X}_{ij} = x_{ij}/y_j$ is the share of the physical output y_j of producer j used by producer i. Specify the vector of productivities A and denote by $\mathcal{Y}(A,\mathcal{X})$ the output Y achieved by this allocation.[13,14] Finally, define $\mathcal{X}_{ij}(A,\mu)$ to be equal to $x_{ij}(A,\mu)/y_j(A,\mu)$ 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.[15]

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

[13]The allocation matrix is admissible if the following conditions are verified: $0\le\mathcal{X}_{ij}\le 1$ for all i and j; $\mathcal{X}_{ij}=0$ for all j and for $N+1\le i\le N+F$; $\sum_{i=1}^{N+F}\mathcal{X}_{ij}\le 1$ for all $1\le j\le N$; $\sum_{i=1}^{N+F}\mathcal{X}_{ij}=1$ for all $N+1\le j\le N+F$; 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 $\mathcal{X}_{ij}=\frac{x_{ij}}{y_j}$.

[14]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 $1\le i\le N$ and $y_{N+f}=L_f$ for $1\le f\le F$. The corresponding level of final consumption of good i is $c_i = y_i(1-\sum_{j=1}^{N+F}\mathcal{X}_{ji})$ and the level of output is $\mathcal{D}(c_1,\ldots,c_N)$.

[15]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 $\partial\log y_j/\partial\log\mu_i = 0$.

characterizing changes in output following changes in wedges.

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

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

and

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

Furthermore,

$$\mathrm{d}\log Y = \underbrace{\tilde{\lambda}'\,\mathrm{d}\log A}_{\frac{\partial \log \mathcal{Y}}{\partial \log A}\,\mathrm{d}\log A} \underbrace{-\tilde{\lambda}'\,\mathrm{d}\log \mu - \tilde{\Lambda}'\,\mathrm{d}\log \Lambda}_{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}}\,\mathrm{d}\,\mathcal{X}} \tag{4}$$

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. 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 $(\partial \log \mathcal{Y}/\partial \log A)\, \mathrm{d} \log A = \tilde{\lambda}' \, \mathrm{d} \log A$.

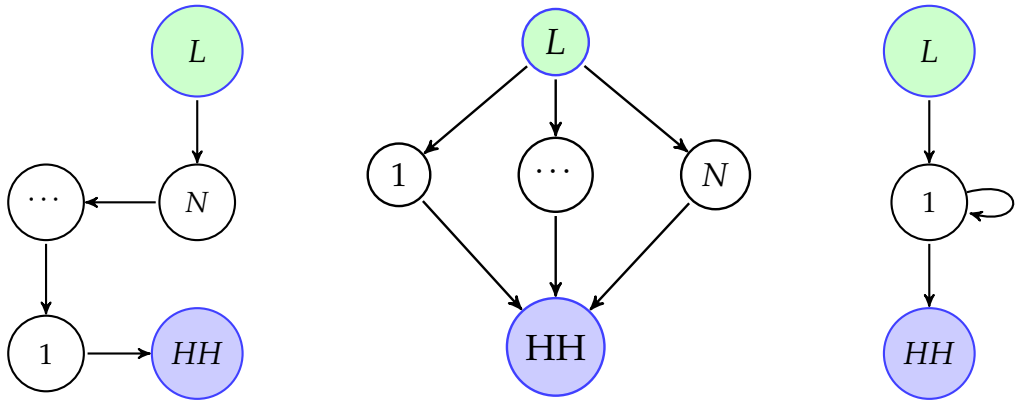
In the case of productivity shocks, Theorem 1 implies that changes in allocative efficiency are given by a simple sufficient statistic: the weighted average of the change in factor income shares $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = -\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$. This 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 examples 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 consequence of downstream double-marginalization which divorces the revenues earned by a producer from that producer's share in the costs 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

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \lambda_k - \lambda_k(\theta_0 - 1)\left(\frac{\mu_k^{-1}}{\sum_j \lambda_j \mu_j^{-1}} - 1\right) \tag{5}$$

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \tilde{\Lambda}_L \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = \lambda_k \theta_0 \left(\frac{\mu_k^{-1}}{\sum_i \lambda_i \mu_i^{-1}} - 1\right). \tag{6}$$

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 (5): 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.[16] The reallocation of labor towards producer k therefore improves allocative efficiency and increases output.[17] 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

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

[17] 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.

of labor remains unchanged, and hence there are no changes in allocative efficiency.[18]

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 $-\tilde{\Lambda}_L \, \mathrm{d} \log \Lambda_L = -\mathrm{d} \log \Lambda_L$ (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

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

[18] 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.

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*

$$\underbrace{\Delta \log Y_t - \tilde{\Lambda}'_{t-1} \Delta \log L_t}_{\Delta \textit{ Aggregate TFP}} \approx \underbrace{\tilde{\lambda}'_{t-1} \Delta \log A_t}_{\Delta \textit{ Technology}} - \underbrace{\tilde{\lambda}'_{t-1} \Delta \log \mu_t - \tilde{\Lambda}'_{t-1} \Delta \log \Lambda_t}_{\Delta \textit{ Allocative Efficiency}}. \tag{7}$$

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

$$\Delta \log Y_t - \hat{\Lambda}'_{t-1} \Delta \log L_t \approx \tilde{\lambda}'_{t-1} \Delta \log A_t - \tilde{\lambda}'_{t-1} \Delta \log \mu_t - \tilde{\Lambda}'_{t-1} \Delta \log \Lambda_t + (\tilde{\Lambda}_{t-1} - \hat{\Lambda}_{t-1})' \Delta \log L_t, \tag{8}$$

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 (7) 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 (7) 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 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.

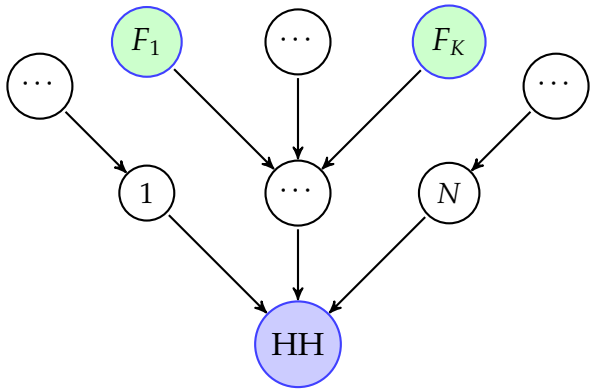


Figure 2: An acyclic economy, where the solid arrows represent the flow of goods. The factors are the green nodes. Each supplier (including factors) have at most one customer, whereas a single customer may have more than one supplier. Economies without cycles can be represented as directed trees with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

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

$$\frac{\mathrm{d} \log Y}{\mathrm{d} \log A_k} = \tilde{\lambda}_k, \quad \frac{\mathrm{d} \log Y}{\mathrm{d} \log \mu_k} = 0. \tag{9}$$

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 $\tilde{\lambda}$.[19,20]

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. Following Hsieh and Klenow (2009), one could compute 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

[19] See Appendix E for a fully worked-out example.

[20] 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.

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 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[21]

$$\begin{aligned}\log\left(\frac{Y(A,1)}{Y(A,\mu)}\right) = \log\left(\frac{Y(A,\hat{\mu}(0))}{Y(A,\hat{\mu}(1))}\right) &= -\int_0^1 \frac{\mathrm{d}\log Y(A,\hat{\mu}(t))}{\mathrm{d}\log\mu}\frac{\mathrm{d}\log\hat{\mu}(t)}{\mathrm{d}\,t}\,\mathrm{d}\,t \\ &= \frac{1}{2}\sum_i \frac{\mathrm{d}\log Y(A,\mu)}{\mathrm{d}\log\mu_i}\left(\frac{1-\mu_i}{\mu_i}\right) + O(\|\mu - 1\|^3). \qquad (10)\end{aligned}$$

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 $\mathrm{d}\log Y/\mathrm{d}\log\mu$ 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 $\mathrm{d}\log Y(A, \hat{\mu}(0))/\mathrm{d}\log\mu = 0$ (from Corollary 1).

The expression $\mathrm{d}\log Y/\mathrm{d}\log\mu_i$ in formula (10) can either be used with the non-parametric results in Theorem 1, or with their parametric counterparts provided in Propositions 3 and 5 in Section 3. As an example, combining equation (6), which is a particular case of Proposition 3, with equation (10), we can 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

[21] 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.

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.[22]

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.[23]

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

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

General Nested CES Economies in Standard Form

A CES economy in standard form is defined by a tuple (ω, θ, μ, F) and a set of normalizing constants $(\overline{y}, \overline{x})$. The $(N + F + 1) \times (N + F + 1)$ matrix ω is a matrix of input-output parameters where the first row and column correspond to the reproducible final good, the next N rows and columns correspond to reproducible goods and the last F rows and columns correspond to non-reproducible factors. The $(N + 1) \times 1$ vector θ is a vector of microeconomic elasticities of substitution. Finally, the $N \times 1$ vector μ is a vector of markups/wedges for the N non-final reproducible goods.[24]

The F factors are modeled as non-reproducible goods and the production function of these goods are endowments

$$\frac{y_f}{\overline{y}_f} = 1.$$

The other $N + 1$ other goods are reproducible, and the production of a reproducible good k can be written as

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

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 final-demand and its production function the final-demand aggregator so that

$$\frac{Y}{\overline{Y}} = \frac{y_0}{\overline{y}_0}, \tag{11}$$

where Y is output and y_0 is the final good.

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 desired.

Consider some initial allocation with productivities $A = 1$ and markups/wedges μ. Normalize the normalizing constants $(\overline{y}, \overline{x})$ to correspond to this initial allocation. Let b and $\tilde{\Omega}$ be the corresponding vector of consumption shares and cost-based input-output matrix. Then we must have $\omega_{0i} = b_i$ and $\omega_{(i+1)(j+1)} = \tilde{\Omega}_{ij}$. From there, all the other cost-based and revenue-based input-output objects can be computed exactly as in Section 2.2.

[24] For convenience we use number indices starting at 0 instead of 1 to describe the elements of ω and θ, but number indices starting at 1 to describe the elements of μ. We impose the restriction that $\omega_{ij} \in [0, 1]$, $\sum_j \omega_{ij} = 1$ for all $0 \leq i \leq N$, $\omega_{fj} = 0$ for all $N < f \leq N + F$, $\omega_{0f} = 0$ for all $N < f \leq N + F$, and $\omega_{i0} = 0$ for all $0 \leq i \leq N$.

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. 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.

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.

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

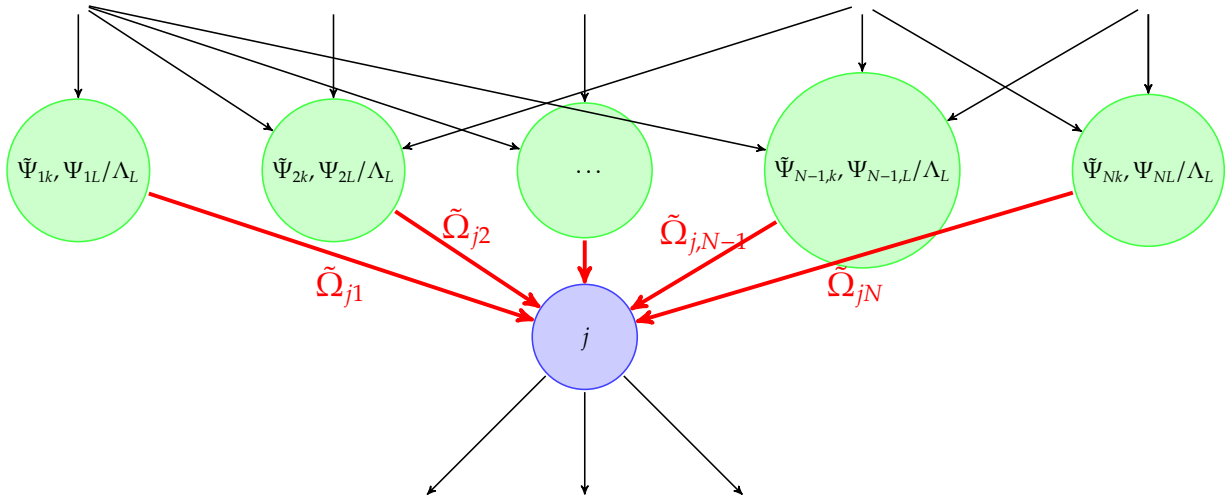


Figure 3: Illustration of the input-output covariance operator defined in equation (12).

input — 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 $\Psi_{(L)}/\Lambda_L$. 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.[25]

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.

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

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

The intuition for Proposition 3 is the following. An increase in the markup/wedge of producer k changes the labor share via two effects. First, there is a mechanical reduction in the labor share, for a given allocation of resources, which is commensurate with $\tilde{\lambda}_k$. Second, there is an effect through the reallocation of resources across producers which reduces the labor share if resources are reallocated to more distorted producers. The change in allocative efficiency is measured by the second effect, and so to isolate it we must net out the first effect $\tilde{\lambda}_k$ from the reduction in the labor share $-\mathrm{d}\log \Lambda_L/\mathrm{d}\log \mu_k$.

Another way to think about it is that a positive markup to producer k shock acts like a negative productivity shock to this producer with the difference that it also releases resources. 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

[25] Baqaee and Farhi (2017a) 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.

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*

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

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

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

We call δ the *factor share impulse matrix*. Its $k$the column encodes the direct or first-round effects of a shock to the productivity of producer k on factor income shares, taking relative factor prices as given. We call Γ the *factor share propagation matrix*. It encodes the effects of changes in relative factor prices on factor income shares, and it is independent of k. When there is only factor, Γ is a zero 1×1 matrix, and we are left with only $\delta_{(k)}$, which allows us to recover Proposition 2 as a special case.

To see the intuition for equation (15), imagine a negative shock $\mathrm{d}\log A_k < 0$ to producer k. For fixed relative factor prices, every producer j will substitute across its inputs in response to this shock. Suppose that $\theta_j < 1$, so that producer j substitutes (in shares) *towards* those inputs i 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 $Cov_{\Omega^{(j)}}\left(\tilde{\Psi}_{(k)}, \Psi_{(f)}/\Lambda_f\right)$, then

substitution by j will increase demand for factor f and hence the income share of factor f.

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

However, when there are multiple factors, this initial change in the demand for factors will affect relative factor prices, and will then set off additional rounds of substitution in the economy that we must account for, which are captured by Γ. For a given set of factor prices, the shock to k affects the demand for each factor, hence factor income shares as measured by the $F \times 1$ vector $\delta_{(k)}$ and in turn relative factor prices. These changes in relative factor prices then cause further substitution through the network, leading to additional changes in factor demands and relative factor prices. The impact of the change in the relative price of factor g on the share 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).[26]

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

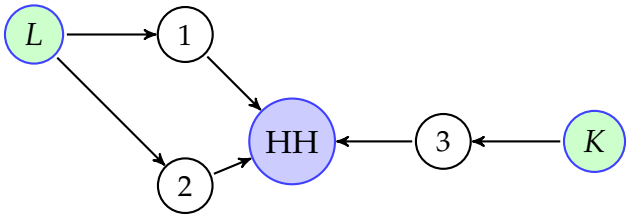


Figure 4: An economy with two factors of production L and K. The subgraph from L to the household contains a cycle, and hence can be subject to misallocation. On the other hand, there is only a unique path connecting K to the household, so there is no misallocation.

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,

[26] Proposition 4 is tightly connected with the results in Baqaee and Farhi (2017a), 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 distorted economy, this information is required even for the first-order impact of shocks.

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.[27] 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 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 Cobb-Douglas economy with a markup shock, where the only source of reallocation comes from the fact that the

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

producer which increases its markup releases some resources. 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:

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

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.[28] So, we only need to observe sales data at the level of disaggregation at which a shock occurs.

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

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 E for a worked-out example.[29]

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 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 modeled 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.[30,31] Beyond actual taxes, these wedges can also implicitly capture frictions preventing the reallocation of resources.

[29] 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).

[30] 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.

[31] 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.

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.[32] 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.

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.[33] 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 refer the reader to the working paper of this version Baqaee and Farhi (2017b) for a treatment of the version of our 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 and characterize the effects of monetary policy on output

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

[33] 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).

from the oft-neglected 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. These applications are meant to be a proof of concept of the usefulness of our theoretical framework.

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 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 assume that all firms within an industry have the same production function. 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 this sample of firms is representative of the overall economy in the following sense: we assume 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. Our representativeness assumption is obviously a strong one, and so our numbers should be interpreted with care.

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, Farhi, and Gourinchas (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.[34]

Despite their differences, all three markup series share some commonalities. Average markups have been increasing over the sample for all three series, as has been documented by Gutierrez (2017). More importantly for us, we find that for all three series, when we decompose this increase in the average markup into an effect between (across) firms and an effect within firms, we find that it is overwhelmingly due to the between (across) effect. In other words, average markups are increasing mostly because high-markup firms are getting larger on average, and not because firms are increasing their markups on average. Figure 5 illustrates this decomposition for the GP markups.[35] To the best of our knowledge, we were the first to document this composition effect for markups.

[34]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.

[35]For the purposes of documentation, in this figure, we have rolled back the data to 1985 because these purely statistical calculations do not require the input-output matrix, which we only have from 1997 onwards.

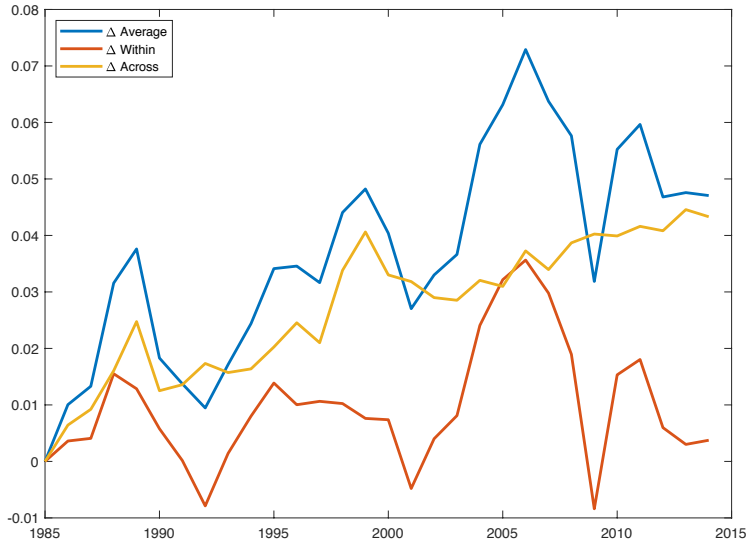


Figure 5: Decomposition of the increase in the harmonic sales-weighted average markup into a between and a within effect, using the Gutiérrez and Philippon (2016) markup data. The change in the average markup is computed as $d\log((\lambda_t'\mu_t^{-1})^{-1})$. The contribution of the within effect is $(\lambda_t'\mu_t^{-1}d\log\mu_t)/(\lambda_t'\mu_t^{-1})$. The contribution of the between (across) effect is the residual. All the changes are cumulated over time

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 (8). The results are plotted in Figure 6 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.

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 fundamental intuition behind these large cumulated improvements in allocative efficiency is that the increase in average markup is largely driven by a composition effect, whereby firms with high markups are getting larger. Given our theoretical results, 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 (8). In Figure 7 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

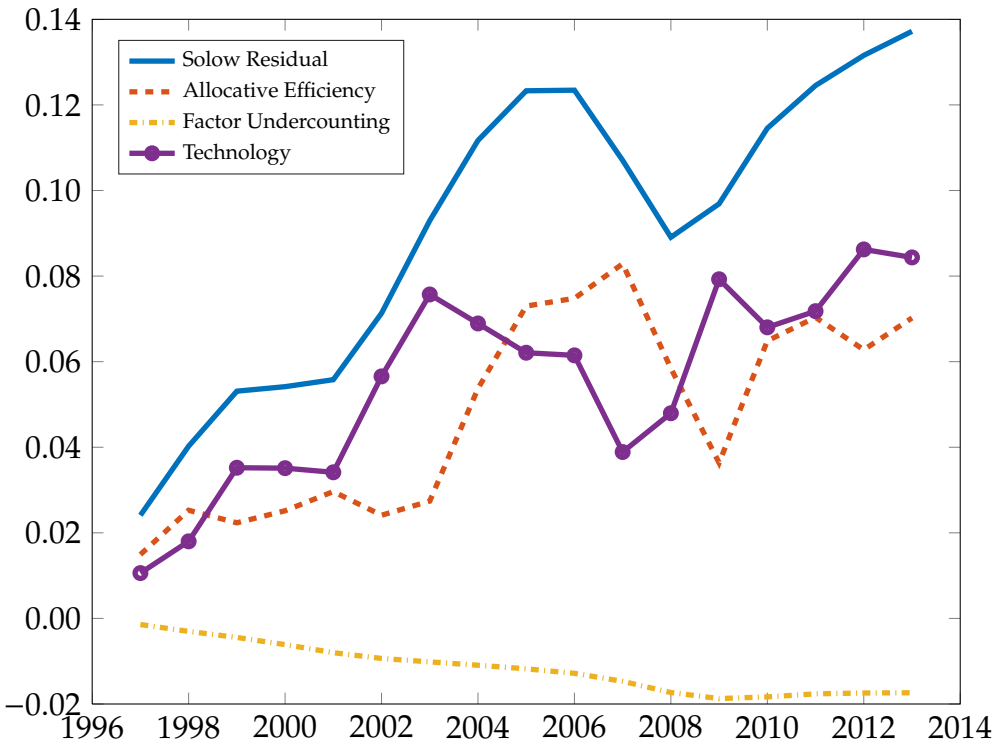


Figure 6: The decomposition in equation (8) using the Gutiérrez and Philippon (2016) markup data.

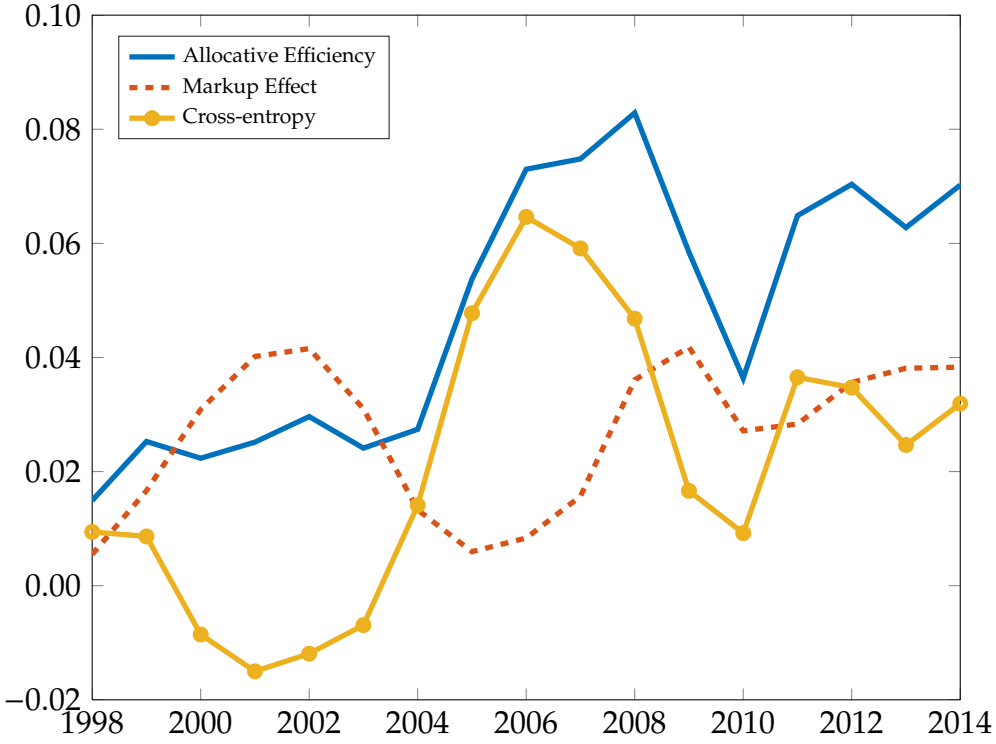


Figure 7: The cumulated contribution of (minus) changes in log markups $-\tilde{\lambda}'_{t-1}\Delta \log \mu_t$ and of changes in cross entropy $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1})$ using the Gutiérrez and Philippon (2016) markup data. The sum of the two components give the overall cumulated change in allocative efficiency.

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

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

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 **?**BaqaeeFarhi, and drawing on estimates from Atalay (2017) and Boehm, Flaaen, and Pandalai-Nayar (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 [36] 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 (10) in Table 1. Using the benchmark GP markups and our second-order approximation, eliminating markups, holding fixed technology, increases aggregate TFP by around 20%.

[36] 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.

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.[37]

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%.

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 Section 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

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

	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.

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.

Volatility of Aggregate TFP

With comparative statics in both productivity and markup shocks at both the firm and industry level[38]

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

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

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

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 "diversified" away in the aggregate. Similarly, $\| D_{\log \mu} \log Y \|$ measures the diversification factor relative to markup shocks.[39]

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

[38] When we consider firm-level shocks, we assess only the contribution of shocks to Compustat firms, i.e. we account for the macro-volatility arising from firm-level shocks when only Compustat firms are being shocked, and not non-Compustat firms. We focus on this exercise because we do not have the data required to compute the contribution of shocks to all firms.

[39] Although Baqaee and Farhi (2017a) 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.

	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 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

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 efficiency.[40] 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.

These intuitions are confirmed in the first two columns of Figure 8, 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 8 plot the output elasticity with respect to markup shocks to specific firms or industries relative to their revenue-based Domar weight.

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

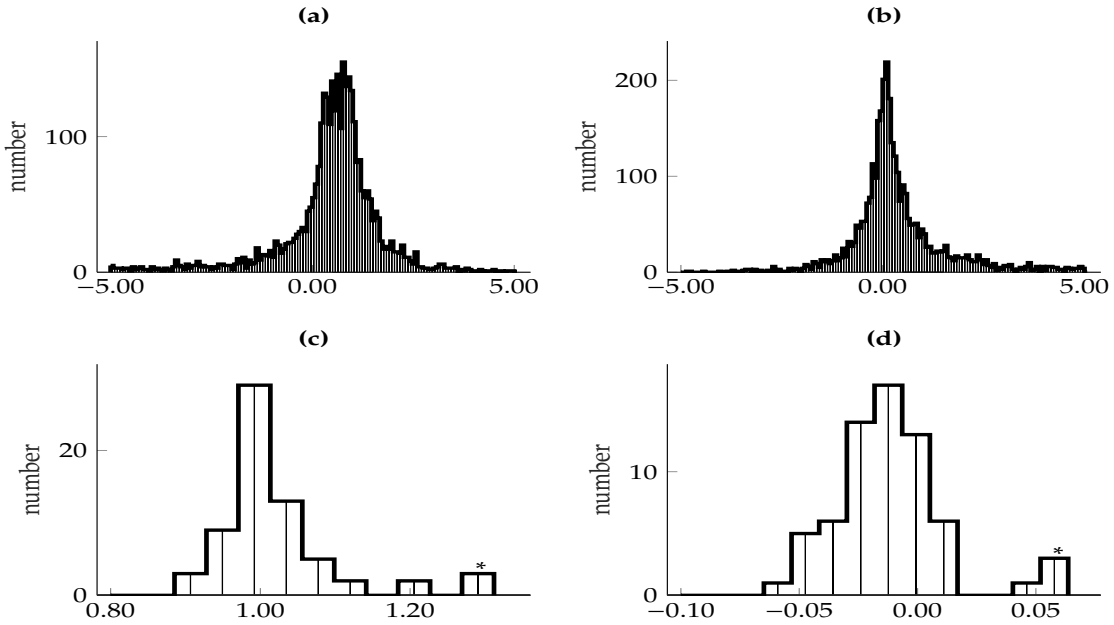


Figure 8: The left column contains histograms of $\mathrm{d}\log Y/\mathrm{d}\log A$ and the right $\mathrm{d}\log Y/\mathrm{d}\log \mu$ relative to λ for firm-level and industry-level shocks respectively. The bunching at the extremes, winsorizing at 4 standard deviations, marked with a star arise solely 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.

6 Robustness and Extensions

Our results can be extended to address some limitations of our analysis. In this section, we fully flesh out one such extension by showing to generalize our analysis to allow for endogenous factor supplies.

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, and nonlinearities. We also refer the reader to the NBER working paper version of this paper Baqaee and Farhi (2017b) for a discussion of how our framework could deal with fixed costs and entry. All these issues introduce additional forces and mechanisms into the model, and we plan to squarely focus on these in future work. However, these discussions show that the intuitions gleaned from the basic framework continue to be useful in analyzing these more complex scenarios.

To model elastic factor supplies, let $G_f(w_f, Y)$ be the supply of factor f, where w_f is the real price of the factor and Y is aggregate income. Let $\zeta_f = \partial \log G_f/\partial \log w_f$ 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 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.[41]

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:

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

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

Equations (18) and (19) can also be applied to frictionless economies with endogenous

[41] 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 $G_L(w, Y) = (w/Y)^\nu$, 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.

factor supplies. They show 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. These results therefore also extend Hulten's theorem to efficient economies with endogenous factor supplies.

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.

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.

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)*.

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 (2017a). The macroeconomic impact of microeconomic shocks: Beyond Hulten's Theorem.

Baqaee, D. R. and E. Farhi (2017b, November). Productivity and Misallocation in General Equilibrium. NBER Working Papers 24007, National Bureau of Economic Research, Inc.

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. 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.

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.

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.

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.

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.

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.

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.

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

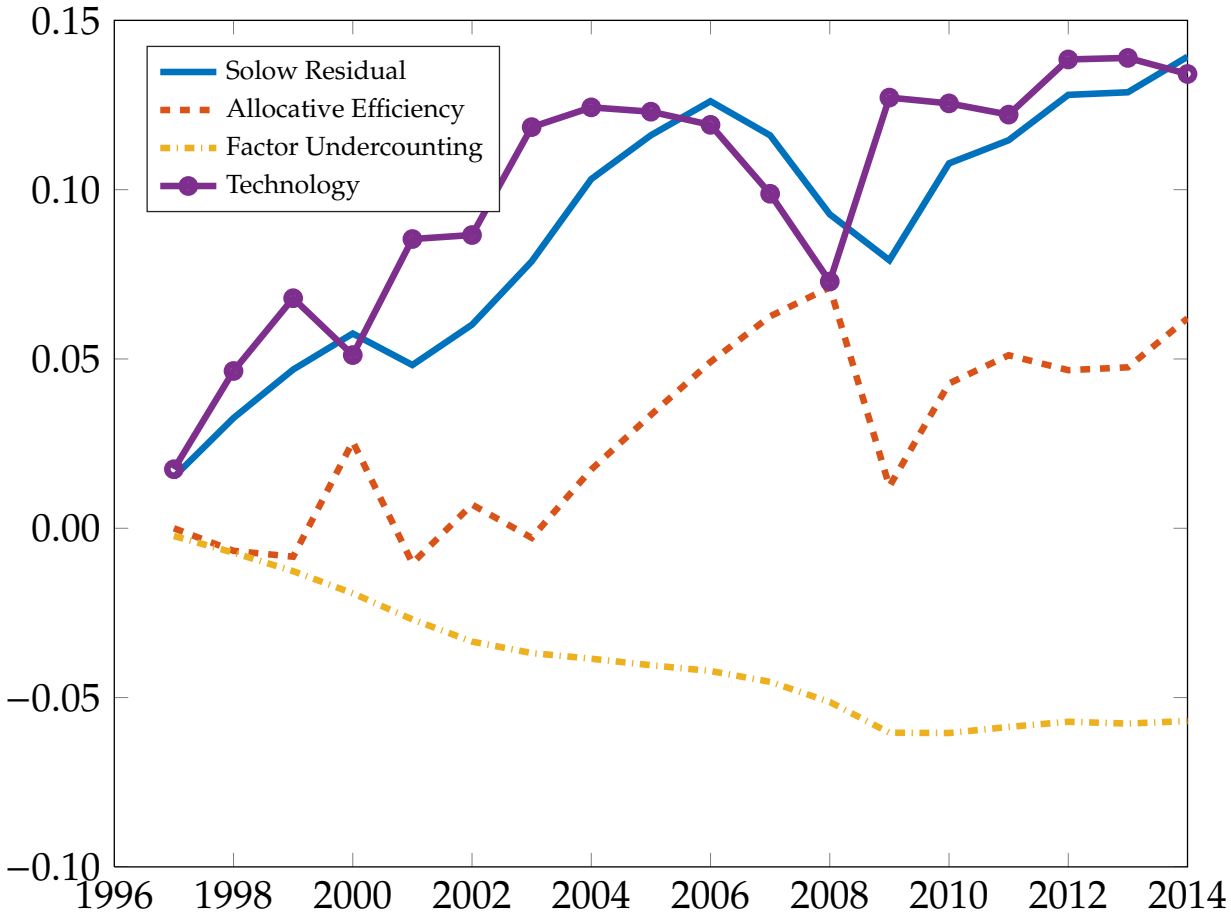


Figure 9: Decomposition of the Solow Residual using LI markups.

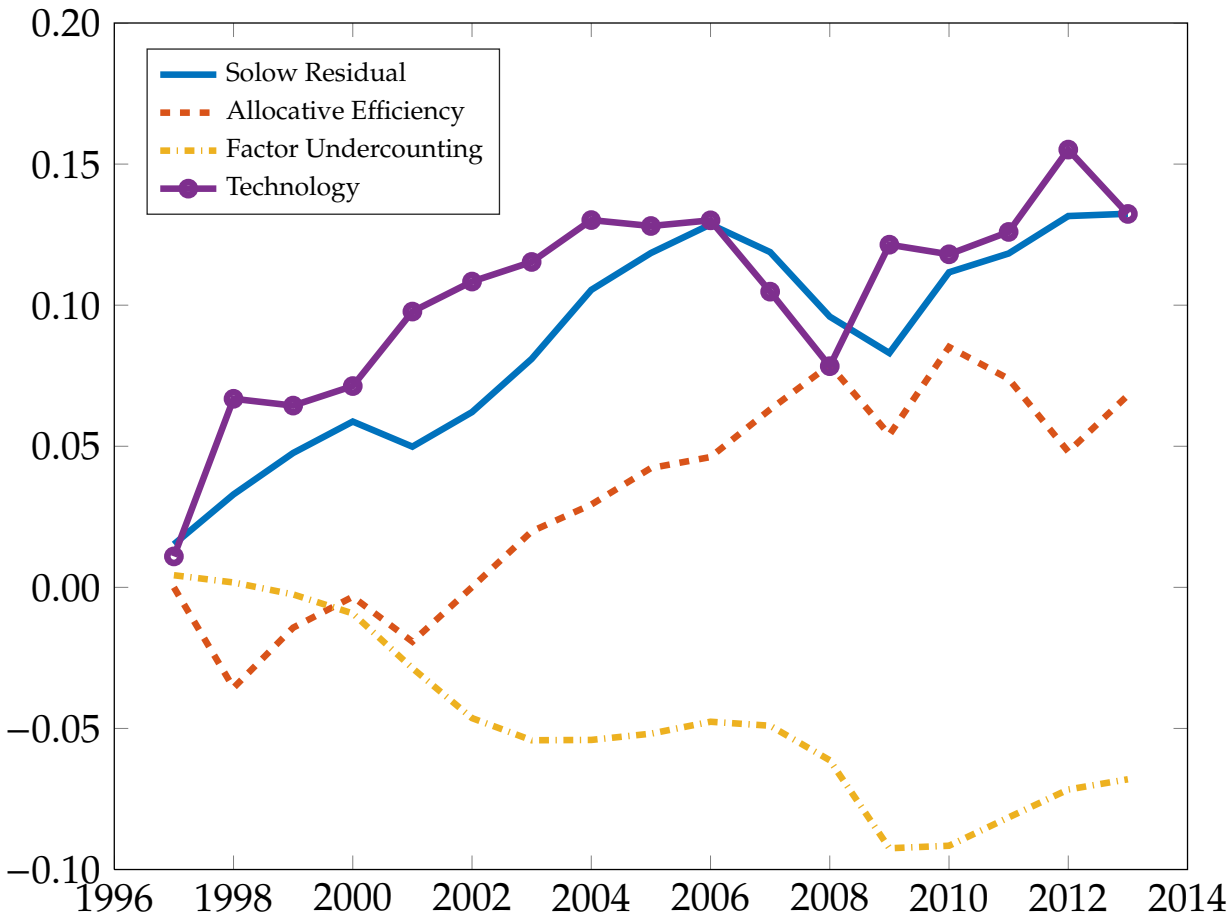


Figure 10: Decomposition of the Solow Residual using DE markups.