Productivity and Misallocation in General Equilibrium

David Rezza Baqaee
UCLA

Emmanuel Farhi*
Harvard

October 9, 2019

Abstract

We provide non-parametric formulas for aggregating microeconomic shocks in economies with distortions such as taxes, markups, frictions to resource reallocation, financial frictions, and nominal rigidities. We allow for arbitrary elasticities of substitution, returns to scale, factor mobility, and input-output network linkages. We show how to separately measure changes in technical and allocative efficiency, thereby generalizing Solow (1957) and Hulten (1978) to inefficient economies. We also show how to compute the social cost of distortions. We pursue applications focusing on firm-level markups in the US. We find that improvement in allocative efficiency, due to the reallocation over time of market share to high-markup firms, accounts for about half of aggregate TFP growth over the period 1997-2015. We also find that eliminating the misallocation resulting from the large and dispersed markups estimated in the data would raise aggregate TFP by about 15%, increasing the economy-wide cost of monopoly distortions by two orders of magnitude compared to the famous 0.1% estimate by Harberger (1954). These exact numbers should be interpreted with care since the data is imperfect and requires substantial imputation.

*Emails: baqaee@econ.ucla.edu, 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 are especially grateful to Natalie Bau for detailed conversations. 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.

1 Introduction

The foundations of macroeconomics rely on Domar aggregation: changes in a constant-returns-to-scale index are approximated by the sales-weighted average of the changes in its components. Hulten (1978), building on the work of Solow (1957) and Domar (1961), provided a rationale for using Domar aggregation to construct measures of aggregate productivity. He showed that in perfectly competitive economies

$$\frac{\Delta Y}{Y} - \sum_f \Lambda_f \frac{\Delta L_f}{L_f} \approx \sum_k \lambda_k \frac{\Delta TFP_k}{TFP_k},$$

where Y is real GDP, L_f is the supply of factor f, Λ_f is its income share in GDP, TFP_k is the TFP of producer k, and λ_k is its sales as a share of GDP, also known as its Domar weight. In other words, aggregate productivity (output growth minus input growth) is equal to a Domar-weighted sum of technology changes. Although Hulten's theorem is most prominent for its use in growth accounting, where it is used 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 multisector models and models with production networks.

The generality of Hulten's theorem comes from exploiting a macro-envelope condition resulting from the first welfare theorem. This means that the theorem requires perfect competition and frictionless markets — without these conditions, Hulten's theorem does not apply. This paper generalizes Hulten's theorem beyond efficient economies, and provides an aggregation result for disaggregated 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.

Hulten's theorem is instrumental for both measuring and predicting the effects of microeconomic shocks on aggregate TFP. Our objective is to provide a workable generalization of Hulten's theorem for economies with distortions, and our results can also be used to both measure changes in productivity and to make predictions about how aggregate TFP will evolve in response to shocks.

Our results suggest a new and structurally interpretable decomposition of changes in aggregate TFP into pure (exogenous) changes in technology and (endogenous) changes in allocative efficiency. 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 to a first order. 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 define a new measure of aggregate TFP growth which nets out the purely technological impact of factor growth from output growth. Our distortion-adjusted Solow residual generalizes the construction of Hall (1988) to disaggregated economies with misallocation. We show that in the presence of misallocation, it is this measure, rather than the traditional Solow residual, which correctly measures aggregate TFP growth.

We then show that aggregate TFP growth can be further decomposed into changes in technical efficiency and changes in allocative efficiency. We explain why our decomposition of aggregate TFP into pure technology effects and changes in allocative efficiency is preferable on conceptual grounds to those of Gollop et al. (1987), Basu and Fernald (2002), and Petrin and Levinsohn (2012). We show that these differences matter quantitatively in empirical applications. For example, when we apply our decomposition to US data, we find that changes in allocative efficiency play a significantly larger role in explaining aggregate productivity growth than in the other decompositions.

Furthermore, we provide an analytical formula for the social cost of distortions, generalizing misallocation formulae like those of Hsieh and Klenow (2009) to economies with arbitrary input-output network linkages, numbers of factors, microeconomic elasticities of substitution, and distributions of distorting wedges. We show that these generalizations matter quantitatively in empirical applications. For example, we find that accounting for these details of the production structure triples the aggregate TFP losses from markups in the US relative to the commonly-used formula in Hsieh and Klenow (2009).

By bringing together the growth-accounting literature, which is concerned with measurement, and the misallocation literature, which is concerned with counterfactuals, we also hope to clarify some potentially confusing subtleties. Namely, the growth-accounting notion of changes in allocative efficiency due to the reallocation of resources to more or less distorted parts of the economy over time, is very different to the misallocation literature's notion of allocative efficiency measured as the distance to the Pareto-efficient

frontier. Our formulas theoretically characterize and relate these two different notions, and our empirical applications show that these subtleties are important for understanding patterns in the data.

We demonstrate the framework's empirical relevance and scope of applicability with some proof-of-concept examples. Specifically, we answer two different questions about the role of firm-level markups in determining aggregate TFP. We focus on markups in light of the accumulating evidence that average markups have increased over the past decades in the US.

1. How have changes in the allocation of resources contributed to TFP growth in the US over the past 20 years?

   We perform a non-parametric decomposition of measured TFP growth into a pure technology effect and an allocative efficiency effect. We implement our decomposition in the US over the period 1997-2014. Focusing on firm-level markups as a source of distortions, we find that the improvements in the allocation of resources across firms accounts for about 50% of the cumulated growth in aggregate TFP.

   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.[1] From a social perspective, these high-markup firms were too small to begin with, and so the reallocation of resources towards them increases measured aggregate TFP over time.

2. What are the gains from eliminating 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 10-25% (depending on the markup series). This increases the estimated cost of monopoly distortions by two orders of magnitude compared to the famous estimates of 0.1% of Harberger (1954).[2]

   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

[1] This finding is consistent with Vincent and Kehrig (2019) and Autor et al. (2019) who argue that the labor share of income has decreased because low labor-share firms have become larger, and not because the labor share has declined within firms. Our finding that this composition effect also holds for markups has since been corroborated by Autor et al. (2019) and De Loecker et al. (2019).

[2] Like Harberger, we measure only the static gains from eliminating markups, holding fixed technology, and abstract away from the possibility that lower markups may reduce entry and innovation. In other words, even though 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.

higher across firms within a sector than across sectors.[3] 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, the use of firm-level data and higher elasticities of substitution is not enough: accounting for the existence of input-output linkages, instead of assuming value-added production functions, almost triples the losses.

We also find that the distance from the Pareto-efficient frontier has increased since 1997. This has happened because the dispersion in markups has increased over time. This finding may appear to contradict our conclusion that reallocation has made a positive contribution to measured TFP growth over the period. The resolution is that these results are conceptually different: the first is about how changes in the allocation of resources relative to the past have changed output, whereas the latter is about how output has changed relative to the optimal allocation of resources. We find that the reallocation of resources over time has contributed positively to output growth, but at the same time, the optimal allocation of output has grown more quickly, increasing the distance to the frontier.

Our empirical results are tentative due to a variety of compromises we are forced to make in mapping data to the theory. First, the theory demands that we observe input-output connections at the level of the distortions. Therefore, if there are heterogenous markups at the firm-level, then the input-output data must, in principle, also be at the firm-level. Since this data does not exist, we impute a firm-level network from sectoral input-output data by assuming that the productions functions within each industry are the same. Second, our firm-level dataset is Compustat, and to the extent that the patterns in Compustat are non-representative of the broader economy, this can contaminate our empirical conclusions about the behavior of economic aggregates. Finally, in our empirical applications, unlike in our theoretical results, we assume that markups are the only distorting wedges. In our view, convincingly measuring other wedges such as financial frictions for a whole economy is even more difficult than measuring markups. Nonetheless, the presence of unobserved wedges which interact with markups in non-trivial ways can contaminate our empirical conclusions.

Despite their generality, our theoretical results also have some important theoretical limitations. First, our basic framework abstracts away from entry and external economy effects such as those studied by Baqaee (2016).[4] Second, in this paper we focus on

[3]See also Asker et al. (2019), who use micro-data from the oil industry to show much higher amounts of economic waste than the classic Harberger analysis suggested.

[4]Our results allow for a form of entry relying on individual demand curves with choke-prices. In this

first-order approximations.[5] Finally, we model frictions using wedges, which we take as primitives. The advantage is that we characterize the response of the 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.[6]

The outline of the paper is as follows. In Section 2, we set up the general model, prove our growth-accounting results, and introduce our decomposition of aggregate TFP changes into pure changes in technology and changes in allocative efficiency. In Section 3, we define a new residual, show that it correctly measures aggregate productivity in inefficient economies, and compare our results to the existing literature, like Gollop et al. (1987), Basu and Fernald (2002), and Petrin and Levinsohn (2012). Section 4 shows explicitly how the components of our decomposition depend on microeconomic primitives. In the body of the paper, we use a parametric version of the general model with nested CES production and consumption functions with an arbitrary number of nests, input-output patterns, returns to scale, factors of production, and elasticities to show this dependence, but the more general results are in Appendix H. In Section 5, we use these results to characterize the distance to the efficient frontier, and we discuss different notions of changes in allocative efficiency which can all be computed with our structural results. In Section 6, we discuss the data requirements of our results and some subtleties in implementing and interpreting them. In Section 7, we apply our results to the data by performing non-parametric decompositions of the sources of growth in the US, as well as exercises measuring the distance from the efficient frontier. In Section 8, we mention some extensions of the basic framework.

Related Literature

Our paper is related to the literature on misallocation, growth-accounting, and production networks. The misallocation literature is perhaps the oldest of the three, tracing its way

case, small shocks can push a producer to enter a market, but at a small scale commensurate with the shocks, so that all producer-level variables evolve smoothly. This is in contrast to models with fixed costs and their associated non-convexities. We refer the reader to Baqaee and Farhi (2019a) for a general treatment of such models.

[5]In Appendix H, we discuss conditions under which the nonlinear analysis of efficient economies in Baqaee and Farhi (2017a) can be leveraged to characterize nonlinearities in inefficient economies.

[6]In Appendix F, we provide an example with endogenous markups using an oligopolistic model à la Atkeson and Burstein (2008). In the NBER working paper version of this paper (Baqaee and Farhi, 2017b), we also showed how to use our results to analyze the effect of monetary policy and productivity shocks in a model with sticky prices captured as endogeneous markups.

from Alfred Marshall and Jules Dupuit to Arnold Harberger. More recently, the literature has been reinvigorated by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). Perhaps closest to us are Edmond et al. (2018) who use a calibrated model to study the costs of markup distortions in the US. They find that the social costs of markups are lower than we do primarily because they focus on size-related markup dispersion, whereas we take into account the total dispersion in markups.

The growth-accounting literature began with Solow (1957), and was extended by Domar (1961), Hulten (1978), Hall (1990), Gollop et al. (1987), Basu and Fernald (2002), Petrin and Levinsohn (2012), and Osotimehin (2019), among others.

Finally, our paper is related to other studies of inefficient production networks. Jones (2013) considers an input-output economy with Cobb-Douglas production and consumption functions, distortionary wedges, and two factors. Bigio and La'O (2019) expand on Jones (2013) by considering input-output economies with Cobb-Douglas production and consumption functions with distortionary wedges and elastic labor supply. They characterize the elasticity of output with respect to productivities and wedges. They present an application to the 2008-09 financial crisis in the US and find that input-output linkages double the output effects of financial frictions. Our non-parametric results show that the Cobb-Douglas case, while tractable, is also special. For example, it implies that the allocation of resources is invariant to productivity shocks and this matters both qualitatively and quantitatively.

Baqaee (2016) also considers input-output economies with distortionary wedges (due to endogenous markups) but allows for CES production and consumption functions with two nests, assuming that the cross-industry elasticities of substitution are the same for all agents. He studies how productivity shocks affect aggregate TFP in this environment, allowing for the possibility that there is free entry and external economies of scale. Grassi (2017) considers input-output economies that are Cobb-Douglas across sectors, CES within sectors, and without entry. He allows for oligopolistic competition within industries and studies the way productivity shocks propagate through the network. Our analysis differs from Baqaee (2016) and Grassi (2017) in that our results are non-parametric, allowing for more general substitution elasticities and multiple factors, but do not allow for free entry or explicitly model oligopolistic competition.

Finally, Liu (2019) considers an input-output economy with one factor, constant returns to scale, and some pre-existing non-tax wedge distortions. Unlike in the previous papers mentioned above, but like in our paper, the results are non-parametric. However, the object of interest of the paper is not GDP but GDP minus the revenues generated by non-tax wedges (taken to be offset by non-pecuniary costs). The main result characterizes the

elasticity of this object with respect to taxes at the zero-tax point. It shows that this elasticity depends on the network and on the wedges but not on the elasticities of substitution. The paper applies this result to conclude that in South Korea and China, industrial policy rightfully targets upstream sectors which are more distorted. Our non-parametric results show that the irrelevance of elasticities of substitution for aggregates, while convenient, is also special. This irrelevance breaks away from the zero-tax point, when there are multiple factors or decreasing returns to scale, or when the object of interest is GDP or aggregate TFP.

Since we apply our results to data on US markups, our analysis is also related to the debate about the recent trends in factor and profit shares in the US. See for example, Elsby et al. (2013), Rognlie (2016), Barkai (2019), Caballero et al. (2017), Gutierrez (2017), Farhi and Gourio (2018), Koh et al. (2019), Vincent and Kehrig (2019), Autor et al. (2019), and De Loecker et al. (2019). Here our contribution is to point out that the changes in the aggregate profit share (and in the aggregate markup) seem primarily to be driven by composition effects within industries, whereby the high profit (and high markup) firms have been getting larger. Our analysis shows that this composition effect has implications for aggregate TFP growth.

2 Ex-Post Reduced-Form Results

In this section, we set up our framework, characterize how shocks to wedges and productivity affect output and TFP, and define a notion of change in allocative efficiency.

2.1 Set up

The model has N producers indexed by i and F factors indexed by f with inelastic supply L_f. Each producer uses intermediate inputs and factors, and sells its output as both an intermediate good to other producers and as a final good.

Final Demand

Real GDP is the maximizer of a constant-returns aggregator of final uses of 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.[7]

Producers

Good i is produced using a constant-returns technology described by the cost function

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

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

General Equilibrium

Given productivities A_i, markups μ_i, wedges τ_{ij} and τ_{ij}^f, 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.

Variable Returns to Scale, Non-Hicks-Neutral Shocks, and Markup-Wedge Equivalence

Without loss of generality, we exploit three simplifications.[8] First, despite specifying constant-returns cost functions, our setup can accommodate variable (increasing or decreasing) returns to scale. This is because, as pointed out by McKenzie (1959), variable returns to scale can be modeled with constant returns to scale and producer-specific fixed

[7] The existence of a constant-returns-to-scale aggregate real GDP function allows us to avoid index-number complications. In follow-up work, Baqaee and Farhi (2018), we show how these results can be extended to environments with heterogeneous final consumers (i.e. non-aggregable final consumption).

[8] In Section 6, we discuss how these simplifications affect the mapping of the theory to the data.

factors. Going forward, we proceed with our constant-returns setup with the understanding that it can be reinterpreted to capture variable returns to scale provided that the original set of factor is expanded to include producer-specific fixed factors.

Second, 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 (biased) productivity shocks to the use of input j by producer i by introducing a fictitious producer buying input j and selling to producer i with a linear technology, and by considering Hicks-neutral shocks to this fictitious producer. Demand shocks can also be modeled in this way by considering combinations of positive and negative non-neutral productivity shocks to the different inputs of producer i.[9]

Third, all the wedges τ_{ij} and τ_{ig}^{f} 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. We do this to simplify the notation.

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/wedges μ_k. Without loss of generality, we normalize the initial productivity levels to one. We now define accounting 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 the budget of the final consumers

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

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

[9] 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.

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.

The revenue-based input-output matrix Ω is 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 zeros.

The cost-based input-output matrix $\tilde{\Omega}$ is the $(N+F)\times(N+F)$ matrix whose ijth element is equal to the elasticity of i's marginal costs with respect 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. 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 by

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

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

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

While the input-output matrices Ω and $\tilde{\Omega}$ record the *direct* exposures of one producer to another, in revenues and in costs respectively, the Leontief inverse matrices Ψ and $\tilde{\Psi}$ record instead the *direct and indirect* exposures through the production network. This can

be seen most clearly by noting that $(\Omega^n)_{ij}$ and $(\tilde{\Omega}^n)_{ij}$ measure the weighted sums of all paths of length n from producer i to producer j.

Domar Weights

The revenue-based Domar weight λ_i of producer i is its sales 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 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. 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.

2.3 Ex-Post Reduced-Form Results

In this section, we derive our comparative-static results stated in terms of ex-post reduced-form sufficient statistics. 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.[10,11] 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 markups/wedges is μ. The level of output at this equilibrium is given by $\mathcal{Y}(A, \mathcal{X}(A, \mu))$.

Now consider shocks $\mathrm{d}\log A$ and $\mathrm{d}\log \mu$ and the associated change in the equilibrium allocation matrix $\mathrm{d}\mathcal{X} = (\mathrm{d}\mathcal{X}/\mathrm{d}\log A)\,\mathrm{d}\log A + (\mathrm{d}\mathcal{X}/\mathrm{d}\log \mu)\,\mathrm{d}\log \mu$. The change in aggregate output in response to these shocks is given by

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

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 $\mathrm{d}\mathcal{X}$, holding technology constant.

Now, we extend Hulten (1978) to cover inefficient economies and provide an interpretation for the result. We also extend Hulten's theorem along another dimension by characterizing changes in output following changes in markups/wedges.

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

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

and

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

Furthermore, the decomposition of output changes into pure changes in technology and changes in

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

[11] To see how to construct this allocation, consider the production functions F_i defined as dual to the cost functions C_i in the usual way. Then the vector of outputs y_i solves the system of equations $y_i = F_i(\mathcal{X}_{1i}y_1, \ldots, \mathcal{X}_{(N+F)i}y_{N+F})$ for $1 \leq i \leq N$ and $y_{N+f} = L_f$ for $1 \leq f \leq F$. The corresponding level of final consumption of good i is $c_i = y_i(1 - \sum_{j=1}^{N+F} \mathcal{X}_{ji})$ and the level of output is $\mathcal{D}(c_1, \ldots, c_N)$.

allocative efficiency is given by

$$\mathrm{d}\log Y = \underbrace{\tilde{\lambda}'\,\mathrm{d}\log A}_{\Delta\textit{Technology}} \underbrace{-\tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}'\,\mathrm{d}\log\Lambda}_{\Delta\textit{Allocative Efficiency}}. \tag{4}$$

Theorem 1 not only provides a formula for the macroeconomic output impact of microeconomic productivity and markup/wedge shocks, but it also provides an interpretable decomposition of the effect. Specifically, the first component $(\partial \log \mathcal{Y}/\partial \log A)\,\mathrm{d}\log A = \tilde{\lambda}'\,\mathrm{d}\log A$ is the pure technology effect: the change in output holding fixed the share of resources going to each user; the second component $(\partial \log \mathcal{Y}/\partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = -\tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}'\,\mathrm{d}\log\Lambda$ is the change in output resulting from the reallocation of shares of resources across users.[12]

Hulten's theorem obtains as a special case of Theorem 1 when there are no markups/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)**.** *If the initial equilibrium is efficient so that there are no markups/wedges* $\mu = 1$ *, 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 $(\partial \log \mathcal{Y}/\partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = -\tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}'\,\mathrm{d}\log\Lambda$, which is nonzero in general. Furthermore, because of wedges between prices and marginal costs, the direct effect of changes in

[12]The term $-\tilde{\Lambda}' d\log\Lambda$ can be interpreted as the change in the gap between the probability distribution defined by $\tilde{\Lambda}$ and Λ, as measured by relative entropy. That is, when there are no shocks to wedges, an increase in the distance between $\tilde{\Lambda}$ and Λ implies an improvement in allocative efficiency due to reallocation. At the efficient allocation, the change in this distance is zero, so reallocation has no effects on efficiency.

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 $-\tilde{\Lambda}' \, \mathrm{d} \log \Lambda = -\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$. This is simply a weighted average of the reductions in factor shares $-\mathrm{d} \log \Lambda_f$ with weights $\tilde{\Lambda}_f$ satisfying $\sum_f \tilde{\Lambda}_f = 1$. A decrease in the weighted average of factor shares $\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f < 0$ indicates that resources are reallocated to the more monopolized or downwardly distorted parts of the economy. Allocative efficiency improves because from a social perspective, these monopolized or downwardly distorted parts of the economy received too few resources to begin with.

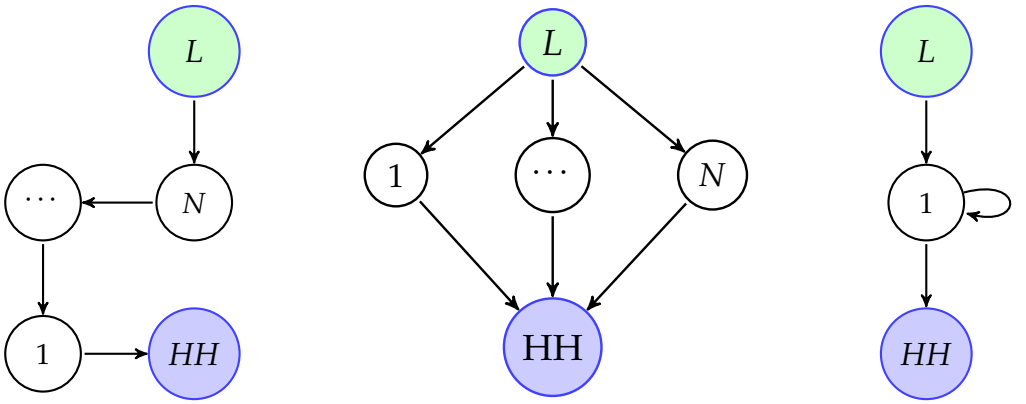
Similarly, in the case of markup/wedge shocks, Theorem 1 implies that changes in allocative efficiency are given by a simple sufficient statistic: $-\tilde{\lambda}' \, \mathrm{d} \log \mu - \tilde{\Lambda}' \, \mathrm{d} \log \Lambda = -\tilde{\lambda}' \, \mathrm{d} \log \mu - \sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$. Now $-\tilde{\Lambda}' \, \mathrm{d} \log \Lambda = -\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f$ reflects both the direct reduction $\tilde{\lambda}' \, \mathrm{d} \log \mu$ of factor shares from increased markups for a given allocation of resources, 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.

It is remarkable that in both cases, 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. In Section 4, we also provide an explicit characterization of $-\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f / \, \mathrm{d} \log A_k$ and $-\sum_f \tilde{\Lambda}_f \, \mathrm{d} \log \Lambda_f / \, \mathrm{d} \log \mu_k$ in terms of the microeconomic elasticities of substitutions of the production functions and final demand, the properties of the input-output network, and the markups/wedges. In the main body of the paper, we focus on nested-CES economies. In Appendix H, we show how to fully generalize the results to general non-CES economies.

2.4 Illustrative Examples

In this section, we introduce some bare-bones examples to illustrate the intuition of Theorem 1. In Section 4, we specialize Theorem 1 to the case of general nested constant-elasticity-of-substitution (CES) economies with arbitrary input-output linkages. The examples here are special cases of these upcoming general results.

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.



(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.

The economy in Figure 1a is a production line where 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.[13] The household purchases the output of the downstream producers according to a CES aggregator with elasticity θ_0. The round-about 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 two ways Hulten's theorem can break down: (1) the equality of revenue-based and cost-based Domar weights (used to weigh the 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

For the economy in Figure 1a, Theorem 1 implies that

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

[13]The terms "vertical" and "horizontal" economies are due to Bigio and La'O (2019), although for our example, we relax the Cobb-Douglas assumption.

and

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

where for this special case, $\tilde{\lambda}_k = 1$, $\tilde{\Lambda}_L = 1$, $\mathrm{d}\log \Lambda_L / \mathrm{d}\log A_k = 0$, and $\mathrm{d}\log \Lambda_L / \mathrm{d}\log \mu_k = -1$.

In this economy, there is only one feasible allocation of resources, so the equilibrium allocation is efficient regardless of the wedges. Therefore, our decomposition detects no changes in allocative efficiency in response to shocks since $-\tilde{\lambda}' \,\mathrm{d}\log \mu - \tilde{\Lambda}_L \,\mathrm{d}\log \Lambda_L = -\tilde{\lambda}' \,\mathrm{d}\log \mu - \mathrm{d}\log \Lambda_L = 0$.

Even though its equilibrium is efficient, Hulten's theorem fails for the vertical economy. This is because for this example, the revenue-based Domar weight $\lambda_k = \prod_{i=1}^{k-1} \mu_i^{-1}$ is not the same as the cost-based Domar weight $\tilde{\lambda}_k = 1$. 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 consider the horizontal economy represented in Figure 1b. The consumption of the household, or final output, 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.

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 markup downstream from producer k and $1 = \tilde{\Lambda}_L \neq \Lambda_L$ since there are markups downstream from labor. The 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 $-\tilde{\lambda}' \, \mathrm{d} \log \mu - \tilde{\Lambda}_L \, \mathrm{d} \log \Lambda_L = -\tilde{\lambda}' \, \mathrm{d} \log \mu - \mathrm{d} \log \Lambda_L$.

Consider equation (5): the effects of a positive technology shock $\mathrm{d} \log A_k$ to producer k. For this economy, the allocation matrix $\mathcal{X}$ determines the share of labor used by each producer. Holding fixed the allocation matrix, 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 this reallocation of workers increases or decreases allocative efficiency and output in turn depends of the comparison of the markup μ_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.[14] The reallocation of labor towards producer k therefore improves allocative efficiency and increases output.[15] The opposite occurs when the markup of producer k is smaller than the average markup. This effect works in the opposite direction when $\theta_0 < 1$, since in that case, the shock would reallocate workers away from producer k. Of course, in the Cobb-Douglas case when $\theta_0 = 1$, the allocation of labor remains unchanged, and hence there are no changes in allocative efficiency.[16]

Finally, note that there are no changes in allocative efficiency if $\mu_k = (\sum_i \lambda_i \mu_i^{-1})^{-1}$, since in this case the initial allocation of resources between k and the rest the economy is efficient, and therefore, we recover Hulten's theorem.

All of this information is summarized by a simple sufficient statistic: the change in allocative efficiency is exactly the reduction in the labor share $-\mathrm{d} \log \Lambda_L$. The labor share of

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

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

[16] This last property is a more general property of Cobb-Douglas economies which we shall encounter in Section 4: 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.

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.

For a markup shock $\mathrm{d}\log\mu_k$ to producer k, as long as $\theta_0 > 0$, 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 $-\tilde{\lambda}_k \,\mathrm{d}\log\mu_k - \mathrm{d}\log\Lambda_L$. Now the reduction in the labor share $-\mathrm{d}\log\Lambda_L$ reflects both the direct reduction $\tilde{\lambda}_k \,\mathrm{d}\log\mu_k$ of the labor share from the increase in the price of good k for a given wage, 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.

When consumption is Leontief $\theta_0 = 0$ the household consumes a fixed quantity of each good regardless of its price. As a result, the allocation of labor does not change in response to $\mathrm{d}\log\mu_k$, and there are therefore no associated changes in allocative efficiency.

To summarize, with productivity shocks, the benchmark elasticity with no changes in allocative efficiency $\mathrm{d}\mathcal{X}/\mathrm{d}\log A$ is Cobb Douglas $\theta_0 = 1$. With markup shocks, the benchmark elasticity with no changes in allocative efficiency $\mathrm{d}\mathcal{X}/\mathrm{d}\log\mu$ is Leontief $\theta_0 = 0$ instead. In Section 4, we show that these are generic properties of Cobb-Douglas and Leontief economies.

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

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

Theorem 1 implies that

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

and

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

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

3 Growth Accounting

In this section, we explain how the results derived in Section 2 can be used for growth accounting. We show that in inefficient economies, the traditional Solow residual, which weighs factor growth according to factor shares, cannot be interpreted as a measure of changes in aggregate TFP. To remedy this problem, we propose a distortion-adjusted Solow residual that weighs factor growth according to cost-based factor shares. The modified Solow residual can be decomposed into changes in pure technology and changes in allocative efficiency. We argue that our decomposition is preferable to the related decompositions of Gollop et al. (1987), Basu and Fernald (2002), and Petrin and Levinsohn (2012).

3.1 Distortion-Adjusted Solow Residual and Aggregate TFP Decomposition

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 $\mathrm{d}\log Y/\mathrm{d}\log L_f = \tilde{\Lambda}_f - \sum_g \tilde{\Lambda}_g \,\mathrm{d}\log \Lambda_g/\mathrm{d}\log L_f$.

The appropriate measure of aggregate TFP growth $\Delta \log Y_t - \tilde{\Lambda}_{t-1}' \Delta \log L_t$ is the part of the growth in aggregate output $\Delta \log Y$ that cannot be attributed to the pure technology effect $\tilde{\Lambda}_{t-1}' \Delta \log L_t$ of the growth of factors. This *distortion-adjusted Solow residual* weighs the change in $L_{f,t}$ by the cost-based Domar weight $\tilde{\Lambda}_{f,t}$ rather than by its share in aggregate income.[17]

A clear example of why the traditional Solow residual is an inappropriate measure of TFP growth in the presence of distortions is the vertical economy in Figure 1a. In that example, since the equilibrium is efficient, output increases one-for-one with an increase

[17] In practice, the traditional Solow residual attributes all non-labor income to capital (and has no room for profit income). The capital share of the traditional Solow residual would therefore be different from our capital share.

in labor. Since $\tilde{\Lambda}_L = 1$, this means that the distortion-adjusted Solow residual does not change in response to changes in labor supply. However, if the profit share is non-zero, then the traditional Solow residual $\Delta \log Y - \Lambda_L \Delta \log L$ increases in response to an increase in the quantity of labor, despite there being no change in either physical or allocative productivity. In other words, the Solow residual detects changes in aggregate TFP even though aggregate TFP has not changed.[18]

We now show how to decompose changes in aggregate TFP, captured by the distortion-adjusted Solow residual, into pure changes in technology and changes in allocative efficiency.

Proposition 1 (Distortion-Adjusted Solow Residual and Decomposition)**.** *To the first order, we can measure aggregate TFP with the distortion-adjusted Solow residual* $\Delta \log Y_t - \tilde{\Lambda}'_{t-1} \Delta \log L_t$ *and decompose it into pure changes in technology and changes in allocative efficiency as*

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

In the case of an efficient economy, the envelope theorem implies that the reallocation terms are zero (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 markups/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

[18] This point relates to Hall (1988, 1990) who considers environments where aggregate output is generated from factors $L = (L_1, \cdots, L_F)$ by a representative producer charging an aggregate markup μ via a structural aggregate production function $Y = AF(L)$. Hall showed that under these restrictive assumptions implying that there is no misallocation given factor supplies, changes in aggregate TFP can be recovered via a modified Solow residual $\Delta \log A = \Delta \log Y - (\mu\Lambda)' \Delta \log L$, where the growth of each factor $\Delta \log L_f$ is weighted by its share of total cost $\mu\Lambda_f$ rather than by its share of total revenue Λ_f. This can be seen as a particular case of our result since under these assumptions, $\mu\Lambda_f = \tilde{\Lambda}_f$ and so Hall's modified Solow residual coincides with our distortion-adjusted Solow residual.

the end of the period.[19,20]

3.2 Gollop et al., Basu-Fernald, and Petrin-Levinsohn

Seminal papers by Gollop et al. (1987), Basu and Fernald (2002), and Petrin and Levinsohn (2012) propose alternative decompositions of aggregate TFP changes into pure technology changes and changes in allocative efficiency for economies with markups/wedges. Just like our decomposition, changes in allocative efficiency seek to isolate the aggregate TFP changes which are due to the reallocation of resources to more or less distorted parts of the economy. In this section, we discuss how these alternative decompositions differ from ours, point out that they suffer from a conceptual problem, and argue that our decomposition is preferable.[21]

Gollop et al. only allow for markups/wedges in factor markets, but not in intermediate input markets. Unlike ours, this decomposition is therefore inapplicable when there are also markups/wedges in intermediate input markets. When there are no markups/wedges in intermediate input markets, it coincides with the decomposition of Petrin and Levin-

[19] We can cumulate the distortion-adjusted Solow residual over time to get

$$\sum_{t=t_0}^{t_1} \underbrace{\Delta \log Y_t - \tilde{\Lambda}'_{t-1} \Delta \log L_t}_{\Delta \text{ Aggregate TFP}} \approx \sum_{t=t_0}^{t_1} \underbrace{\tilde{\lambda}'_{t-1} \Delta \log A_t}_{\Delta \text{ Technology}} + \sum_{t=t_0}^{t_1} \underbrace{-\tilde{\lambda}'_{t-1} \Delta \log \mu_t - \tilde{\Lambda}'_{t-1} \Delta \log \Lambda_t}_{\Delta \text{ Allocative Efficiency}}.$$

This is a (Riemann-sum) approximation to the integral of (7). The approximation is exact in the limit of small time intervals (assuming the shocks are continuous in time) where we get the exact integral formula $\int_{t_0}^{t_1} \mathrm{d} \log Y_t - \int_{t_0}^{t_1} \tilde{\Lambda}'_t \, \mathrm{d} \log L_t = \int_{t_0}^{t_1} \tilde{\lambda}'_t \, \mathrm{d} \log A_t - \int_{t_0}^{t_1} \tilde{\lambda}'_t \, \mathrm{d} \log \mu_t - \int_{t_0}^{t_1} \tilde{\Lambda}'_t \, \mathrm{d} \log \Lambda_t$. Note that the cost-based Domar weights $\tilde{\Lambda}_t$ and $\tilde{\lambda}_t$ are updated over time, just like the factor shares Λ_t are updated over time in the cumulation of the traditional Solow residual. Cumulated pure changes in technology do reflect changes in the allocation of resources across producers over time via changes in $\tilde{\lambda}_t$. But these effects are not first-order, and can only be detected at higher orders of approximation or over time when the cumulated shocks are large. Furthermore, they are also present in efficient economies, and we refer the reader to Baqaee and Farhi (2017a) for a comprehensive analysis of these nonlinearities in this context.

[20] We can use the exact integral formula to give a second-order approximation of changes in real output $\Delta \log Y_t \approx (\tilde{\Lambda}'_t + \Delta \tilde{\Lambda}'_t/2) \Delta \log L_t + (\tilde{\lambda}'_t + \Delta \tilde{\lambda}'_t/2) \Delta \log A_t - (\tilde{\lambda}'_t + \Delta \tilde{\lambda}'_t/2) \Delta \log \mu_t - (\tilde{\Lambda}'_t + \Delta \tilde{\Lambda}'_t/2) \Delta \log \Lambda_t$ via a Törnqvist adjustment.

[21] Like ours, these approaches are economic decompositions based on explicit general equilibrium models. This feature distinguishes them from a popular set of purely statistical decompositions, such as, for example, Baily et al. (1992), Griliches and Regev (1995), Olley and Pakes (1996b), and Foster et al. (2001). These statistical approaches start by defining an ad-hoc aggregate TFP index. They (wrongly) take changes in this index to be changes in aggregate TFP. They then somewhat arbitrarily decompose them into technology effects and reallocation effects. Beyond the fact that the object that they decompose is not aggregate TFP (neither the distortion-adjusted nor the traditional Solow residual), these decompositions are problematic because they detect reallocation effects even in efficient economies and in economies where there are no reallocations of resources. This happens for example in the acyclic economies discussed below. See Osotimehin (2019) for related criticisms. Nevertheless, such decompositions provide moments in the data that can be used to distinguish between competing models (e.g. Bartelsman et al., 2013).

sohn (2012) up to the first order. In what follows, we therefore restrict our discussion to Basu-Fernald, and Petrin-Levinsohn.[22]

Whereas we decompose changes in aggregate TFP as appropriately measured by the distortion-adjusted Solow residual, these alternatives decomposition actually adopt different, and as we have argued inadequate, measures of changes in aggregate TFP: the traditional Solow residual for Petrin-Levinsohn and a Solow residual which strips out profits from capital income to compute the share of capital for Basu-Fernald. But the differences between these approaches and ours go beyond this observation. To make this clear, we conduct the whole discussion under the assumption that factor supplies are fixed. All the aforementioned notions of changes in aggregate TFP then coincide with changes in aggregate output.

At a high level, the difference between these alternative decompositions and our approach is that our approach decomposes output changes into two feasible counterfactual allocations: the allocation where the allocation matrix is held constant but productivities change, and the allocation where productivities are held constant but the allocation matrix changes. The alternative decompositions we discuss in this section do not have a similar interpretation in terms of feasible counterfactual allocations.

Each of these decompositions defines changes in pure technology by weighting microeconomic productivity shocks, and then constructs changes in allocative efficiency as the residual between the pure technology effect and aggregate output growth. To understand the differences between these decompositions, it is enough to compare their pure technology components since the allocative efficiency components are just the residual differences between aggregate output changes and pure technology changes. The pure technology terms for Petrin-Levinsohn are

$$\sum_k \lambda_{kt} \Delta \log A_{kt},$$

while for Basu-Fernald they are

$$\sum_k \lambda_{kt} \frac{1 - s_{Mkt}}{1 - \mu_{kt} s_{Mkt}} \Delta \log A_{kt},$$

where s_{Mkt} is k's expenditures on intermediate inputs as a share of its revenues in period t. These expressions, like ours, are weighted averages of technology changes $\Delta \log A_{kt}$ across

[22] Technically, Basu-Fernald only allows for one type of intermediate input ("materials"), which is restrictive. However, their decomposition can be extended to allow for multiple types of intermediate inputs. We use this extension in our discussion.

producers, but the weights are different. For Petrin-Levinsohn, the weight λ_{kt} attached to producer k is the usual sales share λ_{kt}. With Basu-Fernald, the weight $\lambda_{kt}(1 - s_{Mkt})/(1 - \mu_{kt}s_{Mkt})$ attached to a given producer k is its share in aggregate value added $\lambda_{kt}(1 - s_{Mkt})$ multiplied by a correction $1/(1 - \mu_{kt}s_{Mkt})$ involving its own intermediate input share in revenues s_{Mkt} and its own markup μ_{kt}.

In either case, the weights differ from the cost-based Domar weights $\tilde{\lambda}_{kt}$ of our decomposition. In fact, the (micro) information required to calculate their weights for producer k — its sales share, value added share, intermediate-input share, and markup — is not enough in general to calculate our cost-based Domar weight which requires more (macro) information — the whole input-output matrix and the whole set of markups. This (macro) information is used in our pure technology term precisely because it is corresponds to a counterfactual feasible allocation and not to a grouping of residual terms.[23] Overall, the pure technology components of these alternative decompositions are different from ours, and by implication, so are the allocative efficiency components.

To see the problem with the decompositions of Gollop et al., Basu-Fernald, and Petrin-Levinsohn 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 feasible allocation, simply because there is no option to allocate a given factor or good to different uses. This allocation is necessarily efficient. Markups/wedges have no effect on the allocation of resources, and as a result, there is no misallocation. As consequence, it is unambiguous a priori that there should be no changes in allocative efficiency in response to shocks.

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

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

This corollary follows immediately from the fact that by construction, acyclic economies hold the allocation matrix fixed. An important consequence of this corollary is that for

[23]Relative to the decompositions of Basu-Fernald, and Petrin-Levinsohn, there is a sense in which our decomposition 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 necessary to compute changes in allocative efficiency into changes in the markups/wedges and changes in the primary factor shares. By contrast, to isolate changes in allocative efficiency, the other decompositions require tracking the change of every output and input quantity for every producer (for brevity, we do not reproduce here their exact expressions for changes in allocative efficiency). This requirement leads to another disadvantage: it necessitates the observation of prices and quantities at the micro level, which is typically only possible in selected datasets. By contrast, our approach does not require such information.

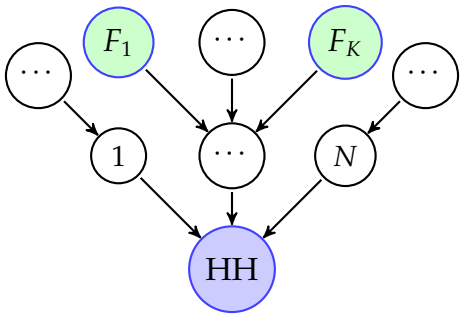


Figure 2: An acyclic economy, where the solid arrows represent the flow of goods. The factors are the green nodes. Each supplier (including factors) have at most one customer, whereas a single customer may have more than one supplier. Economies without cycles can be represented as directed trees with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

acyclic economies, our decomposition correctly identifies that there are no changes in allocative efficiency in response to productivity (or wedge shocks). On the other hand, the decompositions of Basu-Fernald, and Petrin-Levinsohn, which are not based on counterfactual allocations but instead on grouping of residual terms, do detect changes in allocative efficiency in response to productivity shocks. This follows from the fact that, as explained above, their pure changes in technology differ from ours. For a fully worked-out example, we refer the reader to Appendix E.

Of course, acyclic economies are not realistic. They are only useful because they reveal clearly a conceptual problem with these alternative decompositions, but this conceptual problem applies to generic economies.

In this section, we have highlighted the conceptual advantage of our approach over these alternative approaches. It turns out that these approaches also lead to very different empirical results. In Section 7, we empirically implement our decomposition in the US at the firm level in the presence of markups. In Appendix B we do the same for the Petrin-Levinsohn decomposition.[24] The results, displayed in Figures 6b and 8, illustrate material quantitative differences between the results of these two approaches. In particular, changes in allocative efficiency are much lower under the Petrin-Levinsohn approach than under ours.

4 Ex-Ante Structural Results

Our results so far rely on the response of factor shares $\mathrm{d}\log\Lambda$, which are endogenous objects determined in equilibrium. In this section, we show how $\mathrm{d}\log\Lambda$ depends on

[24] We could not perform the same comparison for Basu-Fernald because its implementation would require data on the quantity of intermediate-input changes at the firm level which are not available.

the (exogenous) primitives of the model. This allows us to use Theorem 1, not just for measurement, but also for prediction.

For expositional clarity, in this section, we focus on a special parametric class of models. In particular, we work with nested CES-economies, where every production and consumption function can be written as a nested-CES function (albeit 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. The examples in Section 2.4 (the vertical, horizontal, and round-about economies) are all special cases. In Appendix H, we show how to generalize these results to non-CES economies with arbitrary production functions.

4.1 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. A nested CES economy in standard form is one where every CES aggregator is treated as a separate producer in the input-output matrix. This means that every row of the input-output matrix i, or equivalently every producer i, has associated with it a unique elasticity of substitution parameter θ_i. For more details, see Appendix G.

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 each producer j, we can formally think of this as a covariance (when j corresponds to a factor, the operator always returns 0).

4.2 Ex-Ante Structural Results

Propositions 2 and 3 below characterize the macroeconomic impact of microeconomic productivity and markup/wedge shocks as well as their decomposition into pure changes in technology and changes in allocative efficiency as a function of the properties of the network and the elasticities of substitution.

Productivity Shocks

We first consider the case of productivity shocks.

Proposition 2 (Productivity Shocks)**.** *In response to a productivity shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} - \sum_g \tilde{\Psi}_{(g)}\frac{\mathrm{d}\log\Lambda_g}{\mathrm{d}\log A_k}, \frac{\Psi_{(f)}}{\Lambda_f}). \tag{9}$$

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

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

To gain some intuition for equation (9), it is useful to start with the case where there is a single factor $F = 1$. We call this factor labor L and apply the equation to $f = L$. In this case, the term $\sum_g \tilde{\Psi}_{(g)}\,\mathrm{d}\log\Lambda_g/\mathrm{d}\log A_k$ can be dropped in the covariance on the right-hand side of equation (9) because all the elements of this vector are identical, and so we get[25]

$$\frac{\mathrm{d}\log\Lambda_L}{\mathrm{d}\log A_k} = \sum_{j=1}^{N} (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}). \tag{10}$$

The numbers Ψ_{iL} are the payments to labor as a share of the total revenue 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/wedges are positive, we have $\Psi_{iL} < 1$ and $\Lambda_L < 1$. A low value of Ψ_{iL} indicates that on average, markups/wedges are high along the supply chain of i, and a low value of Λ_L indicates that on average, markups/wedges are high in the economy as a whole. The lower Ψ_{iL}/Λ_L, the more marked up is the supply chain of i relative to the economy as a whole.

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 expenditures across its inputs towards the producers with higher exposure $\tilde{\Psi}_{ik}$ to k, since their relative prices decline by more. If $Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}) < 0$, then those producers also have more marked up supply chains, and substitution by j lowers the labor share. Of course, these effects on the labor share must be summed over all

[25]This follows immediately from the fact that $\tilde{\Psi}_{iL} = 1$ for all i

producers j. If the overall effect is a decline in the labor share, then resources are overall reallocated towards more marked up parts of the economy, and there are positive changes in allocative efficiency which increase output over and above the pure technology effect of the shock.[26]

The complication brought on by having multiple factors is that relative factor prices change in response to substitution, and changes in factor prices can trigger additional substitution. To understand how things work with multiple factors, we rewrite equation (9) as the following linear system

$$\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} = \Gamma\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} + \delta_{(k)}, \tag{11}$$

with

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

and

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

We call δ the *factor share impulse matrix*. Its kth column encodes the 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 is independent of the source of the shock k. When there is only factor, Γ is a zero 1×1 matrix, and we are left with only the first-round effects $\delta_{(k)}$. However, when there are multiple factors, productivity shocks lead to changes in factor prices, which lead to further changes in factor income shares, etc. ad infinitum via higher-round effects captured by Γ.[27]

We now turn to the effects of markups/wedge shocks. We rely on the intuitions developed for the effects of productivity shocks and emphasize only the main differences.

Markup/Wedge Shocks

We now consider the case of markup/wedge shocks. We leverage the intuitions developed for the case of productivity shocks.

[26]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.

[27]See Example J.1 in Appendix J for a simple illustration.

Proposition 3 (Markup/Wedge Shocks)**.** *In response to a markup/wedge shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log\mu_k} = -\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}(\tilde{\Psi}_{(k)} + \sum_g \tilde{\Psi}_{(g)}\frac{\mathrm{d}\log\Lambda_g}{\mathrm{d}\log\mu_k}, \frac{\Psi_{(f)}}{\Lambda_f}) - \lambda_k\frac{\Psi_{kf}}{\Lambda_f}. \tag{12}$$

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

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

To gain some intuition for equation (12), it is once again useful to focus on the case where there is a single factor $F = 1$. We call this factor labor and apply the equation to $f = L$. In this case, exactly like for productivity shocks and for the same reason, the term $\sum_g \tilde{\Psi}_{(g)}\,\mathrm{d}\log\Lambda_g/\mathrm{d}\log\mu_k$ can be dropped in the covariance on the right-hand side of the equation, so we get

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

An increase in the markup/wedge of producer k changes the labor share via two effects. First, it changes relative prices like a negative productivity shock to producer k and triggers corresponding substitution effects, with an overall effect on labor demand and on the labor share captured by the first term on the right-hand side of equation (13). Second, it reduces the wage by reducing input and hence ultimately labor demand, which reduces the labor share according to the second term $-\lambda_k\Psi_{kL}/\Lambda_L$ on the right-hand side of equation (13).[28]

To get the effect on output, we must isolate the part of the reduction in the labor share $-\mathrm{d}\log\Lambda_L/\mathrm{d}\log\mu_k$ which is due to the reallocation of resources to more distorted parts of the economy. To do so, we must net out from the reduction in the labor share the mechanical reduction $\tilde{\lambda}_k$ in the labor share arising, for a given wage, from the increase in prices. Overall, the positive markup shock to producer k acts on output like a negative productivity shock to producer k combined with a release of resources $\lambda_k\Psi_{kL}/\Lambda_L$ (since unlike negative productivity shocks, positive markups do not destroy physical resources).

[28]See Example J.3 in Appendix J for a simple illustration with a Cobb-Douglas example.

4.3 Cobb-Douglas and Leontief

In this section, we isolate two benchmark specifications with no changes in allocative efficiency: Cobb-Douglas with productivity shocks and Leontief with markup/wedges shocks.

Proposition 4 (Cobb-Douglas and Leontief)**.** *If the economy is Cobb-Douglas with $\theta_j = 1$ for all j, then there are no changes in allocative efficiency in response to productivity shocks and we have*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k.$$

If the economy is Leontief with $\theta_j = 0$ for all j, then there are no changes in allocative efficiency in response to markup/wedge shocks and we have

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = 0.$$

The Cobb-Douglas result follows directly from Proposition 2. Basically, when all elasticities are unitary, the allocation matrix is not a function of productivity shocks. Productivity shocks only affect output via pure changes in technology captured by the cost-based Domar weights.

The Leontief result can be obtained by manipulating the expressions in Proposition 3 to obtain the following useful general formula

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = \sum_f \sum_j \lambda_j \mu_j^{-1} \theta_j Cov_{\tilde{\Omega}^{(j)}}\Big(\sum_k \tilde{\Psi}_{(k)} - \sum_g \tilde{\Psi}_{(g)} \frac{\mathrm{d}\log \Lambda_g}{\mathrm{d}\log \mu_k}, \frac{\tilde{\Lambda}_f}{\Lambda_f} \Psi_{(f)}\Big), \tag{14}$$

where $\mathrm{d}\log \Lambda_g / \mathrm{d}\log \mu_k$ is given by equation (12). From this, it is immediate that output does not change all elasticities $\theta_j = 0$. This means that for a Leontief economy, the allocation matrix is not a function of the wedges. Hence, shocks to markups/wedges have no effect on output.

4.4 Endogenous Productivities and Wedges

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.[29] For example, consider some parameter Z which gives rise to

[29] 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 7.2).

some endogenous vector of equilibrium productivities $A(Z)$ and markups/wedges $\mu(Z)$, for example a theory of innovation or market structure. These functions are not primitives of the model, instead they are equilibrium objects, the determination of which could be complex and interesting in and of itself. This is however not the focus of our paper. Our results can be used to understand the comparative statics of output with respect to Z by using the chain rule. The effect of a shock to Z can then be decomposed in two: how A and μ respond to a change in Z, 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 productivity or markup/wedge determination.

Our ex-post reduced form results,

$$\mathrm{d}\log Y = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'\,\mathrm{d}\log\mu - \tilde{\Lambda}'\,\mathrm{d}\log\Lambda, \tag{15}$$

can be used with no modification if the changes in productivities $\mathrm{d}\log A = (\mathrm{d}\log A/\,\mathrm{d}\,Z)\,\mathrm{d}\,Z$ and wedges $(\mathrm{d}\log\mu/\,\mathrm{d}\,Z)\,\mathrm{d}\,Z$ in response to a shock $\mathrm{d}\,Z$ are observed. Our ex-ante structural results

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\,Z} = \sum_k \frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k}\frac{\mathrm{d}\log A_k}{\mathrm{d}\,Z} + \sum_k \frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_k}\frac{\mathrm{d}\log\mu_k}{\mathrm{d}\,Z}, \tag{16}$$

can also be used for counterfactuals since the expressions for $\mathrm{d}\log Y/\,\mathrm{d}\log A_k$ and for $\mathrm{d}\log Y/\,\mathrm{d}\log\mu_k$ are still given by Propositions 2 and 3.

In Appendix F, we follow this approach to fully solve a simple example of an economy with endogenous markups inspired by Atkeson and Burstein (2008). We use our results to characterize by how much incomplete pass-through of productivity shocks to prices mitigates the impact of microeconomic productivity shocks on aggregate output.[30]

5 Distance to the Frontier and Other Efficiency Measures

In this section, we show how the ex-ante structural results of Section 4 can be used to calculate the "distance to the frontier," that is, the increase in aggregate productivity that would result from the elimination of all wedges. We then discuss the relation between our notion of changes in allocative efficiency and alternative measures such as the change in the distance to the frontier. We use the nested-CES setup of Section 4 throughout. We refer the reader to Appendix H for an explanation of how to generalize the results of this section to non-CES economies with arbitrary production functions.

[30] For a different example we refer the reader to the working paper of this version Baqaee and Farhi (2017b) where we show how to handle nominal rigidities. The model can be recast as a model with endogenous markups ensuring that the relevant prices stay constant and can be analyzed via the chain rule.

5.1 Distance to the Frontier

Recall that $Y(A,\mu) = \mathcal{Y}(A, \mathcal{X}(A,\mu))$ is the equilibrium level of output given the productivity vector A and the markup/wedge vector μ. By the first welfare theorem, the efficient level of output is $Y^* = Y(A,1)$ and the distance to the frontier is $\mathcal{L} = \log(Y^*/Y)$.

Second-Order Approximation

We use equations (12) and (14) derived in Section 4 to compute a second-order approximation of the distance to the frontier for small markups/wedges.

Proposition 5 (Distance to the Frontier)**.** *For small markups/wedges $\mu_k = \exp \Delta \log \mu_k$, the distance to frontier up to the second-order in the markups/wedges $\Delta \log \mu_k$ is*

$$\mathcal{L} \approx \frac{1}{2} \sum_j \lambda_j \theta_j Var_{\Omega^{(j)}}(\sum_k \Psi_{(k)} \Delta \log \mu_k) + \frac{1}{2} \sum_j \lambda_j \theta_j Cov_{\Omega^{(j)}}(\sum_g \Psi_{(g)} \Delta \log \Lambda_g, \sum_l \Psi_{(l)} \Delta \log \mu_l), \quad (17)$$

with $\Delta \log \Lambda_g = \sum_k (\mathrm{d} \log \Lambda_g / \mathrm{d} \log \mu_k) \Delta \log \mu_k$, where $\mathrm{d} \log \Lambda_g / \mathrm{d} \log \mu_k$ is given by equation (12)*, applied either at the distorted equilibrium with markups/wedges or at the efficient equilibrium without markups/wedges.*

This approximation gives the distance to the frontier as a function of the markups/wedges, the elasticities of substitution, and the production network. It generalizes the results of Hsieh and Klenow (2009) to arbitrary production networks and patterns of elasticities of substitution, and without imposing distributional assumptions on markups/wedges and productivities.

The approximation simplifies greatly in the case of a single factor $F = 1$. In this case, the terms on the second line of equation (17) drop out and we are left with[31]

$$\mathcal{L} \approx \frac{1}{2} \sum_j \lambda_j \theta_j Var_{\Omega^{(j)}}(\sum_k \Psi_{(k)} \Delta \log \mu_k). \quad (18)$$

Some general lessons are immediately available. First, all the terms scale with the square of the markups/wedges μ. There is therefore a sense in which misallocation increases with the markups/wedges. Second, all the terms scale with the elasticities of substitution θ of the different producers. There is therefore a sense in which higher

[31] Osotimehin and Popov (2018) present related results.

elasticities of substitution magnify the extent of misallocation. For example, the distance to the frontier is always zero when the economy is Leontief. Third, all the terms also scale with the sales shares λ of the different producers and with the square of the Leontief inverse matrix Ψ. There is therefore also a sense in which accounting for intermediate inputs magnifies the extent of misallocation from markups/wedges. Fourth, all the terms mix the markups/wedges, the elasticities of substitution, and of properties of the network. Hence, in general, the distance to the frontier depends on how the markups/wedges are distributed over the network and not simply on their univariate probability distribution, a point to which we will return in the context of a simple example below. Fifth, in general, the distance to the frontier is not zero when markups/wedges are uniform due to double-marginalization (for instance, consider the round-about economy in Figure 1c).

Relation to Hsieh-Klenow in a Simple Example

Applying this formula to the horizontal economy of Section 2.4, we get

$$\mathcal{L} \approx \frac{1}{2}\theta_0 Var_{\Omega^{(0)}}(\sum_k \Psi_{(k)}\Delta \log \mu_k) = \frac{1}{2}\theta_0 Var_\lambda(\Delta \log \mu). \tag{19}$$

In words, the distance to the frontier depends positively on the elasticity of substitution θ_0 and the sales-share weighted variance of the markups/wedges. The specialization of the general formula to this simple example clarifies a potential source of confusion: whether the joint distribution of productivities and wedges matters. Formula (19) shows that in general, the joint distribution of productivities and wedges does matter via the joint distribution of sales shares and markups/wedges.

In the particular case where producers differ only by their productivities and markups, and where productivities and markups are jointly log-normally distributed, Hsieh and Klenow (2009) show that the distance to the frontier can be approximated, up to the second order, by

$$\mathcal{L} \approx \frac{1}{2}\theta_0 Var(\Delta \log \mu). \tag{20}$$

In contrast to formula (19), this formula implies that the joint distribution of (A, μ) is irrelevant, and only the marginal distribution of the markups/wedges matters.

Only under the assumption that A and μ are jointly log-normal is equation (20) consistent with equation (19). In this special case, $Var(\Delta \log \mu)$ is, to a second-order approximation, equal to $Var_\lambda(\Delta \log \mu)$.[32] Formula (19) is a generalization of formula (20) beyond the

[32]The reason for the irrelevance of the joint distribution of markups and sales shares in this lognormal case is best understood using the decomposition $Var_\lambda(\Delta \log \mu) = E_\lambda[Var(\Delta \log \mu|\lambda)] + Var_\lambda[E(\Delta \log \mu|\lambda)]$.

jointly log-normal case. It captures the general dependence of misallocation on the joint distribution of sales and wedges. It sheds light on the results of Restuccia and Rogerson (2008), who do not assume log-normality, and who find that the correlation of wedges and sales matters greatly for their results. See Appendix A for a worked out example.

5.2 Three Different Notions of Changes in Allocative Efficiency

In this section, we define and discuss different notions of changes in allocative efficiency to the one defined in Section 7.2. We discuss their interpretations, differences, and connections.

Definitions and Interpretation

In general, changes in allocative efficiency can defined relative to a benchmark alternative allocation rule. For example, in Theorem 1, we defined changes in allocative efficiency relative to the allocation of resources in the previous period:

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

So, if resources are reallocated in such a way that they yield more output relative to the previous period for given microeconomic technologies and factor supplies, then this represents an improvement in the efficiency of the allocation relative to the previous allocation matrix. We call this *changes in allocative efficiency due to reallocation*.

However, other notions of changes in allocative efficiency can be defined relative to other allocation rules. For example, one could also compute changes in allocative efficiency relative to the allocation where wedges are kept constant:

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

This notion, due to Osotimehin (2019), measures changes in the efficiency of the allocation relative to the allocation where wedges are held constant, and may be called *changes in*

Under joint log-normality, changes in the variance of sales or in its covariance with markups/wedges create exactly-offsetting changes in $E_\lambda[Var(\mu|\lambda)]$ and $Var_\lambda[E(\mu|\lambda)]$ up to the second order in the markups/wedges. For example, increasing the covariance between sales and markups/wedges reduces the average variance of markups/wedges given sales $E_\lambda[Var(\mu|\lambda)]$ but increases the variance of average markups given sales $Var_\lambda[E(\mu|\lambda)]$, but the two effects exactly offset each other up to the second order in the markups/wedges.

allocative efficiency due to changes in wedges.

Finally, an alternative is to define allocative efficiency relative to an optimal allocation:

$$\mathrm{d}\log Y = \underbrace{\mathrm{d}\log Y^*}_{\Delta\text{Technology}} + \underbrace{\mathrm{d}\log Y - \mathrm{d}\log Y^*}_{\Delta\text{Allocative Efficiency}},$$

where Y^* is the efficient level of output. This notion goes back to Debreu (1951) and Farrell (1957). We call this *changes in allocative efficiency relative to the frontier*, because it measures changes in the efficiency of the allocation relative to the optimal allocation.

These three different concepts of changes in allocative efficiency are all admissible. Actually, they coincide when there are only changes in the wedges/markups.[33] By contrast, they typically differ when there are also changes in productivities. They then have different economic interpretations based on different underlying counterfactuals and different data requirements.

Differences in Results in a Simple Example

These three measures of changes in allocative efficiency may easily move in different directions. To see how this could happen, consider a horizontal economy with two producers and $\theta_0 > 1$. Suppose that $\mu_1 > \mu_2$ and $A_1/\mu_1 < A_2/\mu_2$ so that $\lambda_1 < 1/2 < \lambda_2$. Write $\mu_i = \exp(\Delta \log \mu_i)$ and consider for simplicity the case of small markups/wedges with $\Delta \log \mu_i$ close to 0. Now suppose that the first producer receives a positive productivity shock $\mathrm{d}\log A_1 > 0$, but that A_2, μ_1, and μ_2 remain unchanged. Allocative efficiency due to reallocation (the first concept) *increases* by approximately $\lambda_1\lambda_2(\theta_0 - 1)(\Delta \log \mu_1 - \Delta \log \mu_2)\,\mathrm{d}\log A_1$ because workers are reallocated from the lower-markup producer 2 to the higher-markup producer 1. Allocative efficiency due to changes in wedges (the second concept) *stays unchanged* because markups/wedges do not change. Allocative efficiency relative to the frontier (the third concept) *decreases* by approximately $(1/2)\theta_0 Var_\lambda(\Delta \log \mu)(1 - 2\lambda_1)(\theta_0 - 1)\,\mathrm{d}\log A_1$ because the sales-weighted dispersion of markups increases.

[33] It follows from this observation that the first and second concepts can be used to compute the level of the distance to the frontier. Take an economy with productivity vector A and markup vector μ. Consider a transformation of each markup $\hat{\mu}_i(t) = (1-t)\mu_i + t$. When $t = 0$, this transformation leaves markups as they are. On the other hand, all distortions are eliminated at $t = 1$. The distance to the frontier is then the integral along this path of changes in allocative efficiency due to reallocation/wedges:

$$\mathcal{L} = \int_0^1 \frac{\partial \log \mathcal{Y}(A, \mathcal{X}(A, \hat{\mu}(t)))}{\partial \log \mu} \frac{\mathrm{d}\log \hat{\mu}(t)}{\mathrm{d}\,t}\,\mathrm{d}\,t = \int_0^1 \frac{\partial \log Y(A, \hat{\mu}(t))}{\partial \log \mu} \frac{\mathrm{d}\log \hat{\mu}(t)}{\mathrm{d}\,t}\,\mathrm{d}\,t.$$

Differences in Data Requirements

Other than answering different questions, these three different concepts also have different information requirements. The first and second concepts are both local notions computing the effects of local changes along the equilibrium path. The third concept is a global notion comparing local changes in equilibrium output to changes in the globally efficient frontier.[34] The first concept requires the least information: the network and the markups/wedges. The second concept requires more information: the network, the markups/wedges, and the elasticities. The third concept requires the most information: not just the network, the markups/wedges, and the elasticities, but also the full global nonlinear functional forms of the production and final demand functions. However, with small markups/wedges, the full global nonlinear functional forms are not needed and the network, markups/wedges, and elasticities are enough to approximate it. Given the required information, all these notions can be computed using the results in Sections 2 (first concept) and 4 (second and third concepts).

6 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 three important issues that have to be confronted: the identification of the factors of production, the level of aggregation of the data, and the mapping of model wedges to the data frictions. We discuss these issues in turn after reviewing data requirements. We also discuss how to account for issues like 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

[34] For growth accounting purposes, the first two concepts seem more natural since the change in aggregate TFP along the equilibrium path measured by the traditional or distortion-adjusted Solow residual is also a local notion.

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). 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 on 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 $\mathrm{d} \log Y / \mathrm{d} \log A = \sum_{i \in S} \lambda_i$. In other words, the total sales of all producers in S will yield the impact of an aggregate shock to all producers in S.[35] So, we only need to observe sales data at the level of disaggregation at which a

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

shock occurs.

This aggregation property does not hold for distorted economies, even in the Cobb-Douglas or acyclic cases where we do not need to account for changes in allocative efficiency. Unlike sales, cost-based Domar weights $\tilde{\lambda}$ are not directly observable, and instead need to be computed from input-output data *at* the level of disaggregation at which the 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 I for a worked-out example.[36] When we discuss quantitative and empirical applications in Section 7, we return to how aggregation bias can affect our results.

Mapping Model Wedges to the Data Frictions

Throughout, we model distortions via wedges. The wedges act like linear taxes, the revenues of which are rebated lump sum.[37] Beyond actual taxes, these wedges can also implicitly capture frictions preventing the reallocation of resources.

In mapping our results to the data, we assume that expenditures by i on inputs from j are recorded *gross* of the wedge τ_{ij}. This is an accounting convention which does not change anything about the underlying allocation or real GDP. It is the natural convention for markups where sales are recorded inclusive of markups, but for other distortions, it 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.

To make this point very clearly, consider two examples based on different modifications of the horizontal economy in Section 2.4. The first example introduces three modifications to capture an economy with financial frictions: the factor of production is capital K instead of labor L; producers do not charge any markup; and producers incur a non-pecuniary proportional cost $\delta_i \geq 1$ from using capital. The cost δ_i is a wedge which changes the cost of capital perceived (but not paid) by producer i.[38] To apply our formulas, the revenue-based input-output matrix which would be observed would need to be adjusted as follows: for each producer i, introduce a new fictitious producer $\phi(i)$ purchasing capital, selling it to

[36] In Section 7, 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).

[37] 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.

[38] In some cases, financial frictions take the form of credit constraints, in which case they have to be interpreted as the Lagrange multipliers on the constraints in the individual firms' cost minimization problem.

producer i, and charging a markup $\mu_{\phi(i)} = \delta_i$; take the expenditure share of i on $\phi(i)$ to be $1 = \delta_i^{-1}\delta_i$, which is the expenditure share δ_i^{-1} of i on capital in the data *reflated* by a factor δ_i; take the expenditure share of $\phi(i)$ on capital to be equal to δ_i^{-1}. The sales shares λ_i of producer i is the same in the data and in the adjusted model. The capital share in the data is also the same as in the adjusted model $\sum_i \lambda_i \delta_i^{-1} \leq 1$. This is because in the data, the expenditure share on capital of producer i is recorded *net* of the wedge δ_i, and therefore corresponds to the expenditure share on capital of producer $\phi(i)$ in the adjusted model.

The second example introduces the following modification to capture an economy with compensating differentials: producers do not charge any markup; and labor incurs a non-pecuniary proportional cost δ_i when employed by producer i. The cost $\delta_i \geq 1$ is a wedge which changes the cost of labor w_i paid by producer i, with $w_i/w_j = \delta_i/\delta_j$. To apply our formulas, the revenue-based input-output matrix which would be observed would need to be adjusted as follows: for each producer i, introduce a new fictitious producer $\phi(i)$ purchasing labor, selling it to producer i, and charging a markup $\mu_{\phi(i)} = \delta_i$; take the expenditure share of i on $\phi(i)$ to be 1, which is the expenditure share of i on labor in the data; and take the expenditure share of $\phi(i)$ on labor to be δ_i^{-1}. The sales shares λ_i of producer i is the same in the data and in the adjusted model. But now the labor share in the data $\sum_i \lambda_i = 1$ is not the same as in the adjusted model $\sum_i \lambda_i \delta_i^{-1} \leq 1$. This is because in the data, the expenditure share on labor of producer i is recorded *gross* of the wedge δ_i, and is therefore higher by a factor δ_i than the expenditure share on labor of producer $\phi(i)$ in the adjusted model. The labor share in the data therefore needs to be *deflated* by a factor $(\sum_i \lambda_i \delta_i^{-1})^{-1} \geq 1$.

7 Applications

In this section, we pursue some quantitative applications of our results, focusing on markups as the source of inefficiency. These applications are proofs of concept for the usefulness of the theoretical framework laid out above.

In Section 7.1, we describe our data and its mapping to the theory. In Section 7.2, we measure changes in allocative efficiency in the US over time, and decompose the distortion-adjusted Solow residual into changes in pure-technology and changes in allocative efficiency. In Section 7.3, we use our structural results to measure the distance to the frontier in a version of the parametric model calibrated to match firm-level size and markup data as well as industry-level input-output data.[39]

[39] In Appendix K, we also use the structural model to estimate the amount of macroeconomic volatility arising from microeconomic shocks. We find that markups materially affect the impact of microeconomic

7.1 Data

We apply our results to US data assuming markups are the only wedges in the economy. The details of how we clean and map the data to the model are in Appendix C. Here, we give a brief account of how we proceed.

To apply our formulas, we need a measure of the input shares and markup of each producer. This data is not available at the firm level and so we are forced to make compromises and rely on a large degree of imputation.

We use the annual US input-output data from the BEA, dropping the government, non-comparable imports, and second-hand scrap industries. The dataset contains industrial output and inputs from 1997 to 2015 with 66 industries. We also use sales and markup data from publicly listed firms in Compustat. We assume that each firm produces a single output. We explain below the different methods that we use to estimate markups for Compustat firms.

We make the following assumptions. Compustat firms are assigned to BEA industries. The outputs of the different firms in a given industry are aggregated into an industry bundle using a homothetic aggregator. The inputs used by firms and by households are these industry bundles. All the firms in a given industry have the same production functions up to Hicks-neutral productivity shifters. Finally, we assume that within each industry, the sales-share-weighted distribution of markups and its transition matrix can be extrapolated from Compustat. These assumptions imply that we can construct both Ω and $\tilde{\Omega}$ using industry-level input-output data from the BEA and firm-level data from Compustat.

With a slight abuse of notation, we denote industry variables with capital letters (e.g. I) and firm-level variables with lower-case letters (e.g. i), where $i \in I$ to means firm i is in industry I. In the data, we observe: the industry-level sales shares λ_I and input-output entries Ω_{IJ} for industries I and J; the sales shares of the Compustat firms i in industry I, which we rescale so that $\sum_{i \in I} \lambda_i = \lambda_I$; and the markup μ_i of Compustat firm i (see below). We then compute: industry-level markups as $\mu_I = (\sum_{i \in I} (\lambda_i / \lambda_I) \mu_i^{-1})^{-1}$; industry-level cost-based input-output matrix entries as $\tilde{\Omega}_{IJ} = \mu_I \Omega_{IJ}$; firm-level cost-based input-output matrix entries as $\tilde{\Omega}_{ij} = (\lambda_j / \lambda_J) \tilde{\Omega}_{IJ}$ when firms i and j belong to industries I and J; and firm-level revenue-based input-output matrix entries as $\Omega_{ij} = \mu_i^{-1} \tilde{\Omega}_{ij}$.

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

For the estimates of firm-level markups, we use three different measures: (1) the user-cost approach (UC), (2) the production function estimation approach (PF), and (3) the accounting profits approach (AP). Our benchmark estimates use the UC approach, which we compute following Gutiérrez and Philippon (2016) and Gutierrez (2017). This markup estimate relies on measuring each firm's total costs and computing the markup by comparing total costs to sales. This method is also used by Foster et al. (2008), and (as a robustness check) by De Loecker et al. (2019). Measuring total costs requires estimating each firm's user-cost of capital and capital stock. We compute the user-cost of capital in the spirit of Hall and Jorgenson (1967) where the rental price takes into account, not just the riskless rate and industry-specific depreciation, but also industry-level risk premia.

The second measure comes from estimates of the production function similar to the benchmark estimates from De Loecker et al. (2019) and following the methodology of De Loecker and Warzynski (2012). These markups are given by the ratio of the elasticity of the production function to a variable input (estimated at the industry-year level) to the share of that input in revenues (estimated at the firm-year level). This measure bypasses having to estimate a firm's total costs by instead relying on the firm's first-order condition that equates the output elasticity with respect to a variable input to the expenditures on that input as a share of costs (in our case, the variable input is the "cost of goods sold").

The last set of estimates are the accounting profits approach, where we assume that operating income is equal to profits, and hence markups are equal to sales divided by costs (sales minus profits). Although quite simplistic, this approach has the virtue of requiring very little manipulation of the raw data and is similar to the method employed by Harberger (1954). Since this accounting approach yields roughly similar conclusions as the more sophisticated approaches, it suggests that firms' reports of their operating income (which is all the AC markups rely on) is the important source of the variation across the different approaches.

Each markup series comes with its own pros and cons. The UC markups require industry-level estimates of the risk premium and depreciation, both of which are notoriously difficult to measure, and assumes that all inputs are flexible in production. The PF markups on the other hand, rely on more parametric methods, and face tough identification challenges. On the other hand, while the UC and AP markups capture "average" markup margins (by stripping out expenses from revenues), the PF 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 on these grounds.

We use the UC markups for our benchmark numbers, and we report numbers for the

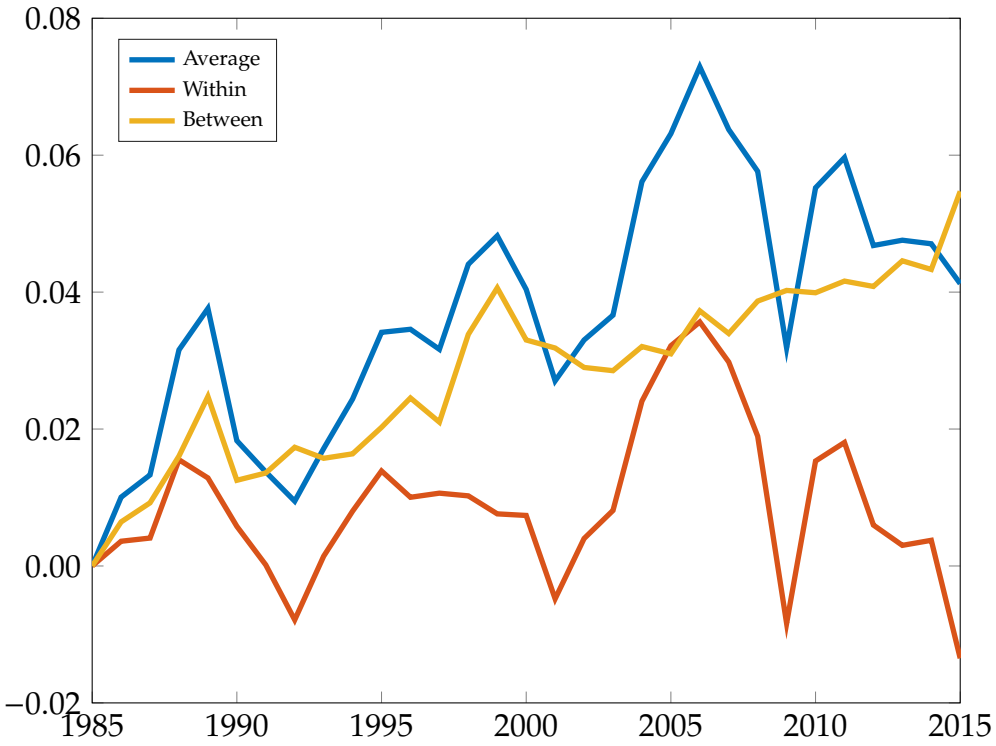


Figure 3: Decomposition of the increase in the average markup into a between and a within effect, using the user-cost approach markup data. All the changes are cumulated over time.

other two markup series in the tables and in Appendix A.[40] Our hope is that by reporting estimates using three different measures, we uncover patterns that are reasonably robust.

The three markup series give different levels of markups: the UC markups are the smallest (and average around 5%), the AP markups are higher (averaging around 10%), and the PF markups are the largest (averaging around 16%).

Despite their differences, all three markup series share some commonalities. Average markups (as measured by the profit share) 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, this increase in the average markup is driven by compositional effects. Namely, when we decompose changes in the average markup into a within-firms effect and a between firms effect, we find that the trend is overwhelmingly due to the between 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 3 illustrates this decomposition for the UC markups.[41]

[40]Note that our method allows for capital-biased technical change. In particular, we 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 6.

[41]The average markup is computed as the harmonic sales-weighted average $(\hat{\lambda}_t' \mu_t^{-1})^{-1}$ of firm markups, where the vector of weights $\hat{\lambda}_t$ is proportional to the vector λ_t of firm sales shares, with the constant of

This accords with the contemporaneous findings of Autor et al. (2019) and Vincent and Kehrig (2019) who find a similar composition effect in firms' labor shares. It has also been corroborated by De Loecker et al. (2019), who find a similar pattern in US markups. As we shall see shortly, this composition effect has profound implications for the evolution of aggregate TFP, since it implies positive contributions of changes in allocative efficiency due to reallocation.

7.2 Decomposing Aggregate TFP Growth

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

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 as well as from variable capacity utilization which matter more at business-cycle frequencies.

Conditional on markups and the input-output matrix at a given point in time t, we calculate $-\tilde{\Lambda}'_{t-1}\Delta \log \Lambda_t$ from $t-1$ to t using the change in observed factor income shares. Then we can decompose aggregate TFP using Theorem 1. The results are plotted in Figure 4 using the UC markups. The red line (allocative efficiency) is measured as $-\tilde{\lambda}'_{t-1}\Delta \log \mu_t - \tilde{\Lambda}'_{t-1}\Delta \log \Lambda_t$, and the yellow line (pure technology) is the residual difference between the red line and the distortion-adjusted Solow residual.

Since the start of the sample, cumulated changes in allocative efficiency due to reallocation account for about 50% of aggregate TFP growth as measured by the cumulated Solow residual. This implies that pure changes in technology, which are computed as a residual, also account for about 50% of aggregate TFP growth.

The fundamental intuition behind these large cumulated improvements in allocative efficiency is that the increase over time in average markup is largely driven by a composition effect, whereby firms with high markups, which were too small to begin with have been getting larger.

In Figure 5, we repeat our exercise but instead using data that has been aggregated to the industry-level. That is, we assume that each industry contains a representative firm

proportionality given by the inverse of the sum of the firm sales shares, so that the weights sum up to one, and μ_t is the vector of firm markups. This is the correct way to aggregate markups to match the aggregate profit share. The change in the average markup is computed as $d\log((\hat{\lambda}'_t\mu_t^{-1})^{-1})$. The contribution of the within effect is $(\hat{\lambda}'_t\mu_t^{-1}d\log\mu_t)/(\hat{\lambda}'_t\mu_t^{-1})$. The contribution of the between effect is the residual. 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.

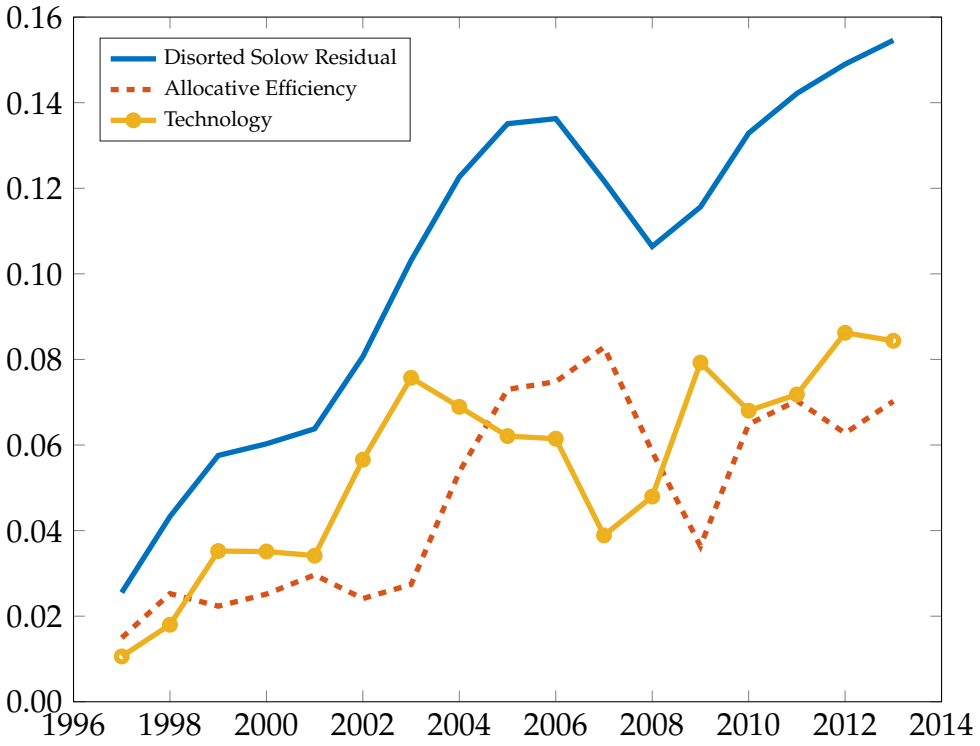


Figure 4: Cumulative decomposition of changes in aggregate TFP (distortion-adjusted Solow residual) into pure changes in technology and changes in allocative efficiency along the lines of equation (7), with markups obtained from the user-cost approach.

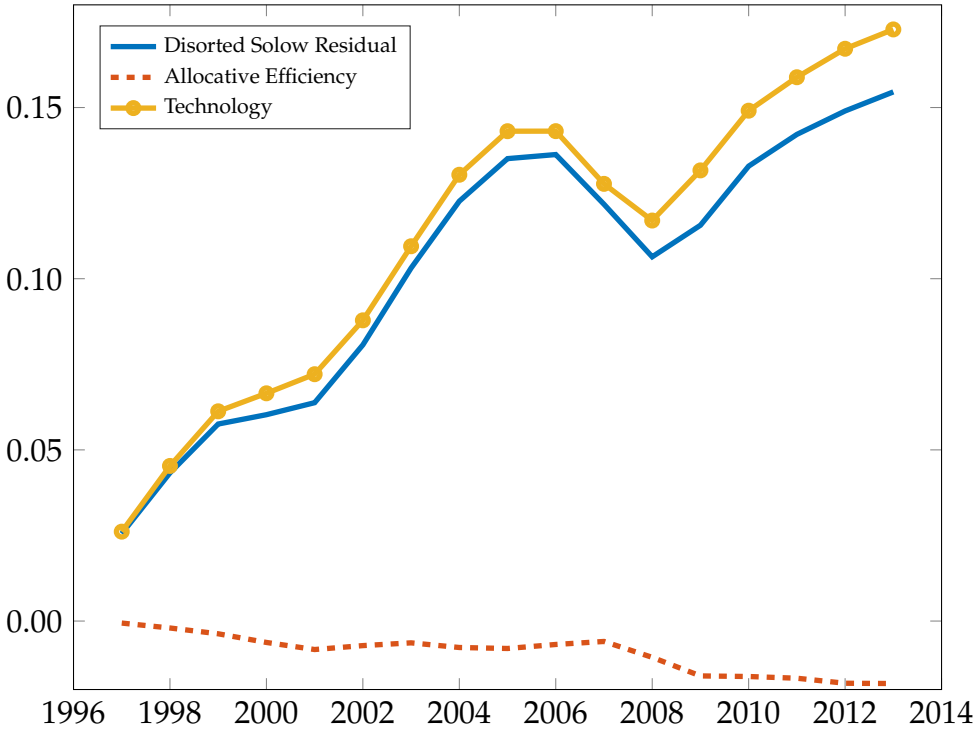


Figure 5: The same decomposition as in Figure 4 but the user-cost markup data was first aggregated to the industry-level. In effect, this decomposition assumes that each industry contains a representative firm (ruling out reallocations within the industry).

charging a markup equal to the harmonic industry average. In this case, the cumulated contributions of changes in allocative efficiency due to reallocation are drastically reduced and even flip signs. Basically, the compositional effects in Figure 3, and the reallocation effects in Figure 4, have occurred across firms within industries and not across industries. Once we aggregate up to the industry-level the pattern disappears, underscoring the importance of disaggregation in inefficient economies, as discussed in Section 6.

These patterns are also borne out when we use the PF and AP markups, although the magnitudes are somewhat different (see Figures 7a- 6b in Appendix A). The contribution of changes in allocative efficiency due to reallocation to aggregate TFP growth is large at the firm level, and becomes negligible at the industry level. In Appendix B, we use the Petrin and Levinsohn (2012) decomposition on the firm-level data and also find that it vastly undermeasures the importance of reallocation effects.

7.3 Gains From Eliminating Markups

In this section, we calibrate a simplified version of the parametric model presented in Section 4 and use our structural results to compute the gains to aggregate TFP from eliminating markups. Our calculation in this section quantitatively illustrates the intuition from Proposition 5, namely that accounting for intermediate inputs and higher elasticities of substitution (say, across firms within an industry as opposed to across industries) magnifies the losses from misallocation.

To calibrate the model, we need estimates for industry-specific firm-level and industry-level structural elasticities of substitution. Unfortunately, detailed disaggregated estimates of these elasticities do not exist. We consider a nested CES structure where each firm i in industry j produces using a CES aggregator of value-added and intermediate inputs with an elasticity of substitution θ. The value-added component is a CES aggregator of labor and capital with an elasticity of substitution η. The intermediate input component is a CES aggregator of inputs from other industries with an elasticity of substitution ε .Finally, inputs purchased by firms in industry j from industry k are a CES aggregate of all varieties in that industry with elasticity of substitution ξ.

Following our previous work in Baqaee and Farhi (2017a), and drawing on estimates from Atalay (2017) and Boehm et al. (2014), we set $\theta = 0.5$, $\theta_0 = 0.9$, and $\varepsilon = 0.2$. We set $\eta = 1$ which is a focal point in the literature about the micro-elasticity of substitution between labor and capital. 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. We also experiment with $\xi = 4$, reflecting a lower elasticity of substitution

	User Cost (UC)	Accounting (AP)	Production Function (PF)
2015	13%	11%	25%
1997	3%	5%	23%

Table 1: Gains from eliminating markups for the beginning and end of our sample.

across firms within each industry.

We use the calibrated model to approximate the gains to aggregate TFP from eliminating markups following the result in Proposition 5. This corresponds to the sum of the areas of the deadweight loss triangles in our general equilibrium model. The results are reported in Table 1.

Using the benchmark UC markups, we find that eliminating markups, holding fixed technology, would increase aggregate TFP by around 13%. The AP markups imply somewhat smaller gains of 11% and the PF markups imply larger gains of 25%.

Our estimate that eliminating markups in the US economy in 2015 would increase TFP by between 11% and 25% raises the estimated cost of monopoly distortions by two orders of magnitude compared to the famous estimates of 0.1% of Harberger (1954). The reasons for this dramatic difference can be traced back to Proposition 5. First, we use firm-level data, whereas Harberger only had access to sectoral data, and the dispersion of markups is higher across firms within a sector than across sectors. Second, a key elasticity of substitution is higher in our exercise than in Harberger's since it applies across firms within a sector rather than across sectors. Finally, we properly take into account the input-output structure of the economy to aggregate the numbers in all industries whereas Harberger only focused on manufacturing and did not offer a full treatment of 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. But even if markups do play an important role in incentivizing innovation, they also distort the allocation of resources and our calculation shows that this effect is quantitatively large.

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 UC markups, we find that the gain from eliminating markups is 3% in 1997, much smaller than the 13% in 2014. As we described in Section 5.2, this finding is entirely consistent with our other finding that changes in allocative efficiency due to reallocation have contributed positively to aggregate TFP growth since the start of the sample. Intuitively, the

	Benchmark	CD + CES	$\xi = 4$	Cobb-Douglas	No IO	Sectoral
UC	13%	14%	8%	3%	5%	0.7%
AP	11%	12%	6%	3%	5%	1%
PF	25%	29%	14%	10%	14%	4%

Table 2: Gains in aggregate TFP from eliminating for different markup series and different calibrations. The first column are the benchmark numbers in Table 1. The third column sets all elasticities except ξ equal to one. The third column lowers the elasticity of substitution ξ between firms within each industry. The fourth column sets all elasticities in the model to be equal to one. The fifth column recalibrates the model using value-added production functions, ignoring the input-output connections, similar to Hsieh and Klenow (2009). The final column uses data aggregated to the sectoral level rather than firm-level similar to Harberger (1954).

distance to the frontier has increased because markups have become more dispersed over time. By contrast, changes in allocative efficiency have contributed positively to measured aggregate TFP growth because high-markup firms have been getting larger. The frontier has shifted faster, resulting in a simultaneous increase in the distance to the frontier.

In Table 2, we repeat the markup reduction exercise for some alternative specifications of the structural model. The first column repeats the benchmark numbers from Table 1. The second column considers a situation where all elasticities of substitution are set equal to one except the one across firms within industries (which is kept at 8). The distance to the frontier is not very sensitive to this change, which is in line with Proposition 5 since most of the dispersion in markups is across firms within a sector and not across sectors or factors, and since the change keeps the elasticities across industries and factors relatively low (from values below one to one).

The third column shows the gains implied by a much lower elasticity of substitution across firms within industries $\xi = 4$ rather than $\xi = 8$. Halving this elasticity more or less halves the losses from misallocation, which is in line with Proposition 5 since most of the dispersion between markups is across firm within industries.

The fourth column shows the gains implied by a Cobb-Douglas specification of the model which imposes that all elasticities of substitution are equal to 1, including those across firms within industries. This model conforms to the assumption in Harberger (1954) that all demand curves are unit elastic, though we use a more detailed model with multiple factors, intermediate inputs, and disaggregated at the firm level. Once again, moving to the Cobb-Douglas specification significantly reduces the losses from misallocation.

The fifth column shows the gains implied by using value-added production functions which ignore the role of the production network. We find that working with value-added

productions can cut the estimated gains from reducing markups by more than half, which is in line with Proposition 5 given that the distance to the frontier scales in both the sales shares and the Leontief inverse. 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 distortions.

Finally, the last column shows the losses from using a sectoral version of the model — one which aggregates each industry to contain a single firm charging a markup equal to the harmonic average of markups in that industry. This model is similar to the one used by Harberger (1954), although we stray from his unit-elastic demand assumption, model the whole economy rather than just manufacturing, and properly take input-output linkages into account. But at any rate, as expected, the numbers in this case are much smaller, and not drastically dissimilar to Harberger's 0.1% estimate. All in all, Table 2 shows the importance of modeling micro-level heterogeneity, the input-output network, and realistic elasticities when using structural models to measure economic waste.

8 Robustness and Extensions

In Appendix H, we extend our results to address some limitations. We discuss, and in some cases fully characterize, how our basic framework can be adapted to handle the following complications: arbitrary non-CES production functions, elastic factor supplies, capital accumulation, adjustment costs, variable capacity utilization, and nonlinearities. All these issues introduce additional forces 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.

9 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 how, locally, the impact of a shock can be decomposed into a pure technology effect and an allocative efficiency effect. We also show how to compute the distance to the efficient frontier. 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 apply our findings to the firm-level markups in the US. We find that from 1997-2015, allocative efficiency in the US accounts for about 50% of aggregate TFP growth. We also find that the gains from reducing markups have increased since 1997, and that eliminating markups would increase aggregate TFP by around 15%. These numbers are two orders of magnitude higher than classic estimates like those of Harberger (1954).

References

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

Asker, J., A. Collard-Wexler, and J. De Loecker (2019). (mis) allocation, market power, and global oil extraction. *American Economic Review 109*(4), 1568–1615.

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

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

Autor, D., D. Dorn, L. Katz, C. Patterson, and J. Van Reenen (2019). The fall of the labor share and the rise of superstar firms.

Baily, M. N., C. Hulten, and D. Campbell (1992). Productivity Dynamics in Manufacturing Plants. *Brookings Papers on Economic Activity 23*(1992 Micr), 187–267.

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.

Baqaee, D. R. and E. Farhi (2018). Macroeconomics with heterogeneous agents and input-output networks. Technical report, National Bureau of Economic Research.

Baqaee, D. R. and E. Farhi (2019a). Aggregating with Scale: Entry, Exit, and Selection. Working paper, Harvard.

Baqaee, D. R. and E. Farhi (2019b). Networks, Barriers, and Trade. Technical report.

Barkai, S. (2019). Declining labor and capital shares.

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 (2019). 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.

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.

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

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

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica: Journal of the Econometric Society*, 273–292.

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

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

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.

Farhi, E. and F. Gourio (2018). Accounting for macro-finance trends: Market power, intangibles, and risk premia. Technical report, National Bureau of Economic Research.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society: Series A (General) 120*(3), 253–281.

Foster, L., J. Haltiwanger, and C. Syverson (2008). Reallocation, firm turnover, and efficiency: Selection on productivity or profitability? *American Economic Review 98*(1), 394–425.

Foster, L., J. C. Haltiwanger, and C. J. Krizan (2001, September). Aggregate Productivity Growth: Lessons from Microeconomic Evidence. In *New Developments in Productivity Analysis*, NBER Chapters, pp. 303–372. National Bureau of Economic Research, Inc.

Gollop, F. M., B. M. Fraumeni, and D. W. Jorgenson (1987). Productivity and us economic growth.

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

Griliches, Z. and H. Regev (1995, January). Firm productivity in Israeli industry 1979-1988. *Journal of Econometrics 65*(1), 175–203.

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. (1988). The relation between price and marginal cost in us industry. *Journal of political Economy 96*(5), 921–947.

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

Hall, R. E. and D. W. Jorgenson (1967). Tax policy and investment behavior. *American economic review 57*(3), 391–414.

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

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 (2019). Labor share decline and intellectual property products capital.

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

Liu, E. (2019). Industrial policies and economic development. Technical report.

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

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

Olley, G. S. and A. Pakes (1996b, November). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica 64*(6), 1263–1297.

Osotimehin, S. (2019, April). Aggregate productivity and the allocation of resources over the business cycle. *Review of Economic Dynamics 32*, 180–205.

Osotimehin, S. and L. Popov (2018). Misallocation and intersectoral linkages. Technical report.

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

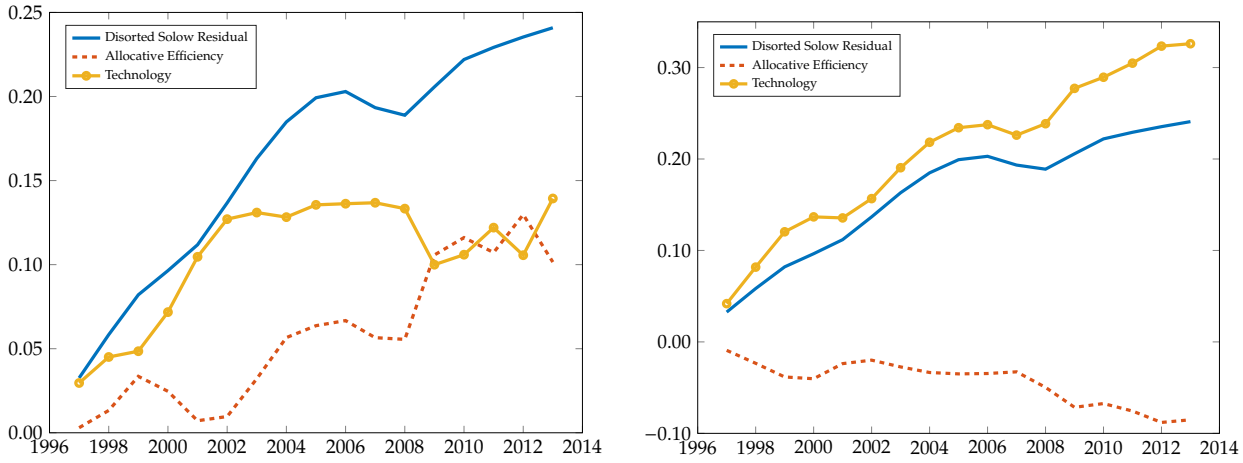
Restuccia, D. and R. Rogerson (2008). Policy distortions and aggregate productivity with heterogeneous establishments. *Review of Economic dynamics 11*(4), 707–720.

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.

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

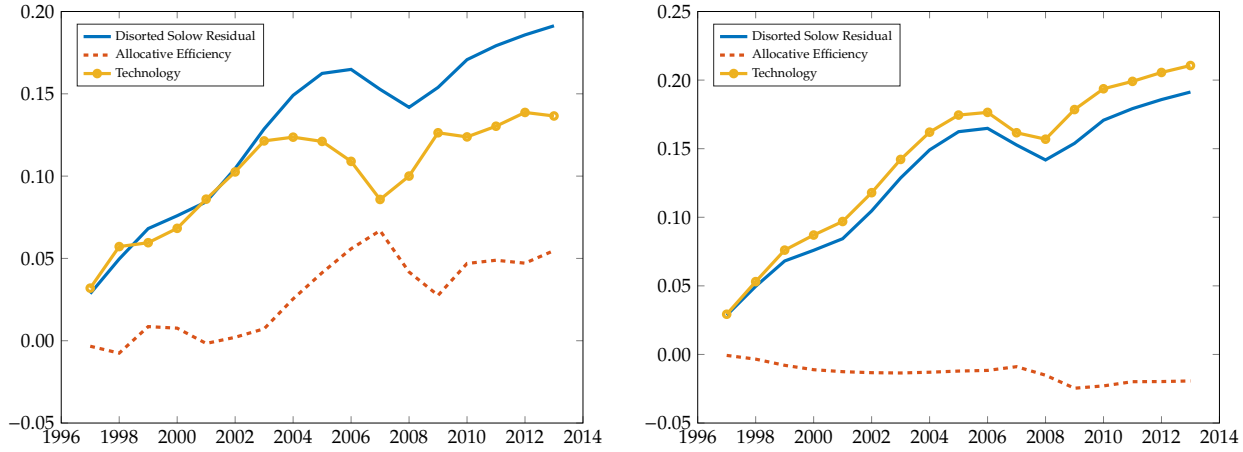
Vincent, N. and M. Kehrig (2019). The Micro-Level Anatomy of the Labor Share Decline. Nber working papers, National Bureau of Economic Research, Inc.

Appendix A Additional Figures



(a) Firm Level (b) Sectoral Level

Figure 6: Cumulative decomposition of changes in aggregate TFP into pure changes in technology and changes in allocative efficiency following equation (7), with markups obtained from the production-function approach, at the firm level (left panel) and at the sectoral level (right panel).



(a) Firm Level (b) Sectoral Level

Figure 7: Cumulative decomposition of changes in aggregate TFP into pure changes in technology and changes in allocative efficiency following equation (7), with markups obtained from the accounting profits approach, at the firm level (left panel) and at the sectoral level (right panel).