Productivity and Misallocation in General Equilibrium

David Rezza Baqaee
LSE

Emmanuel Farhi*
Harvard

December 24, 2017

Abstract

We provide a general non-parametric formula for aggregating microeconomic shocks in general equilibrium economies with distortions such as taxes, markups, frictions to resource reallocation, and nominal rigidities. We show that the macroeconomic impact of a shock can be boiled down into two components: its "pure" technology effect; and its effect on allocative efficiency arising from the associated reallocation of resources, which can be measured via changes in factor income shares. We also derive a formula showing how these two components are determined by structural microeconomic parameters such as elasticities of substitution, returns to scale, factor mobility, and network linkages. Overall, our results generalize those of Solow (1957) and Hulten (1978) to economies with distortions. To demonstrate their empirical relevance, we pursue different applications, focusing on markup distortions. For example, we operationalize our non-parametric results and show that improvements in allocative efficiency account for about 50% of measured TFP growth over the period 1997-2015. We also implement our structural results and conclude that eliminating markups would raise TFP by about 40%, increasing the economywide cost of monopoly distortions by two orders of magnitude compared to the famous 0.1% estimates of Harberger (1954).

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

1 Introduction

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

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

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

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

The non-parametric power of Hulten's theorem comes from exploiting a macro-envelope condition resulting from the first welfare theorem. It requires that the equilibrium allocation be Pareto-efficient. When it is not, Hulten's theorem generally fails.[3]

Our paper generalizes Hulten's theorem beyond efficient economies, and provides aggregation results for economies with arbitrary neoclassical production functions, input-output networks, and distortion wedges. Rather than relying on a macro-envelope condition like the first welfare theorem, our results are built on micro-envelope conditions: namely that all producers are cost minimizers. As a byproduct, they suggest a new and structurally interpretable decomposition of the Solow residual into "pure" changes in technology and changes in allocative efficiency. Our results provide a unified framework

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

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

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

for analyzing the effects of distortions and misallocation in general equilibrium economies, the study of which is the subject of a vibrant literature, recently reinvigorated by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009).[4]

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

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

To demonstrate the empirical relevance and the scope of applicability of our framework, we put it to use to answer four different questions. The first three questions focus on markups as a source of distortions, which we find particularly interesting in light of

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

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

the accumulating evidence that average markups have increased over the past decades in the US.[6] The fourth question studies nominal rigidities.

1. How have changes in allocative efficiency contributed to measured TFP growth in the US over the past 20 years?

   We use our ex-post reduced-form results to propose a new decomposition of measured TFP growth as captured by the Solow residual into a "pure" technology effect and an allocative efficiency effect.[7] Our decomposition has a structural interpretation as a local counterfactual, since we measure changes in allocative efficiency as the gap that opens up between the equilibrium allocation and a passive allocation which does not allow for the reallocation of resources. Although our decomposition shares the same objectives as the one provided by Basu and Fernald (2002), it is different. We compare these two decompositions in Section 2.7 and argue why we find ours preferable.

   We implement our Solow residual decomposition in the US over the period 1997-2014. Focusing on markups as a source of distortions, we find that the improvement in allocative efficiency accounts for about 50% of the cumulated Solow residual. This occurs despite the fact that average markups have been increasing. A rough intuition for this surprising result is that average markups have been increasing primarily due to an across-firms composition effect, where firms with high markups have been getting larger, and not a within-firm increase in markups.[8] From a social perspective, these high-markup firms were too small to begin with, and so the reallocation of factors towards them has improved allocative efficiency and TFP.

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

   Using our ex-ante structural results, we find that in the US in 2014, eliminating markups would raise aggregate TFP by about 40%. This increases the estimated cost of monopoly distortions by two orders of magnitude compared to the famous

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

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

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

estimates of 0.1% of Harberger (1954).[9] Essentially, the reasons for this dramatic difference are that we use firm-level data, whereas Harberger only had access to sectoral data, and that the dispersion of markups is higher across firms within a sector than across sectors. Moreover, the relevant elasticity of substitution is higher in our exercise than in Harberger's since it applies across firms within a sector rather than across sectors. Finally, we properly take into account the input-output structure of the economy to aggregate the numbers in all industries whereas Harberger focused on manufacturing.

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

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

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

   Our ex-ante structural results allow us to conclude that markups materially affect the impact of microeconomic productivity and markup shocks on output, both at the sector and at the firm level. They amplify some shocks and attenuate others. On the whole, we find that output is more volatile than in a competitive model, especially with respect to firm-level shocks.

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

4. What are the macroeconomic impact of monetary shocks and microeconomic shocks in a model with sticky prices?

   We can use our framework to answer this question by leveraging the observation that one can always model sticky prices as endogenous markups which adjust to ensure that nominal prices stay fixed. We characterize the effects of aggregate monetary shock and microeconomic productivity shocks. Focusing here on a money shock, we decompose the impact into the traditional demand effects and the oft-neglected allocative-efficiency effects on TFP.

   Typically, models with sticky prices are linearized around an efficient steady state, which ensures that reallocation terms disappear. We use our framework to study the model's behavior away from the efficient steady-state using empirically estimated steady-state markups, and with a realistic microeconomic production structure featuring input-output connections, complementarities in production, and substitutability among heterogenous firms within an industry.

   The size and direction of the TFP effects of monetary shocks depend crucially on the correlation pattern between price-stickiness and the level of markups. They can be large and positive if goods with higher markups have stickier prices.

Despite their generality, our results have two important limitations. First, our basic framework accommodates neoclassical production with decreasing or constant returns to scale. It can also easily handle fixed costs, as long as production has constant or increasing marginal cost. However, it is unable to deal with non-neoclassical production featuring increasing returns such as those studied by Baqaee (2016), where by *increasing returns*, we refer to a situation where marginal variable costs are decreasing in output. Towards the end of the paper, we sketch how our results can be extended to cover such cases, when we discuss entry and exit. Second, in this paper we focus on first-order approximations. We show that under some conditions, the nonlinear analysis of efficient economies in Baqaee and Farhi (2017) can be leveraged to characterize nonlinearities in the sort of inefficient economies studied in this paper.

The outline of the paper is as follows: in Section 2, we set up the general model, and we prove our main non-parametric results. We also discuss how to interpret these results, and the data required to implement our formulae. In Section 3, we introduce a parametric version of the general model and present our structural results. We use a model with CES production and consumption functions, with an arbitrary number of nests, input-output

patterns, returns to scale, and factors of production. In Section 4, we apply our results to the data by performing non-parametric ex-post decompositions of the sources of growth in the US, as well as structural exercises measuring the gains from markup reductions, macroeconomic volatility arising from microeconomic shocks, and macroeconomic impact of microeconomic shocks, in a calibrated model. As a final application, we show how our results can be used to study the effects of monetary policy and productivity in a model with nominal rigidities. In Section 5, we consider how our results can be extended to deal with fixed costs, entry, and nonlinearities.

2 General Framework and Non-Parametric Results

We begin by setting up our general framework, and characterize how shocks to wedges and productivity affect equilibrium output and TFP. We define our notion of change in allocative efficiency and discuss its implementation and data requirements. We explain how it leads to a new decomposition of the Solow residual into changes in "pure" technology and changes in allocative efficiency. We end by discussing the generality of our setup and in particular its ability to handle elastic factor supply and demand shocks.

2.1 General Model and Competitive Benchmark

The baseline model consists of a representative consumer and heterogeneous producers of different goods. Throughout, we model distortions via monopoly markups and wedges. The wedges act like linear taxes, the revenues of which are rebated lump sum.[10,11] Beyond actual taxes, these wedges can also implicitly capture frictions preventing the reallocation of resources. For example, they can capture credit constraints and they must then be interpreted as the Lagrange multipliers on the constraints in the individual firms' cost minimization problem.

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

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

[10] 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.

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

subject to the budget constraint

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

where p_i is the price of good i, τ_i^c is the consumption wedge on good i, π_i is the profits of the producer of good i, τ is a net government transfer, and L_f is a non-reproducible factor f with wage w_f.[12] The existence of a constant-returns-to-scale aggregate final demand function allows us to unambiguously define real GDP using the corresponding ideal price index.[13]

For now, we assume that factor supply is inelastic. The output effects identified in this section can therefore also be interpreted as aggregate TFP effects, and we use both terminologies interchangeably. At the end of this section, we generalize our results to cover the case with elastic factor supply and characterize the output effects when they differ from aggregate TFP effects.

We assume that each good i is produced by producer i according to a constant-returns technology described by the constant-returns cost function

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

where A_i is a Hicks-neutral shock, τ_{ij} is an input-specific tax wedge, τ_{ij}^f is a factor-specific tax wedge, and y_i is the total output. The function $\mathbf{C}_i$ is the industry's marginal cost function. Although we assume that the cost function is constant-returns-to-scale, this assumption is without loss of generality, since we can model decreasing returns to scale by using producer-specific fixed factors.[14] The producer charges an exogenous markup

[12] In mapping this set-up to the data, there are two ways to interpret this model: either we could interpret final demand as a per-period part of a larger dynamic problem, or we could interpret final demand as an intertemporal consumption function where goods are also indexed by time à la Arrow-Debreu. When we interpret the model intertemporally, then output is the net present-value of consumption streams. When we interpret the model intratemporally, the output function encompasses demand for consumption goods and for investment goods, and treats the two as perfect substitutes. Our results actually generalize to the case where they are not. See Section 2.8 for more details.

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

[14] To see how decreasing returns can be modeled via fixed-factors, see, for example, page 16 in Varian (1992). The case of increasing returns to scale is not nested by this setup; see Baqaee (2016), and the last section of this paper, for an analysis of increasing returns to scale in production networks.

μ_i over its marginal cost.[15]

Remark (Accounting Convention). We assume that expenditures by i on inputs from j, and the revenues of i, are recorded *gross* of taxes and markups. In the case where these wedges are reduced-form representations of frictions like credit constraints, we adopt the convention of writing expenditures gross of these implicit wedges.[16] This is purely a convention which does not change anything to the economics of the problem. Our convention need not coincide with the accounting convention for expenditures adopted in the data. In that case, the data must be converted into the format required by our theory. This conversion is completely straightforward. For example, in the case of a credit constraint which increases the rental rate of capital perceived by a firm but not its true rental rate, the conversion requires inflating the firm's expenditure on capital measured in the data by a percentage equal to the equivalent implicit tax on capital given by the Lagrange multiplier on the constraint in the cost minimization problem of the firm.

Remark (Markup-Wedge Equivalence). In this economy, we can always relabel the input-output matrix in such a way as to represent a wedge as a markup or vice versa. Therefore, going forward, we work with shocks to markups *only* with the understanding that this is done without loss of generality. Any pattern of wedges can be represented as markups by relabelling each input of each industry to be a new industry which charges a markup. Similarly, any pattern of markups can be represented via output wedges.

Our goal in this paper is to understand how to aggregate microeconomic shocks. Before stating our results, it is helpful to define some notation.

Definition 2.1 (Input-Output Matrix)**.** Let Ω be the $N \times N$ matrix whose ijth element is equal to i's expenditures on inputs from j gross of taxes as a share of its total revenues

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

Let $\tilde{\Omega}$ be the $N \times N$ matrix whose ijth element is equal to the elasticity of i's marginal costs

[15] With one exception, we do not attempt to endogenize markups in this paper. We see it as an important direction for future research to marry the sort of input-output models that we analyze here with industrial organization models of firm competition which generate endogenous markups. See for example Baqaee (2016) and Grassi (2017) for steps in this direction. In these models, our framework can be used to trace out the implications of these endogenous markups. The aforementioned exception arises in Section 4.3 when we consider nominal rigidities and model them as endogenous variable markups required to keep certain prices constant.

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

relative to the price of j

$$\tilde{\Omega}_{ij} \equiv \frac{\partial \log \mathbf{C}_i}{\partial \log p_j} = \frac{p_j x_{ij}}{\sum_k p_k x_{ik} + \sum_f w_f L_{if}}.$$

The second equality follows from Shephard's lemma, where x_{ij} are inputs from j and L_{if} is factor type f used by the ith producer. Let

$$\tilde{\Psi} \equiv (I - \tilde{\Omega})^{-1}$$

and

$$\Psi \equiv (I - \Omega)^{-1}$$

denote the Leontief inverse of $\tilde{\Omega}$ and Ω. Let α and $\tilde{\alpha}$ denote the $N \times F$ matrices whose ifth element is equal to i's expenditures on factor f as a share of its total revenues and costs respectively. Finally, let b be the $N \times 1$ vector whose ith element is equal to the household's expenditures on inputs from i as a share of GDP

$$b_i = \frac{p_i c_i}{\sum_j p_j c_j}.$$

The matrices Ω, α, and b are directly observable from input-output data (at the industry level, and sometimes even at the firm-level using value-added-tax data). Unlike revenues, however, we do not typically directly observe costs, so the matrices $\tilde{\Omega}$ and $\tilde{\alpha}$ are not readily observable. The link between Ω and $\tilde{\Omega}$ is given by

$$\Omega = \mu^{-1}\tilde{\Omega},$$

where μ is the diagonal matrix whose ith diagonal element corresponds to the ith markup or wedge.

Throughout the paper, we find it convenient to consider each factor of production as an industry in the input-output table. This means that the rows of the matrix $\tilde{\Omega}$ either sum to 1 (if that row corresponds to a good) or 0 (if the that row corresponds to a factor). Given this normalization, the matrix of factor uses satisfies the simple relation $\alpha = \tilde{\alpha}$.

The following accounting market-clearing identity

$$p_i y_i = p_i c_i + \sum_j p_i x_{ij} = b_i GDP + \sum_j \Omega_{ij} p_j y_j,$$

implies that

$$\lambda' = \left[\frac{p_i y_i}{GDP}\right]_i = b'(I - \Omega)^{-1} = b'\Psi, \tag{1}$$

where the sales of i as a share of GDP, denoted by λ_i, is called the *Domar weight* of i. Although the Domar weight of i is directly observable as the sales over GDP, it can also be computed from the input-output matrix Ω and final demand b. For expositional convenience, for a non-reproducible (factor) industry f we use Λ_f instead of λ_f to denote its Domar weight, which is then simply the corresponding income share of this factor.

We can also define

$$\tilde{\lambda}' = b'(I - \tilde{\Omega})^{-1} = b'\tilde{\Psi}$$

to be the vector of *cost-based Domar weights*. We choose the name cost-based Domar weight for $\tilde{\lambda}$ to contrast it with the traditional revenue-based Domar weight λ. For a non-reproducible (factor) industry f we use uppercase $\tilde{\Lambda}_f$ instead of $\tilde{\lambda}_f$. Intuitively, $\tilde{\lambda}_k$ measures the importance of k as a supplier to the household, both directly, and indirectly as an intermediate input. It only depends on k's role as a supplier rather than its role as a consumer.[17] This can be seen most clearly by writing

$$\tilde{\lambda}' = b'I + b'\tilde{\Omega} + b'\tilde{\Omega}^2 + b'\tilde{\Omega}^3 + \dots,$$

where the nth term in the geometric sum computes the set of paths of length n from each producer to the household. When the economy is efficient ($\mu = I$) — there are no markups or wedges – it must be the case that $\tilde{\lambda} = \lambda$. Hence, for an efficient economy, we observe $\tilde{\lambda}$ directly as sales over GDP. In the presence of wedges or markups, $\tilde{\lambda} \neq \lambda$, and so the cost-based Domar weights are *not* directly observable from gross output data.

We can now state Hulten's theorem for productivity shocks in efficient economies. In fact, we can slightly generalize it to allow for the case where only the initial pre-shock equilibrium is efficient. We also derive a simple extension for shocks to wedges.

Theorem 2.1 (Hulten 1978)**.** *When the initial pre-shock allocation is efficient with $\mu = I$, we have*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \lambda_k = \tilde{\lambda}_k, \quad \frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = 0. \tag{2}$$

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

Our goal in this paper is to extend this result to cover the case when the initial pre-shock equilibrium is inefficient. We begin by considering the impact of microeconomic productivity shocks, and then extend the results to cover wedge/markup shocks.

To state our results, we will use the notion of cross-entropy, which loosely speaking, is a measure of distance between distributions from information theory.[18]

Definition 2.2 (Cross Entropy)**.** For two probability distributions $\tilde{P}$ and P over the set of outcomes $\mathcal{S}$, the cross-entropy between $\tilde{P}$ and P is

$$H(\tilde{P}, P) = -E_{\tilde{P}}(\log P) = -\sum_{s \in \mathcal{S}} \tilde{P}_s \log(P_s).$$

For a given change dP in the probability distribution for P, we denote by $dH(\tilde{P}, P) = H(\tilde{P}, P + dP) - H(\tilde{P}, P)$ the change in relative entropy relative to the fixed distribution $\tilde{P}$. Similarly if P is indexed by x, then we denote by by $dH(\tilde{P}, P)/dx = dH(\tilde{P}, P_x)/dx$ the derivative of the relative entropy of P_x relative to the fixed distribution $\tilde{P}$.

We will apply these definitions using the set of non-reproducible (factor) industries for $\mathcal{S}$, $\tilde{\Lambda}$ as the fixed probability distribution $\tilde{P}$, and Λ as the variable distribution P.[19]

2.2 Productivity Shocks

In this section, we extend Hulten (1978) for productivity shocks to cover the case when the economy is inefficient.

Theorem 2.2 (Productivity Shocks)**.** *For productivity shocks, we have*[20]

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

[18] Cross-entropy between two distributions is minimized when the two distributions are the same. This definition is due to Claude Shannon (1948). Note that the Kullback and Leibler (1951) divergence is cross-entropy plus a constant, so that $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) = \mathrm{d}\, D_{KL}(\tilde{\Lambda} \| \Lambda)$ in our context.

[19] Note that $\sum_f \tilde{\Lambda}_f = 1$ so that we can indeed think of $\tilde{\Lambda}$ as a probability distribution. By contrast Λ is typically not a probability distribution since $\sum_f \Lambda_f \neq 1$ in general. However, we can always supplement Λ with the pure profit share $\Lambda_{f^*} = 1 - \sum_f \Lambda_f$ accruing to an extra factor f^* for which the cost-based share is zero $\tilde{\Lambda}_{f^*} = 0$.

[20] In the proof in the appendix, we also provide an explicit characterization of $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)/d \log(A_k)$ in terms of the structural characteristics of the production and consumption functions. We present this characterization in the main body of the paper for the more special parametric version of the model in Section 3.

Intuitively, in response to a productivity shock, two things can happen: first, given the share of resources allocated across producers, more output is produced in response to productivity; second, the share of resources allocated across producers can change in response to changing productivity.

The first term $\tilde{\lambda}_k$ corresponds to the change in output in response to the productivity shock, holding fixed the share of resources allocated to different producers. The second term corresponds to the change resulting from movement in resources.

When the economy is efficient, the first term corresponds to the Domar weight of industry k, and the second term is always equal to zero. The latter fact follows from $\mathrm{d}(\tilde{\Lambda}, \Lambda) = \mathrm{d}(\tilde{\Lambda}, \tilde{\Lambda}) = \sum_F \tilde{\Lambda}_F \,\mathrm{d} \log \tilde{\Lambda}_F = 0$. In this case, this formula collapses to Hulten (1978). In other words, when the economy is efficient, we can ignore the endogenous changes in the reallocation of resources arising from the productivity shock. To a first order, we can treat the (endogenous) share of resources allocated across producers to be constant. This is a manifestation of the first welfare theorem, or equivalently a macroeconomic envelope theorem. This logic fails when the equilibrium is inefficient. A nonzero second term indicates that the first welfare theorem has failed, and measures the resulting change in aggregate allocative efficiency.

Hence, Theorem 2.2 shows that we can boil down changes in misallocation arising from microeconomic productivity shocks in an economy with an arbitrary neoclassical production structure into an appropriately weighted average of the changes in factor income shares.[21] In particular, it is not necessary to track how the allocation of every single good is changing across its users. Instead, it suffices to to track how factor income shares change.

Theorem 2.2 implies that output increases when the distribution of factor income shares $\tilde{\Lambda}$ gets further away from the cost-share of the factors Λ. This may seem surprising given that $\Lambda = \tilde{\Lambda}$ at the efficient equilibrium, which maximizes output. However, the intuition is simple: that Λ gets further away from $\tilde{\Lambda}$ means that the more (less) monopolistic parts of the economy are receiving more (fewer) resources. This *improves* allocative efficiency, since from a social perspective, the corresponding firms or sectors absorb too few (too much) resources to begin with. For example, ceteris paribus, a decrease in all factor income shares necessarily implies an improvement in allocative efficiency. To see an explicit demonstration of this phenomenon, see Example 3.2.

[21] Since the landmark work of Theil (1967), changes in entropy have been used as a measure of changing income inequality. Our results show that changes in the cross-entropy of Λ relative to $\tilde{\Lambda}$ can track changes in allocative efficiency.

2.3 Markup/Wedge Shocks

So far, we have focused on productivity shocks, but it turns out a very similar intuition holds for shocks to markups, and hence also for other wedges given the general possibility of capturing wedges as markups under relabelling discussed above.

Theorem 2.3 (Markup/Wedge Shocks)**.** *For markup/wedge shocks, we have*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \sum_f \tilde{\Lambda}_f \frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log \mu_k}. \tag{4}$$

A markup shock acts much like the combination of a fictitious negative productivity shock combined with a fictitious shock to available factors. This is because a positive markup shock, like a negative productivity shock, increases the price of the corresponding good, which absent reallocation effects, increases the GDP deflator by the associated cost-adjusted Domar weight. But compared to a negative productivity shock, a markup shock also releases some resources which are reallocated to other parts of the economy, thereby generating further changes in allocative efficiency. A remarkable feature of this theorem is that all these changes in allocative efficiency are captured by the changes in the factor income shares.

Although stated in terms of markups, this result also characterizes the response of output to other distortion shocks τ. This follows from the observation that markup shocks and wedge shocks are equivalent, up to a relabelling. Hence, Theorem 2.3 allows to connect with the broader literature on misallocation like Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and in particular Jones (2013).[22]

[22]For example, for a factor wedge shock to factor k for producer l, we have

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log(1+\tau^f_{lk})} = -\tilde{\lambda}_l \alpha_{lk} + \frac{\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)}{\mathrm{d}\log(1+\tau^f_{lk})} = -\tilde{\lambda}_l \alpha_{lk} - \sum_{f'} \tilde{\Lambda}_{f'} \frac{\mathrm{d}\log \Lambda_{f'}}{\mathrm{d}\log(1+\tau^f_{lk})}. \tag{5}$$

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

2.4 Ex-post Decompositions: Aggregate TFP and Solow Residual

In this section, we show how to decompose time-series changes in aggregate TFP and the Solow residual into "pure" technology changes and allocation efficiency changes. We put together Theorems 2.2 and 2.3. For the purpose of this section, we introduce a small but simple modification to allow for changes in factor supplies in order to separate aggregate output and aggregate TFP. We denote the supply of factor f by L_f and by L the vector of factor supplies. The impact of a shock to the supply of a factor is given by $\mathrm{d}\log Y/\mathrm{d}\log L_f = \tilde{\Lambda}_f + \mathrm{d}\,H(\tilde{\Lambda},\Lambda)/\mathrm{d}\log L_f = \tilde{\Lambda}_f - \sum_{f'}\tilde{\Lambda}_{f'}\,\mathrm{d}\log\Lambda_{f'}/\mathrm{d}\log L_f$.

Proposition 2.4 (TFP Decomposition)**.** *To the first order, we can decompose aggregate TFP as*

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

The left-hand side of this expression, which we define to be aggregate TFP growth, differs from the Solow residual since it weighs the change in $L_{f,t}$ by the cost-based Domar weight $\tilde{\Lambda}_{f,t}$ rather than the revenue-based Domar weight $\Lambda_{f,t}$. This is consistent with Hall (1990), who showed that for an aggregate production function, aggregate TFP should weigh changes in factor inputs by their share of total cost rather than their share of total revenue. In our context unlike in Hall's, the equilibrium can be distorted given factor supplies and there is no structural aggregate production function. We must weigh factors by their cost-based Domar weight. Proposition 2.4 therefore unifies the approach of Hulten (1978), who eschews aggregate production functions, but maintains efficiency, with that of Hall (1990) who does not require efficiency but maintains aggregate production functions and therefore ignores the allocative efficiency issues that most concern us.

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

Proposition 2.4 also allows us to connect our results to the "revenue-based" Solow

residual defined by Basu and Fernald (2001):

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

The first three summands on the right hand side are the same as in Proposition 2.4 The last summand corrects for the undercounting of the contribution of factors growth to output growth, given that $\tilde{\Lambda}_{f,t} \geq \Lambda_{f,t}$.[23]

It is important to note that Proposition 2.4 can be used in contexts where productivity or wedges are endogenous to some more primitive fundamental shocks, because these endogenous changes are actually observed in the data.

2.5 Constant Misallocation

An important advantage of Proposition 2.4 is that, despite its low information requirement, the decomposition it provides has a structural interpretation as a local counterfactual. Specifically, we can interpret the changes in allocative efficiency term in equation (6) as measuring the gap that opens up between the equilibrium allocation and a passive allocation where the distribution of resources is not changed in response to a shock. The passive allocation, which has constant allocative efficiency, isolates the "pure" technology effect of the shock.

In this section, we define this passive allocation and establish its aforementioned property. We also provide two examples where the passive allocation coincides with the equilibrium allocation, and we identify the importance of cyclic graphs in generating misallocation.

Passive Allocation

To define the passive allocation rule, consider an initial equilibrium with some distortions, and imagine this economy is hit with some shocks to productivities $\mathrm{d} \log A$ and markups

[23]The left-hand side of equation (7) is not the traditional Solow residual defined by Solow (1957). The traditional Solow residual, unlike the version used by Basu and Fernald (2002), attributes all non-labor income to capital (and has no room for profit income). Therefore, with only labor (L) and capital (K) as factors, the traditional Solow residual would be

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

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

$\mathrm{d}\log\mu$. Suppose that if the physical output of some industry k increases by $\mathrm{d}\log y_k$, then that new output is proportionally divided amongst k's customers according to their shares in the initial equilibrium. In other words, $\mathrm{d}\log x_{lk} = \mathrm{d}\log c_k = \mathrm{d}\log y_k$.[24] Call the change in output under this passive rule for allocating new resources $\mathrm{d}\log Y^p$. We can show that

$$\mathrm{d}\log Y^p = \tilde{\lambda}'\,\mathrm{d}\log A. \tag{9}$$

When the initial equilibrium is efficient, the passive and equilibrium output move by the same amount $\mathrm{d}\log Y^p = \mathrm{d}\log Y$. Intuitively, as a consequence of the Envelope theorem, the rule by which resources are reallocated is irrelevant up to a first order. On the other hand, when the initial equilibrium is distorted, we know that

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

Hence, the gap between the general equilibrium allocation and the passive allocation is exactly what we call the change in allocative efficiency

$$\mathrm{d}\log Y - \mathrm{d}\log Y^p = -\tilde{\lambda}'\,\mathrm{d}\log\mu + \mathrm{d}\,H(\tilde{\Lambda},\Lambda) = -\tilde{\lambda}'\,\mathrm{d}\log\mu - \sum_f \tilde{\Lambda}_f\,\mathrm{d}\log\Lambda_f. \tag{10}$$

We view this property as building considerable support for our definition of allocative efficiency. Essentially, the passive allocation just scales the initial allocation proportionately, without allowing any other form of reallocation through substitution. In this sense, it constitutes a benchmark without changes in allocative efficiency, and so it stands a useful yardstick against which to measure changes in allocative efficiency in the equilibrium allocation.

Moreover, as we shall now see, there are cases where it should be unambiguous a priori that there are no changes in allocative efficiency: acyclic economies, and Cobb-Douglas economies. And these cases are correctly diagnosed by our definition: the equilibrium allocation and the passive allocation coincide, and our measure of change in allocative efficiency is zero.

[24] More formally, these proportional adjustments lead to a system of linear equations in $\mathrm{d}\log y_k$ and $\mathrm{d}\log x_{lk}$ with forcing variables given by the shocks, and the passive allocation is the solution of this system.

Acyclic Economies – No Misallocation

A case where misallocation plays no role is the case where the production network is an acylic graph, as illustrated in Figure 1. The term acyclic here means that any two industries are connected to one another by exactly at most one undirected path, so that each factor and each good has a unique consumer. This implies that markups and wedges have no effect on the allocation of resources, simply because there is no option to allocate a given factor or good to different uses. In other words, misallocation requires cycles (non-trivial undirected paths that connect a node back to itself) in the production network.

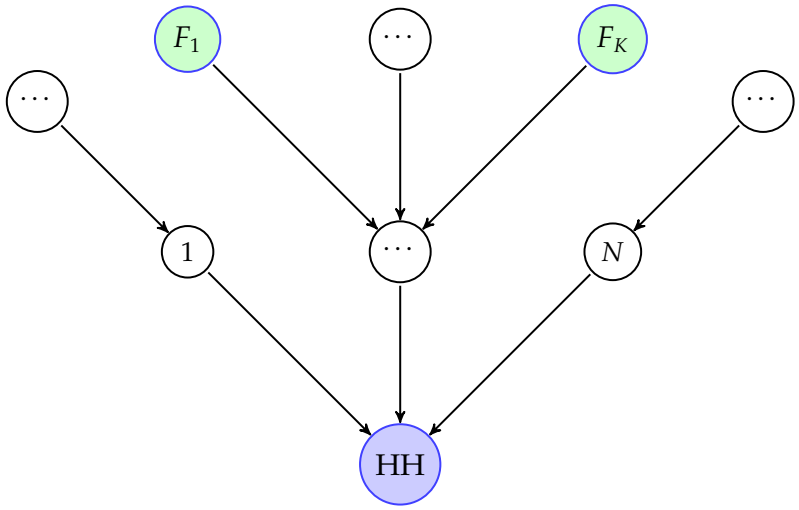


Figure 1: An acyclic economy, where the solid arrows represent the flow of goods. The factors are the green nodes. Each supplier (including factors) have at most one customer, whereas a single customer may have more than one supplier. Economies without cycles can be represented as directed trees with the household being the root.

Proposition 2.5 (Acyclic Economies)**.** *If the production network of the economy is an acyclic graph, then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k, \quad \frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = 0. \tag{11}$$

An important consequence of this proposition, anticipated above, is that in acyclic economies in which it is unambiguous a priori that there is no misallocation, our definition of changes in allocative efficiency indeed identifies that there are no changes in allocative efficiency in response to shocks.

Note that although the equilibrium allocation in this economy is efficient, Hulten's theorem still fails because the observed sales shares do not coincide with $\tilde{\lambda}$. In fact, due to

double-marginalization (or, more generally, piling up of wedges), we can write examples where λ and $\tilde{\lambda}$ can be arbitrarily different from one another. A stark illustration of the gap between $\tilde{\lambda}$ and λ, drawn from Baqaee (2016), is a vertical economy shown in Figure 2, to which we now turn.

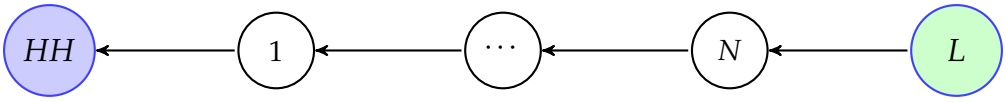


Figure 2: Vertical economy where the solid arrows represent the flow of goods. The flow of profits and wages from firms to households has been suppressed in the diagram. The sole factor for this economy is indexed by L.

Example 2.1 (Vertical Economy)**.** Consider the economy depicted in Figure 2: only the final industry N uses labor. Since the economy is acyclic, there is no reallocation term to account for. Indeed, applying Propositions 2.6, we get

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k = 1,$$

while sales as a share of GDP are

$$\lambda_k = \frac{p_k y_k}{Y} = \prod_{i=1}^{k-1} \mu_i^{-1}. \tag{12}$$

For simplicity, assume that $\mu_i < 1$ for all i. Then $\tilde{\lambda}_k = 1 > \lambda_k$. The gap between the cost-based Domar weight $\tilde{\lambda}_k$ and the revenue-based Domar weight λ_k can be arbitrarily large. Indeed, with high enough markups or long enough chains, we can drive the revenue-based Domar weight λ_k to 0, while keeping $\tilde{\lambda}_k$ constant at 1, which represents an extreme failure of Hulten's theorem.

Cobb-Douglas Economies – Constant Misallocation

A second case where there are no changes in allocative efficiency is the case of a Cobb-Douglas economies with productivity shocks. In this case, in response to a shock, all expenditure shares are constant and the equilibrium allocation coincides with the passive allocation. Hence, although unlike in acyclic economies, resources can be misallocated, the extent of misallocation does not change in response to shocks.

Proposition 2.6 (Cobb-Douglas Economies)**.** *Consider the case where the final-demand and all cost functions are Cobb-Douglas. Then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k. \tag{13}$$

This result follows immediately from Theorem 2.2, and the fact that $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)/\,\mathrm{d}\log A_k = 0$, since for a Cobb-Douglas economy, factor income shares do not respond to productivity shocks.[25]

Without markups or wedges, $\tilde{\lambda}_k = \lambda_k$ and we recover Hulten's theorem. However, in the presence of wedges, $\tilde{\lambda}_k$ can be arbitrarily different to the Domar weight λ_k. The gap between $\tilde{\lambda}_k$ and λ_k opens up due to double-marginalization (or, more generally, piling up of wedges). Even with symmetric uniform markups, the economy will typically feature misallocation, since asymmetries in the input-output network will result in different industries being differentially exposed to markups. [26]

Although in a Cobb-Douglas economy, the equilibrium allocation responds to a productivity shock exactly as the passive allocation, it does not in response to markup/wedge shocks, and so these latter shocks do generate changes in allocative efficiency. We return to how output responds to markup/wedge shocks in the Cobb-Douglas special case in Section 3.1.

2.6 Implementing the Results

Implementing our formulas using actual data requires more care than when handling efficient economies. Over and above the difficulties involved with coming up with reliable empirical measures of distortions, there are two important issues that have to be confronted: (1) identification of the factors of production; (2) and aggregation of the data at hand. We discuss them in turn.

[25] Liu (2017) shows that a formula like the one in Proposition 2.6 can hold even when the consumption and production functions are not necessarily Cobb-Douglas. This result follows from the fact that in his model, $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda) \equiv 0$, since there are no profits and labor's share of income is always equal to 1, so that allocative efficiency does not change. Intuitively, in that model, distortions dissipate resources (through excess entry, or technological transactions costs), so that they behave similarly to productivity shocks where the sales shares are mismeasured (since expenditures are recorded net of the destruction, they need to be adjusted to take the cost of the destroyed resources into account).

[26] The observation that for Cobb-Douglas economies with markups or wedges, $\mathrm{d}\log Y/\,\mathrm{d}\log A_k = \tilde{\lambda}_k \neq \lambda_k$ is not new to this paper. For instance, see Jones (2013) for generic wedges in a Cobb-Douglas economy, or Bigio and La'O (2016), who study a Cobb-Douglas economy with credit constraints that manifest as wedges, and the Cobb-Douglas special case of Baqaee (2016), who studies a Cobb-Douglas economy with markups.

Identification of the Factors

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

In other words, when the equilibrium is inefficient, we need to take a stance about whether factors are "stuck" due to technological restrictions or market imperfections. Intuitively, if labor does not flow from mining into secretarial services, treating miners and secretaries as two separate factors or a single factor will affect the implied measure of allocative efficiency. 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 zero to the first order.

Data Aggregation Level

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

This aggregation property does not hold for distorted economies, even in the Cobb-Douglas or acyclic cases where we do not need to account for changes in allocative efficiency. Unlike sales, cost-based Domar weights $\tilde{\lambda}$ are not directly observable, and instead need to be computed from input-output data *at* the level of disaggregation at which the markups and wedges appear. If wedges apply at the firm or establishment

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

level, then firm or establishment-level input-output data is in general necessary. See Appendix D for a worked-out example.[28]

2.7 Comparison with the Basu-Fernald Decomposition

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

We have already independently argued the merits of our approach. We also think that it better captures the distinction between "pure" technology and misallocation than Basu and Fernald's. One extreme but illuminating example regarding this comparison is the case of productivity shocks in acyclic economies with markups. In this case, the Basu-Fernald decomposition detects "pure" changes in technology *and* changes in misallocation, even though these economies feature efficient equilibria and have no misallocation.[29] By contrast, as already explained above, our approach only detects changes in "pure" technology and finds no change in allocative efficiency. See Appendix D for a worked-out example.

[28] In Section 4, 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).

[29] More precisely, in this case, the change in misallocation detected by their decomposition shows up in their "materials-misallocation" term.

2.8 Extensions to Basic Framework

In this section, we discuss some extensions. Readers may want to to skip it on first reading as it will not affect their understanding of the rest of the paper.

The basic set up of the model abstracts away from biased technical change, demand shocks, elastic factor supplies, and capital accumulation, as well as adjustment costs and capacity utilization. In this section, we explain how to apply or modify our results to take these features into account. We also explain how to use our results in cases where markups/wedges and productivities are endogenous.

Biased Technical Change and Demand Shocks

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

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

Capital Accumulation, Adjustment Costs, and Capacity Utilization

Our benchmark model treats accumulable factors such as capital in the economy as being exogenously determined. One possibility, already mentioned, is to treat Y as capturing per-period final demand, which consists of final consumption and investment demands. By specifying how investments augment the accumulable factors, one would obtain an evolution equation for accumulable factors over time. Our results generalize immediately in the case where consumption and investment goods are perfect substitutes. When they are not, then microeconomic shocks to productivity or markups/wedges could in principle

lead to endogenous changes in the composition of final demand across investment and consumption: these changes in the composition of final demand would not have pure technology effects, but they could affect TFP by changing allocative efficiency, and this would be captured in our decomposition.[30,31]

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

Elastic Factor Supplies

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

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \varrho\left(\tilde{\lambda}_k - \frac{1}{1+\zeta}\frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log A_k}\right) = \varrho\left(\tilde{\lambda}_k - \sum_f \frac{1}{1+\zeta_f}\tilde{\Lambda}_f\frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log A_k}\right), \tag{14}$$

[30] Another way to see this is to separate the change in the composition of final demand across consumption and investment as a separate shock which changes TFP only through a change in allocative efficiency. This shock can then be expressed as a function of the underlying productivity and markups/wedges disturbances, and the chain rule can be applied. When these demand composition changes are taken into account, the same formula applies, but the corresponding changes in factor shares are different, and hence the changes in allocative efficiency are different.

[31] Things are a bit different In Section 3, where we derive structural results by characterizing the movements in factor shares and hence allocative efficiency triggered by a shock to productivity or markups/wedges. These results apply to the case where consumption and investment goods are perfect substitutes, or when they are not but the composition of final demand across consumption and investment remains constant. Separate structural results could be derived to characterize the impact of final demand composition changes on factor shares and hence allocative efficiency; and then by expressing the final demand composition shock as a function of the underlying disturbances and applying the chain rule, one could characterize the full effect on factor shares or allocative efficiency of the underlying disturbances, including the effects arising from endogenous final demand composition changes. In the interest of space, we do not report these results in the paper.

and

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = \varrho\left(-\tilde{\lambda}_k - \frac{1}{1+\zeta}\frac{\mathrm{d}\,H(\tilde{\Lambda},\Lambda)}{\mathrm{d}\log \mu_k}\right) = \varrho\left(-\tilde{\lambda}_k - \sum_f \frac{1}{1+\zeta_f}\tilde{\Lambda}_f \frac{\mathrm{d}\log \Lambda_f}{\mathrm{d}\log \mu_k}\right), \tag{15}$$

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

With inelastic factors, a decline in factor income shares, ceteris paribus, *increases* output since it represents a reduction in the misallocation of resources and an increase in aggregate TFP. With elastic factor supply, the output effect is dampened by the presence of $1/(1+\zeta_f) < 1$. This is due to the fact that a reduction in factor income shares, while increasing aggregate TFP, reduces factor supply, which in reduces output. Hence, when factors are elastic, increases in allocative efficiency from assigning more resources to more monopolistic producers are counteracted by reductions in factor supplies due to the associated suppression of factor demand.[32]

For the rest of the paper, unless explicitly specified, we focus on TFP effects and so we work with the version of the model where factors are inelastically supplied. This is primarily for clarity: whenever the crux of the argument does not hinge on having elastic factor supplies, we assume it away. In some specific cases, like when discussing the effects of monetary policy, elastic factor supply is important, and then we reintroduce it.

Endogenous Productivities and Markups

Our results are comparative statics on the macroeconomic impact of a microeconomic productivity or markup/wedge shock, holding fixed all other markups/wedges and productivity levels. Of course, these results can then be used to study situations where several microeconomic shocks occur at the same time. They can also be used in contexts where productivity or wedges are endogenous to some more primitive fundamental shock.

In the latter case, our results can be operationalized in two different ways. First, they can be used ex post without any modification because these endogenous changes are actually observed in the data, as explained in Section 2.4. Second, they can combined ex ante with an application of the chain rule characterizing the derivatives of the vectors

[32] In the limit where factor supplies become infinitely elastic, the influence of the allocative efficiency effects disappear from output, since more factors can always be marshaled on the margin at the same real price. To see this, consider the case with a single factor called labor, and factor supply function $G_L(w/P_y, Y) = \left(w/P_y Y\right)^{\nu}$, which can be derived from a standard labor-leisure choice model. In this case, $\gamma_L = -\zeta_L = \nu$, and so equation (14) implies that $\mathrm{d}\log Y/\mathrm{d}\log A_k = \tilde{\lambda}_k + 1/(1+\nu)\,\mathrm{d}\,H(\tilde{\Lambda}_L, \Lambda_L)/\mathrm{d}\log A_k$. When labor supply becomes infinitely elastic $\nu \to \infty$, this simplifies to $\mathrm{d}\log Y/\mathrm{d}\log A_k = \tilde{\lambda}_k$, so that changes in allocative efficiency have no effect on output, even though they affect TFP.

of microeconomic productivities and markups to these more fundamental shocks arising from some additional structure imposed on the model. We refer the reader to Section 3.4 for a detailed discussion of our ex-post reduced-form and of our ex-ante structural results in such contexts.

3 Parametric Model and Structural Results

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

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

3.1 Parametric Model Setup

Any CES economy with a representative consumer, an arbitrary numbers of nests, elasticities, and intermediate input use, can be re-written in *standard form*, which turns out to be more convenient to study. In this section, we first define the notion of standard form and clarify the mapping of any CES economy to its standard form. We then fully characterize the behavior of such economies.

Throughout this section, variables with over-lines are normalizing constants equal to the values in steady-state. Since we are interested in log changes, the normalizing constants are irrelevant.[33]

An economy in standard form is defined by a tuple $(\Omega, \boldsymbol{\theta}, \mu, F)$ where Ω is the $(N+1)\times(N+1)$ input-output matrix whose ijth element is equal to ω_{ij}, the vector of elasticites $\boldsymbol{\theta}$ is

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

an $(N+1-F)\times 1$ vector whose ith element is θ_i, the vector of markups is μ, and F is the number of factors.

The F factors are modeled as non-reproducible goods, which we denote with uppercase letters, and the production function of the corresponding industries is an endowment

$$\frac{Y_f}{\overline{Y}_f} = 1. \tag{16}$$

For the non-reproducible industries, we set their $\omega_{fj} = 0$ for all j. The other $N+1-F$ other goods are reproducible, and the production of a reproducible good k can be written as

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

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

$$\frac{Y}{\overline{Y}} = \frac{c_0}{\overline{c}_0}, \tag{17}$$

where c_0 is the final good. We set $\omega_{i0} = 0$ for all i.

Through a relabelling, this structure can represent any CES economy with an arbitrary pattern of nests and wedges and elasticities. Intuitively, by relabelling each CES aggregator to be a new industry, we can have as many nests as we like. Wedges on any specific input, say inputs from industry i sold to industry j, can be achieved by relabelling that input to be a new industry and assigning a markup to that industry. This relabelling amounts to changing the way the input-output table is written, and since we work with arbitrary input-output tables, we can do this without loss of generality (see Appendix C for a concrete example).

Given the revenue-based input-output matrix Ω and the vector of markups μ, we can define, exactly as in Section 2.1, the cost-based input-output matrix $\tilde{\Omega}$, the revenue-based Leontief inverse matrix Ψ, the cost-based Leontief inverse matrix $\tilde{\Psi}$, the vector of revenue-based Domar weights or sales λ, the vector of cost-based Domar weights $\tilde{\lambda}$. As above, we flag the Domar weights for factors by using uppercase variables, denoting the vector of factor shares in income by Λ, and the vector of factor shares in costs by $\tilde{\Lambda}$.

In order to state our results, it will also be helpful to introduce the following 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.[34] Since the rows of $\tilde{\Omega}$ always sum to one for a reproducible (non-factor) industry j, we can formally think of this as a covariance, and for a non-reproducible industry, the operator just returns 0.

3.2 Single Factor

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

3.2.1 Productivity Shocks

Proposition 3.1 (Productivity Shocks with One Factor)**.** *Suppose there is only one factor, denoted by L. Then*

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

where

$$\frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log A_k} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}\right), \tag{18}$$

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

With variable misallocation, we keep the first term as before: namely, the effect of the shock if the new resources were allocated passively, with the cost-based Domar weight $\tilde{\lambda}$ correcting for the double-marginalization happening downstream from k. If the economy is acyclic or Cobb-Douglas, this is all we need. Otherwise, the allocation of labor across different producers *depends* on the productivity levels. Hence, a productivity shock

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

can trigger reallocation of resources across producers, and thereby change the allocative efficiency of the economy. This effect is captured by the change in cross-entropy, characterized by equation (18). The result is a centrality measure which mixes networks and elasticities of substitution.

We can think of Ψ_{iL} as the payments to labor as a share of the total revenue of producing i — what Baqaee (2015) calls the *network-adjusted labor share* of i — and of Λ_L as the share of labor in the revenues of the economy. These quantities are related to the payments to labor $\tilde{\Psi}_{iL}$ as a share of the total cost of producing i — taking into account the entire supply chain — and to the share of labor $\tilde{\Lambda}_L$ in the costs of the economy. In an efficient economy with one factor, we have $\Psi_{iL} = \tilde{\Psi}_{iL} = 1$ and $\Lambda_L = \tilde{\Lambda}_L = 1$. By contrast, in an inefficient economy we still have $\tilde{\Psi}_{iL} = 1$, and $\tilde{\Lambda}_L = 1$ but we no longer necessarily have $\Psi_{iL} = 1$ or $\Lambda_L = 1$. For example, if all markups are positive, we have $\Psi_{iL} < 1$ and $\Lambda_L < 1$. A low value of Ψ_{iL} indicates that on average, markups are high along the supply chain of industry i, and a low value of Λ_L indicates that on average, markups are high in the economy along as a whole. The lower Ψ_{iL}/Λ_L, the more distorted is the supply chain of industry i relative to the economy as a whole. In other words, we can think of the Lth column of the Leontief inverse $\Psi_{(L)}$ as measuring the degree of double-marginalization along the supply chain of each industry. The economy's labor content as a whole is given by labor's share of income Λ_L. Industries with low values of $\Psi_{(L)}/\Lambda_L$ have too few workers in their supply chain, relative to the economy as a whole, due to the presence of markups.

With this interpretation, Proposition 3.1 becomes very intuitive. In response to a positive productivity shock to industry k, the relative prices of all industries change according to their exposure to k, measured by $\tilde{\Psi}_{(k)}$. If $\theta_j > 1$, the jth industry substitutes across its inputs towards the industries with higher exposure $\tilde{\Psi}_{(k)}$ to k, since their relative prices decline by more. If those industries also happen to have lower $\Psi_{(L)}$, then these industries are inefficiently too small in the initial pre-shock equilibrium. In this case, there is negative covariance between $\tilde{\Psi}_{(k)}$ and $\Psi_{(L)}$. This means that substitution, due to the productivity shock, lowers overall misallocation, sending more workers to produce goods which are inefficiently receiving too few workers. In this case, the changing allocation of workers boosts the impact of the productivity shock on output. Of course, j is not the only industry whose expenditure shares change and the same logic applies to all industries, so we sum over all j. If the elasticities are less than one, or the covariance is negative, the reallocation forces work against the positive impact of the technology shock.[35]

[35]Baqaee and Farhi (2017) show that for an economy like the one in Proposition 3.1, if the economy is

To further demonstrate the intuition of Proposition 3.1, we work through two examples: the horizontal economy depicted in Figure 3, and the Cobb-Douglas economy.

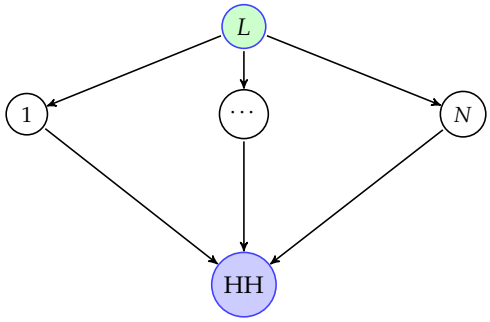


Figure 3: An "horizontal" economy where the solid arrows represent the flow of goods. Note that this economy has a cycle.

Example 3.1 (Cobb-Douglas)**.** A particular case of Proposition 3.1 is that of a Cobb-Douglas economy as a special case. As mentioned earlier, when $\theta_i = 1$ for each $i = \{0, \dots N\}$, we have

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k. \tag{19}$$

This example takes a very simplified form because it does not feature changes in allocative efficiency, because it coincides the with the passive rules for the allocation of new resources. Next, we look at the simplest example which can feature such changes.

Example 3.2 (Horizontal Economy)**.** Consider the economy depicted in Figure 3. Let i index final goods and let L index the labor input.[36] Applying Proposition 3.1, we have

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \lambda_k \left(1 - (\theta_0 - 1)\left(\frac{\mu_k^{-1}}{\sum_i \lambda_i \mu_i^{-1}} - 1\right)\right), \tag{20}$$

noting that $\tilde{\lambda}_k = \lambda_k$ and $\Lambda_L = \sum_i \lambda_i \mu_i^{-1}$.

As long as $\theta_0 \neq 1$, a productivity shock changes the fraction of workers employed by each industry. If markups are heterogenous, then this reallocation of workers could improve or worsen the amount of misallocation in this economy, thereby amplifying or

efficient, then the output response to a shock to industry k depends *only* on k's role as a supplier. Proposition 3.1 shows that this fails if the equilibrium is inefficient. In particular, $\Psi_{(L)}$ — which captures information about how distorted the supply chain of each industry is (i.e. it depends on the industry's role as a consumer of inputs), also matters, since it affects the response of misallocation.

[36] Note that for this simple example, $\lambda_i = \tilde{\lambda}_i = b_i$ for $i \in \{1, \dots, N\}$.

mitigating the effect of the shock. When $\theta_0 > 1$: the shock is amplified when k charges a markup μ_k higher than the average markup $\mu_k^{-1} < \sum_i b_i \mu_i^{-1} = \Lambda_L$. On the other hand, the shock is attenuated if k charges a lower markup. These patterns are reversed if $\theta_0 < 1$.

All of this information is summarized by the change $\mathrm{d}\, H(\tilde{\Lambda}, \Lambda)/\,\mathrm{d}\log A_k$ in the cross-entropy between the revenue and cost-based labor income shares, which in this simple case is simply $-\,\mathrm{d}\log \Lambda_L/\,\mathrm{d}\log A_k$. If in response to a productivity shock, the labor share of income decreases, then this implies that allocative efficiency has improved because industries that were too small because they were charging markups above the economy's average. The converse happens when the labor share of income increases in response to the shock.[37,38]

Since misallocation is the reason for this complex behavior, a shock which does not reallocate workers across industries is much simpler to analyze. For example, these forces do not show up for a shock to labor (or equivalently, a TFP shock to industry L), since that will simply scale up employment across all industries by the same amount:

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log L} = \tilde{\Lambda}_L - \frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log L} = \tilde{\Lambda}_L = 1.$$

3.2.2 Markup/Wedge Shocks

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

Proposition 3.2 (Makup/Wedge Shocks with One Factor)**.** *Suppose that there is only one factor, denoted by L. Then*

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

where

$$\frac{\mathrm{d}\log \Lambda_L}{\mathrm{d}\log \mu_k} = -\sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}, \frac{\Psi_{(L)}}{\Lambda_L}\right) - \lambda_k \frac{\Psi_{kL}}{\Lambda_L}. \tag{21}$$

[37] When $\theta_0 > 1$, the reduction in allocative efficiency can be so extreme that a positive productivity shock can actually reduce output. A positive productivity shock can reduce TFP if k is significantly more competitive than the average firm $\mu_k^{-1} > \frac{\theta_0}{\theta_0 - 1}\sum_i b_i \mu_i^{-1}$

[38] This intuition only applies for productivity shocks — for a shock to markups, the change in the labor share of income is no longer sufficient to diagnose the change in misallocation. We come back to this point later when we examine markup shocks.

Proposition 3.2 implies that the effects of a positive markup shock are analogous to the effects of a negative productivity shock. In response to a change in markup, there is a direct effect $-\tilde{\lambda}_k$ on consumer prices from higher markups, and an indirect effect from the changing factor income shares captured by $-\operatorname{d}\log\Lambda_L/\operatorname{d}\log\mu_k$.

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

We consider the same two examples as before: the Cobb-Douglas economy, and the horizontal economy. The Cobb-Douglas example helps to isolate the importance of the new term in Proposition 3.2. For a Cobb-Douglas economy, the only source factor reallocation comes from the fact that the industry which increases its markups releases some labor. We also add an example illustrating how our results can be applied to compute the gains from removing distortions, and we relate this popular measure of misallocation to our notion of changes in allocative efficiency.

Example 3.3 (Cobb Douglas)**.** Let $\theta_j = 1$ for every j, which is the Cobb-Douglas special case. Now, applying Proposition 3.2, we get

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

As before, Ψ_{kL}/Λ_L is a measure of how distorted the supply chain of k is relative to the economy as a whole. If $\Psi_{kL}/\Lambda_L < 1$, then this means that for each dollar k earns, a smaller share reaches workers than it would if that dollar was spent by the household. In other words, industry k's supply chain has inefficiently too few workers. On the other hand, $\lambda_k/\tilde{\lambda}_k$ is a measure of how distorted the demand of chain of k is. If $\lambda_k/\tilde{\lambda}_k < 1$, this implies that k is facing double-marginalization. When the product of the downstream and upstream terms is less than one, this means industry k is inefficiently starved of demand and workers. Hence, an increase in the markups of k reduces the allocative efficiency of

the economy. On the other hand, when the product of these two terms is greater than one, the path connecting the household to labor via industry k is too large. Therefore, an increase in the markups of k reallocates resources to the rest of the economy where they are more needed and increaes allocative efficiency.

We also revisit the horizontal economy of Figure 3, but this time, instead of productivity shocks, we examine the effect of markup shocks. Furthermore, we now allow for the possibility that the economy does not have a Cobb-Douglas structure.

Example 3.4 (Horizontal Economy)**.** Consider the horizontal economy example, but now suppose that markups are shocked instead. By Proposition 3.2 we know that this is just the negative of the effect for the productivity shock plus a correction for the fact that the markup shock releases some labor, in proportion to the share of workers it employs, to the rest of the economy:

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = -\tilde{\lambda}_k - \frac{1-\theta_0}{\Lambda_L}\lambda_k\left(\frac{\mu_k^{-1}}{\Lambda_L} - 1\right) + \frac{\lambda_k \mu_k^{-1}}{\Lambda_L}.$$

The first two terms are just the negative of what we had for a productivity shock and the final term adjusts for the difference between productivity and markup shocks. Since this economy does not have intermediate inputs, we have $\tilde{\lambda}_k = \lambda_k$. We can therefore simplify this further to

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log \mu_k} = \theta_0\left[\frac{\mu_k^{-1}}{\Lambda_L} - 1\right]\lambda_k = \theta_0\left[\frac{\mu_k^{-1}}{\sum_i \lambda_i \mu_i^{-1}} - 1\right]\lambda_k. \tag{22}$$

Unlike the case with productivity shocks, where Cobb-Douglas was the case with constant misallocation where workers would not be reallocated in response to productivity shocks, here $\theta_0 = 0$, or Leontief production, is the case where misallocation is constant. This is due to the fact that with perfect complementarity, the minimum amount of goods needed to consume 1 unit aggregate consumption does not depend on the markups. Hence, in this case, changes in the markup have no effect on the extent of misallocation. However, if $\theta_0 \neq 0$, then labor is reallocated across producers in response to the shock. In particular, labor is reallocated from k to the other industries. Hence, depending on whether k is inefficiently too small or too large, the shock can have a positive or negative impact. The higher is θ_0, the larger the magnitude of this impact since reallocation is

monotonically increasing in the elasticity of substitution in this case.[39,40]

Finally, we consider an example illustrating how our results can be applied to compute the gains from removing distortions, and we relate this popular measure of misallocation to our notion of changes in allocative efficiency.

Example 3.5 (Measuring Allocative Efficiency)**.** Changes in allocative efficiency are marginal in nature: we define them as the gap that opens up between the competitive allocation and the passive allocation in response to shocks, assuming both start at the same initial allocation. In this sense, allocative efficiency can be interpreted locally as a counterfactual conditional. Our measure is especially convenient because it can be measured directly *without* any knowledge about the production functions over and above expenditure data.

There are other measures of allocative efficiency in the literature. For example, Restuccia and Rogerson (2008), Hsieh and Klenow (2009), and many others adopt the following measure: by how much would output increase if all wedges were eliminated. The larger the number, the more misallocated the economy. And then one can try to look at changes in misallocation as changes in this measure over time.

This measure captures the change in the distance from the unobserved efficient frontier of the economy. Our notion of changes in allocative efficiency is an entirely different concept: it measures the contribution to TFP along the equilibrium path, rather than the change in the distance of the economy from the efficient production frontier. The former concept relies on a tightly parameterized model, with full information about the underlying production functions, as this is the only way of placing the unobserved global efficient production frontier. Our concept, by contrast, is local, and can be measured without strong parametric assumptions.

[39] As discussed earlier, relabelling allows us to turn markup shocks into shocks to tax wedges and vice versa. For example, consider the horizontal economy with N homogenous industries who produce using labor as their only input, and sell to the household. These industries are perfectly competitive, but their labor inputs are subject to an industry specific tax τ_i for each industry i. The revenues generated from these taxes is rebated lump sum to the household. From our results above, we can write

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log(1+\tau_k)} = \theta_0\left[\frac{(1+\tau_k)^{-1}}{\Lambda_L} - 1\right]\lambda_k.$$

Once again, in this second best world, a reduction in a wedge can decrease output and welfare by worsening misallocation.

[40] This result relates to Epifani and Gancia (2011), who show that in a horizontal economy, output losses from markup dispersion are monotonically increasing in the elasticity of substitution (when compared to social planner's allocation).

To see how these two could move in opposite directions, consider a horizontal economy with two producers and $\sigma > 1$. Suppose $A_1/\mu_1 = A_2/\mu_2$, with $A_1 > A_2$. Then, at steady-state, the two firms split demand and workers evenly. Now, if the first firm receives a positive productivity shock, workers are reallocated from 2 to 1, and allocative efficiency improves. However, locally, this can increase dispersion in $1/\mu_i$, and thereby increase the gains from eliminating the markups. In general, the relationship between our measure of changes in allocative efficiency and changes in the gains from eliminating distortions do not need to move together. Of course, our ex-ante results can be used to allow us to think about the gains from reducing wedges.

As an example, consider the horizontal economy in Figure 3. Using formula (22), we deduce the first-order impact on output from shrinking all relative markups towards 1, by considering a transformation of each markup $\hat{\mu}_i = t\mu_i + (1 - t)$. When $t = 1$, this transformation leaves markups as they are. On the other hand, $t = 0$ eliminates all distortions in the economy. In this case, we can write

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\,t} = \sum_k \frac{\mathrm{d}\log Y}{\mathrm{d}\log\mu_k}\frac{\mathrm{d}\log\mu_k}{\mathrm{d}\,t},$$

which, substituting in formula (22), can be re-written as

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\,t} = \theta_0 \frac{Var_\lambda(\mu^{-1})}{E_\lambda(\mu^{-1})}.$$

The variance of a random variable divided by its mean is called the index of dispersion. Hence, the gain to output from shrinking all markups depends positively on the elasticity of substitution θ_0 and the expenditure-share weighted index of dispersion in $1/\mu_i$. Using the dispersion of wedges to measure misallocation is common in the literature following Hsieh and Klenow (2009), and this example shows how our formulas, in these simple cases, will deliver such results.

3.3 Multiple Factors

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

be the aggregate income share of factor f.

3.3.1 Productivity Shocks

Proposition 3.3 (Productivity Shocks with Multiple Factors)**.** *In response to a productivity shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log A_k} = \sum_j (\theta_j - 1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)} - \sum_{f'}\tilde{\Psi}_{(f')}\frac{\mathrm{d}\log\Lambda_{f'}}{\mathrm{d}\log A_k}, \frac{\Psi_{(f)}}{\Lambda_f}\right). \tag{23}$$

Given $\mathrm{d}\log\Lambda_f/\mathrm{d}\log A_k$*, we know, from Theorem 2.2 that*

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

Proposition 3.1 can be seen immediately as a special case of Proposition 3.3. When there is only one factor, $\sum_f \tilde{\Psi}_{(f)}\,\mathrm{d}\log\Lambda_f/\mathrm{d}\log A_k$ is a constant vector and drops out of the covariance term, which allows us to recover the result for the case with a single factor.

As a bonus, this Proposition 3.3 also determines how factor income shares for different factors move in response to productivity shocks in a distorted economy. This is a question of independent interest for analyses of inequality and growth. We can rewrite equation (23) as the following linear system

$$\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} = \Gamma\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A_k} + \delta^{(k)}, \tag{24}$$

with

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

and

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

When there is only factor, $\Gamma = \mathbf{0}$, and we are left with only $\delta^{(k)}$, which is just equation (21). To see intuition for equation (24), imagine a negative shock $\mathrm{d}\log A_k < 0$ to industry k. For fixed factor prices, every industry i will substitute across its inputs in response to this shock. Suppose that $\theta_i < 1$, so that industry i substitutes *towards* those inputs that

are more reliant on industry k, captured by $\tilde{\Psi}_{ik}$. Now, if those inputs are also more reliant on factor f, captured by a high $Cov_{\Omega^{(i)}}\left(\tilde{\Psi}_{(k)}, \Psi_{(f)/\Lambda_f}\right)$, then substitution by i will increase demand for factor f.

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

However, when there are multiple factors, the change in demand for factors will affect relative factor prices, and the change in relative factor prices will set off additional rounds of substitution in the economy that we must account for, and this is the role Γ plays. Crucially, the matrix Γ does not depend on which industry k has been shocked, since it encodes how changes in factor income shares affect factor income shares.

Indeed, for a given set of factor prices, the shock to k affects demand for each factor, and hence the factor income shares, and this is measured by the $F \times 1$ vector $\delta^{(k)}$. This change in the factor income shares then causes further substitution through the network, leading to additional changes in factor demands and prices. The impact of the change in the relative price of factor f' on the demand for factor f is measured by the ff'th element of the $F \times F$ matrix Γ. The movements in factor shares are the fixed point of this process, i.e. the solution of equation (24).

Proposition 3.3 is tightly connected with the results in Baqaee and Farhi (2017), which characterize the change in sales shares (for goods and factors) in a general efficient multisector economy. However, whereas for an efficient economy, changes in factor income shares determine the second-order impact of shocks on output, for a distorted economy, this information is required even for the first-order impact of shocks.

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

Example 3.6 (Horizontal Economy with Multiple Factors)**.** We have

$$\Gamma = -(\theta_0 - 1)\begin{pmatrix} Cov_b(\tilde{\Psi}_{(L)}, \Psi_{(L)}) & Cov_b(\tilde{\Psi}_{(K)}, \Psi_{(L)}) \\ Cov_b(\tilde{\Psi}_{(L)}, \Psi_{(K)}) & Cov_b(\tilde{\Psi}_{(K)}, \Psi_{(K)}) \end{pmatrix}, \tag{25}$$

and

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

Substituting in the values and solving the system of equations (24), using Proposition 3.3,

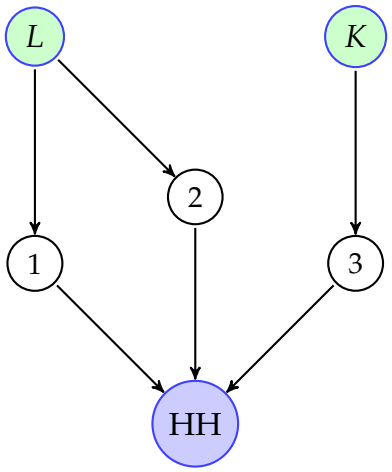


Figure 4: An economy with two factors of production L and K. The subgraph from L to the household contains a cycle, and hence can be subject to misallocation. On the other hand, there is only a unique path connecting K to the household, so there is no misallocation.

and noting that $\lambda_i = \tilde{\lambda}_i$ for all i, we find that

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

but

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

The details for this example are in Appendix B.[41] A lesson is that changes in allocative efficiency are only present for shocks to industries 1 and 2 which share a factor of production, but not for industry 3 which has its own factor of production. Moreover, the changes in allocative efficiency for shocks to industries 1 and 2 only depends on the markups in these two industries and not on the markup in industry 3.

We can also use Proposition 3.4 to analyze models with decreasing-returns-to-scale (or equivalently, limited factor reallocation).

Example 3.7 (Decreasing-returns-to-scale/limited reallocation)**.** Suppose there are N goods, each good i is produced using a specific factor f_i and a generic non-specific factor L with a Cobb-Douglas production function. The weight of specific factor is α for every i. This could capture limited factor reallocation, but it is also equivalent to letting each industry

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

have decreasing returns to scale. The household has uniform preferences over all the final goods with weights $b_i = 1/N$ at steady-state. Then,

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \frac{1}{N}\left(\frac{(1-\alpha)^2(1-\theta_0)+\Lambda_L\mu_k\theta_0}{\Lambda_L\mu_k(1-\alpha+\alpha\theta_0)}\right),$$

where $\Lambda_L = (1-\alpha)/N\sum_i \mu_i^{-1}$. When $\alpha = 1$, the economy is acyclic, there are no changes in allocative efficiency and $\mathrm{d}\log Y/\mathrm{d}\log A_k = 1/N$. But when $\alpha = 0$, we recover the horizontal economy with only one factor of production, which in general features changes in allocative efficiency so that $\mathrm{d}\log Y/\mathrm{d}\log A_k \neq 1/N$. Intermediate values of α interpolate between these two extremes.

Our last example is Cobb-Douglas, which, as per our earlier discussion, features constant misallocation even with multiple factors

Example 3.8 (Cobb-Douglas)**.** Let $\theta_i = 1$ for every i. Then

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log A_k} = \tilde{\lambda}_k. \tag{26}$$

We end this section but showing how our results can be extended to cover the impact of shocks to markups/wedges, in a manner similar to Proposition 3.2. The intuition for this result is the same as it was in the single factor case: a markup shock has the same effect as a negative productivity shock, with the additional fact that we must account for the fact that compared to a negative productivity shock, a markup shock leads the corresponding industry to release some resources to the rest of the economy. These released resources can eventually be translated into released factor uses.

3.3.2 Markup Shocks

Proposition 3.4 (Markup/Wedge Shocks with Multiple Factors)**.** *In response to a markup shock, the following linear system describes the change in factor income shares:*

$$\frac{\mathrm{d}\log\Lambda_f}{\mathrm{d}\log\mu_k} = -\sum_j(\theta_j-1)\mu_j^{-1}\lambda_j Cov_{\tilde{\Omega}^{(j)}}\left(\tilde{\Psi}_{(k)}+\sum_{f'}\tilde{\Psi}_{(f')}\frac{\mathrm{d}\log\Lambda_{f'}}{\mathrm{d}\log\mu_k},\frac{\Psi_{(f)}}{\Lambda_f}\right)-\lambda_k\frac{\Psi_{kf}}{\Lambda_f}. \tag{27}$$

Given $\mathrm{d}\log\Lambda_f/\mathrm{d}\log\mu_k$*, we know, from Theorem 2.3 that*

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

To isolate the importance of this new term, we go back to the the Cobb-Douglas economy with a markup shock, where the only source factor reallocation comes from the fact that the industry which increases its markups releases some factors.

Example 3.9 (Cobb Douglas)**.** Let $\theta_j = 1$ for every j, which is the Cobb-Douglas special case. Now, applying Proposition 3.2, we get

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

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

3.4 Endogenous Productivities and Markups/Wedges

Our results can be used in contexts where markups/wedges or productivities are endogenous in order to characterize the effects of some more primitive fundamental shocks. Imagine that some additional structure imposed on the model gives rise to endogenous productivities and markups. Consider some more fundamental disturbance θ. For any such structure, the vector of equilibrium productivities and markups μ can be expressed as a vector functions $A(\theta)$ and $\mu(\theta)$ of the vector θ.

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. What our results can be used for is to understand the consequences of these endogenous movements in markups and productivities. Indeed, using our results in conjunction with the chain rule, we can write

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\,\theta} = \tilde{\lambda}'\frac{\mathrm{d}\log A}{\mathrm{d}\log\theta} - \tilde{\lambda}'\frac{\mathrm{d}\log\mu}{\mathrm{d}\log\theta} - \tilde{\Lambda}'\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log A}\circ\frac{\mathrm{d}\log A}{\mathrm{d}\,\theta} - \tilde{\Lambda}'\frac{\mathrm{d}\log\Lambda}{\mathrm{d}\log\mu}\circ\frac{\mathrm{d}\log\mu}{\mathrm{d}\,\theta}, \tag{28}$$

where $\circ$ denotes the element-by-element (Hadamard) product of two matrices, and where the expressions $\mathrm{d}\log\Lambda/\mathrm{d}\log A$ and $\mathrm{d}\log\Lambda/\mathrm{d}\log\mu$ are given as a function of the structural microeconomic parameters of the model by Propositions 3.3 and 3.4.

We provide a fully worked out example along these lines in Section 4.3 when we consider a model with nominal rigidities. The model can be recast as a model with endogenous markups ensuring that the relevant prices stay constant. In this case, we actually explicitly solve these endogenous markups as a function of the underlying productivity and monetary policy shocks. We then apply the chain rule in conjunction with our results exactly as in equation (28) to characterize the effects of these shocks.

4 Applications

In this section, we pursue some quantitative applications of our results. First, we use our reduced-form results to measure changes in allocative efficiency in the US over time, and to decompose the Solow-residual into changes in pure-technology and changes in allocative efficiency. Next, we we calibrate a simplified version of our parametric model to match firm-level markup and size data, as well as input-output data. We compute output elasticities with respect to firm-level and industry-level shocks to productivity and markups, and we compare these elasticities to those implied by the perfectly competitive and Cobb-Douglas models. We also use these elasticities to approximate aggregate volatility. Finally, we end the section by showing how our results can be used to study models with nominal rigidities that have inefficient steady-states and input-output networks. All of our results emphasize that reallocation forces that we emphasize play a quantitatively significant role in determining aggregate output and TFP.

We work with the annual US input-output data from the BEA, dropping the government, noncomparable imports, and second-hand scrap industries. The dataset contains industrial output and inputs from 1997 to 2015 with 66 industries. We calibrate the expenditure share parameters to match the input-output table, and we use three alternative measures of markups estimated for Compustat firms. We only have markups and sales data at the microeconomic level for publicly listed firms in the US from Compustat. To

extrapolate to the whole economy, we therefore make the assumption that Compustat firms are representative of the overall economy in the sense that the sales-weighted distributions of markups by industry and their transition matrices for the overall economy are the same as for Compustat. We then combine these data with input-output data at the industry level from the BEA to aggregate the economy.

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

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

Each markup series comes with its own pros and cons. The GP markups require measurements of the capital stock and industry-level estimates of the equity risk premium, both of which are notoriously difficult to measure. The DE markups on the other hand, rely on more parametric methods, and their changes over time are potentially biased in the presence of capital-biased technical change. We use the GP markups for our benchmark numbers, and we report numbers for the other two markup series in the tables and in Appendix A.[42] Despite their differences, all three markup series show an increasing

[42]Note that our method allows for capital-biased technical change. In particular, we can measure changes

average markup over the sample.

4.1 Decomposing the Solow Residual

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

Conditional on markups and the input-output matrix at a given point in time t, we can approximate $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1}) = -\tilde{\Lambda}'_{t-1}\Delta \log \Lambda_{t-1}$ from $t-1$ to t using the change in observed factor income shares. Then we can decompose the Solow residual using equation (8). The results are plotted in Figure 5 using the GP markups. The sum of the red (allocative efficiency), yellow (factor under-counting), and purple ("pure" technology) lines add up to give the cumulative change in the Solow residual. Since we are interested in long-run trends, we assume that the only factors are labor and capital, and we abstract away from barriers to reallocation of factors like adjustment costs and variable capacity utilization.

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

As documented by Gutierrez (2017), average markups have been increasing in all three of our markup series. Given the increase in the average markup, and the growing profit share in the economy, how then can we claim that allocative efficiency has increased over the same period? The key lies in realizing that markups on average have increased primarily because firms that charge large markups have gotten larger. While on average, markups are trending upwards, the average change in log markups has not been increasing. If on average, markups are not increasing, but firms with high markups are getting larger, then this implies that allocative efficiency in the economy *must* be increasing. Of course, to quantify and weigh the various changes correctly, we need to use the weights in equation (8).

in allocative efficiency independently of the nature of productivity shocks (Hicks neutral or factor biased). For more details see the discussion in Section 2.8.

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

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

4.2 A Quantitative Structural Model

In this section, we use our structural results to explore quantitatively the importance of markup distortions. We calibrate a simplified version of the parametric model in Section 3.1.

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

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

Value-added consists of labor and capital inputs

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

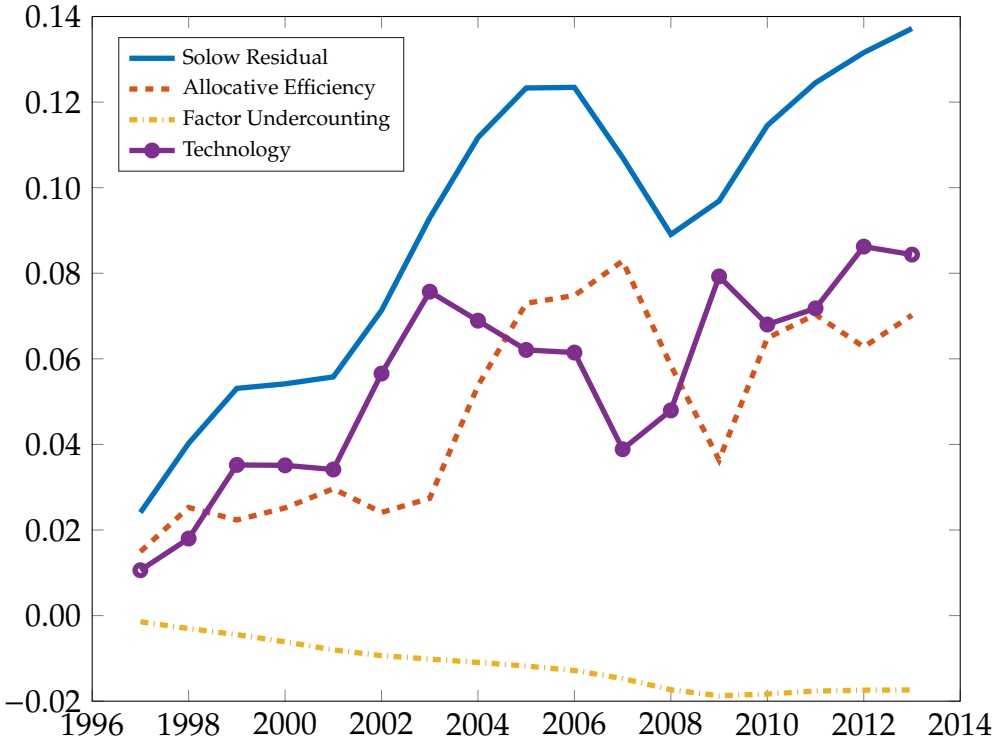


Figure 5: The decomposition in equation (8) using the Gutiérrez and Philippon (2016) markup data.

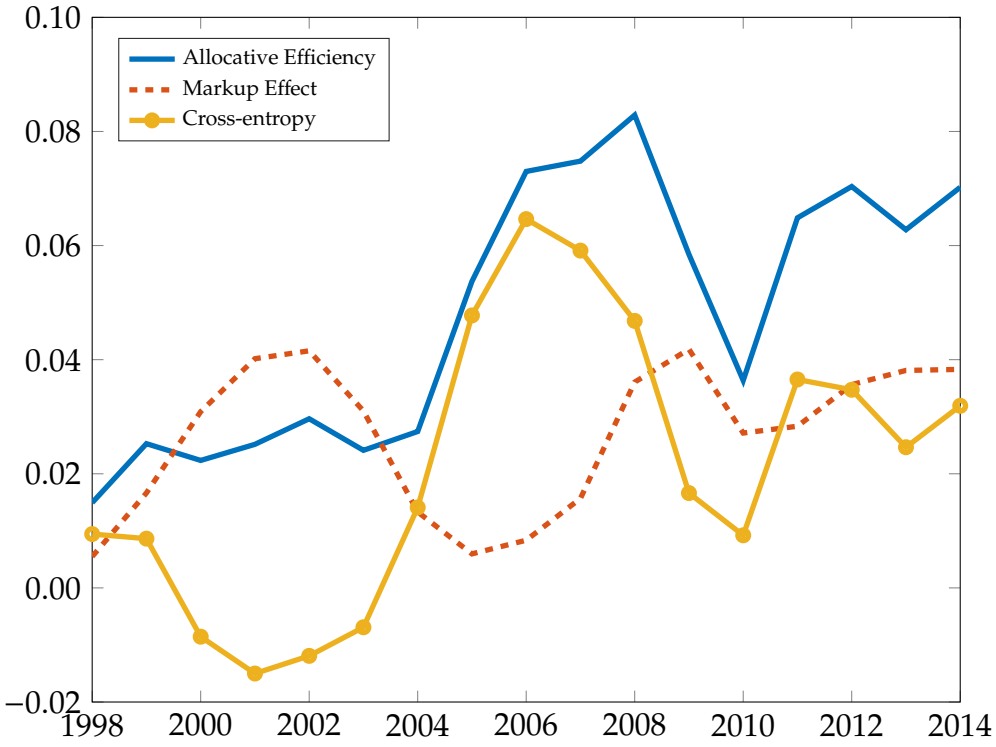


Figure 6: The cumulated contribution of (minus) changes in log markups $-\tilde{\lambda}'_{t-1}\Delta \log \mu_t$ and of changes in cross entropy $\Delta H(\tilde{\Lambda}_{t-1}, \Lambda_{t-1})$ using the Gutiérrez and Philippon (2016) markup data. The sum of the two components give the overall cumulated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

and the intermediate input consists of inputs from other industries

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

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

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

Following our previous work (Baqaee and Farhi, 2017), and drawing on estimates from Atalay (2017) and Boehm et al. (2014), we set $\theta = 0.3$, $\theta_0 = 0.4$, and $\varepsilon \approx 0$. We set $\eta_i = 1$ which is a focal point in the literature about the micro-elasticity of substitution between labor and capital [43] Finally, we set $\xi = 8$, which is within the range of estimates of the variety-level elasticity of substitution from the industrial organization and international trade literatures. An elasticity of $\xi = 8$ is also consistent with our measure of average markups in the benchmark model, assuming a Dixit and Stiglitz (1977) market structure.

Gains from Reducing Markups

We first use the model to calculate the gain to aggregate TFP from eliminating markups.To do this, consider changing markups $\mu_{i,k}$ for producer i in industry k to be $t\overline{\mu}_{i,k} + (1-t)$ where $\overline{\mu}_{i,k}$ is the original markup. This transformation shifts all markups towards unity by t percentage points. Table 1 reports the elasticity of the TFP gains to the markup reduction. Using the benchmark GP markups, the implied aggregate TFP gains from a 1% reduction in markups are about 0.73% points.

To extrapolate the aggregate TFP gains from eliminating markups, we use a second-order approximation. We have just computed the elasticity of aggregate TFP gains to a reduction in markups at the observed inefficient allocation. We also know from Theorem 2.1 that *at* the efficient equilibrium, the elasticity of aggregate TFP to a reduction in markups is zero. Hence, using the insights of Hotelling (1938) and Theil (1967), we can construct a second-order estimate of the gains from reducing markups by averaging the first order gains at the initial and terminal (efficient) allocation. This means that to

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

a second order, the gains from eliminating markups are given by $e^{0.73/2+0/2} - 1 \approx 44\%$ of aggregate TFP. This corresponds to the equivalent of the area of the Harberger deadweight loss triangle in our general equilibrium model, which can also be interpreted as a second-order approximation, as shown by Hotelling (1938).

The LI markups imply somewhat smaller gains, and the DE markups imply the largest gains. Indeed, with LI markups, the aggregate TFP gains from a 1% reduction in markups is 0.65%, and the gains from eliminating markups is 38%. With DE markups, these numbers are 0.79% and 48%.

Interestingly, we find that the gains from reducing markups have increased substantially since the start of the sample for all three series. For example, using our benchmark GP markups, we find that the gains from eliminating markups are $e^{0.09/2+0/2} - 1 \approx 4.5\%$ in 1997, much smaller that the corresponding number of 44% in 2014. As we described in example 3.5 in Section 3.1, this finding is logically consistent with our finding that allocative efficiency has improved since the start of the sample, since the counterfactual comparison in the two scenarios is conceptually different. Our concept of change in allocative efficiency measures contributions of changes in allocative efficiency to aggregate TFP along the equilibrium path, whereas changes in gains from eliminating markups measure changes in the distance from the unobserved efficient frontier of the economy.

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

Our estimate that eliminating markups in the US economy in 2014 would increase TFP by about 40% raises the estimated cost of monopoly distortions by two orders of magnitude compared to the famous estimates of 0.1% of Harberger (1954). Essentially, the reasons for this dramatic difference is that we use firm-level data, whereas Harberger only had access to sectoral data, and that the dispersion of markups is higher across firms within

	Gutiérrez-Philippon	Lerner Index	De Loecker-Eeckhout
2014	0.73	0.65	0.79
1997	0.09	0.16	0.55

Table 1: Gains from shrinking all markups towards 1 by 1% as given by $\exp(\sum(\sum((\mathrm{d}\log Y/\,\mathrm{d}\log\mu_i(\mu_i-1)/\mu_i)))(0.99-1))-1$ for the various markup series for the beginning and end of our sample.

	Benchmark	CD+CES	CD+CD	VA Benchmark	VA CD + CES	VA CD + CD
GP	0.73%	0.76%	0.47%	0.31%	0.30%	0.19%
LI	0.65%	0.67 %	0.43 %	0.31%	0.30%	0.19%
DE	0.79%	0.84%	0.27%	0.40 %	0.41 %	0.13%

Table 2: Gains in aggregate TFP from shrinking all markups towards 1 by 1% for different markup series, and different structural models. CD+CES preserves the input-output structure, but sets all elasticities except ξ equal to one. CD + CD sets all elasticities to one. VA specifications eliminate the input-output matrix and use value-added production functions, a la Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). For VA, all elasticities except ξ equal to one, VA CES uses the same elasticities as the benchmark model, and VA CD sets all elasticities of substitution equal to one.

a sector than across sectors. Moreover, the relevant elasticity of substitution is higher in our exercise than in Harberger's since it applies across firms within a sector rather than across sectors. Finally, we properly take into account the input-output structure of the economy to aggregate the numbers in all industries whereas Harberger only focused on manufacturing. Of course, both our estimate and Harberger's are static, taking as given the level of productivity in the economy. Of course, markups may be playing an important role in incentivizing innovation and entry, so that exogenously eliminating markups may harm productivity. In Section 5, we discuss how one might try to account for these forces. Briefly, even if markups do play an important role in incentivizing innovation, they also distort the allocation of resources and our calculation is aimed at quantifying this latter effect.

Volatility of Aggregate TFP

We perform comparative statics in both productivity and markup shocks at both the firm and industry level[44]

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

Assuming productivity shocks and markup shocks are independent and identically distributed, we can approximate the volatility of output using

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

Hence, the Euclidean norm $||D_{\log A} \log Y||$ of the Jacobian of $\log Y$ with respect to $\log A$ gives the degree to which microeconomic productivity shocks are not "diverisified" away in the aggregate. Similarly, $||D_{\log \mu} \log Y||$ measures the diversification factor relative to markup shocks.[45]

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

First, consider the case of productivity shocks: as mentioned previously, the benchmark model is more volatile than the competitive model for both sets of shocks. However, the more interesting comparison is with respect to the distorted Cobb-Douglas economy. As explained in Section 2.5, the allocation of factors is invariant to productivity shocks in the Cobb-Douglas model. Hence, the Cobb-Douglas model lacks the reallocation channel,

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

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

	Benchmark	Competitive	Cobb-Douglas	Passive
Firm Productivity Shocks (GP)	0.0476	0.0376	0.0396	0.0396
Firm Markup Shocks (GP)	0.0451	0.0000	0.0345	0.0000
Industry Productivity Shocks (GP)	0.3162	0.3109	0.3261	0.3261
Industry Markup Shocks (GP)	0.0279	0.0000	0.0297	0.0000
Firm Productivity Shocks (LI)	0.0502	0.0372	0.0415	0.0415
Firm Markup Shocks (LI)	0.0602	0.0000	0.0413	0.0000
Industry Productivity Shocks (LI)	0.3188	0.3079	0.3377	0.3377
Industry Markup Shocks (LI)	0.0356	0.0000	0.0382	0.0000
Firm Productivity Shocks (DE)	0.0621	0.0346	0.0398	0.0398
Firm Markup Shocks (DE)	0.0663	0.0000	0.0295	0.0000
Industry Productivity Shocks (DE)	0.3299	0.3133	0.3618	0.3618
Industry Markup Shocks (DE)	0.0648	0.0000	0.1166	0.0000

Table 3: Diversification factor for different productivity and markup shocks at firm and industry level for different specifications of the model. A diversification factor of 1 means that the variance of microeconomic shocks moves aggregate variance one-for-one. A diversification factor of 0 means that microeconomic shocks are completely diversified away at the aggregate level. GP corresponds to the Gutiérrez and Philippon (2016) markups, LI is markups according to the Lerner Index, and DE is using markup data from De Loecker and Eeckhout (2017).

and hence can tell us in which direction the reallocation force is pushing. In the case of industry-level shocks, the benchmark model is slightly less volatile than the Cobb-Douglas model, whereas in the case of firm-level shocks, the benchmark model is significantly more volatile.

A partial intuition here relates to the elasticities of substitution: whereas industries are complements, firms within an industry are strong substitutes. Recall that loosely speaking, changes in allocative efficiency scale with the elasticity of substitution minus one. Firm-level shocks cause a considerable amount of changes in allocative efficiency whereas industry-level shocks cause much milder changes. At both levels of aggregation, these changes in allocative efficiency amplify some shocks and mitigate some others compared to the Cobb-Douglas model with no change in allocative effiency.[46] On the whole, at the firm level, the changes in allocative efficiency are so large that they dwarf the "pure" technology effects picked up by the Cobb-Douglas model and amplify the volatility of these shocks. By contrast, at the industry level, changes in allocative efficiency are more moderate and turn out to slightly mitigate the volatility of these shocks.

This intuition is confirmed in the first two rows in Figure 7, where we plot the output elasticity with respect to productivity shocks to specific firms or industries relative to their cost-based Domar weight (i.e. relative to the Cobb-Douglas model) and to their revenue-based Domar weight (i.e. relative to the competitive model). We find considerable dispersion in the response of the model relative to both, but much more so at the firm level than at the industry level.

Next, consider the effects of markup shocks. In this case, the distorted Cobb-Douglas economy is not necessarily a very natural benchmark since even with Cobb-Douglas, shocks to markups will reallocate factors across producers. Nonetheless, it is still instructive to compare the benchmark model to the Cobb-Douglas one to find that a similar lesson applies as with productivity shocks. The volatility of firm-level shocks is amplified relative to Cobb-Douglas while the volatility of industry-level shocks is attenuated relative to Cobb-Douglas. This follows from the fact that industries are more complementary than firms, and hence, in line with the intuition from example 3.9, the effect of the shock are monotonically increasing in the degree of substitutability.

The last two rows of Figure 7 plot the output elasticity with respect to markup shocks to specific firms or industries relative to their cost-based Domar weight or revenue-based

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

Domar weight. In the case of markup shocks and in contrast to productivity shocks, these ratios can no longer be interpreted as comparisons with counterfactuals, and instead should be taken as sensible normalizations of the way to report the results.

4.3 Nominal Rigidities

Finally, we apply our general framework to study the effects of sticky prices in economies with arbitrary production structures.[47] In general, sticky prices can be modeled via variable markups: markups which move to ensure that the relevant nominal prices stay constant. This is the point of connection with our framework. We show how to solve for these endogenous markups, and then to trace their impact on the economy.

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

These exercises are useful demonstrations of how to apply our results more generally in cases where markups are endogenous or variable. They also draw attention to the fact that the typical loglinearization of New Keynesian models around undistorted steady states is potentially misleading.

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

$$P_y Y = M,$$

where M is the instrument of monetary policy.

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

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

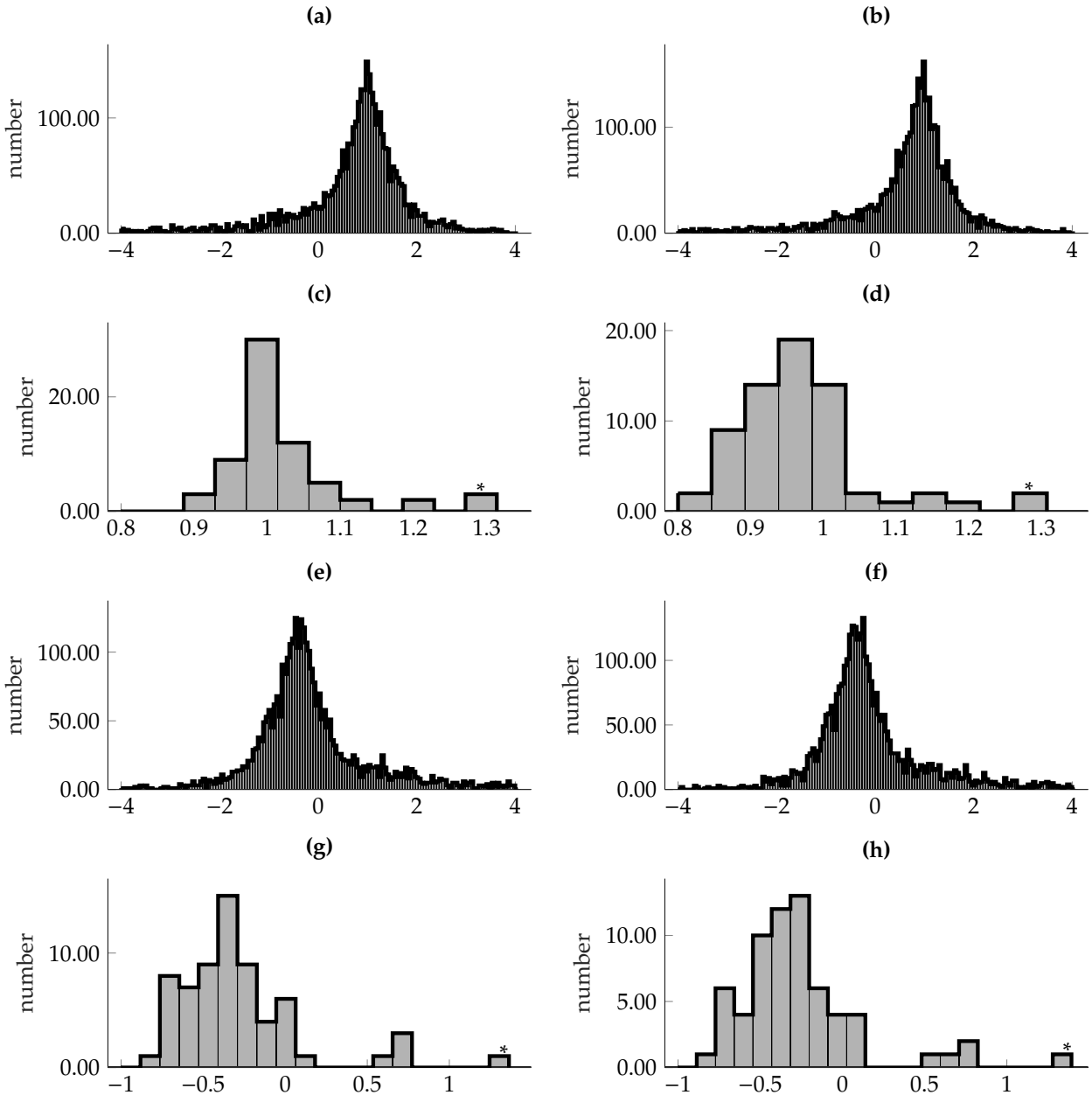


Figure 7: The top two rows are histograms of $\mathrm{d}\log Y/\mathrm{d}\log A$ relative to λ and $\tilde{\lambda}$ for firm-level and industry-level shocks respectively. The bottom two rows are $\mathrm{d}\log Y/\mathrm{d}\log \mu$ relative to λ and $\tilde{\lambda}$ for firm-level and industry-level shocks respectively. The bunching at the extremes marked with a star arise from a truncation performed solely for displaying purposes. In all cases, the degree of dispersion around the response implied by the competitive model or the size of the producer is substantial. Distributions have been truncated at 4 standard deviations.

shocks $\mathrm{d}\log A$ the vector of changes in factor prices $\mathrm{d}\log w$:

$$\mathrm{d}\log\mu = (e_s'\tilde{\Psi}e_s)^{-1}e_s'\tilde{\Psi}(\mathrm{d}\log A - \tilde{\alpha}\,\mathrm{d}\log w). \tag{29}$$

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

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

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

$$\mathrm{d}\log Y - \tilde{\Lambda}'\,\mathrm{d}\log L = \tilde{\lambda}'\,\mathrm{d}\log A - \tilde{\lambda}'e_s\,\mathrm{d}\log\mu + \mathrm{d}\,H(\tilde{\Lambda},\Lambda), \tag{30}$$

where $\mathrm{d}\log\mu$ is now determined according to (29).[48]

To finish our characterization, we need to make an assumption about labor supply, since equation (30) takes the change in factor supply as given. To fix ideas, we follow convention in the New Keynesian literature, and assume labor is the only factor of production and that the household utility function takes the form:

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

where ν is the Frisch elasticity of labor supply. Under these conditions, we can combine equations (30), (14), and (15) to explicitly solve out the labor supply decision.

Proposition 4.1 (Nominal Rigidities)**.** *Suppose that labor is the only factor of production. Then*

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

with

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

[48] See Proposition B.1 in the Appendix for a formal statement and proof of these results.

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

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

Equation (30) gives the change in aggregate TFP in response to either monetary or technology shocks in a New Keynesian type environment and decomposes it into a "pure" change in technology component and change in allocative efficiency component. Proposition 4.1 characterizes the corresponding changes in output.

Typically, New Keynesian models are log-linearized around the efficient steady state, and in those cases, the changes in allocative efficiency are second-order and are therefore neglected.[49] In these cases, our results deliver the same conclusion.[50] Outside of these special cases, for inefficient steady states, changes in allocative efficiency are not zero, and our results then permit us to isolate these effects.

In their important study, Pasten et al. (2016) characterize the response of output to shocks in a model with Calvo frictions and production networks. They write the input-output matrix at the industry level, and suppose that some fraction δ_i of firms in industry i have flexible prices. Their sharpest analytical result is for the case with log utility in consumption and an infinite Frisch elasticity of labor supply. In this special case, the response of output to shocks takes a very simple form — sticky prices act like shock absorbers to productivity shocks. Using Proposition 4.1, we can recover their result and shed light on why this happens.

Proposition 4.2. *[Pasten et al. 2016] Suppose that utility is log in consumption and the Frisch elasticity of labor supply is infinite. Then Proposition 4.1 implies*

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

[49] See, for example, Galí (2008).

[50] The case of the efficient steady-state is immediate in Proposition 4.1, since at the efficient steady-state $\Psi_{(L)}$ is a vector of all ones, and the covariances are all zero. Hence, at the efficient steady state, Proposition 4.1 implies that the change in allocative efficiency, to a first order, is

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

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

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

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

Proposition 4.2 is simple to interpret: for productivity shocks, the impact of a shock is the same as one in a model with flexible prices, but where productivity shocks affect *only* some fraction δ_i of each industry i's costs, or in other words, productivity shocks are attenuated by some weight δ_i at each industry. The impact of monetary policy shocks on output is given by $1 - b'(I - \delta\tilde{\Omega})^{-1}\delta\tilde{\alpha}$, or the total share of value-added which is sticky in the economy.

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

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

which is nonzero.[51]

We now turn our attention to an application of our results. We calibrate a version of our quantitative model from Section 4, but augmented with a labor-leisure choice, and a Frisch elasticity of labor supply of $\nu = 1/2$, which is broadly consistent with the recommendation of Chetty et al. (2011). We create two copies of each firm in our sample, one copy has sticky prices while the other has flexible prices. We then use Proposition 4.1 to compute the impact of monetary policy shocks and firm-level productivity shocks.

Monetary Policy Shocks

In Table 4, we show the response of output and aggregate TFP for a shock to the money supply. We consider different specifications of the model with different elasticities of substitution. For each specification, we also consider some speculative scenarios where

[51]The fact that changes in allocative efficiency are not required to compute the changes in output is a generic property of infinitely elastic labor supply, and does not depend on parametric assumptions about the production functions, see footnote (32) for more details.

the flexible firms have slightly higher or lower markups than the sticky firms (keeping the average constant).

We find that the movements in aggregate TFP, which are purely caused by changes in allocative efficiency, can become very large if the elasticity of substitution across firms is high, and if average markups are not the same between the flexible and sticky firms. The size and sign of these movements depend crucially on the correlation between price rigidity and markups. These results suggest that empirical work on understanding the correlation between microeconomic price rigidities and the levels of markups could be of great importance.[52] They also suggest that the literature on the New Keynesian model, by assuming that the steady-state is efficient, and by assuming away correlations between price rigidities and levels of markups, could potentially be missing important first-order effects.[53]

Comparing the benchmark model to the one-sector model, we also recover the famous insight by Basu (1995) that intermediate goods can increase stickiness. If intermediate inputs are sticky, then flexible firms adjust their prices less in response to shocks. The degree of amplification caused by the intermediate-input share is hump-shaped in the fraction of firms δ that have sticky prices. In the limit, as all firms become sticky, the intermediate input share becomes irrelevant, and the same occurs when all firms become flexible. Here we show that these effects are stronger when the fraction of sticky prices is 20% (roughly corresponding to a horizon of 5 quarters) than if it is 50% (roughly corresponding to a horizon of 2 quarters).[54]

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

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

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

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

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

Productivity Shocks

In general, we can decompose the effect on output, from a productivity shock, into three components: (1) the "pure" technology effect captured by the cost-based Domar weight, (2) changes in allocative efficiency, and (3) changes in quantity of factors supplied. In the context of a model with sticky prices, the second effect is subtle, since markups adjust in response to a productivity shock to ensure that sticky prices do not adjust. In Figure 8, we plot the allocative efficiency component of the response relative to the total output response for firm-level productivity shocks, as a histogram. We color the flexible and sticky priced firms differently. We see that changes in allocative efficiency, which are typically neglected by the positive literature on nominal rigidities, are sizable as a fraction of the total output response. Furthermore, we find that the sign on this term changes. For flexible firms, the "pure" technology impact of the productivity shock is amplified by improvements in allocative efficiency, while for sticky priced firms, the opposite is true.

In Figure 9, we plot the histogram of the output responses to firm-level productivity shocks relative to the sales shares of the affected firms. Since the sales share gives the output response in the competitive model, we can think of this as measure the degree of amplification or attenuation of productivity shocks relative to a model with perfectly flexible firms and perfect competition. We see that the impact of shocks is more attenuated when the firm is sticky than when it is flexible, but the effect is not necessarily zero, even when the firm has sticky prices. This is because of the upstream reallocation that happens in the case when a firm with sticky prices receives a productivity shock. On the whole, the model with nominal rigidities attenuates shocks relative to the competitive flexible model (most of the mass in the histogram is to the left of 1), but this is by no means universal, and there are some firms for whom shocks are amplified.

Finally, in Figure 10, we plot the response of the shock on output when the firm is sticky relative to when the firm is flexible. Since all the mass it to the left of 1, we can conclude that the flexible firm always affects output more than the sticky firm, but of course, the degree of attenuation is highly dispersed.

5 Robustness and Extensions

In this section, we discuss how our results could be extended to address some limitations of our analysis: the absence of fixed costs, the lack of entry, and the importance of nonlinearities. These issues introduce additional forces and mechanisms into the model,

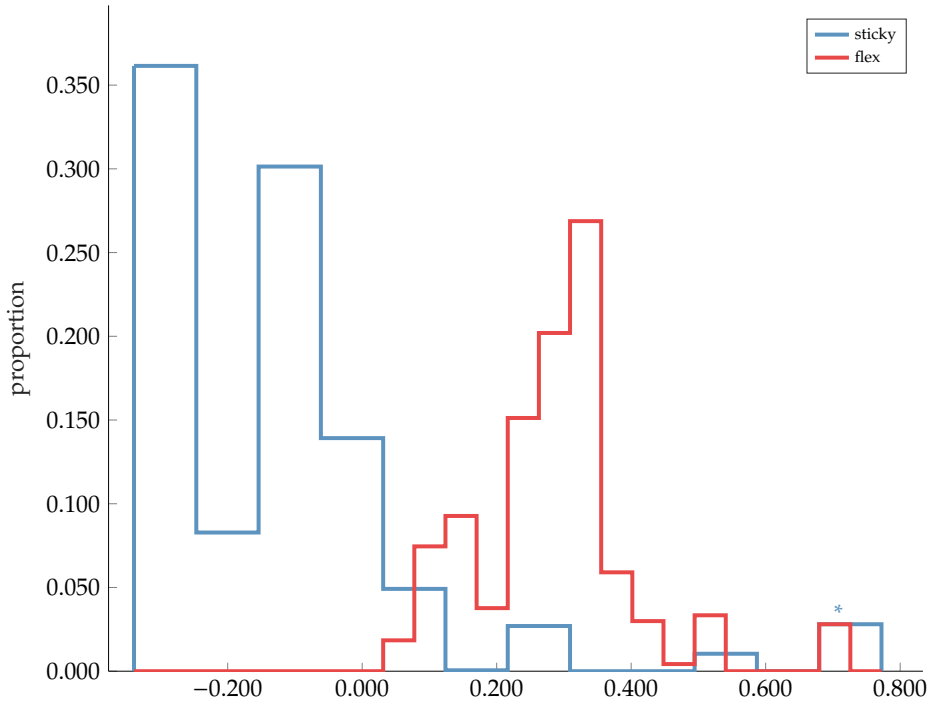


Figure 8: The change in allocative efficiency relative to the change in output for productivity shocks to the firms in Compustat with sticky and flexible prices. We set $\nu = 1/2$, and use GP markups.

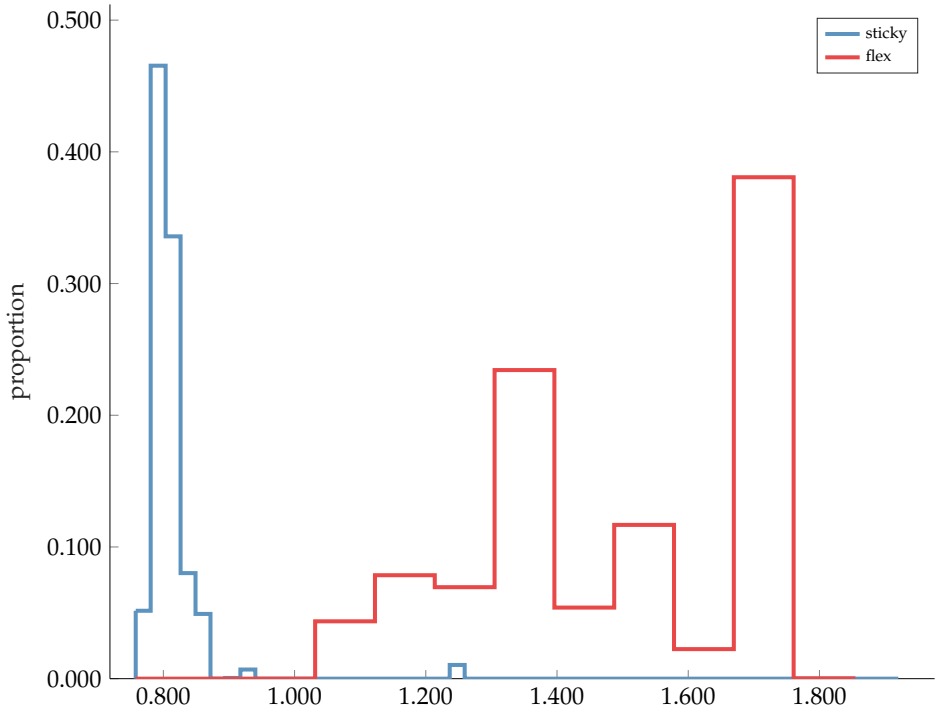


Figure 9: Response of output for productivity shocks to the sticky and flexible firms in Compustat relative to the response in a fully competitive model. We set $\nu = 1/2$, and use GP markups.

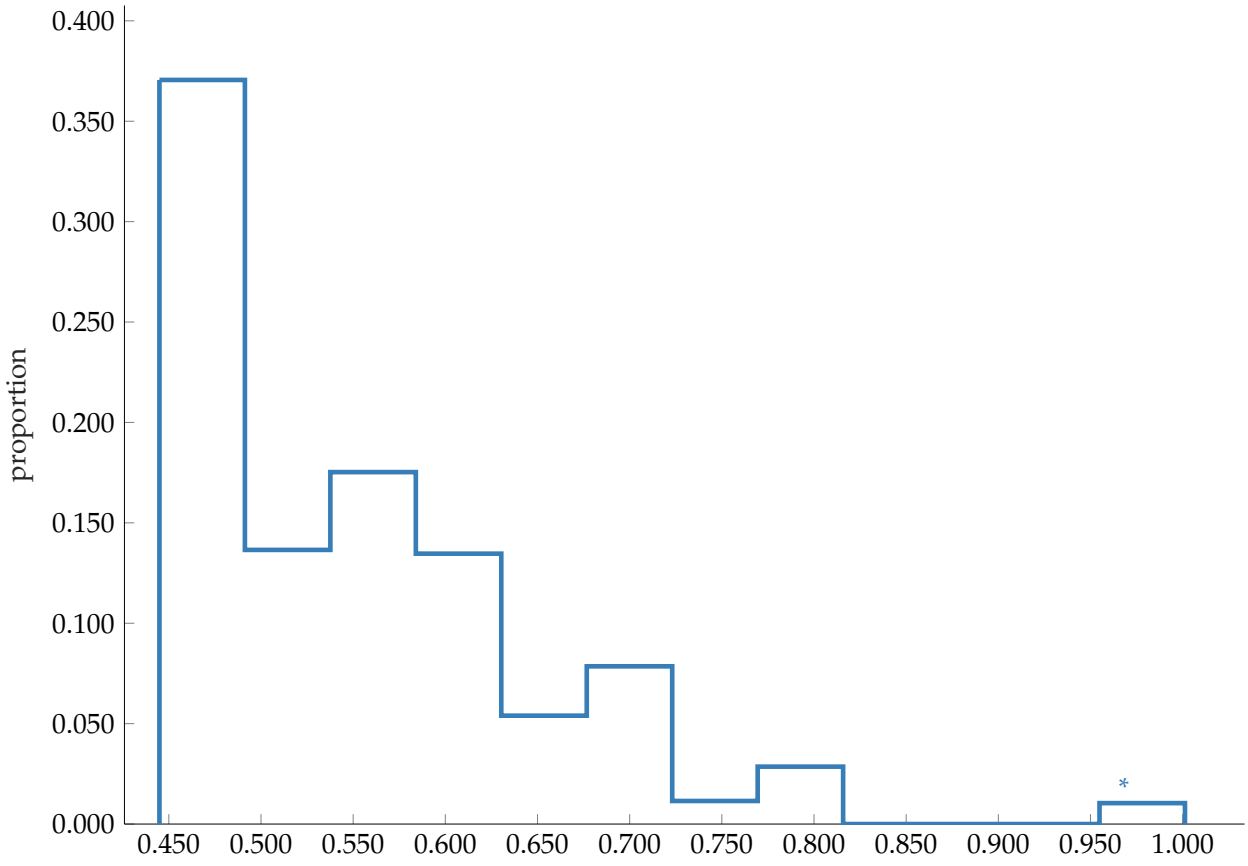


Figure 10: Response of output relative to a firm-level productivity shock for the sticky firm relative to the flexible firm. We set $\nu = 1/2$, and use GP markups.

and which we plan to squarely focus on in future work. However, we show here that the intuitions gleaned from the basic framework continue to be useful in analyzing these more complex models.

5.1 Fixed Costs

To add fixed overhead costs to the model, we need to separate variable cost from total cost. Under these conditions $\tilde{\Omega}$ becomes total-variable-cost based rather than total-cost based. It is the matrix whose ijth element is

$$\tilde{\Omega}_{ij} = \frac{\mathrm{d}\log \mathbf{C}_i}{\mathrm{d}\log p_j} = \frac{p_j x_{ij}}{VC_i}, \tag{31}$$

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

$$\mathrm{d}\log Y = \tilde{\lambda}' \,\mathrm{d}\log A - \tilde{\lambda}\,\mathrm{d}\log \mu + \mathrm{d}\,H(\tilde{\Lambda}, \Lambda). \tag{32}$$

where Λ is the share of income going to each factor including the payments to the infra-marginal fixed costs, but the cost-based Domar weights $\tilde{\lambda}$ and $\tilde{\Lambda}$ are constructed using the variable-cost based $\tilde{\Omega}$.[55]

5.2 Entry

If firms are earning positive economic profits, we might expect that this would induce entry and competition. In this section, we extend our basic results to cover the case where there is free entry. First, we establish that our results can easily be applied to cases where entry is "wasteful." Typically, entry can be economically meaningful due to several reasons: (1) it increases product variety; (2) it reduces markups; (3) it selects

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

$$\mathrm{d}\log Y^p - \mathrm{d}\log Y = (\hat{\lambda} - \tilde{\lambda})' \,\mathrm{d}\log A - \tilde{\lambda}' \,\mathrm{d}\log \mu + \mathrm{d}\,H(\tilde{\Lambda}, \Lambda), \tag{33}$$

where $\hat{\lambda} = b'(I - \varphi^{-1}\tilde{\Omega})^{-1}$ and φ is a diagonal matrix whose iith element is the share industry i's output that is not used for fixed costs in the initial allocation. If the equilibrium is efficient, then (33) is always zero. For reference, $\tilde{\lambda} = b'(I - \tilde{\Omega})^{-1}$ and $\lambda = b'(I - \Omega)^{-1}$.

the most productive firms; (4) it counters decreasing returns to scale at the firm level. If we assume that firms have constant-returns-to-scale production functions, are ex-ante identical, markups/wedges are exogenous, and there is no returns to product variety, then entry is entirely socially wasteful.[56] Under these conditions, our results survive unchanged.

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

No External Economies

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

Let industry k output be given by

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

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

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

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

Proposition 5.1 (No External Economies)**.** *Suppose that there is free entry subject to fixed costs,*

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

that there are no external economies of scale, and that firms have constant returns to scale. Then

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

and

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

This set up allows us to span the basic framework in, for example, Autor et al. (2017), who argue that the decline in the labor share in the US in recent times is due to an increase in the size of low-labor-share firms.[57]. They consider a model with entry and where labor intensity falls with the scale of operation because of overhead labor costs. The mechanism Autor et al. (2017) emphasize is an increase in the industry-level elasticity of substitution, whereby more productive firms are able to capture more demand over time.[58] Our interpretation, consistent with their empirical evidence, is that low-labor share firms charge higher markups, and that these markups are part of the reason why they have a low labor share. Our results in Section 4 indicate that these changes in the composition of firms within industries have increased aggregate productivity by improving allocative efficiency.

Constant External Economies

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

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

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

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

Proposition 5.2 (Constant External Economies)**.** *Suppose that there is free entry subject to fixed entry costs. In addition, suppose that entrants in each industry charge the same markup and have access to the same homothetic production function, and each entrant pays a fixed cost of entry in units of the industry's input. Then*

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

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

Increasing External Economies

Finally, we sketch how our framework would relate to models with richer heterogeneity and entry properties by turning on mechanisms (1), (3), and (4).[59] In particular, we allow for the possibility that entry can induce increasing returns to scale at the industry level, where the industry becomes more productive as more firms enter.[60]

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

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

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

Index producers in industry k by i in such a way that their idiosyncratic productivity $z_k(i)$ is weakly increasing, and suppose that i is distributed according to the distribution

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

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

$\phi_k(i)$. We assume that the price of the composite industry k good can be written as

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

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

Proposition 5.3 (Increasing External Economies)**.** *Suppose that there is free entry, fixed entry costs, and fixed overhead costs paid after productivity and markup draws conditional on operating, and external economies of scale in the form of love for variety effects. Then*

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

and

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

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

The model of Baqaee (2016) becomes a special case of this setup. Proposition 5.3 shows that by incorporating "fictitious" factors, and hence separating profits due to distortions from competitive rents from decreasing returns, we can extend our results to a much more general class of models going even beyond neoclassical production by allowing for increasing returns to scale at the macroeconomic level.

Importantly, allowing for increasing returns makes the "pure" technology effect of a shock endogenous, and this endogenous response of productivity is what the final (and new) term in Proposition 5.3 accounts for. There are two reasons why technology becomes endogenous: first, an increase in the mass of entrants improves the productivity of firms through its effect on product variety. Second, an increase in the mass of entrants can change the distribution of productivity in each industry, since the change in cutoff value at which firms enter the market changes. Interestingly, we see that these cut-off values for the productivity and their derivatives are sufficient statistics for understanding the first-order impact of the shocks. Analyzing this model in any more detail and expressing

these cutoff values and their derivatives as a function of the structural microeconomic parameters of the model is well beyond the scope of this paper, but we pursue it in ongoing work.

5.3 Nonlinear Impact of Shocks

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

Consider the quantitative parametric model in Section 4. Let $\delta_k(i)$, $\mu_k(i)$, and $A_k(i)$ denote firm i in industry k's share of industry sales, markups, and productivity. Define

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

and

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

Define the *efficiency* of each firm i in industry k to be $e_k(i) = 1/\mu_k(i)$. Consider a transformation $e_k(i) = t_k + (1 - t_k)\overline{e_k(i)}$ which shrinks dispersion in markups relative to its steady-state value $\overline{\mu_k(i)} = 1/\overline{e_k(i)}$. This transformation keeps $\mu_k = 1/e_k$ constant. Define the revenue-based Domar weight of industry k by λ_k.

Proposition 5.4 (Competitive Isomorphism)**.** *Consider an economy where $\mu_k = 1$ for every k, then*

$$\frac{\mathrm{d}\log Y}{\mathrm{d}\log t_k} = \lambda_k \frac{\mathrm{d}\log A_k}{\mathrm{d}\log t_k} \tag{34}$$

and

$$\frac{\mathrm{d}^2\log Y}{\mathrm{d}\log t_k^2} = \lambda_k \frac{\mathrm{d}\log \lambda_k}{\mathrm{d}\log A_k}\left(\frac{\mathrm{d}\log A_k}{\mathrm{d}\log t_k}\right)^2 + \lambda_k \frac{\mathrm{d}^2\log A_k}{\mathrm{d}\log t_k^2}, \tag{35}$$

with $\mathrm{d}\log A_k/\mathrm{d}\log t_k \geq 0$.[61]

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

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

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

6 Conclusion

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

We apply our findings to the US, where our measure of wedges correspond to markups. We find that from 1997-2015, allocative efficiency in the US accounts for about half of aggregate TFP growth. We also find that the gains from reducing markups have increased since 1997, that eliminating markups will increase aggregate TFP by around 40% (up to a second order approximation). These numbers are substantially higher than classic estimates like those of Harberger (1954).

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

is endogenous). In particular, we show how to characterize the oft-neglected potential effects of monetary policy shocks on TFP.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Basu, S. and J. Fernald (2001). Why is productivity procyclical? why do we care? In *New developments in productivity analysis*, pp. 225–302. University of Chicago Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dixit, A. K. and J. E. Stiglitz (1977). Monopolistic competition and optimum product diversity. *The American Economic Review*, 297–308.

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

Edmond, C., V. Midrigan, and D. Y. Xu (2015). Competition, markups, and the gains from international trade. *The American Economic Review 105*(10), 3183–3221.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Varian, H. R. (1992). *Microeconomic Analysis* (3 ed.). WW Norton & Company.

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

A Additional Figures

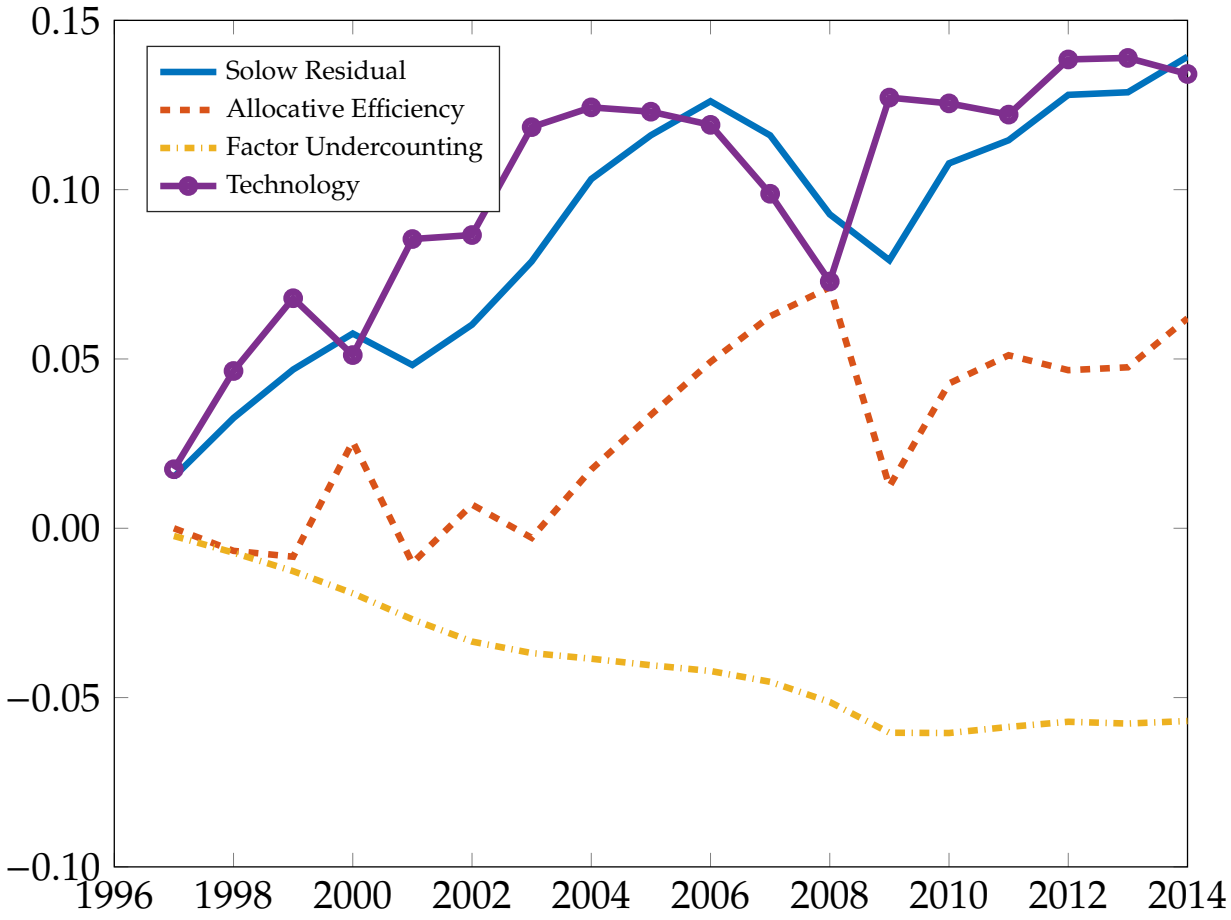


Figure 11: Decomposition of the Solow Residual using LI markups.

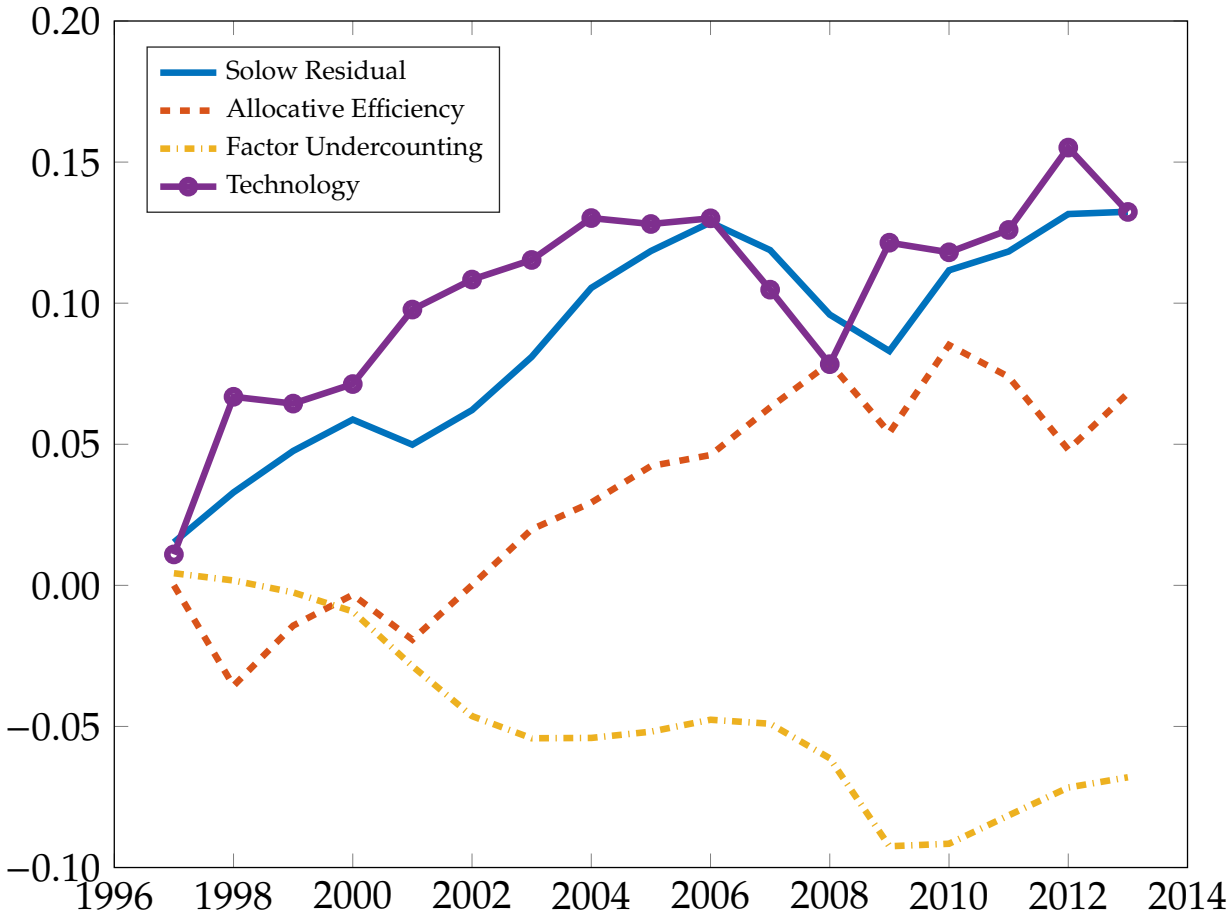


Figure 12: Decomposition of the Solow Residual using DE markups.