NBER WORKING PAPER SERIES

A SHORT NOTE ON AGGREGATING PRODUCTIVITY

David Baqaee
Emmanuel Farhi

Working Paper 25688
http://www.nber.org/papers/w25688

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2019, Revised July 2019

We are grateful to Natalie Bau for comments. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

A Short Note on Aggregating Productivity
David Baqaee and Emmanuel Farhi
NBER Working Paper No. 25688
March 2019, Revised July 2019
JEL No. E0,L0

ABSTRACT

This paper discusses two simple decompositions for aggregate productivity analysis in the presence of distortions and in general equilibrium. The first is a generalization of Baqaee and Farhi (2017) and the second is due to Petrin and Levinsohn (2012). In the process, we propose a new "distorted" Solow residual which, contrary to the traditional Solow residual, accurately measures changes in aggregate productivity in disaggregated economies with distortions. These formulas apply to any collection of producers ranging from one isolated producer to an industry or to an entire economy. They can be useful for empiricists and theorists alike. Potential applications of these formulas include: (1) decomposing aggregate productivity into its microeconomic sources, separating technical and allocative efficiency;(2) aggregating microeconomic estimates (for example, from natural experiments) to assess macroeconomic effects; (3) constructing and interpreting aggregate counterfactuals. Despite their simplicity, the formulas are general, allowing for production networks, multi-product firms, and non-constant returns. They are also entirely nonparametric. They only assume market clearing and cost minimization.

David Baqaee
UCLA
315 Portola Plaza
Los Angeles, CA
baqaee@econ.ucla.edu

Emmanuel Farhi
Harvard University
Department of Economics
Littauer Center
Cambridge, MA 02138
and NBER
emmanuel.farhi@gmail.com

A Short Note On Aggregating Productivity

David Rezza Baqaee UCLA

Emmanuel Farhi* Harvard University

July 19, 2019

This paper discusses two simple decompositions for *aggregate* productivity analysis in the presence of distortions and in general equilibrium. The first is a generalization of Baqaee and Farhi (2017) and the second is due to Petrin and Levinsohn (2012). In the process, we propose a new "distorted" Solow residual which, contrary to the traditional Solow residual, accurately measures changes in aggregate productivity in disaggregated economies with distortions. These formulas apply to any collection of producers ranging from one isolated producer to an industry or to an entire economy. They can be useful for empiricists and theorists alike. Potential applications of these formulas include: (1) decomposing aggregate productivity into its microeconomic sources, separating technical and allocative efficiency; (2) aggregating microeconomic estimates (for example, from natural experiments) to assess macroeconomic effects; (3) constructing and interpreting aggregate counterfactuals. Despite their simplicity, the formulas are general, allowing for production networks, multi-product firms, and non-constant returns. They are also entirely nonparametric. They only assume market clearing and cost minimization.

1 Introduction

A group of producers receives external resources (that it does not produce) and produces final resources (that it does not use). Changes in final outputs produced by the collective depend not only on the amount of external inputs, but also on the technology and the distribution of resources amongst the firms in the group. At any point in time, the allocation

*We are grateful to Natalie Bau for comments.

of resources across producers is determined through general equilibrium with arbitrary distortions captured as wedges.

In this context, we discuss two formulas for changes in aggregate output in terms of microeconomic sufficient statistics. The first formula is a generalization of Baqaee and Farhi (2017), and the second is a re-expression of Petrin and Levinsohn (2012). In the process, we propose a new "distorted" Solow residual which, contrary to the traditional Solow residual, accurately measures changes in aggregate productivity in disaggregated economies with distortions.

These formulas can be used to measure and decompose the sources of aggregate productivity. They can also be used to go from microeconomic information up to the behavior of the aggregate. Finally, they can be used to derive and gain intuition for counterfactuals in structural models. Despite their simplicity, the formulas are general, allowing for production networks, multi-product firms, and non-constant returns. They are also entirely nonparametric. They only assume market clearing and cost minimization.

First, and foremost, this paper relates to the literature on growth accounting. The measurement of aggregate productivity dates back at least to Solow (1957), who used an aggregate production function and perfectly competitive markets to measure changes in technical efficiency over time. Domar (1961) and Hulten (1978) extended Solow's result to perfectly competitive economies with many producers and intermediate inputs. On the other hand, Hall (1990) extended Solow's result to imperfectly competitive economies with a single producer and an aggregate production function. The closest to this paper is Baqaee and Farhi (2017), who extended the Solow-Hulten-Hall results to closed economies with multiple producers allowing for both intermediate inputs and imperfect markets, and Petrin and Levinsohn (2012), who provided a decomposition of the Solow residual in inefficient economies. This paper goes beyond the results in Baqaee and Farhi (2017) by generalizing that analysis to subsets of the economy (for instance, a single industry, country, region, time period, etc.) and not just to the economy as a whole. This paper also relates to the growing literature on production networks in closed and open economies, see for example Carvalho and Tahbaz-Salehi (2018) and the references therein.

Second, these results are related to classic results in welfare economics, like Hotelling (1938), Hicks (1965), and Meade (1962), who derived measures of societal welfare. In particular, the second decomposition relates to the formula proposed by Harberger (1964) for cost-benefit analysis. In these papers, to measure the effects of a policy change on societal welfare, Kaldor-Hicks transfers are assumed to operate that neutralize the distributional

consequences of the policy change. The second decomposition validates and nests the formula proposed by Harberger (1964), and shows that it can be used to study purely *positive* questions rather than *normative* ones — questions like how should we measure and predict aggregate productivity growth in an industry.

The outline of the paper is as follows. In Section 2, we describe the environment and define our notation. In Section 3, we define the notion of aggregate productivity, which can be broken down into technical productivity and allocative efficiency. In Section 4, we state and discuss the first decomposition, showing how it can be used to measure and decompose aggregate productivity. In Section 5, we state and discuss the second decomposition, showing how it can be used to combine producer-level estimates to estimate changes in aggregate output and the Solow residual. The two decompositions have very different data requirements, which means that each can be useful depending on the availability of data and the context. We conclude in the final section.

2 Setup

In this section, we introduce the setup and define the primitives of the model.

Group: Let $\mathcal{I}$ be a set of firms producing different goods and interacting with the rest of the economy and with each other. For convenience, we refer to $\mathcal{I}$ as a *group* of producers. Each producer $i \in \mathcal{I}$ produces gross output y_i using intermediate inputs y_{ij} with $j \in \mathcal{I}$ produced *inside* the group as well as *external* inputs l_{if} with $f \in \mathcal{F}$, where $\mathcal{F}$ denotes the set of external inputs. The technology of producer i is described by a constant-returns gross production function $F_i(\cdot, A_i)$ indexed by a productivity shifter A_i. Without loss of generality, we assume that $\partial \log F / \partial \log A_i = 1$ at the initial equilibrium. As we describe below, the assumption of constant-returns is also made without loss of generality.

The gross output y_i of a producer $i \in \mathcal{I}$ is used as an intermediate input by other producers $j \in \mathcal{I}$ in the group and as by "final" users outside the group in respective amounts $\sum_{j \in \mathcal{I}} y_{ji}$ and c_i. The the total quantity of external inputs used by all firms in $\mathcal{I}$ by L_f for each $f \in \mathcal{F}$.

We keep the amount of structure to a minimum. We impose two fundamental assumptions. The first assumption is that for all $i \in \mathcal{I}$, producer i minimizes its cost, taking prices as given, and charges a price p_i given by a markup μ_i times its marginal cost. The second assumption is market clearing: $y_i = c_i + \sum_{j \in \mathcal{I}} y_{ji}$ for all $i \in \mathcal{I}$, and $L_f = \sum_{i \in \mathcal{I}} l_{if}$

for each $f \in \mathcal{F}$.

For example, for the neoclassical growth model, in a given period, external inputs are labor and capital; for an open economy, they are labor, capital and imported intermediate goods; for an industry, they are labor, capital, and intermediates from other industries, and so on.[1]

Generality: This setup is very general. It does not require that technical change be Hicks-neutral and allows for arbitrary biased technical change at the producer level. The setup allows for decreasing returns at the producer level via the introduction of producer-specific quasi-fixed factors earning each producer's decreasing-returns profits.[2] It can handle multi-product producers by choosing, for each producer, an arbitrary product as the output, and treating the other products negative inputs.

Finally, the setup allows for arbitrary wedges other than markups. Such wedges can always be modeled as markups via the introduction of "fictitious producers" as explained in Baqaee and Farhi (2017). For example a wedge τ_{ij} in the use of some good or factor j by producer i can be captured by introducing a fictitious producer transforming good or factor j into good or factor j for producer i and charging a markup τ_{ij}. We use the terminology markups/wedges throughout to remind the reader of this observation.

Shocks: We consider a perturbation of equilibrium outcomes. This perturbation changes productivities $\mathrm{d}\log A$, external inputs $\mathrm{d}\log L$, markups/wedges, $\mathrm{d}\log \mu$, and the composition of final demand. These changes may arise due to some deep endogenous mechanism or they may be exogenous. Our two theorems will hold regardless of the origin or cause of the perturbation. An example would be a change in foreign preferences that increases the flow of external inputs $\mathrm{d}\log L$ a small open economy $\mathcal{I}$, and that the flow of foreign inputs endogenously changes $\mathrm{d}\log A$ the productivity of producers in $\mathcal{I}$, and these firms then adjust their markups $\mathrm{d}\log \mu$.

Aggregate (group) output: The final output c_i of each producer i is the quantity of goods produced by i which leaves $\mathcal{I}$. We denote nominal aggregate group output by $PY =$

[1] We can also consider dynamic closed economies whose only external input is labor (for instance, by including capital accumulation as part of the group).

[2] Increasing returns can also be accommodated, but only to some limited extent, by allowing these quasi-fixed factors to be local "bads", i.e. to receive negative payments over some range. However, care must be taken because increasing returns introduce non-convexities in the cost minimization over variable inputs, and our formulas only apply when variable input demand changes smoothly.

$\sum_{i\in\mathcal{I}} p_i c_i$. Final expenditure shares are given by $b_i = p_i c_i/(PY)$. For a perturbation of an initial equilibrium, we define the log change in aggregate output at constant prices as the final-expenditure-weighted log change in final output:

$$\mathrm{d}\log Y = \sum_{i\in\mathcal{I}} b_i \,\mathrm{d}\log c_i. \tag{1}$$

This is a first-order approximation (in the shocks) to the chain-weighted log change in aggregate output. Since our results will provide first-order approximations (for small shocks) to the log change in group output at constant prices, they will by implication also provide first-order approximations to the chain-weighted log change in group output.[3]

Accounting definitions: The sales shares of producer i is $\lambda_i = p_i y_i/(PY)$ and the sales share of external input f is $\Lambda_f = p_f L_f/(PY)$. Sales shares are also called revenue-based Domar weights. The revenue-based input-output matrix is denoted by Ω, where $\Omega_{ij} = (p_j y_{ij})/(p_i y_i)$ is the expenditure of i on j as a fraction of the revenues of i, where j can be a good in $\mathcal{I}$ or an external input from $\mathcal{F}$. The revenue-based Leontief-inverse matrix is define by $\Psi = (I-\Omega)^{-1}$, where Ψ_{ij} captures the direct and indirect exposures of i to j in revenues. It follows from market-clearing identities that the vector of sales shares (or Domar weights) is given by $\lambda' = b'\Psi$.

We define the cost-based input-output matrix $\tilde{\Omega}$ by $\tilde{\Omega}_{ij} = \mu_i \Omega_{ij}$, the cost-based Leontief-

[3]To deal with chain-weighted log changes in output, we can proceed as follows. All the primitives of the economy and all the equilibrium variables are indexed by a shifter/shock t. This dependence is recorded with a superscript t. The initial equilibrium corresponds to $t = 0$, for which the superscripts are suppressed for expositional convenience. An equilibrium with $t > 0$ is a perturbed equilibrium. For example, then an infinitesimal shock ds shifts group wedges by $d\log\mu_i^s$, group productivities by $d\log A_i^s$, external input uses by $d\log L_f^s$, group final uses by $d\log c_i^s$, etc. In general, group output cannot be defined in levels, but only in changes along a path $s\in[0,t]$ of intermediate equilibria connecting the initial equilibrium indexed by $s=0$ to the final equilibrium indexed by $s=t$:

$$\Delta\log Y = \int_0^t \sum_{i\in\mathcal{I}} b_i^s d\log c_i^s. \tag{2}$$

When final demand is homothetic, group output can be defined in levels. The change in group output $\Delta\log Y$ in equation (2) is then independent of the specific integration path between the initial and final equilibria. When final demand is not homothetic, group output cannot be defined in levels but only in changes via equation (2). The change in group output $\Delta\log Y$ in equation (2) then depends on integration path. As is well known, this undesirable property is unavoidable. All the results in the paper can be interpreted as first-order approximations in the shifter/shock t, assuming that the equilibrium changes smoothly with it. The path-dependence considerations that we just discussed do not pose problems for these approximations. For example equation (1) is a first-order approximation of equation (2).

inverse matrix by $\tilde{\Psi} = (I - \tilde{\Omega})^{-1}$, and the cost-based Domar weights by $\tilde{\lambda}' = b'\tilde{\Psi}$. The quantities $\tilde{\Omega}_{ij}$ and $\tilde{\Psi}_{ij}$, capture the direct and direct and indirect exposures of i to j in costs. They measure the direct and direct and indirect elasticity of the cost of i to the price of j, keeping the price of external inputs constant. The cost-based Domar weight $\tilde{\lambda}_i$ captures the direct and indirect exposure of final users to i.

The difference between the cost- and revenue-based Domar weights of a good or external input is due to multiple marginalization of markup/wedges in supply chains leading to the final sales.

Allocation Rule: We define the allocation matrix $\mathcal{X}$, where $\mathcal{X}_{ij} = y_{ij}/y_j$ is the share of the physical output of producer j used by producer i. Unlike the input-output and Leontief inverse matrices, this matrix does not make any use of prices and simply describes how the output of each producer j is allocated across the different producers i. It is a compact way of summarizing the physical allocation of resources in the group.

The level of aggregate output (at constant prices) can be written as a function $\mathcal{Y}(L, A, \mathcal{X})$ of the vector L of external input quantities, of the vector A of group productivities, and of the allocation rule $\mathcal{X}$. By varying the allocation matrix, we can map out every feasible allocation the set of producers $\mathcal{I}$ can attain.

3 Decomposing Changes in Aggregate Output

Following Baqaee and Farhi (2017), we decompose the change in output as

$$\mathrm{d}\log Y = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log L}\,\mathrm{d}\log L}_{\Delta\text{Inputs}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log A}\,\mathrm{d}\log A}_{\Delta\text{Technical Efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}}\,\mathrm{d}\,\mathcal{X}}_{\Delta\text{Allocative Efficiency}} \quad . \tag{3}$$

The first term corresponds to the increase in output due to changes in external inputs $\mathrm{d}\log L$ which are being brought into $\mathcal{I}$. The second term is the change in output due to changes in the production functions $\mathrm{d}\log A$ inside $\mathcal{I}$. The final term corresponds to changes in output due to a reallocation of resources across producers in $\mathcal{I}$. In equilibrium, the allocation rule may change due to a variety reasons, including changes in the composition of final demand, changes in productivities, changes in wedges, or changes in external inputs.

The decomposition in equation (3) leads to a natural definition of changes in aggregate

productivity, denoted by $\mathrm{d}\log A_{\mathcal{I}}$ as[4]

$$\mathrm{d}\log A_{\mathcal{I}} = \mathrm{d}\log Y - \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log L}\,\mathrm{d}\log L}_{\Delta\text{Inputs}} = \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \log A}\,\mathrm{d}\log A}_{\Delta\text{Technical Efficiency}} + \underbrace{\frac{\partial \log \mathcal{Y}}{\partial \mathcal{X}}\,\mathrm{d}\,\mathcal{X}}_{\Delta\text{Allocative Efficiency}} \quad .$$

This measures changes in aggregate output that are not directly accounted for by purely technological impact of the change in external inputs. In other words, the change in aggregate productivity is the combination of the effects of changes in technical and allocative efficiency.

4 First Decomposition: Extending Baqaee and Farhi (2017)

We can now state the first decomposition and discuss its relationship to equation (3).

Theorem 1. *The following first-order approximation holds:*

$$\mathrm{d}\log Y = \underbrace{\sum_{f\in\mathcal{F}} \tilde{\Lambda}_f\,\mathrm{d}\log L_f}_{\Delta\text{Inputs}} + \underbrace{\sum_{i\in\mathcal{I}} \tilde{\lambda}_i\,\mathrm{d}\log A_i}_{\Delta\text{Technical Efficiency}} \underbrace{- \sum_{i\in\mathcal{I}} \tilde{\lambda}_i\,\mathrm{d}\log \mu_i - \sum_{f\in\mathcal{F}} \tilde{\Lambda}_f\,\mathrm{d}\log \Lambda_f}_{\Delta\text{Allocative Efficiency}},$$

where the different terms correspond to the partial derivatives of the function $\mathcal{Y}(L, A, \mathcal{X})$, so that $\partial \log \mathcal{Y}/\partial \log L_f = \tilde{\Lambda}_f$, $\partial \log \mathcal{Y}/\partial \log A_i = \tilde{\lambda}_i$, and $(\partial \log \mathcal{Y}/\partial \mathcal{X})d\mathcal{X} = -\sum_{i\in\mathcal{I}} \tilde{\lambda}_i\,\mathrm{d}\log \mu_i - \sum_{f\in\mathcal{F}} \tilde{\Lambda}_f\,\mathrm{d}\log \Lambda_f$.

Theorem 1 generalizes Baqaee and Farhi (2017) to open systems. In other words, the decomposition offered there can be applied not just to entire closed economies (closed systems), but also to open subparts of such economies (open systems): a region in a country, a country in the world, an industry within a country, or even the set of producers in a given period of time or state of nature.

Theorem 1 can be used to separate changes in technical productivity from changes in allocative efficiency. For instance, Baqaee and Farhi (2017) use it to decompose changes in technical and allocative efficiency for the US using firm-level markup data. Theorem 1 also shows that the effect of endogenous reallocation $\mathrm{d}\,\mathcal{X}$ on output can be deduced purely from changes in the expenditure share on external inputs $\mathrm{d}\log \Lambda$.

[4]Note that just like for aggregate output, we only define the changes in aggregate productivity. In general, the level of aggregate productivity cannot be defined unambiguously except in special cases with a particular structure.

To unpack the intuition for Theorem 1, it is useful to start by considering the special cases when there are no markups/wedges in the group $\mathcal{I}$ so that the allocation of resources within the group is efficient. Importantly, there could still be markups/wedges outside of the group.

No Markups/Wedges: When there are no markups/wedges inside $\mathcal{I}$, there are no changes in allocative efficiency. Indeed, it follows from the envelope theorem that $\partial \log \mathcal{Y} / \partial \mathcal{X} = 0$, so that changes in the allocation of resources $\mathrm{d}\,\mathcal{X}$ have no first-order impact on the group's output $(\partial \log \mathcal{Y} / \partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = 0$. This in turn implies that output changes are only due to changes in external inputs and changes in technical efficiency $(\partial \log \mathcal{Y} / \partial \log L)\,\mathrm{d} \log L + (\partial \log \mathcal{Y} / \partial \log A)\,\mathrm{d} \log A$. Moreover, the group's output elasticities to external input quantities and productivities are given by the Domar weights $\partial \log \mathcal{Y} / \partial \log L_f = \Lambda_f$ and $\partial \log \mathcal{Y} / \partial \log A_i = \lambda_i$, so that

$$\mathrm{d} \log Y = \sum_{f \in F} \Lambda_f \,\mathrm{d} \log L_f + \sum_{i \in \mathcal{I}} \lambda_i \,\mathrm{d} \log A_i.$$

This is, of course, the decomposition due to Solow (1957), Domar (1961), and Hulten (1978). Although, in this case, we have a slight generalization of these results, since the equation above implies that shocks to markups/wedges $\mathrm{d} \log \mu_i$ have no effect on aggregate output when we start from a position with no markups/wedges.

Markups/Wedges: In general, in the presence of markups/wedges, there are nonzero changes in allocative efficiency. Indeed, the envelope theorem logic fails and $\partial \log \mathcal{Y} / \partial \mathcal{X} \neq 0$. Changes in the allocation of resources $\mathrm{d}\,\mathcal{X}$ have a first-order impact on industry output given by $(\partial \log \mathcal{Y} / \partial \mathcal{X})\,\mathrm{d}\,\mathcal{X} = -\sum_{i \in \mathcal{I}} \tilde{\lambda}_i \,\mathrm{d} \log \mu_i - \sum_{f \in \mathcal{F}} \tilde{\Lambda}_f \,\mathrm{d} \log \Lambda_f$.

Changes in allocative efficiency are positive when resources are reallocated to producers that are "too small" from a social perspective. Producers are "too small" when there are relatively higher markups/wedges placed on their goods, either directly by the producer itself or indirectly via double-marginalization. Such virtuous reallocation patterns are detected by average reductions in the revenue shares of the external inputs so that $-\sum_{f \in \mathcal{F}} \tilde{\Lambda}_f \,\mathrm{d} \log \Lambda_f > 0$. Changes in markups/wedges induce direct changes in the group profit share (the share of revenues collected by the markups/wedges) $\sum \tilde{\lambda}_i \,\mathrm{d} \log \mu_i$ which are unrelated to these patterns of reallocation and must therefore be netted out.

Furthermore, the purely technical output elasticities with respect to inputs and pro-

ductivity are the cost-based Domar weights $\partial \log \mathcal{Y} / \partial \log L_f = \tilde{\Lambda}_f$ and $\partial \log \mathcal{Y} / \partial \log A_i = \tilde{\lambda}_i$ and not the revenue based Domar weights Λ_f and λ_i that would be predicted by Hulten's theorem.

We refer the reader to Baqaee and Farhi (2017) for more detailed intuitions, illustrative examples, and precise guidance for mapping the theoretical results to the data.

"Distorted" Solow Residual: These considerations have important implications for the measurement of aggregate productivity since

$$\mathrm{d} \log A_{\mathcal{I}} = \mathrm{d} \log Y - \sum_{f \in \mathcal{F}} \tilde{\Lambda}_f \, \mathrm{d} \log L_f,$$

where the right-hand side is a generalization to open systems of the "distorted" Solow residual introduced in Baqaee and Farhi (2017).

Because the traditional Solow residual $\mathrm{d} \log Y - \sum_{f \in \mathcal{F}} \Lambda_f \, \mathrm{d} \log L_f$ uses revenue-based Domar weights Λ_f instead of cost-based Domar weights $\tilde{\Lambda}_f$ for the change $\mathrm{d} \log L_f$ of external inputs $f \in \mathcal{F}$, it does not correctly net out the pure technological impact of the change in external inputs. It is therefore not an appropriate measure of the change in aggregate productivity.

Hall (1990) showed that when the economy consists of a representative producer, in order to correctly pick up the change in the exogenous productivity shifter, the Solow residual should weigh factor growth by factors' share of costs rather than share of revenues. Under these restrictive assumptions which do not allow for any misallocation of resources given factor supplies, our proposed "distorted" Solow residual coincides with Hall's residual. However, in the more general disaggregated environments that we consider where there is misallocation given factor supplies, aggregate productivity is endogenous and instead captured by the "distorted" Solow residual and not by Hall's residual.[5]

5 Second Decomposition: Petrin and Levinsohn (2012)

The downside of Theorem 1 is that it leans heavily on knowledge of the input-output network structure inside $\mathcal{I}$ (except in the case where there are no wedges). Unfortunately, this information is often unavailable or of poor quality. Theorem 2 instead decomposes

[5]Basically, in disaggregated environments, it is possible to show that the distorted Solow residual only coincides with Hall's residual when there are no intermediate goods.

output changes in a way that does not require knowledge of the input-output structure, but does require detailed knowledge of the gross quantity of production. This theorem is due to Petrin and Levinsohn (2012), although we express it differently here.

Theorem 2. *The following first-order approximation holds:*

$$\mathrm{d}\log Y = \sum_{f\in\mathcal{F}} \Lambda_f \,\mathrm{d}\log L_f + \sum_{i\in\mathcal{I}} \lambda_i \,\mathrm{d}\log A_i + \sum_{i\in\mathcal{I}} \lambda_i (1-\mu_i^{-1})(\mathrm{d}\log y_i - \mathrm{d}\log A_i).$$

Theorems 1 and 2 have different advantages and disadvantages at a conceptual level and in terms of their data requirements. Depending on the question at hand and the available data, Theorem 1 or 2 may be more convenient or relevant.

On a conceptual level, Theorem 1 has the advantage of isolating aggregate productivity and of producing the decomposition into external inputs, technical efficiency, and allocative efficiency defined in equation (3). These desirable properties are not shared by Theorem 2 except in special cases where either (1) there are no distortions, or (2) there are distortions, but no (micro) productivity changes. If there are both distortions and changes in productivities, then this decomposition cannot be interpreted in terms of changes in technical and allocative efficiency. We discuss these issues in more detail below.

The theorems have different data requirements, and so depending on data availability, one or the other may be easier to operationalize. The main drawback of Theorem 1 is that it requires information on input-output linkages. The main drawbacks of Theorem 2 are that it requires knowing the change in every quantity $\mathrm{d}\log y_i$ for every producer $i \in \mathcal{I}$ and that it requires the separation of prices from quantities.[6]

By rearranging Theorem 2, we can write

$$\mathrm{d}\log Y - \sum_{f\in\mathcal{F}} \Lambda_f \,\mathrm{d}\log L_f = \sum_{i\in\mathcal{I}} \lambda_i \,\mathrm{d}\log A_i + \sum_{i\in\mathcal{I}} \lambda_i (1-\mu_i^{-1})(\mathrm{d}\log y_i - \mathrm{d}\log A_i). \qquad (4)$$

The left-hand side of this equation is the traditional Solow residual, which, even though it is not an appropriate measure of aggregate productivity in the presence of wedges, remains an interesting and prominent object. Equation (4) can be used to aggregate the effects of a shock in inefficient environments using firm-level data from knowledge of the level of wedges μ, sales shares λ, and changes in output $\mathrm{d}\log y_i$.

[6]Even when price and quantity data are available, the "price" is often measured by unit cost and the "quantity" is measured by the ratio of sales to unit costs. However, in Theorem 2, the relevant notion of quantity $\mathrm{d}\log y_i$ needs to be quality-adjusted, which, for non-homogeneous goods and services, can be challenging.

For example, Bau and Matray (2019) use our derivations to estimate the effect of FDI liberalization in India on the Solow residual of different industries. Briefly, they use reduced-form firm-level regressions to identify $\mathrm{d}\log y_i$ and $\mathrm{d}\log A_i$ due to the policy change, and aggregate firm-level outcomes using equation (4).

A good way to understand Theorem 2 is to consider some special cases of it.

No Markups/Wedges: When there are no markups/wedges inside $\mathcal{I}$, it is immediate that

$$\mathrm{d}\log Y = \sum_{f\in\mathcal{F}} \Lambda_f \,\mathrm{d}\log L_f + \sum_{i\in\mathcal{I}} \lambda_i \,\mathrm{d}\log A_i,$$

which again coincides with Hulten (1978). In the case where there are no wedges, the first and second decompositions coincide.

No Productivity/External Input Changes: Consider the case where the physical environment is not changing, so that neither changes in external inputs nor in technologies $\mathrm{d}\log L = \mathrm{d}\log A = 0$. Then

$$\mathrm{d}\log Y = \sum_{i\in\mathcal{I}} \lambda_i \left(1-\mu_i^{-1}\right) \mathrm{d}\log y_i. \tag{5}$$

Intuitively, output changes are entirely due to changes in allocative efficiency. The envelope theorem logic fails and changes in the allocations of resources $\mathrm{d}\,\mathcal{X}$ in response to shocks to markups/wedges, or due to changes in final demand, have a non-trivial first-order impact on group output $\mathrm{d}\log Y = (\partial\log\mathcal{Y}/\partial\mathcal{X})\,\mathrm{d}\,\mathcal{X}$. These changes in allocative efficiency are expressed as the weighted sum of the output changes $\mathrm{d}\log y_i$ of the different producers, where the weight $\lambda_i(1-\mu_i^{-1})$ on producer i is increasing in the size λ_i and in the markup/wedge μ_i of the producer.

For each $i \in \mathcal{I}$, we have $\lambda_i(1-\mu_i^{-1})\,\mathrm{d}\log y_i = [p_i(1-\mu_i^{-1})/(PY)]\,\mathrm{d}\,y_i$, which measures the gap between the social marginal benefit p_i of an additional unit of good i to its users and the social marginal cost $p_i\mu_i^{-1}$ of producing this unit to its producer, expressed in the numeraire of nominal group output PY, multiplied by the change $\mathrm{d}\,y_i$ in the units of good i.[7]

[7]When there are no markups/wedges ($\mu = 1$), social marginal benefits and social marginal costs are equalized for all producers, there are no changes in allocative efficiency, and no first-order changes in group output.

The formula fully accounts for the presence of joint distortions. The output $\log y_i$ of producer i changes not only because μ_i changes but because all markups/wedges as well as final demand change. Because there is a markup/wedge μ_i on producer i, these changes in output translate into changes in aggregate output and productivity.

Harberger (1964) uses a version of equation (5), without proof, to study how the *unobservable* change in total welfare $\mathrm{d}\log W$ changes in response to policy changes in a closed system where Kaldor-Hicks transfers are made across consumers to neutralize distributional effects. Equation (5) validates Harberger (1964) and shows that it can be used to predict or unpack *observable* changes in aggregate output in open or closed-systems in general equilibrium. This justifies using formulas from welfare economics in a positive, rather than a normative, way.

Productivity/External Input Changes with Markup/Wedges: Note that when there are simultaneous changes in productivities $\mathrm{d}\log A$, external inputs $\mathrm{d}\log L$, and markups/wedges $\mathrm{d}\log \mu$, then technical efficiency changes due to $\mathrm{d}\log A$ and $\mathrm{d}\log L$ must be weighted by their cost-based Domar weights and not the revenue-based ones.

This means that we cannot interpret $\sum_{i\in\mathcal{I}} \lambda_i \,\mathrm{d}\log A_i$ as the change in technical efficiency, or $\sum_{i\in\mathcal{I}} \lambda_i(1-\mu_i^{-1})(\mathrm{d}\log y_i - \mathrm{d}\log A_i)$ as the change in allocative efficiency due to reallocation. Baqaee and Farhi (2017) contains examples of economies that, despite having wedges, are efficient because they have only one feasible allocation of resources. In such economies, there should not be any changes in allocative efficiency due to reallocation, and yet, the second summand on the right-hand side of (4) is nonzero.

Single Producer: Another interesting case is when the group consists of a singleton $\mathcal{I} = \{i\}$ which does not use its own output as an input. Then $\mathrm{d}\log Y = \mathrm{d}\log y_i$, and the aggregate Solow residual coincides with the Solow residual of producer i

$$\mathrm{d}\log Y - \sum_{f\in\mathcal{F}} \Lambda_f \,\mathrm{d}\log L_f = \mathrm{d}\log y_i - \sum_{f\in\mathcal{F}} \Omega_{if} \,\mathrm{d}\log l_{if}.$$

Theorem 2 becomes

$$\mathrm{d}\log y_i - \sum_{f\in\mathcal{F}} \Omega_{if} \,\mathrm{d}\log l_{if} = \mathrm{d}\log A_i + (1-\mu_i^{-1})(\mathrm{d}\log y_i - \mathrm{d}\log A_i).$$

This equation replicates the formula derived by Hall (2018) for estimating markups. Hall's

method for estimating the markup supposes that we are in possession of an instrument Z which is correlated with output $\mathrm{d}\log y_i$ (relevance), but not with productivity $\mathrm{d}\log A_i$ (exclusion). For example, supply or demand shocks external to producer i might satisfy these conditions. Then, if we regress the Solow residual $\mathrm{d}\log y_i - \sum_{f\in\mathcal{F}} \Omega_{if}\,\mathrm{d}\log l_{if}$ on gross output $\mathrm{d}\log y_i$ instrumented by Z, the coefficient will recover $(1-\mu_i^{-1})$.

Theorem 2 shows how Hall's observation can be aggregated across a set of producers (say, an industry). Theorem 2 also shows that once aggregated over several producers, then the identification strategy outlined above can fail. Intuitively, once there are wedges inside the industry, changes in the allocation of resources $\mathrm{d}\,\mathcal{X}$ inside the industry caused by supply and demand shocks outside the industry can lead to endogenous changes in the industry's aggregate productivity $\mathrm{d}\log A_{\mathcal{I}}$. This violates the exclusion restriction at the industry level, creating correlation between the instrument and the industry's productivity, even if the instrument is uncorrelated with every individual producer's productivity $\mathrm{d}\log A_i$.

6 Conclusion

In this paper, we propose a definition of productivity for a group of producers in imperfect markets that we call the "distorted" Solow residual. We discuss two decompositions of aggregate output into its microeconomic sources. The first decomposition, generalizing Baqaee and Farhi (2017), can be used to measure and separate the "distorted" Solow residual into contributions due to technical efficiency and allocative efficiency. The second decomposition, due to Petrin and Levinsohn (2012), can be used to aggregate up firm-level outcomes up to group-level aggregate output without knowledge of the input-output table. The two decompositions, by having different information requirements, complement one another, and give two different theoretical perspectives on the determinants of aggregate output in imperfectly competitive general equilibrium settings.

References

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

Bau, N. and A. Matray (2019). Misallocation and capital market integration: Evidence from India.

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

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

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

Hall, R. E. (2018). New evidence on the markup of prices over marginal costs and the role of mega-firms in the us economy. Technical report, National Bureau of Economic Research.

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

Hicks, J. R. (1965). Value and capital.

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.

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

Meade, J. (1962). *Trade and Welfare*. The Theory of International Economic Policy. Oxford University Press.

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

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

7 Appendix

Proof of Theorem 1. The following equations are to be interpreted as exact differentials. They can then be used to derive first-order approximations for small discrete changes. We have

$$d\log Y = d\log PY - d\log P,$$

where PYdenotes nominal industry output and $d\log P$ is defined in changes by

$$d\log P = \sum_{i\in\mathcal{I}} b_i d\log p_i.$$

We use

$$d\log p_i = d\log \mu_i - d\log A_i + \sum_{j\in\mathcal{I}\cup\mathcal{F}} \tilde{\Omega}_{ij} d\log p_j,$$

to get

$$d\log p_i = \sum_{j\in\mathcal{I}} \tilde{\Psi}_{ij}(d\log \mu_j - d\log A_j) + \sum_{f\in\mathcal{F}} \tilde{\Psi}_{if} d\log p_f.$$

Using

$$d\log p_f = d\log \Lambda_f + d\log PY - d\log L_f,$$

we can rewrite this as

$$d\log p_i = \sum_{j\in\mathcal{I}} \tilde{\Psi}_{ij}(d\log \mu_j - d\log A_j) + \sum_{f\in\mathcal{F}} \tilde{\Psi}_{if}(d\log \Lambda_f + d\log PY - d\log L_f).$$

Since $\sum_{f\in\mathcal{F}} \tilde{\Psi}_{if} = 1$, this can be rewritten as

$$d\log p_i = d\log PY + \sum_{j\in\mathcal{I}} \tilde{\Psi}_{ij}(d\log \mu_j - d\log A_j) + \sum_{f\in\mathcal{F}} \tilde{\Psi}_{if}(d\log \Lambda_f - d\log L_f).$$

We then replace this expression in the expression for $d\log P = \sum_{i\in\mathcal{I}} b_i d\log p_i$ to get

$$d\log P = d\log PY + \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{I}} b_i\tilde{\Psi}_{ij}(d\log \mu_j - d\log A_j) + \sum_{i\in\mathcal{I}}\sum_{f\in\mathcal{F}} b_i\tilde{\Psi}_{if}(d\log \Lambda_f - d\log L_f).$$

We then use $\tilde{\lambda}_j = \sum_{i\in\mathcal{I}} b_i\tilde{\Psi}_{ij}$ and $\tilde{\Lambda}_f = \sum_{i\in\mathcal{I}} b_i\tilde{\Psi}_{if}$ to rewrite this as

$$d\log P = d\log PY + \sum_{j\in\mathcal{I}} \tilde{\lambda}_j(d\log \mu_j - d\log A_j) + \sum_{f\in\mathcal{F}} \tilde{\Lambda}_f(d\log \Lambda_f - d\log L_f).$$

We plug this back in the expression for $d\log Y = d\log PY - d\log P$ to get

$$d\log Y - \sum_{f\in\mathcal{F}} \tilde{\Lambda}_f d\log L_f = \sum_{j\in\mathcal{I}} \tilde{\lambda}_j (d\log A_j - d\log \mu_j) - \sum_{f\in\mathcal{F}} \tilde{\Lambda}_f d\log \Lambda_f.$$

■

Proof of Theorem 2. We prove Theorem 2. The results follow. The following equations are to be interpreted as exact differentials. They can then be used to derive first-order approximations for small discrete changes. We have:

$$d\log c_i = \frac{y_i}{c_i} d\log y_i - \sum_{j\in\mathcal{I}} \frac{y_{ji}}{c_i} d\log y_{ji},$$

$$\mu_j^{-1}(d\log y_j - d\log A_j - \sum_{f\in\mathcal{F}} \frac{p_f y_{jf}}{p_j y_j} \mu_j d\log y_{jf}) = \sum_{i\in\mathcal{I}} \Omega_{ji} d\log y_{ji} = \sum_{i\in\mathcal{I}} \frac{p_i y_{ji}}{p_j y_j} d\log y_{ji}.$$

The first equation is simply an accounting identity. The second equation is a direct implication of cost minimization and constant returns to scale. We now use these two equations to manipulate the expression for output growth and obtain the result. We then have

$$\begin{aligned}
d\log Y &= \sum_{i\in\mathcal{I}} \frac{p_i c_i}{PY} d\log c_i, \\
&= \sum_{i\in\mathcal{I}} \frac{p_i y_i}{PY} d\log y_i - \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} \frac{p_i y_{ji}}{PY} d\log y_{ji}, \\
&= \sum_{i\in\mathcal{I}} \frac{p_i y_i}{PY} d\log y_i - \sum_{j\in\mathcal{I}} \frac{p_j y_j}{PY} \mu_j^{-1}(d\log y_j - d\log A_j - \sum_{f\in\mathcal{F}} \frac{p_f y_{jf}}{p_j y_j} \mu_j d\log y_{jf}), \\
&= \sum_{f\in\mathcal{F}} \Lambda_f d\log L_f + \sum_{i\in\mathcal{I}} \lambda_i \mu_i^{-1} d\log A_i + \sum_{i\in\mathcal{I}} \lambda_i (1-\mu_i^{-1}) d\log y_i.
\end{aligned}$$

■