Mark-Up Distortions and Endogenous Misallocation

Michael Peters*

February 11, 2010

Abstract

Income differences across countries are to a large degree driven by differences in aggregate TFP. Recently it has been argued that part of these differences are due to misallocation of factor across firms. In this paper we propose a structural model of this degree of misallocation. Our model is one where monopolistic competition across firms generates non-constant mark-ups, which reduce aggregate TFP relative to the competitive benchmark. Equilibrium mark-ups depend only on the distribution of firm-level productivity, which evolves endogenously according to a Schumpeterian process of creative destruction. This provides the link between the economy's innovation environment and the equilibrium degree of misallocation. In particular, the economy's entry intensity is a sufficient statistic for the invariant distribution of mark-ups and their TFP consequences. If entry is more intense, aggregate TFP is higher, as product market competition reduces the distorting effect of mark-ups. This provides a new channel how impediments to firm entry, like entry costs, reduce allocative efficiency, aggregate TFP and income per capita.

1 Introduction

There is a general consensus that productivity differences across countries are an important determinant of cross-country income differences and living standards (Banerjee and Duflo, 2005; Caselli, 2005; Klenow and Rodriguez-Clare, 1997). However, assuming that productivity differences are simply exogenous country characteristics is deeply unsatisfactory both from a theoretical as well as from a policy perspective. Hence, there is a broad literature aiming to explain these productivity differences. To name a few, Hall and Jones (1999) and Acemoglu, Johnson, and Robinson (2001) argue in favor for the importance of social capital or institutions, Lucas (1988, 1990) stresses the importance of human capital

*eMail: mipeters@mit.edu. I thank Daron Acemoglu, Abhijit Banerjee, Marios Angeletos, Chang-Tai Hsieh, Ufuk Akcigit, Thomas Chaney and Joaquin Blaum for many helpful comments.

externalities, Parente and Prescott (1994) focus on barriers of technology adoption and Ganica and Zilibotti (2009) identify an important role for the direction of technological progress.

Whereas these contributions take an aggregate production function as the primitive of the economy, recently various authors focused their attention explicitly on the behavior of individual firms and their interaction. By carefully modelling the microstructure of the economy, these papers aim to shed light on the micro mechanisms causing the productivity differences observed at the aggregate. Acemoglu, Antras, and Helpman (2007) for example analyze a model of technology adoption under incomplete contracting. They show that contractual incompleteness among suppliers of intermediary inputs reduces the incentives to invest into technology and via this channel reduces aggregate productivity. Lagos (2006) considers a labor market with search frictions and shows that aggregate TFP depends on the primitives of the underlying search technology. Buera, Kaboski, and Shin (2009) and Moll (2009) consider dynamic models of entrepreneurship and sectoral choice with borrowing constraints. They find that capital market imperfections reduce total factor productivity by both distorting the allocation of talent across occupations and the allocation of capital across firms.[1]

Finally, Hsieh and Klenow (2009) and Restuccia and Rogerson (2008) argue that distortions on the firm-level can have sizable negative effects on aggregate TFP as they cause a misallocation of resources in that marginal products are not equalized. In contrast to the other papers mentioned above, this "distortion literature" is agnostic about the nature of those distortions but models and empirically identifies them using the approach of exogenous wedges introduced by Chari, Kehoe, and McGrattan (2007).

This paper aims to bridge the latter two approaches by proposing a structural model of firm-level distortions, where misallocation across firms arises endogenously. The source of misallocation is monopolistic price setting which in our economy generates non-constant mark-ups. This cross-sectional heterogeneity in mark-ups reduces aggregate TFP, because the economy's resources are not allocated toward the most productive units but rather to the ones who set a low mark-up. This reasoning is of course similar to the one put forward in the distortion literature - in these papers not the most productive units but the ones facing low firm-level taxes will employ more resources. The important difference however is, that our mark-ups are not exogenous but are determined endogenously by firms' equilibrium pricing decision.

Despite this non-trivial pricing rule, our model is still remarkably tractable. In fact from an aggregate perspective, our economy looks exactly like the single-sector neoclassical growth model with Cobb-Douglas technology and an aggregate TFP term. We show that aggregate TFP is the product of two terms, where the first one depends only on the distribution of firm-level productivity and the second one is entirely determined by the distribution of mark-ups. It

[1] Midrigan and Xu (2010) however argue that financial constraints have only small effects on aggregate TFP as agents will accumulate capital and hence the constraint not be binding. See also Banerjee and Moll (forthcoming).

is this second term (and only this second term) that is not present in the efficient allocation and it is hence a sufficient statistic for the static TFP losses of non-constant mark-ups. We also show that this pricing-induced misallocation affects equilibrium factor prices above and beyond this reduction of TFP and that different moments of the underlying mark-up distribution have different impacts on aggregate TFP and factor prices. Whereas TFP is only affected by the dispersion of mark-ups, but insensitive with respect to first-moment shifts, factor prices are only dependent on the distribution's first moment but invariant to higher order properties.

So how does the equilibrium distribution of mark-ups look like? In our model, a firm's mark-up depends only on its productivity relative to its competitors. Hence, the source of market power is productivity differences across firms, where the most productive firms have a higher degree of monopolistic leverage as they face less competition from their competitors. This implies that any theory of the evolution of firm-level productivity also has implications for the endogenous determination of mark-up distortions and their implications on aggregate TFP and factor prices. In this paper we consider a particular theory, namely an endogenous growth model in the Schumpeterian tradition (Aghion and Howitt, 1992; Grossman and Helpman, 1991; Klette and Kortum, 2004). This allows to relate the distribution of mark-ups (and hence the misallocation component of TFP and factor prices) to the primitives of the innovative environment. In our model innovations will either come from currently producing firms or from new entrants. We show that along the balanced growth path (BGP) of this economy, we can characterize the endogenous distribution of mark-ups in closed form and we show that this distribution depends only on a single statistic. This statistic is the economy's entry intensity, which is given by the ratio of the entrants' versus the incumbents' innovation rates. If an economy's productivity growth is largely accounted for by new entrants, the entry intensity is high and mark-up distortions are low because entering firms keep the productivity distribution compressed and hence monopoly power limited. In such an environment, aggregate TFP and factor prices are high. If on the other hand most of productivity growth is generated by current producers, the efficiency costs of mark-up distortions are high as most of the firms will have a large degree of monopoly power as their competitors are relatively unproductive. This causes TFP and factor prices to be low.

The existence of this sufficient statistic is convenient because it provides a transparent link between the economy's primitives and aggregate TFP. In order to use this model for a cross-country productivity comparison, it therefore has to be true that rich and poor countries have different entry intensities. The aspect we are mostly focusing on are entry costs. Djankov, Porta, Lopez-De-Silvaes, and Shleifer (2002) and Porta and Shleifer (2009) for example claim that entry of new firms is particularly inefficient in underdeveloped economies and Barseghyan (2008) and Barseghyan and DiCecio (2009) provide empirical evidence for the negative relation between aggregate TFP and entry costs. When viewed through the lens of our model, higher entry costs will lower TFP and factor prices by reducing the equilibrium entry intensity and increasing equilibrium mark-ups. Besides these aggregate predictions, the model also has implications for the equilibrium firm-size distribution. We show that firms are bigger in entry-intensive, developed economies and that the size dispersion is lower. This is consistent with the findings in Alfaro, Charlton, and Kanczuk (2008).

The structure of the paper is as follows. In the next section we describe the model, solve for the static equilibrium allocations and derive the two sufficient statistics for the impact of mark-up distortions on factor prices and aggregate TFP. We also discuss the relationship of our model with the models featuring exogenous firm-specific taxes as in Hsieh and Klenow (2009) or Restuccia and Rogerson (2008). We then solve for the dynamic allocations and characterize the BGP of the economy. We show that in the dynamic equilibrium the endogenous entry intensity parametrizes the invariant distribution of distortions and hence the degree of misallocation affecting factor prices and TFP. We also show how this entry intensity is related to the underlying primitives of the economy, in particular the entry costs. We then offer some concluding remarks and an appendix gathers the mathematical details.

2 The Model

Our model will be a model of monopolistic competition, where the distribution of firm-level productivity evolves endogenously according to a process of Schumpeterian creative destruction. Both from an analytical as well as from a conceptual point of view it is convenient to first consider the static allocation for a given distribution of productivity. This will also allow us to concisely characterize the effect of monopolistic mark-ups on factor prices and aggregate TFP and hence relate our approach to the recent literature modelling firm-level distortions as firm-specific taxes. Then we will turn to the dynamic equilibrium where the evolution of productivity and mark-ups is endogenously determined.

We consider a continuous-time economy which produces a unique final good using a continuum of intermediary services according to

$$Y(t) = exp\left(\int_0^1 ln\left(\sum_{j\varepsilon S_\nu} y_j\left(\nu,t\right)\right)d\nu\right), \tag{1}$$

where $y_j(\nu,t)$ is the quantity of the intermediary input ν bought from producer j who is active in sector ν. Hence S_ν denotes the number of firms active in sector ν. It is clear from (1) that varieties of different intermediaries ν and ν' are imperfect substitutes, whereas there is perfect substitutability between different brands within a variety.[2] The production of intermediaries is conducted by heterogeneous firms and requires capital and labor. The only source of heterogeneity across firms is their productivity. In particular, a firm in sector ν

[2]Nothing substantial hinges on our assumption that the aggregate production is Cobb-Douglas in different varieties. We could consider a general CES production function but it would make the analysis a little more tedious. See also the discussion in footnote 3.

with current productivity q produces $\nu-$output according to a standard Cobb-Douglas production function

$$f(k,l;q) = qk^{\alpha}l^{1-\alpha}. \tag{2}$$

Labor for intermediary production is provided inelastically by a measure L of production workers. Besides these production workers, there is a measure L_S of scientists who can improve firm productivity and will be employed in the innovation sector. We will provide the details about the innovation technology and market structure below, so for now let it suffice that scientists get paid a wage $w_I(t)$ for their services. We assume that the preferences of the household consisting of workers and scientists can be represented by

$$U = \int_0^{\infty} exp(-\rho t)\, ln(c(t))\, dt,$$

and that the budget constraint of the household is given by

$$c(t) + \dot{A}(t) + \dot{K}(t) = (1 + r_K(t))\, K(t) + (1 + r_A(t))\, A(t) + Lw(t) + L_S w_I(t). \tag{3}$$

Note that this economy provides two instruments to save, namely capital $K(t)$, which is rented out to intermediary producers, and equity $A(t)$, where intermediary firms pay out their profits as dividends. The return on these savings instruments is $r_K(t)$ and $r_A(t)$, which in equilibrium will of course be equalized, and $Lw(t) + L_S w_I(t)$ is the total labor income generated by production workers and scientists. All factor markets and the market for shares is assumed to be competitive. We will now characterize the static equilibrium allocation for a given level of capital $K(t)$ and a given distribution of productivity $\left[\{q_j(\nu,t)\}_{j\in S_\nu}\right]_\nu$.

2.1 Static Allocations

Monopolistic price setting We assume that the market for intermediary goods is monopolistically competitive, i.e. in sector ν there are $S_\nu(t)$ competing firms (and $S_\nu(t)$ will evolve endogenously as outlined below). Given that production takes place with a constant returns to scale technology and different brands of variety ν are perceived as perfect substitutes, in equilibrium only the most productive intermediary firm will be active. Hence, aggregate output is simply given by $exp\int_0^1 \ln(y(\nu,t))\, d\nu$, where now $y(\nu,t)$ is the number of intermediaries produced by the most productive producer in sector ν. However, the presence of potentially competing producers (even though they are less efficient) imposes a constraint on the quality leader's price setting. In particular, final good firms buy from firm j in sector ν as long as $p_j(\nu,t) \leq p_{j'}(\nu,t)$ for all $j' \in S_\nu, j' \neq j$. Given intermediary prices $p(\nu,t)$, the demand for intermediaries is given by

$$y(\nu,t) = \frac{Y(t)}{p(\nu,t)}. \tag{4}$$

Factor markets are assumed to be competitive, i.e. all intermediary firms take wages $w(t)$ and the rental rate of capital $R(t)$ as given. From (2) it then follows immediately that the marginal cost of production are given by

$$MC(q,t) = \left(\frac{w(t)}{1-\alpha}\right)^{1-\alpha} \left(\frac{R(t)}{\alpha}\right)^{\alpha} \frac{1}{q} \equiv \frac{\psi(R(t), w(t))}{q}. \tag{5}$$

As the demand function (4) has unitary elasticity, the leading firm has to resort to limit pricing where it has to price its products equal to the marginal costs of the second most productive firm (the follower). Hence,

$$p(\nu,t) = \frac{\psi(R(t), w(t))}{q_F(\nu,t)}, \tag{6}$$

where $q_F(\nu,t)$ is the follower's productivity.[3] The equilibrium mark-up is therefore given by

$$\xi(\nu,t) = \frac{p(\nu,t)}{MC(\nu,t)} = \frac{q(\nu,t)}{q_F(\nu,t)}. \tag{7}$$

Note in particular that the mark-up is not constant but depends on the productivity advantage of the producing firm relative to its closest competitor - a higher quality advantage shields the current producer from competition and allows him to post a higher mark-up. Equilibrium sales are therefore (see (4) and (6))

$$y(\nu,t) = \frac{Y(t)\, q_F(\nu,t)}{\psi(R(t), w(t))} = \frac{1}{\xi(\nu,t)} \frac{Y(t)\, q(\nu,t)}{\left(\frac{w(t)}{1-\alpha}\right)^{1-\alpha} \left(\frac{R(t)}{\alpha}\right)^{\alpha}}, \tag{8}$$

so that the cost-minimizing factor demands are

$$k(\nu,t) = \frac{y(\nu,t)}{q(\nu,t)} \left(\frac{\alpha}{1-\alpha}\frac{w(t)}{R(t)}\right)^{1-\alpha} = \frac{1}{\xi(\nu,t)} \frac{\alpha Y(t)}{R(t)} \tag{9}$$

$$l(\nu,t) = \frac{y(\nu,t)}{q(\nu,t)} \left(\frac{1-\alpha}{\alpha}\frac{R(t)}{w(t)}\right)^{\alpha} = \frac{1}{\xi(\nu,t)} \frac{(1-\alpha) Y(t)}{w(t)}. \tag{10}$$

The quality leader's profits are therefore given by

[3] It is at this point where our assumption of the aggregate production function being Cobb-Douglas simplifies the exposition because the producing firm will always be forced to use limit pricing. If the demand elasticity was positive, the firm might want to set the unconstrained monopoly price in case its productivity advantage over its closest competitor is big enough. To see this suppose that $Y = \left(\int_0^1 y(\nu)^{\frac{\sigma-1}{\sigma}} d\nu\right)^{\frac{\sigma}{\sigma-1}}$. The unconstrained monopoly price is given by $p^M = \frac{\sigma}{\sigma-1} MC = \frac{\sigma}{\sigma-1}\frac{\psi}{q}$, so that the optimal price is given by $p = min\left(\frac{\sigma}{\sigma-1}\frac{\psi}{q}, \frac{\psi}{q_F}\right) = \frac{\psi}{q_F} min\left(\frac{\sigma}{\sigma-1}\frac{q_F}{q}, 1\right)$. Hence, for $q >> q_F$, the firm will not have to use limit pricing. In the limit where $\sigma \to 1$, we get $min\left(\frac{\sigma}{\sigma-1}\frac{q_F}{q}, 1\right) = 1$. See also Bernard, Eaton, Jensen, and Kortum (2003).

$$\pi(\nu,t|q) = (p(\nu,t) - MC(\nu,t))\, y(\nu,t) = \left(1 - \frac{q_F(\nu,t)}{q(\nu,t)}\right) Y(t) \quad (11)$$

$$= \frac{\xi(\nu,t)-1}{\xi(\nu,t)} Y(t) \quad (12)$$

The expressions above show that the decision-relevant sector-specific variables are the leaders' and follower's productivity $q(\nu,t)$ and $q_F(\nu,t)$ and that the cross-sectional variation in profits and factor demands can be entirely traced back to firm-specific mark-ups. Furthermore it is precisely these varying mark-ups which will induce an inefficient allocation of resources across plants and hence have detrimental effects on aggregate TFP. Before we characterize these expressions from the general equilibrium of this economy, we briefly discuss the close relation of this model to the distortions literature, where firm-level distortions are exogenously imposed.

Monopolistic Mark-ups and the Distortion Debate To see this relation formally, consider the approach developed in Foster, Haltiwanger, and Syverson (2008) and for example extensively employed in Hsieh and Klenow (2009). They argue that if there was no misallocation of resources, revenue total factor productivity, which is defined as $TFPR = \frac{py}{K^{\alpha}L^{1-\alpha}}$, should be equalized across firms. In fact, Hsieh and Klenow (2009, p. 1410) show that in their framework $TFPR$ is proportional to $\frac{(1+\tau_K(\nu,t))^{\alpha}}{1-\tau_Y(\nu,t)}$, where $\tau_K(\nu,t)$ and $\tau_Y(\nu,t)$ are the exogenous firm-specific taxes on capital and output. In our economy, $TFPR$ is given by

$$TFPR(\nu,t) = \frac{p(\nu,t)\, y(\nu,t)}{k(\nu,t)^{\alpha}\, l(\nu,t)^{1-\alpha}} = \frac{q(\nu,t)/q_F(\nu,t)}{\left(\frac{\alpha}{R(t)}\right)^{\alpha}\left(\frac{(1-\alpha)}{w(t)}\right)^{1-\alpha}} \propto \xi(\nu,t). \quad (13)$$

Hence, $TFPR$ is also not equalized across firms but the monopolistic mark-up takes the role of the firm-level distortion terms.[4] So if the data was generated by the model above and a researcher would measure factor inputs and firm-level revenue to identify distorted firms, this researcher would conclude that those firms that enjoy the most monopoly power face the most severe distortions, because their marginal productivity is too high (or their production level is too low given their productivity). To see this, consider for example two firms in sector ν' and ν'' with the same level of productivity q. Efficiency requires that that these firms should produce the same quantity y. In the equilibrium however, their relative output levels are given by (see (4))

$$\frac{y(\nu'',t)}{y(\nu',t)} = \frac{\alpha Y q_F(\nu'',t)/\psi(R,w)}{\alpha Y q_F(\nu',t)/\psi(R,w)} = \frac{\xi(\nu',t)}{\xi(\nu'',t)}, \quad (14)$$

[4] In fact, (13) shows that there would be no misallocation across firms if mark-ups were constant as generated by a standard CES demand system.

i.e. the relative firm-size is inversely related to the difference in mark-ups. If we were to stop here, the only progress made was a relabeling of what the literature calls "distortions". However, by interpreting the distribution of mark-ups $[\xi(\nu,t)]$ through the lens of a structural model, these are determined endogenously and we can therefore trace them back to the underlying primitives of the economy. Before we turn to this however, we characterize the general equilibrium of this economy and characterize the effect of mark-ups on factor prices and aggregate TFP.

The General Equilibrium (7) shows that equilibrium mark-ups depend only on the relative productivity between firms. To reduce the state-variable even further, we will model firm productivity by a standard quality-ladder specification, which is often used in Schumpeterian growth models (see Aghion and Howitt (1992); Grossman and Helpman (1993)). In particular we assume that if a firm has had $n(\nu,t)$ innovations in the interval $[0,t]$, its productivity is given by $q(\nu,t)=\lambda^{n(\nu,t)}$, where $\lambda>1$. This productivity specification implies that we can express mark-ups as

$$\xi(\nu,t)=\frac{q(\nu,t)}{q_F(\nu,t)}=\frac{\lambda^{n_L(\nu,t)}}{\lambda^{n_F(\nu,t)}}=\lambda^{\Delta(\nu,t)}, \tag{15}$$

where $\Delta(\nu,t)\equiv n_L(\nu,t)-n_F(\nu,t)\geq 0$ is the quality gap between the leader and the follower. Note that with (15) we can express equilibrium profits (11) and factor demands (9)-(10) as a function of the gap $\Delta(\nu,t)$ only. In particular, profits are increasing in $\Delta(\nu,t)$ as being technological advanced reduces competitive pressure and factor demands are decreasing in $\Delta(\nu,t)$ as (for given $q_F(\nu,t)$) the leading firm does not increase its quantity sold when its productivity improves (see (8)) but rather produces these units with less factor use.

Let us now turn to the general equilibrium. Equilibrium on the factor markets requires that $K(t)=\int_0^1 k(\nu,t)\,d\nu$ and $L(t)=\int_0^1 l(\nu,t)\,d\nu$. Using factor demands ((9) and (10)) and (15), equilibrium factor prices are given by

$$\begin{aligned} R(t) &= \frac{\alpha Y(t)}{K(t)}\left[\int_0^1 \lambda^{-\Delta(\nu,t)}d\nu\right]=\frac{\alpha Y(t)}{K(t)}\Lambda(t), && (16)\\ w(t) &= \frac{(1-\alpha)Y(t)}{L}\left[\int_0^1 \lambda^{-\Delta(\nu,t)}d\nu\right]=\frac{(1-\alpha)Y(t)}{L}\Lambda(t), && (17)\end{aligned}$$

where we defined the *factor price distortion index*

$$\Lambda(t)=\left[\int_0^1 \lambda^{-\Delta(\nu)}d\nu\right]. \tag{18}$$

From (1) and (4) we get that $Y(t)=exp\int_0^1 ln\left(\frac{Y(t)q_F(\nu,t)}{\psi(R,w)}\right)d\nu$. Substituting equilibrium interest rates (16) and wages (17) and using (15) shows that aggregate output can be written as

$$Y(t) = Q(t)\,M(t)\,K(t)^{\alpha}L^{1-\alpha}, \tag{19}$$

where

$$
\begin{aligned}
Q(t) &= exp\left(\int_0^1 ln\left(q\left(\nu,t\right)\right)d\nu\right) \\
M(t) &= \frac{\lambda^{-\int_0^1 \Delta(\nu,t)d\nu}}{\int_0^1 \lambda^{-\Delta(\nu,t)}d\nu}.
\end{aligned}
\tag{20}
$$

Hence this economy looks exactly like the canonical neoclassical growth model

$$Y(t) = TFP(t)\,K(t)^{\alpha}\,L^{1-\alpha},$$

where $TFP(t)$ is determined both by physical productivity $Q(t)$ and the monopolistic mark-ups summarized in the *TFP distortion index* $M(t)$. Furthermore it can easily be verified, that the Pareto optimal allocation in the economy implies that aggregate output is given by $Y^{Eff}(t) = Q(t)\,K(t)^{\alpha}\,L^{1-\alpha}$. In fact, this would also be the expression for aggregate output if mark-ups were constant like in the typical case of a CES demand system. For a given level of capital $K(t)$ and a given distribution of leading qualities $[q(\nu,t)]_{\nu}$, $M(t)$ summarizes therefore precisely the static TFP loss induced by non-constant mark-ups. In particular, this shows that $M(t)$ and $\Lambda(t)$ are sufficient statistics for the TFP and factor price consequences of heterogeneous mark-ups. Hence we also refer to $M(t)$ and $\Lambda(t)$ as measuring the degree of misallocation. Proposition 1 contains some properties with regard to the underlying distribution of productivity gaps.

Proposition 1. *Let F and G be two cross-sectional distributions of productivity gaps $\Delta(\nu,t)$ and let $M_Z(t)$ and $\Lambda_Z(t)$ be defined as in* (20) *and* (18) *if $\Delta(\nu,t)$ is distributed according to Z. Then:*

1. *Dispersion in mark-ups entails an efficiency loss, i.e.*

$$M_Z(t) \leq 1 \text{ for all } Z, \tag{21}$$

   and (21) *holds with equality if and only if $\Delta(\nu,t) = \Delta$ for all ν.*

2. *If G is a mean preserving spread of F, then $\Lambda_G(t) > \Lambda_F(t)$ and $M_G(t) < M_F(t)$.*

3. *If $G(x+\kappa) = F(x)$ for all x and $\kappa > 0$, i.e. G represents an "outward shift" of F, then $M_G(t) = M_F(t)$ and $\Lambda_G(t) < \Lambda_F(t)$.*

Proof. Let $[\Delta(\nu,t)_{\nu}]$ be given. For $M(t)$ to be maximal, we need that

$$\frac{\partial log\left(M(t)\right)}{\partial \Delta(\hat{\nu},t)} = ln(\lambda)\left(-1 + \frac{\lambda^{-\Delta(\hat{\nu},t)}}{\int_0^1 \lambda^{-\Delta(\nu,t)}d\nu}\right) = 0 \text{ for all } \hat{\nu}.$$

This however is only the case if $\Delta(\nu,t) = \Delta$ for all ν so that $M(t) \leq 1$. That $\Lambda_G(t) > \Lambda_F(t)$ if G is a mean-preserving spread of F follows from the fact that

λ^{-x} is convex in x and Jensen's Inequality. That $M_G(t) < M_F(t)$ follows from the definition of $M(t)$. Part 3 is also immediate from (18) and (20), because if $G(x+\kappa) = F(x)$, then $\int \Delta dF(\Delta) = \int \Delta dG(\Delta+\kappa) = \int \Delta dG(\Delta) - \kappa$ and $\int \lambda^{-\Delta} dF(\Delta) = \int \lambda^{-\Delta} dG(\Delta+\kappa) = \lambda^{\kappa} \int \lambda^{-\Delta} dG(\Delta)$. $\square$

Proposition 1 has a direct corollary for the economic quantities.

Corollary 2. *Consider the equilibrium factor prices* (16) *and* (17) *and aggregate $TFP(t) = Q(t) M(t)$ of this economy. Let G and F be two distributions of productivity gaps. Then:*

1. If G is a mean preserving spread of F, then $R_G(t) = R_F(t)$, $W_F(t) = W_G(t)$ and $TFP_G(t) < TFP_F(t)$.
2. If $G(x+\kappa) = F(x)$ for all x and $\kappa > 0$, i.e. G represents an "outward shift" of F, then $R_G(t) < R_F(t)$, $W_G(t) < W_F(t)$ and $TFP_G(t) = TFP_F(t)$.

Corollary 2 shows that the location and the dispersion of the productivity gap distribution has very different impacts on the economy. In particular, factor prices are entirely insensitive with respect to a higher dispersion of the underlying productivity gap distribution determining mark-ups - they depend only on the mean. A shift in the level of productivity gaps will therefore reduce factor prices and increase equilibrium profits. For the case of aggregate TFP, it is exactly the opposite: whereas a higher dispersion of mark-ups will show up as a lower TFP for the aggregate economy, TFP is not affected by level shifts.[5] The reason is the following: Monopolistic distortions affect TFP as intermediary producers restrict their quantity sold. For given factor prices, a shift δ in the distribution of productivity gaps increases mark-ups by a constant proportion and hence will lower aggregate TFP by a factor $\lambda^{-\delta}$.[6] However, factor prices are not constant. In fact, both interest rates and wages will also be lower by a factor $\lambda^{-\delta}$ (see (16) and (17)). As intermediary production takes place under constant returns, this will lower the marginal costs of production and hence boost the quantity sold by exactly the same amount $\lambda^{-\delta}$. Hence, the two effects cancel exactly out and the shift in the economy's mark-ups is solely distributive: factor prices are lower and corporate profits are higher. Note that this precisely that happens in the usual CES case there mark-ups are constant. TFP will be identical to its efficient competitive counterpart but monopolistic power drives down equilibrium factor prices.

Note also that that varying mark-ups reduce equilibrium factor-prices for two

[5] A similar results is also found in Midrigan and Xu (2010).

[6] To see this, note that (16) and (17) imply that $Y(t) = \left(\frac{R(t)}{\alpha}\right)^{\alpha} \left(\frac{w(t)}{1-\alpha}\right)^{1-\alpha} \Lambda(t)^{-1} K(t)^{\alpha} L^{1-\alpha}$. Hence, for given factor prices, TFP is simply equal to $\Lambda(t)^{-1}$. And this term decreases by a factor of $\lambda^{-\delta}$ if the distribution of distortions is shifted. Similarly, the change in mark-ups is given by $dln(\xi) = ln(\lambda) d\Delta = ln(\lambda) \delta$.

reasons (see (16) and (17)).[7] Not only is aggregate TFP (and hence the marginal product) lower compared to the competitive economy, but even conditional on aggregate TFP, mark-up distortions reduce the equilibrium factor prices below their marginal products as

$$\frac{R(t)}{MPK(t)} = \frac{w(t)}{MPL(t)} = \int_0^1 \lambda^{-\Delta(\nu,t)} d\nu = \Lambda(t) < 1. \tag{22}$$

Note in particular that equilibrium interest rates will be depressed relative to the interest rate as imputed from the usual growth accounting calibration. More precisely: A researcher using the data generated by this economy but not taking account of the mark-up distortions would equate the interest rate with the marginal product of capital (net of depreciation), i.e. from (19) she takes the interest rate to be $r(t) = \frac{\partial F(K,L)}{\partial K} - \delta = \frac{\alpha Y(t)}{K(t)} - \delta$. However, the actual equilibrium interest rate is multiplied by the term $\Lambda(t)$ and hence lower. This is a desirable feature compared to the standard Cobb-Douglas specification, because the mark-up distortions break the link between the average and marginal product of capital, which usually causes implied interest rates to be counterfactually high (see for example Banerjee and Duflo (2005); Lucas (1990)).[8]

2.2 Dynamics: The Endogenous Evolution of Mark-Ups

Proposition 1 and Corollary 2 showed that the shape of the productivity gap distribution determines aggregate TFP and equilibrium factor prices via its effect on mark-ups. In contrast to the previous literature, misallocation is not exogenously imposed but results from competition between intermediary producers of different quality. Hence, *any* theory about the determination of firm-level productivity will also characterize the degree of misallocation reflected in the terms $M(t)$ and $\Lambda(t)$. This allow us to link the efficiency costs of misallocation to the economy's primitives.

In this paper will assume that the evolution of productivity is governed by a Schumpeterian model of creative destruction. The endogenous distribution of productivity will therefore depend on various characteristics of the innovation environment, in particular the entry costs, the efficiency of the innovation technology or the amount of innovative resources the economy posses (the number of "scientists"). We will show however, that the entire stationary distribution of productivity gaps is parametrized by a single endogenous variable capturing

[7] In fact, these distorted factor prices have a third, dynamic effect in that capital accumulation is not efficient. In particular, as the equilibrium return to capital is lower than its efficient counterpart, capital accumulation will be inefficiently low. We will discuss these dynamic aspects below.

[8] Although we here mainly allude to the cross-country comparison, very similar problems occur in the comparison across time, when we try to match aggregate data with the neoclassical growth model. It is instructive to note that the discussion in King and Rebelo (1993) (concerning the time variation in the data) is surprisingly similar to the one in Banerjee and Duflo (2005), concerning the cross-county variation. Both calibrations run into the problem that interest rates would need to be implausible high in order for capital-labor ratios to match data on per-capita output.

the *entry intensity* of the economies innovation process. Although we provide the formal connection between the efficiency loss through distorting mark-ups and the entry intensity in more detail below, but we can give the main intuition right now. The equilibrium mark-up $\xi = \lambda^{\Delta}$ in increasing in the productivity of the quality leader relative to its most efficient competitor. Hence, an innovative environment where there is little dispersion in the type of technologies different firms use, will *cause* distortions to be small. One powerful mechanism to limit technology dispersion is the extent of entry of new firms.[9] In fact, for the case of Mexico, Turkey and Morocco, Roberts and Tybout (1996, p. 2) note that "the phenomena of entry, exit and reallocation of market share constitute a source of competitive pressure and can improve allocative efficiency" and that "trade exposure does appear to exert additional competitive pressure on markups in Mexico, Morocco and Turkey ... because it "disciplines" noncompetitive pricing behavior. The finding that openness reduces markups the most among large plants and concentrated industries suggests that market power is at least part of the story" (Roberts and Tybout, 1996, p. 12). So *if* we believe that entry is more efficient in developed countries or that technological diffusion is faster, it is natural to expect that the equilibrium degree of misallocation is lower in developed economies (see Barseghyan (2008) for evidence about the relation between entry costs and aggregate TFP). Our model will have this feature.[10] In the model, we take a very reduced-form approach to model differences in the entry technology by assuming the existence of exogenous entry costs. These could either be thought of as actual differences in the monetary costs to open a new firm (Djankov, Porta, Lopez-De-Silvaes, and Shleifer, 2002; Porta and Shleifer, 2009)) or as the reduced-form representation of a more structural model of entry.

The Process of Innovation We assume that innovations can stem from two sources: either current producers (i.e. leading quality firms) can invest in technological improvements, or new entrants can come up with new blueprints. We assume that both type of innovations are brought about by scientists (or innovators). The innovation technology is the following. Consider first the case of innovation by incumbents. Above we showed that the relevant state variables for the static part of the firm's problem is the firms' productivity q and productivity advantage Δ.[11] Hence we will assume that if a firm with quality q and quality advantage Δ wants to achieve a flow rate of I, it has to hire $G(I, q, \Delta)$ innovators. We will be more specific about the functional form of G below. If the current wage rate of innovators is $w_I(t)$, it thus costs $w_I(t)\, G(I, q, \Delta)$ to achieve a flow rate of innovation of I. In case the innovation is successful, the new productivity will be given by λq.

Now consider the process of entry. New blueprints are also brought about

[9] Similarly, better modes of technology diffusion across firms will have a similar effect.

[10] In the model the causality runs of course both ways: precisely because the economy is characterized by more efficient entry (which is endogenous and depends on the state of development), it is a developed economy, i.e. has high per-capita income and high levels of TFP, and has only little distortions.

[11] See for example (6) which shows that profit maximizing price is dependent on $q_F = q\lambda^{-\Delta}$.

by innovators. We will adopt the usual institutional structure of Schumpeterian growth models: whereas there is patent protection for the usage of technologies (i.e. no other firm can use technology q), the leading technologies are common knowledge for the process of innovation, i.e. innovators can try to improve upon them. More specifically: if successful, the innovator gets a blueprint of quality $\lambda q(\nu, t)$, where $q(\nu, t)$ is the leading quality in sector ν. The innovation technology of entrants is linear: each innovator generates a flow rate of innovation of η. Conditional on success, the innovator can use the generated blueprint after paying the entry costs $(1-\chi)V$, where V is the (endogenously determined) value of producing with the new blueprint. Hence, the innovator earns χV in case the innovation is successful. The parameter $\chi < 1$ parametrizes the entry costs and will be our main parameter for the comparative statics exercises.

To solve for the equilibrium innovation and entry rates, we have to solve for the equilibrium value of a firm. Let $V(q, \Delta, t)$ denote the value of a firm with quality q and quality gap Δ at time t. Consider first a quality leader, i.e. $\Delta > 0$. The value function solves the Hamilton-Jacobi-Bellman (HJB) equation

$$\begin{array}{rc} & r(t) V(q, \Delta, t) - \dot{V}(q, \Delta, t) \\ = & \pi(q, \Delta, t) - z(q, \Delta, t)(V(q, \Delta, t) - V(q, -1, t)) \\ & + \max_I \{I(V(\lambda q, \Delta+1, t) - V(q, \Delta, t)) - w_I(t) G(I, q, \Delta)\} \end{array},$$

where $z(q, \Delta, t)$ is the (endogenous) entry rate for a sector with leading quality q and quality gap Δ, which current incumbents take as given. As usual the return on the "asset", $r(t) V(q, \Delta, t)$, consists of the per-period dividends $\pi(q, \Delta, t)$ and the asset's appreciation $\dot{V}(q, \Delta, t)$. Additionally, with flow rate $z(q, \Delta, t)$ the current leader ceases to be in that position and instead is now one quality step behind the new entrant. Hence, with flow rate $z(q, \Delta, t)$ a value of $V(q, \Delta, t) - V(q, -1, t)$ is destroyed. Similarly, the possibility of investing in technological improvements represents an option value. By spending $w_I(t) G(I, q, \Delta)$, the firm generates a surplus of $V(\lambda q, \Delta+1, t) - V(q, \Delta, t)$ with flow rate I. As usual, the occurrence of both a successful incumbent innovation and entry is of second order.

As lagging firms do not produce, the value of not being the leading-edge firms is simply given by option of entry, so that free entry implies that $V(q, -1, t) = 0$. Furthermore, we assume that

$$G(I, q, \Delta) = \lambda^{-\Delta} g(I), \tag{23}$$

where g is increasing and convex and satisfies $\varepsilon(I) > 1$, where $\varepsilon(I) = g'(I) \frac{I}{g(I)}$ is the elasticity of g.[12] The latter condition is needed for there to exist a positive interior solution to the incumbents' innovation problem. The functional form imposed in (23) implies that innovations are easier the farther you are ahead and that the costs of research do not depend on actual productivity q. This assumption is similar in spirit to the assumption of knowledge capital made in Klette and Kortum (2004) and formalizes the notion that firms build on the

[12]Note that if g had constant elasticity, this latter condition would be implied by convexity.

shoulders of those giants, which they themselves built. Note that the only way how a firm can have a productivity advantage of Δ is if it had Δ innovations *and* there has not been any entry in between. Hence, we assume that innovation gets easier if the firm itself had multiple innovations *in a row*, which basically assumes that there is an externality between successful research conducted yesterday and the research technology today. As in Klette and Kortum (2004), this externality makes the model consistent with Gibrat's Law, i.e. with the fact the firms' growth rates are independent of size. Substituting (23) and the profit function (11) into the HJB equation yields

$$\begin{aligned} & (r(t) + z(q,\Delta,t)) V(q,\Delta,t) - \dot{V}(q,\Delta,t) \\ = \quad & Y(t)\left(1-\lambda^{-\Delta}\right) + \max_I \left\{ I \left(V(\lambda q, \Delta+1, t) - V(q,\Delta,t)\right) - w_I(t) \lambda^{-\Delta} g(I) \right\} \quad . \end{aligned} \tag{24}$$

This concludes the characterization of the model so that we can now define an equilibrium in this economy.

Definition 3. *Consider the economy described above. An equilibrium consists of sequences of consumption, output and capital* $[C(t), Y(t), K(t)]_{t=0}^{\infty}$*, distributions of leading qualities and productivity gaps* $[[q(\nu,t), \Delta(\nu,t)]_{\nu}]_{t=0}^{\infty}$*, wages and interest rates* $[w(t), w_I(t), r(t)]_{t=0}^{\infty}$*, intermediary demands, prices and factor demands* $[[y(\nu,t), p(\nu,t), l(\nu,t), k(\nu,t)]_{\nu}]_{t=0}^{\infty}$*, innovation and entry flow rates* $[[I(\nu,t), z(\nu,t)]_{\nu}]_{t=0}^{\infty}$ *and value functions* $[V(q,\Delta,t)]_{t=0}^{\infty}$ *such that*

- *consumers and intermediary firms behave optimally*
- *the value functions are consistent with free entry*
- *all markets clear*
- *the evolution of capital is consistent with consumers' savings decision*
- *the evolution of firm productivity is consistent with incumbents' innovation decision and the process of entry.*

A Balanced growth path (BGP) is an equilibrium where interest rate are constant and aggregate output grows at a constant rate g*.*

We conjecture that along the BGP, both innovator wages $w_I(t)$ and value functions $V(q,\Delta,t)$ also grow at rate g and that the rate of entry is constant and equal across all sectors, i.e. $z(q,\Delta,t) = z$. We will now characterize the BGP. The proofs some of the steps are contained in the Appendix. As z does not depend on q, the only state variables (besides time) is the quality gap Δ. Along the BGP, the HJB equation therefore reads

$$\begin{aligned} & (r + z - g) V(\Delta,t) \\ = \quad & \alpha Y(t)\left(1-\lambda^{-\Delta}\right) + \max_I \left\{ I \left(V(\Delta+1,t) - V(\Delta,t)\right) - w_I(t)\lambda^{-\Delta} g(I) \right\}. \end{aligned}$$

We show in the appendix, that the value function is given by

$$V(\Delta,t) = \frac{Y(t)}{\rho+z^*} - \frac{w_I(t)\,g(I^*)+Y(t)}{\rho+z^*+I^*\frac{\lambda-1}{\lambda}}\lambda^{-\Delta}, \tag{25}$$

where I^* is the optimal innovation rate, which is implicitly defined by

$$g'(I^*) = \frac{g(I^*)+\frac{Y(t)}{w_I(t)}}{\rho+z^*+I^*\frac{\lambda-1}{\lambda}}\frac{\lambda-1}{\lambda}, \tag{26}$$

and z^* is the equilibrium entry rate. As $\frac{Y(t)}{w_I(t)}$ is constant along the BGP, (36) shows that I^* is also constant. Furthermore, I^* is increasing in $Y(t)$ as higher demand increases monopolistic profits, and is decreasing in $w_I(t)$ and z^*, where the latter reflects Arrow's replacement effect: more entry reduces the innovation incentives because it is less likely that the innovator can reap the benefits of his investment in the future.

We now characterize the equilibrium distribution of productivity gaps Δ, which will be stationary. Let $\mu(\Delta,t)$ be the measure of firms with quality gap Δ. As incumbents innovate at flow rate I^* and new firms enter at flow rate z^*, the distribution $[\mu(\Delta,t)]_\Delta$ solves the flow equations

$$\dot{\mu}(\Delta,t) = \begin{cases} -z^*\mu(\Delta,t)+I^*\mu(\Delta-1,t)-I^*\mu(\Delta,t) & \text{if } \Delta\geq 2 \\ z^*(1-\mu(1,t))-I^*\mu(1,t) & \text{if } \Delta=1 \end{cases}. \tag{27}$$

To understand these flow equations, consider first the upper branch. There are two ways to leave the state (Δ,t): the current producer could have an innovation or there could be entry. The only way to get into this state is by being in state $\Delta-1$ and then having an innovation. Similarly, all firms in state $(1,t)$ exit this state if they have an innovation and all sectors where entry occurs will enter the state $(1,t)$. Hence, given the endogenous (and constant) innovation and entry flow rates I^* and z^*, standard arguments show that this stochastic process will settle into a stationary distribution $[\mu(\Delta)]_\Delta$, which is given by

$$\mu(\Delta) = \left(\frac{I^*}{z^*+I^*}\right)^{\Delta}\frac{z^*}{I^*} = \left(\frac{1}{1+x}\right)^{\Delta}x, \tag{28}$$

where the second equality shows that the distribution of distortions is only dependent on the *entry intensity* defined by $x=\frac{z}{I}$. Hence, along the BGP, both the cumulative distribution of productivity gaps F and the two sufficient statistics summarizing the effect of distortions M and Λ are also only dependent on the entry intensity. In particular, these terms are given by[13]

$$F^{BGP}(\Delta; x) = \sum_{i=1}^{\Delta} \mu(i) = 1 - \left(\frac{1}{x+1}\right)^{\Delta+1} \tag{29}$$

$$\Lambda^{BGP}(x) = \frac{x}{(\lambda-1)+\lambda x} \tag{30}$$

$$M^{BGP}(x) = \lambda^{-\frac{x+1}{x}} \frac{(\lambda-1)+\lambda x}{x}. \tag{31}$$

Using (29)-(31) we can gather the main effects of a higher entry intensity in the following

Proposition 4. *Consider the economy described above. For a given distribution of productivity $[q(\nu,t)]_{\nu}$ and level of capital $K(t)$, a higher entry intensity x*

1. *reduces equilibrium mark-ups in a first order dominance sense, as $F^{BGP}(\Delta, x)$ is increasing in x for all Δ*
2. *increases aggregate TFP and aggregate income $Y(t)$, as $M^{BGP}(x)$ is increasing in x*
3. *increases equilibrium factor prices as $M^{BGP}(x)\Lambda^{BGP}(x) = \lambda^{-\frac{x+1}{x}}$ is increasing in x*
4. *reduces the wedge between factor prices and their marginal products as $\Lambda^{BGP}(x)$ is increasing in x (see (22))*
5. *Increases average firm size and reduces the dispersion in firm size as measured by employment or sales.*

Proof. Claims 1. to 4. follow immediately from (29)-(31) and Proposition (1) and Corollary (2).[14] To see part 5., recall that log employment is given by (see (10))

$$ln(l(\nu,t)) = ln(L) - ln\left(\Lambda^{BGP}(x)\right) - \Delta(\nu,t)\, ln(\lambda).$$

Hence, mean employment and size dispersion are given by

$$E[ln(l(\nu,t))] = ln\left(L^{1-\alpha}K(t)^{\alpha}\right) + ln(q) + ln\left(M^{BGP}(x)\right)$$

$$var(ln(l(\nu,t))) = (ln(\lambda))^2\, var(\Delta(\nu,t)) = (ln(\lambda))^2 \frac{1+x}{x^2},$$

so that mean employment is increasing and employment dispersion is decreasing in the entry intensity of the economy. The analysis is identical for sales. □

[13]See Appendix for the derivation.

[14]To see part 2., define $s = x^{-1}$ so that $M^{BGP}(s) = \lambda^{-(1+s)}((\lambda-1)s+\lambda)$. We will show that M^{BGP} is decreasing in s. This is the case if

$$\frac{\partial ln\left(M^{BGP}(s)\right)}{\partial s} \equiv h(\lambda, s) = \frac{(\lambda-1)}{(\lambda-1)s+\lambda} - ln(\lambda) < 0.$$

As h is decreasing in both s and λ, it is sufficient to note that $h(\lambda,s) < h(1,0) = 0$ and $\frac{\partial h(\lambda,0)}{\partial \lambda} = \frac{1}{\lambda^2}(1-\lambda) < 0$ for all $\lambda > 1$. Hence, $M^{BGP}(x)$ is increasing in x.

Proposition 4 shows the importance of the allocation of innovative resources between entrants and incumbents and especially the efficiency-enhancing role of entry: In this economy both types of innovation bring about the same productivity gain (the leading quality increases by the factor λ) but entry reduces factor misallocation by limiting monopolistic pricing. This formalizes the intuition that both TFP and equilibrium factor prices (even relative to the already low marginal product) are especially low in economies which are characterized by an innovative environment where the entry intensity x is low. The main empirical implication for firm-level data concerns the size distribution where the mark-up reducing role of entry both compresses the cross-sectional variation and increases the average size as the average firm has less monopoly power. These implications are consistent with findings from Alfaro, Charlton, and Kanczuk (2008) and Bartelsman, Haltiwanger, and Scarpetta (2009).

Proposition 4 also has implications for the incentives to accumulate capital. Standard arguments show that the BGP level of capital in efficiency units is given by

$$\tilde{k} = \frac{k(t)}{Q(t)^{1/(1-\alpha)}} = \left(\frac{\alpha \Lambda^{BGP}(x)\, M(x)^{GBP}}{g+\delta+\rho}\right)^{\frac{1}{1-\alpha}} = \left(\frac{\alpha}{g+\delta+\rho}\right)^{\frac{1}{1-\alpha}} \lambda^{-\frac{x+1}{(1-\alpha)x}}.$$

Hence: conditional on the growth rate g, the BGP level of capital (in efficiency units) depends on the distortion terms $\Lambda^{BGP}(x)$ and $M^{BGP}(x)$, because the BGP interest rate does (see (**??**)). Note in particular, that for a given growth rate, the BGP level of (normalized) capital is increasing in the entry intensity. Intuitively: if the entry component to generate a growth rate of g is higher, there will be more competitive pressure. This in turn increases the demand for capital, hence also the equilibrium price of capital (the rental rate) and therefore the incentive to accumulate it.

The Equilibrium Entry Intensity To determine the equilibrium entry intensity x, we have to consider the allocation of researchers between incumbents and entrants. When working for an existing firm, researchers can earn a wage $w_I(t)$. If they instead try to develop a new blueprint to leapfrog the current quality leader, they are successful with flow rate η in which case they get paid the value of the firm net of entry costs, $\chi V(1,t)$. Along the BGP (where $I > 0$ and $z > 0$) innovators are indifferent where to work[15] and we therefore have that

$$\frac{w_I(t)}{\eta} = \chi V(1,t) = \chi \left(\frac{\alpha Y(t)}{\rho + z^*} - \frac{w_I(t)\, g(I^*) + \alpha Y(t)}{\rho + z^* + I^* \frac{\lambda-1}{\lambda}} \frac{1}{\lambda}\right), \tag{32}$$

where the second equality substitutes the value function from (25). (32) not only shows that $\frac{Y(t)}{w_I(t)}$ is constant as required, but we can also use it to express

[15] Note that we anticipated this result already in the formulation of the households' budget constraint (3), where total labor income generated by scientists is given by $w_I(t)\, L_S$.

the incumbents' optimality condition (36) as

$$g'(I) = \frac{g(I^*) + \frac{\rho + z^*}{\eta\chi}}{(\rho + z^* + I^*)}. \tag{33}$$

(33) gives us a first relationship between the innovation intensity I^* and the entry rate z^*. The second relationship comes from the market for innovators. Equilibrium market clearing requires that

$$\begin{aligned} L_S &= \int_0^1 \lambda^{-\Delta(\nu,t)} g(I^*)\, d\nu + L_E = g(I)\, \Lambda^{BGP}\left(\frac{z^*}{I^*}\right) + L_E \\ &= g(I^*)\, \frac{z^*}{(\lambda - 1)\, I^* + \lambda z^*} + L_E, \end{aligned}$$

where L_E is the measure of entrants and $\int_0^1 \lambda^{-\Delta(\nu,t)} g(I^*)\, d\nu$ is the aggregate innovator demand by incumbents (see (23)). As each entrant generates a flow rate of innovation η, it follows that $\eta L_E = z^*$, so that the market clearing condition reduces to

$$g(I^*)\, \frac{z^*}{(\lambda - 1)\, I^* + \lambda z^*} + \frac{z^*}{\eta} = L_S. \tag{34}$$

(34) and (33) are two equations in the two unknowns (z^*, I^*). Hence, to show existence and uniqueness of the equilibrium, we simply have to establish that(34) and (33) have a unique solution. That this is the case is contained in the following Proposition, which is proved in the Appendix.

Proposition 5. *Consider the economy described above. There exists a unique BGP and innovation and entry rates (I^*, z^*) are constant along this BGP. In particular, equilibrium entry z^* and innovation rates I^* are increasing in the supply of innovative resources (the no of scientists) L_S and in the size of an innovation λ. Furthermore, the equilibrium entry intensity $x = \frac{z^*}{I^*}$ is decreasing in the entry costs $1 - \chi$, as higher entry costs reduce the extent of entry z^* and increase the equilibrium innovation rate I^*.*

Proof. See Appendix. □

Together with the results in (29)-(31) and the discussion afterwards, Proposition 5 establishes the efficiency consequences of higher entry costs: Higher entry costs on future entrants shift the innovative resources of the economy towards working for current incumbents. This in turn reduces the allocative efficiency of the economy, as product markets get less competitive and the equilibrium distribution of mark-ups increases (in a distributional sense). In particular, Proposition 4 shows that both the dispersion and the level of mark-ups increases in the entry intensity so that a distorted entry margin reduces both aggregate TFP and equilibrium factor prices.

It is worth noting that the economic channel is very different from the ones stressed in Barseghyan and DiCecio (2009) or Buera, Kaboski, and Shin (2009). In these papers, inefficient entry is also at the heart of aggregate TFP losses but the mechanism is distinct. Barseghyan and DiCecio (2009) use a model a la Hopenhayn (1992) and show that higher entry costs lower aggregate TFP because it reduces the productivity of the marginal entrant as lower factor prices make production profitable even for relatively unproductive firms. Similarly, Buera, Kaboski, and Shin (2009) argue that credit market imperfections distort the allocation of talent across sectors if sectors are characterized by different entry costs. Hence, both of these papers link the costs of entry to the resulting equilibrium talent distribution. This paper supplements these ideas by showing that lower entry intensities will have negative TFP consequences *conditional on the distribution of productivity and capital* (see Proposition (4)) simply by affecting the stationary distribution of mark-ups.

3 Conclusion

Explaining the persistently large income and TFP differences across countries is a question of first-order importance in economics. In this paper we propose a model that combines two methodologically very different approaches considered in the literature so far. Following the distortion literature as exemplified by Restuccia and Rogerson (2008) and Hsieh and Klenow (2009) we argue that misallocation of production factors across firms is important to understand differences in aggregate TFP. In stark contrast to this literature however, we propose a structural model, where we can link aggregate TFP to objects of the economic environment, in our case the costs of entry. By doing so, we can study the equilibrium distribution of firm-level distortions and hence understand what underlying factors generate misallocation across firms.

In our model, distortions are equated with monopolistic mark-ups. We show that both the stationary distribution of mark-ups is parametrized by a single variable, namely the entry intensity of the economy. We also show that it is this entry intensity that summarizes the adverse effects on TFP and factor prices (the degree of misallocation). We can therefore study the static TFP effects of entry distortions via their effects on the economy's entry intensity. In particular: Low entry costs cause the equilibrium entry intensity to be high, which in turn will make product markets more competitive as the productivity dispersion between producing firms and their competitors is reduced. This lowers equilibrium mark-ups, which - from an aggregate perspective - show up in higher TFP and higher factor prices.

The model we consider is very stylized to make the analysis as tractable as possible. By doing so we tried to show how it is possible to embed the recent literature on firm-level distortions in a structural model and hence to understand what those distortions are economically. The model considered here presents one example - monopolistic mark-ups, which are not constant across firms. The obvious next step is of course a careful empirical analysis using micro data to study if this mechanism is important in practice.

References

Acemoglu, D., P. Antras, and E. Helpman (2007): "Contracts and Technology Adoption," *American Economic Review*, 97, 916–943.

Acemoglu, D., S. Johnson, and J. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91(5), 1369–1401.

Aghion, P., and P. Howitt (1992): "A Model of Growth through Creative Destruction," *Econometrica*, 60(2), 323–351.

Alfaro, L., A. Charlton, and F. Kanczuk (2008): "Plant-Size Distribution and Cross-Country Income Differences," NBER Working Paper 14060.

Banerjee, A., and E. Duflo (2005): "Growth Theory through the Lens of Development Economics," in *Handbook of Economic Growth*, ed. by P. Aghion, and S. Durlauf, vol. 1A, chap. 7, pp. 473–552. Elsevier B.V.

Banerjee, A., and B. Moll (forthcoming): "Why does misallocation persist," *American Economic Journal: Macroeconomics*.

Barseghyan, L. (2008): "Entry Costs and cross-country differences in productivity and output," *Journal of Economic Growth*, 13(145-167).

Barseghyan, L., and R. DiCecio (2009): "Entry Costs, Misallocation and Cross-Country Income and TFP Differences," Working paper, Feder Reserve Bank of St. Louis.

Bartelsman, E. J., J. C. Haltiwanger, and S. Scarpetta (2009): "Cross-Country Differences in Productivity: The Role of Allocation and Selection," Working paper 15490, NBER.

Bernard, A., J. Eaton, J. B. Jensen, and S. Kortum (2003): "Plants and Productivity in International Trade," *American Economic Review*, 93, 1268–1290.

Buera, F., J. Kaboski, and Y. Shin (2009): "Finance and Development: A Tale of Two Sectors," Working papers.

Caselli, F. (2005): "Accounting for Cross-Country Income Differences," in *Handbook of Economic Growth*.

Chari, V., P. J. Kehoe, and E. R. McGrattan (2007): "Business Cycle Accounting," *Econometrica*, 75(3), 781–836.

Djankov, S., R. L. Porta, F. Lopez-De-Silvaes, and A. Shleifer (2002): "The Regulation of Entry," *Quarterly Journal of Economics*, 117(1), 1–37.

Foster, J., J. C. Haltiwanger, and C. Syverson (2008): "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?," *American Economic Review*, 98(1), 394–425.

Ganica, G., and F. Zilibotti (2009): "Technological Change and the Wealth of Nations," *Annual Review of Economics*, 1, 93–120.

Grossman, G., and E. Helpman (1991): "Quality Ladders in the Theory of Growth," *Review of Economic Studies*, 58, 43–61.

——— (1993): *Innovation and Growth in the Global Economy*. MIT Press.

Hall, R., and C. Jones (1999): "Why Do Some Countries Produce So Much More Ouput Per Worker Than Others?," *Quarterly Journal of Economics*, 114(1), 83–116.

Hopenhayn, H. (1992): "Entry, Exit and Firm dynamics in Long-Run Equilibrium," *Econometrica*, 60, 1127–1150.

Hsieh, C.-T., and P. Klenow (2009): "Misallocation and Manufacturing TFP in China and India," *Quarterly Journal of Economics*, 124(2), 771–807.

King, R. G., and S. T. Rebelo (1993): "Transitional Dynamics and Economic Growth in the Neoclassical Model," *American Economic Review*, 83(4), 908–931.

Klenow, P., and A. Rodriguez-Clare (1997): "The Neoclassical Revival in Growth Economics: Has it Gone Too Far?," in *NBER Macroeconomics Annual*, vol. 12, pp. 73–103.

Klette, T. J., and S. Kortum (2004): "Innovating Firms and Aggregate Innovation," *Journal of Political Economy*, 112(5), 986–1018.

Lagos, R. (2006): "A Model of TFP," *Review of Economic Studies*, 73(4), 983–1007.

Lucas, R. E. (1988): "On the Mechanics of Economic Development," *Journal of Monetary Economics*, 22, 3–42.

——— (1990): "Why Doesn't Capital Flow From Rich to Poor Countries," *American Economic Review*, 80(2), 92–96.

Midrigan, V., and D. Y. Xu (2010): "Finance and Misallocation: Evidence From Plant-Level Data," NBER Working Paper 15647.

Moll, B. (2009): "Productivity Losses from Financial Frictions: Can Self-financing Undo Capital Misallocation?," .

Parente, S. L., and E. C. Prescott (1994): "Barriers to Technology Adoption and Development," *Journal of Political Economy*, 102(2), 298–321.

Porta, R. L., and A. Shleifer (2009): "The Unofficial Economy and Economic Development," in *Brookings Papers on Economic Activity*, pp. 275–352.

Restuccia, D., and R. Rogerson (2008): "Policy distortions and aggregate productivity with heterogenous establishments," *Review of Economic Dynamics*, 11, 707–720.

Roberts, M., and J. Tybout (eds.) (1996): *Industrial Evolution in Developing Countries. Micro Patterns of Turnover, Productivity, and Market Structure*. Oxford University Press.

4 Appendix

4.1 Characterization of the BGP

As z does not depend on q, the only state variables (besides time) is the quality gap Δ, so along the BGP (24) reads

$$\begin{aligned} & (r+z-g)\,V(\Delta,t) \\ = \quad & \alpha Y(t)\left(1-\lambda^{-\Delta}\right)+\max\nolimits_I\left\{I\left(V(\Delta+1,t)-V(\Delta,t)\right)-w_I(t)\,\lambda^{-\Delta}g(I)\right\}. \end{aligned}$$

We now conjecture that the value function takes the form of

$$V(\Delta,t)=\kappa(t)-\phi(t)\,\lambda^{-\Delta}. \tag{35}$$

Using this conjecture, we get

$$V(\Delta+1,t)-V(\Delta,t)=\phi(t)\,\frac{\lambda-1}{\lambda}\lambda^{-\Delta}.$$

If V grows at rate g, (35) implies that both $\kappa(t)$ and $\phi(t)$ grow at rate g too. Hence let us normalize all variables by $exp(gt)$ so that they become stationary (and hence we drop the time argument) and t ceases to be a state variables. The Bellman equation therefore reads

$$(r+z-g)\,V(\Delta) \quad = \quad Y\left(1-\lambda^{-\Delta}\right)+\lambda^{-\Delta}\max_I\left\{I\phi\frac{\lambda-1}{\lambda}-w_I g(I)\right\}.$$

The optimal innovation flow rate I^* is implicitly defined by the necessary and sufficient FOC

$$g'(I^*)=\frac{\phi}{w_I}\frac{\lambda-1}{\lambda}. \tag{36}$$

Hence, I^* is independent of time and the same for all firms. For I^* to be positive we need that

$$0 \quad < \quad I\phi(t)\,\frac{\lambda-1}{\lambda}\lambda^{-\Delta}-w_I(t)\,\lambda^{-\Delta}g(I)=\lambda^{-\Delta}w_I(t)\,g(I)\left[\varepsilon(I)-1\right],$$

where $\varepsilon(I)$ is the elasticity of the innovation production function. This is satisfied by assumption. Substituting our conjecture (35), we get that

$$\begin{aligned}(r+z-g)\kappa-(r+z-g)\phi\lambda^{-\Delta} &= Y\left(1-\lambda^{-\Delta}\right)+\lambda^{-\Delta}\left[I^*\phi\frac{\lambda-1}{\lambda}-w_I g\left(I^*\right)\right] \\ &= Y+\left[I^*\phi\frac{\lambda-1}{\lambda}-w_I g\left(I^*\right)-\alpha Y\right]\lambda^{-\Delta}.\end{aligned}$$

As this has to hold for all Δ, it is clear that

$$-(r+z-g)\phi = I^*\phi\frac{\lambda-1}{\lambda}-w_I g\left(I^*\right)-Y,$$

which yields

$$\phi=\frac{w_I g\left(I^*\right)+Y}{r+z-g+I^*\frac{\lambda-1}{\lambda}}. \tag{37}$$

Similarly we get $\kappa=\frac{\alpha Y}{r+z-g}$, so that the stationary version of the value function is given by

$$V(\Delta)=\frac{Y}{r+z-g}-\frac{w_I g\left(I^*\right)+Y}{r+z-g+I^*\frac{\lambda-1}{\lambda}}\lambda^{-\Delta}.$$

Hence,

$$V(\Delta,t) = V(\Delta)e^{gt}=\frac{Y(t)}{r+z-g}-\frac{w_I(t)\,g\left(I^*\right)+Y(t)}{r+z-g+I^*\frac{\lambda-1}{\lambda}}\lambda^{-\Delta}.$$

This verifies our conjecture and shows that the value function is increasing in concave in Δ.

For $V(\Delta,t)$ to grow at a constant rate, we need to show that $w_I(t)$ and output $Y(t)$ grow at a constant rate g. This follows from 32 in the main text. That the growth rate along the BGP is constant is shown in Section 4.3.

4.2 Proofs of Proposition 4

Too see this, note that

$$\begin{aligned}F(\Delta) &= \sum_{i=1}^{\Delta}\mu(i)=x\sum_{i=1}^{\Delta}\left(\frac{1}{1+x}\right)^i=\frac{x}{1+x}\sum_{i=1}^{\Delta}\left(\frac{1}{1+x}\right)^i \\ &= \frac{x}{1+x}\frac{1-\left(\frac{1}{1+x}\right)^{\Delta+1}}{1-\frac{1}{1+x}}=1-\left(\frac{1}{1+x}\right)^{\Delta+1},\end{aligned}$$

that

$$\begin{aligned}\Lambda^{BGP}(x) &= x\sum_{i=1}^{\infty}\left(\frac{1}{\lambda(x+1)}\right)^{\Delta}=\frac{x}{\lambda(x+1)}\sum_{i=0}^{\infty}\left(\frac{1}{\lambda(x+1)}\right)^{\Delta} \\ &= \frac{x}{\lambda(x+1)}\left(\frac{1}{1-\frac{1}{\lambda(x+1)}}\right)=\frac{x}{\lambda-1+\lambda x},\end{aligned}$$

and that

$$\int_0^1 \Delta(\nu,t)\,d\nu = \sum_{i=1}^{\infty} i\mu(i) = x\sum_{i=1}^{\infty} i\left(\frac{1}{1+x}\right)^i = x\frac{\frac{1}{1+x}}{1-\frac{1}{1+x}}\sum_{i=0}^{\infty}\left(\frac{1}{1+x}\right)^i$$
$$= \frac{1}{1-\frac{1}{1+x}} = \frac{1+x}{x},$$

so that $M^{BGP}(x) = \frac{1}{\Lambda^{BGP}}\lambda^{-\int_0^1 \Delta(\nu,t)d\nu} = \lambda^{-\frac{1+x}{x}}\frac{\lambda-1+\lambda x}{x}$.

4.3 BGP Growth Rate

To derive the BGP growth rate, note that

$$\begin{aligned} ln(Q(t+\delta)) &= \int_0^1 ln(q(\nu,t+\delta))\,d\nu \\ &= \int_{\nu\in I} ln(q(\nu,t+\delta))\,d\nu + \int_{\nu\in E} ln(q(\nu,t+\delta))\,d\nu \\ &\quad + \int_{\nu\in N} ln(q(\nu,t+\delta)) + d\nu \int_{\nu\in B} ln(q(\nu,t+\delta))\,d\nu, \end{aligned}$$

where I, E, N and B denote the set of sectors which incur an innovation, entry, neither or both. Let $\zeta(I,\delta)$ be the measure of sector who have an innovation in the interval of length δ. Then

$$ln(Q(t+\delta)) - ln(Q(t)) = ln(\lambda)\left[\zeta(I,\delta) + \zeta(E,\delta) + \zeta(B,\delta)\right].$$

With constant equilibrium flow rates I and z, the respective measures are given by

$$\zeta(I,\delta) = I\delta \text{ and } \zeta(E,\delta) = z\delta \text{ and } \zeta(B,\delta) = zI\delta^2.$$

The BGP growth rate of productivity is therefore given by

$$\begin{aligned} g_Q &= \lim_{\delta\to 0}\frac{ln(Q(t+\delta)) - ln(Q(t))}{\delta} = \lim_{\delta\to 0}\frac{ln(\lambda)\left(I\delta + z\delta + zI\delta^2\right)}{\delta} \\ &= ln(\lambda)(I+z). \end{aligned}$$

4.4 Uniqueness and Existence of BGP

We have to establish that the two equations

$$g'(I) = \frac{g(I^*) + \frac{\rho+z}{\eta\chi}}{(\rho+z+I^*)} \tag{38}$$

$$L_S = g(I^*)\frac{z^*}{(\lambda-1)I^* + \lambda z^*} + \frac{z^*}{\eta} \tag{39}$$

have a unique solution. Consider first (39). Totally differentiating yields

$$0 = \left(\frac{g(I^*)(\lambda-1)I^*}{((\lambda-1)I^*+\lambda z^*)^2} + \frac{1}{\eta}\right)dz^* + \frac{z^* g(I^*)(\lambda-1)(\varepsilon(I^*)-1) + (z^*)^2 g'(I^*)\lambda}{((\lambda-1)I^*+\lambda z^*)^2}dI^*$$

so that

$$\frac{dz^*}{dI^*} = -\eta \frac{z^* g(I^*)(\lambda-1)(\varepsilon(I^*)-1) + (z^*)^2 g'(I^*)\lambda}{g(I^*)(\lambda-1)I^*\eta + \eta((\lambda-1)I^* + \lambda z^*)^2} < 0.$$

This is intuitive: for a given supply of scientists, if more entry takes place, the incumbents' innovation rate has to decrease. Now consider (38). Totally differentiating yields

$$\begin{aligned}
g''(I^*)dI^* &= \frac{g'(I^*)(\rho + z^* + I^*) - \left(g(I^*) + \frac{\rho+z^*}{\eta\chi}\right)}{(\rho + z + I^*)^2} dI^* + \\
&\quad \frac{\frac{1}{\eta\chi}(\rho + z^* + I^*) - \left(g(I^*) + \frac{\rho+z^*}{\eta\chi}\right)}{(\rho + z^* + I^*)^2} dz^*. \\
&= \frac{\frac{1}{\eta\chi} - \frac{g(I^*) + \frac{\rho+z^*}{\eta\chi}}{\rho+z^*+I^*}}{(\rho + z + I^*)} dz, \\
&= \frac{\frac{1}{\eta\chi} - g'(I^*)}{(\rho + z + I^*)} dz^*
\end{aligned}$$

where the second equality uses (38). Hence,

$$\frac{dz^*}{dI^*} = \frac{(\rho + z^* + I^*)\, g''(I^*)}{\frac{1}{\eta\chi} - g'(I^*)}.$$

The numerator is clearly positive by the convexity of g. Hence, consider the denominator. From the first order condition (36) we get that

$$\frac{1}{\eta\chi} - g'(I^*) = \frac{1}{\eta\chi} - \frac{\phi(t)}{w_I(t)} \frac{\lambda-1}{\lambda}.$$

From the free entry condition for entrants we get $w_I(t) = V(1,t) = \kappa(t) - \phi(t)\frac{1}{\lambda}$ so that

$$\frac{1}{\eta\chi} - g'(I^*) = \frac{1}{\eta\chi}\left(1 - \frac{\phi(t)\frac{\lambda-1}{\lambda}}{\left(\kappa(t) - \phi(t)\frac{1}{\lambda}\right)}\right) = \frac{1}{\eta\chi}\left(\frac{\kappa(t) - \phi(t)}{\kappa(t) - \phi(t)\frac{1}{\lambda}}\right).$$

But now note that

$$\kappa(t) - \phi(t) = \lambda\left(\kappa(t) - \phi(t)\frac{1}{\lambda} - \frac{\lambda-1}{\lambda}\kappa(t)\right) = \lambda\left(V(1,t) - \frac{\lambda-1}{\lambda}\kappa(t)\right) > 0,$$

where the last inequality follows from

$$\begin{aligned}
V(1,t) &= \frac{1}{r+z-g}\left(\alpha Y(t)\left(1 - \frac{1}{\lambda}\right) + \max_I\left\{I\phi(t)\frac{\lambda-1}{\lambda}\lambda^{-\Delta} - w_I(t)\lambda^{-\Delta}g(I)\right\}\right) \\
&> \frac{\lambda-1}{\lambda}\frac{\alpha Y(t)}{r+z-g} \\
&= \frac{\lambda-1}{\lambda}\kappa(t).
\end{aligned}$$

Hence, $\frac{1}{\eta\chi} - g'(I^*) > 0$ so that

$$\frac{dz^*}{dI^*} = \frac{(\rho + z^* + I^*)\, g''(I^*)}{\frac{1}{\eta\chi} - g'(I^*)} > 0.$$

To conclude the proof note that (39) implies that $lim_{I\to 0} z = \eta L_E$ and that $lim_{I\to\infty} z = 0$. Hence, (39) describes a decreasing function starting at ηL_E which is approaching 0. But (38) implies that $lim_{I\to 0} z = -\rho < \eta L_E$ and $lim_{I\to\infty} z = \infty$,[16] so that (38) describes an increasing function starting at $-\rho$ and approaching infinity. As both are continuous, there is a unique intersection. To see that the entry intensity decreases in the entry costs, note that higher entry costs (which correspond to a lower level of χ) reduce z^* for given I^* from (38) but leave (39) unchanged. Hence, $\frac{dz^*}{d\chi} > 0$ and $\frac{dI^*}{d\chi} < 0$ so that $\frac{dx}{d\chi} = \frac{d(z^*/I^*)}{d\chi} > 0$.

[16] To see former, write (38) as

$$z = \frac{(\rho + I)\, g'(I) - g(I) - \frac{\rho}{\eta\chi}}{\frac{1}{\eta\chi} - g'(I)} \to -\rho,$$

as $lim_{I\to 0} g'(I) = lim_{I\to 0} g'(I) = 0$. To see the latter, rewrite (38) as

$$\varepsilon(I) = \frac{g'(I)\, I}{g(I)} = \frac{I}{g(I)} \frac{g(I^*) + \frac{\rho + z}{\eta}}{(\rho + z + I^*)} = \frac{1 + \frac{\rho + z}{\eta g(I)}}{1 + \frac{\rho + z}{I}},$$

which implies that $lim_I \varepsilon(I) = 1$ if z was bounded. But we assumed that $\varepsilon(I) > 1$.