# Firm Microstructure and Aggregate Productivity

Hugo A. Hopenhayn

9th March 2010

## 1 Introduction

I was asked to consider the impact of Industrial Organization on recent research in macroeconomics. This is a difficult question for at least two reasons. In the first place, macroeconomics is probably one of the broadest areas in economics so its boundaries are hard to define. Research ranges from abstract theory to very applied work, traditional topics such as Monetary economics, Business Cycles and Growth, to others in labor economics, public finance, education, health, international, development, contract theory and many others. Modern macro is defined both on the basis of some central questions as by the application of a common methodology to many different areas, combining dynamics and some general equilibrium concept.

IO is not an easy field to define either, ranging also from abstract theory to very applied work and a broad set of questions relating to the behavior of firms and markets: pricing decisions, determinants of market structure, entry and exit, productivity, research and development, advertising,etc. Most research has taken the form of partial equilibrium analysis in game theoretic settings, combined with important advances in applied econometrics and a recent tendency to focus on narrowly defined empirical studies that do not lend easily to aggregation.[1]

[1] [Peltzman, 1991] famous critique pointed to "the seeming inability of the recent theory to lead to any powerfuld generarlizaton...an almost interminable series of special cases...", suggesting as an explanation the "gulf between theory and empirical work." The response of recent new empirical inducstrial organization has also generated skepticism about the possibility of generalization. As pointed out by [Sutton, 1995], this recent literature Many

As someone that that lingers in the boundaries of these two broad areas, you run the risk -as happens to me frequently- to be considered "*not IO enough*" by some and "*not macro enough*" by others. I will make no attempt here to convice others that I am *IO enough* or *Macro enough* but instead take a view through this narrower window where I stand and focus on Firm Dynamics and some of its implications for Macroeconomics. This work is at the nexus of a central interest to IO -the behavior of firms and markets- and one central to Macroeconomics -the determinants of aggregate productivity.

In his discussion about the state of Industrial Organization, [Sutton, 1995] argues the need for:

> ... sharp empirical regularities arising over a wide cross-section of industries. That such regularities appear in spite of the fact that every industry has many idiosyncratic features suggests that they are molded by some highly robust competitive mechanisms – and if this is so, then these would seem to be mechanisms that merit careful study. If ideas from the IO field are to have relevance in other areas of economics, such as International Trade or Growth Theory, that relevance is likely to derive from mechanisms of this robust kind...

It is thus not surprising that one of the areas of greater intersection between IO and macro -as indicated in the analysis in Appendix 9- is precisely an area relating to *Market structure and Size distribution of firms.* I will describe in this paper a body of theoretical and empirical work that focuses on regularities suggested by Sutton that provides the *firm-microstrucutre foundations* of the aggregate production function and aggregate productivity.

Economists have long been interested in firm dynamics and the size distribution of firms. The turnover of firms and reallocation of factors of production has been an inseparable part of the history of industrialization and the development of a market economy. In the early part of the century, [Viner, 1932] emphasized the role of economies of scale as a determinant

researchers ....natural response .... focus on some specific market ... ‘ultra-micro’ work ....led to a growing skepticism about the value of searching for statistical regularities that hold across a broad run of different industries..." This concern is also echoed by [Schmalensee, 1989], "...inter-industry research in industrial organization should generally be viewed as a search for empirical regularities, not as a set of exercises in structural estimation."

of firm size. This came out also as an important determinant of market structure, together with sunk costs, in the extensive empirical work by [Bain, 1951][Bain, 1954][Bain, 1956]) and others. [Lucas Jr, 1978] classic paper was the first to derive size distribution from an economic model as the solution to the problem of optimally allocating resources to managers with different talents. As in Viner, this theory relies on economies of scale given by the indivisibility of managers and decreasing returns on capital and labor for a given firm, resulting in the analogue to the classic U-shaped average cost curve. In contrast to Viner that focused at the industry level, Lucas interprets his results as applying to the size distribution of firms in general. Lucas' model -or variants of this type- are now the standard way of modeling the firm size distribution.

The interest of applied economists on the size distribution of firms was accompanied by a parallel interest in firm dynamics: stochastic growth, entry and exit of firms. [Gibrat, 1931] law of proportionate growth stating that firm growth is independent of firm size, is a cornerstone of firm dynamics and is used frequently as an assumption in the building of economic models.[2] [Hart and Prais, 1956],[Simon and Bonini, 1958] and [Adelman, 1958] derived with great success size distribution as the stationary distribution of a process of idiosyncratic shocks to firm size. [Jovanovic, 1982] was the first paper to develop an economic model of firm dynamics that is motivated by these and other facts. His model of *selection* provides a very elegant and persuasive darwinian story to explain firm dynamics. Firms learn about their underlying value, as productivity shocks are drawn from a distribution with unknown and firm-idiosyncratic mean. Firm size is determined by the posterior mean of this shock and as it changes over time, firms adjust their sizes. Among other implications, the model predicts a decline in the variance of growth rates and the hazard rates for exit of firms as a function of age, which has been extensively confirmed by a large body of empirical work.[3]

Jovanovic's model gives a non-stationary process for firm size which in the limit converges to a fixed value, and in the long run there is no entry

[2] The large accumulated evidence on firm dynamics suggests that firm size is a highly persistent stochastic process, but with some degree of mean reversion. See [Sutton, 1997].

[3] See [Dunne *et al.*, 1988],[Dunne *et al.*, 1989b],
[Dunne *et al.*, 1989a],[Davis *et al.*, 1998],[Davis and Haltiwanger, 1992],
[Evans, 1987b],[Leonard, 1988],
[Evans, 1987a],[Hall, 1987],[Geroski, 1995],
[Caves, 1998]

and exit. In spite of the model's appeal, it lacks the tractability to be used as a microfoundation for aggregate productivity, in the way Lucas' model has been utilized. [Hopenhayn, 1992a] incorporates entry and exit and firm dynamics in a tractable way in a Lucas-type model, assuming productivity shocks follow a Markov process with a well defined and non-degenerate limiting distribution. The stochastic process assumed on the shocks is in the spirit of the early empirical papers cited above. Firm size is determined each period in a similar fashion to Lucas, as the solution to the problem of allocating resources to firms to maximize total output. Entry and exit are also determined as part of a stationary equilibrium. This model and other variants have been used extensively as firm-microstructure in many recent papers in macroeconomics.

The work by Davis and Haltiwanger and others, had a tremendous impact by considerably extending the interest on job creation and destruction and the importance of firm microstructure among macroeconomists. This work and the work of followers, greatly contributed to an established view on the economic relevance of reallocation as a determinant of aggregate productivity and in this way established the relevance of firm micro-structure in macroeonomics. [Hopenhayn and Rogerson, 1993] was the first paper to assess the quantitative importance of reallocation for aggregate productivity by analyzing the cost of policies that restrict labor mobility. Firing costs introduce a wedge between the marginal products of labor across firms and have thus an impact on aggregate productivity. A calibrated general equilibrium version of [Hopenhayn, 1992a] is used to provide a quantification. The methodology developed to incorporate firm heterogeneity in a general equilibrium macro model had probably a more widespread impact in macroeconomics than the reallocation analysis itself.

Firing costs in [Hopenhayn and Rogerson, 1993] introduce a wedge between the marginal products of different firms that explains productivity losses. More recently, there is a growing literature in macro/development that evaluates the impact of wedges to marginal products, taking a more agnostic view of where these wedges come from. The aim of this literature is to learn of the potential effects of firm level misallocation on aggregate productivity. [Restuccia and Rogerson, 2008] was the first paper to address this question by evaluating the impact of different distributions of wedges -modelled as implicit taxes/subsidies- on TFP. [Hsieh and Klenow, 2009] develop an empirical methodology to compute this distribution and used it to perform a decomposition of TFP differences between India, China and the US, establishing the quantitative importance of these distortions.[4],[5]

The rest of the paper is structured as follows. Section 2 describes a simplified version of a Lucas-type model and provides the key link between the aggregate production function and the firm microstructure. Section 3 generalizes this approach to an economy with firm dynamics, entry and exit. Section 4 provides a summary of empirical regularities on firm dynamics and the size distribution of firms. Section 5 considers an economy with distortions. Section 6 reviews the results on the quantitative analysis of distortions discussed above. Section 7 considers implications of firm-microstructure on the macro response to aggregate shocks. Section 8 concludes.

# 2 A simplified Lucas model

There is a collection of firms $i = 1, ...M$, with production functions

$$y_i = e_i n_i^{\eta}.$$

The only input is labor, and the total endowment in the economy is $N$. As in [Lucas Jr, 1978], each firm has decreasing returns to scale.

An optimal allocation solves

$$\begin{aligned} & \max_{n_i} \sum_i e_i n_i^{\eta} \\ \text{subject to} \quad & : \quad \sum n_i \leq N. \end{aligned}$$

The first order conditions for this problem imply that

$$\ln n_i = a + \frac{1}{1-\eta} \ln e_i \tag{1}$$

where $a$ is a constant that depends on $\eta$, $N$ and the vector of firm level

[4] This is a growing literature and there are many papers. As an example, [Guner *et al.*, 2008] and [Alfaro *et al.*, 2007].

[5] A related research has developed in international economics. A series of recent papers (see [Melitz, 2003]), [Eaton and Kortum, 2002], [Bernard *et al.*, 2003] and [Alvarez and Lucas, 2007] ) consider the effect of tariffs as barriers to the efficient allocation of resources across the world. The spirit of the exercise and overall methodology is very similar to [Hopenhayn and Rogerson, 1993].

productivities. Substituting in the production function,

$$\begin{aligned} \ln y_i &= \ln e_i + \eta \left( a + \frac{1}{1-\eta} \ln e_i \right) \\ &= \eta a + \frac{1}{1-\eta} \ln e_i \end{aligned} \tag{2}$$

is also proportional to $e_i$, implying that at the efficient allocation $y_i/n_i = y/n$ for all $i$. Finally, using the aggregate resource constraint to substitute for $a$, it follows that

$$y = \left( \sum_i e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} N^{\eta}.$$

This is an aggregate production function of the same class as the underlying firm-level production function, with TFP parameter given by $\left( \sum_i e_i^{\frac{1}{1-\eta}} \right)^{1-\eta}$. This technology exhibits decreasing returns in the aggregate, as firms here are treated as a fixed factor. This can be more clearly seen, dividing the first term by $M^{1-\eta}$

$$y = \left( E e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} M^{1-\eta} N^{\eta}. \tag{3}$$

This aggregate production function has constant returns to scale in firms and other inputs (in our example, labor), where aggregate TFP is a geometric mean of firm level productivity.

## 2.1 Endogenizing entry

Suppose that $c_e$ workers are need to create a firm with productivity that is randomly drawn from a cdf $G$, independently for all entrants. A competitive equilibrium is defined as follows. In the first stage, a large mass of identical potential entrants decide whether to enter or not. An entrant must pay the cost of entry given by $c_e$ units of labor and then draws its productivity $e$ according to a cdf $G$. Assuming there is a large number of entrants and that draws of potential entrants are independent, the distribution of realizations is approximately given by $G$. Entry decisions are driven by the expected profits of a firm $E\pi\left(w\right) = \int \pi\left(e,w\right) G\left(de\right)$, where$\pi\left(e,w\right) = max_n e n^{\eta} - wn$. In equilibrium, $E\pi\left(w\right) = c_e w$.

In this simple economy, the welfare theorems hold so equilibrium and Pareto optimal allocations -those that maximize total output-coincide. For fixed number of firms we constructed an equilibrium and optimal allocation in the previous section. The optimal choice of number of firms solves:

$$\begin{aligned} y &= \max_{M,N_e} \left( E e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} (M)^{1-\eta} (N_e)^{\eta} . \\ & \text{subject to: } c_e M + N_e \leq N \end{aligned}$$

The solution to this problem is $N_e = \eta N$, the number of firms $M = \frac{(1-\eta)N}{c_e}$ and the equilibrium wage is the multiplier of the constraint. The corresponding production function in terms of the labor endowment is given by:

$$\begin{aligned} y &= \left( E e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} \left( \frac{(1-\eta) N}{c_e} \right)^{1-\eta} (\eta N)^{\eta} \qquad (4) \\ &= \left[ \left( E e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} \left( \frac{(1-\eta)}{c_e} \right)^{1-\eta} \eta^{\eta} \right] N. \end{aligned}$$

Interestingly, it turns out that in this case the number of firms is independent of the distribution of productivity shocks.

Given the aggregate production function above, total output will be split between wages and firm profits with shares $\eta$ and $(1-\eta)$ , respectively, and the equilibrium entry condition $E\pi (w) = \int \pi (e, w) G (de) = w c_e$ is verified.

Results here generalize naturally to the case where firms have production function

$$y = e_i \left( k_i^{1-\alpha} n_i^{\alpha} \right)^{\eta}$$

just by substituting $K^{1-\alpha} N^{\alpha}$ for $N$ in the above expressions. For the endogenous entry case, this assumes that the relevant input for entry is a composite of capital and labor with weights $\alpha$ and $1-\alpha$.

## 2.2 Connection to monopolistic competition

In a monopolistically competitive economy [Dixit and Stiglitz, 1977], output $y$ is produced by aggregating a continuum of intermediate inputs $y_i$ with production function

$$y = \left( \int y_i^{\eta} di \right)^{\frac{1}{\eta}}$$

and each intermediate good is produced with a linear function of labor

$$y_i = \tilde{e}_i n_i$$

where $\tilde{e}_i$ is the productivity of intermediate producer $i$. As it is well known, in equilibrium firms choose a constant markup over marginal cost such that $p_i = \frac{1}{\eta}\left(w/\tilde{e}_i\right)$ . Letting $e_i = \tilde{e}_i^{\eta}$ it follows that $y_i^{\eta} \propto e_i^{\frac{1}{1-\eta}}$ and $n_i \propto e_i^{\frac{1}{\eta-1}}$. Comparing to the derivations in the first section, it follows that output under monopolistic competition $y^m = y^{\frac{1}{\eta}}$. Hence we can rewrite the expression for equation (3) as

$$\begin{aligned} y &= \left(Ee_i^{\frac{1}{1-\eta}}\right)^{\frac{1-\eta}{\eta}} M^{\frac{1-\eta}{\eta}} N \qquad (5) \\ &= \left(E\tilde{e}_i^{\frac{\eta}{1-\eta}}\right)^{\frac{1-\eta}{\eta}} M^{\frac{1-\eta}{\eta}} N \end{aligned}$$

These are the familiar equations (see [Melitz, 2003]) for the monopolistically competitive case. Note that for fixed $M$ aggregate $TFP$ is the same (given the transformation of the $e_i^{\prime}s$) as in the perfectly competitive case, and the only difference remains in the increasing returns to scale in $M, N$.

# 3 A sequence economy

In this section we add dynamics to the evolution of firms. Suppose that all firms productivities follow a Markov Process with transition function given by the conditional cdf $F\left(ds'; s\right)$. Again assuming that the stochastic processes faced by firms are independent, repeated application of the transition function on the distribution of entrants generates a sequence of probability measures $\tilde{\mu}_s$ for firms of age $s$. We also assume that there is a fraction $\delta$ of firms die exogenously every period.[6]At time $t$, if the sequence of entries of firms has been $\{m_0, ..., m_t\}$, there will be $M_t = \delta^t m_0 + \delta^{t-1} m_1 + ...\delta m_{t-1} + m_t$ firms producing in period $t$ with probability distribution

$$\mu_t = M_t^{-1}\left(m_t \tilde{\mu}_0 + \delta m_{t-1}^{t-1} \tilde{\mu}_1 + ... + \delta^t m_0 \tilde{\mu}_t\right). \qquad (6)$$

The aggregation results from the previous sections apply, so

$$y_t = \left(\int e^{\frac{1}{1-\eta}} d\mu_t\left(e\right)\right)^{1-\eta} M_t^{1-\eta} N^{\eta}.$$

[6] A more satisfying procedure is to have endogenous exit, as in several papers in the literature. The exogenous death rate simplifies considerably the analysis and is commonly used to obtain a well defined steady state with entry and exit of firms.

The planner's problem for this economy is:

$$\begin{aligned} max \qquad \textstyle\sum_{t=0}^{\infty} \qquad & \beta^t \left( \int e^{\frac{1}{1-\eta}} d\mu_t(e) \right)^{1-\eta} M_t^{1-\eta} N_t^{\eta} \\ \text{subject to: } \quad & M_t = \delta^t m_0 + \delta^{t-1} m_1 + ...\delta m_{t-1} + m_t \\ & N_t = N - c_e m_t \end{aligned}$$

To define a competitive equilibrium, let $v_t(e;w)$ denote the value for a firm at time $t$ for a given sequence of wages $w = \{w_s\}_{s=0}^{\infty}$. This value satisfies the following recursive equation:

$$v_t(e;w) = \max_n e n^{\eta} - w_t n + \beta\delta E v_{t+1}(e';w|e).$$

Let $v_t^e = \int v_t(e;w)\, dG(e) - w_t c_e$ denote the expected value for an entrant.

**Definition 1** *A competitive equilibrium is a sequence* $\{m_t, n_t(e), v_t\}$ *and wages* $\{w_t\}$ *that satisfy the following conditions:*

1. *Employment decisions are optimal given wages*

2. *The value functions are as defined above*

3. $v_t^e \leq 0$ *and* $m_t v_t^e = 0$

4. $m_t c_e + \int n_t(e)\, \mu_t(de) = N$

Condition 3 is the free entry condition and assumes that there is an unlimited number of ex-ante identical potential entrants. If there is positive entry, then the value of entry has to be exactly equal to zero. Equation 4 is the labor market clearing condition. The analysis can be easily extended to a growing population.

### 3.0.1 Stationary equilibrium

Firms in this model can be thought as pieces of capital with idiosyncratic productivity. In the same spirit as a steady state (or balanced growth path) in a Solow type model, we can define here a stationary equilibrium. In a stationary equilibrium all allocations are stationary and in particular the entry flow $m_t = m$ for all $t$. This implies that the total mass of firms $M = \frac{m}{1-\delta}$ and the probability distribution over types is given by:

$$\mu = (1-\delta)\sum_{s=0}^{\infty}\delta^s \tilde{\mu}_s,$$

that is, a weighted mixture of the distribution of cohorts of different ages, weighted by the corresponding survival probability. In a stationary equilibrium wages are constant and the value of firms given by:

$$v(e) = \max_n en^{\eta} - wn + \beta\delta \int v(e') F(de'|e)$$

which gives also an employment decision rule $n(e)$. The definition of a stationary equilibrium follows immediately from the general conditions of an equilibrium plus the requirement of stationarity in prices and allocations.

**Proposition 2** *There exists a unique stationary equilibrium.*

The calculation is very straightforward using the condition for an entrant $v^e(w) = 0$ that pins down a unique equilibrium wage. Total labor demand is given by:

$$mc_e + M\int n(e)\, d\mu(e) = mc_e + m\int n(e)\, d\mu(e)$$

so there is a unique value $m$ that clears the market.

**Endogenous Exit** The model given above is a simplified version of [Hopenhayn, 1992a], where exit is exogenous Assume instead that there is no exogenous death ($\delta = 0$) but firms must pay every period a fixed cost $f$ in units of labor. Hopenhayn shows that the stationary equilibrium will be characterized by an exit threshold $\underline{e}$ such that all firms with $e_{it} < \underline{e}$ exit the market. The aggregation procedure given above can be carried out with the simple modification that now $\tilde{\mu}_s$ is interpreted as the distribution of shocks for age $s$ cohort -that is conditional on survival for $s$ periods- and total mass equal to the survival rate. The exit rule described above gives a stopping time $\tau$, that corresponds to the age at exit. [Hopenhayn, 1992b] shows that a stationary equilibrium exists if and only if the expectation $E\tau < \infty$, and in that case the rate of entry/exit in the economy is simply $1/E\tau$ and thus only depends on the exit threshold $\underline{e}$. [Hopenhayn, 1992a] and [Hopenhayn, 1992b] show that the exit threshold is decreasing in the cost of entry $c_e$ and, under some mild regularity conditions increasing in the fixed cost $f$.

### 3.0.2 General equilibrium or partial equilibrium

The standing framework in macroeconomics is general equilibrium, while in Industrial Organization it is partial equilibrium. It turns out that within the model described above, there is not much difference. Indeed, consider an industry where firms produce an homogenous good according to the technology given above in a perfectly competitive setting. Suppose wage is exogenous to the industry and normalize it to one. The zero profit condition for entrants gives a price $p$ which is the inverse of the wage rate $w$ obtained in the general equilibrium version. The mass of firms is chosen to equate demand and supply. All implications for firm dynamics and aggregate productivity are exactly the same.

# 4 Empirical regularities and calibration

Over the last twenty years there has been an abundance of empirical papers documenting firm and employment dynamics. This section provides a very brief summary of findings.

1. The size distribution of firms and establishments is highly skewed. In particular, the size distribution for the US follows approximately a Pareto distribution with coefficient minus one, i.e.

$$(1 - F(n)) = An^{-1}.$$

   This is consistent with Zipf's law, by which the size of the $n^{th}$ firm in ranking is proportional to $1/n$. As shown in the Figure 1 from Rossi-Hansberg and Wright, the actual distribution has a smaller tail than the Pareto and a larger number of very small firms. A Pareto distribution of sizes can be generated as the limiting distribution of a process for firm size where firm dynamics exhibits *scale independence.* Such a process was postulated by [Gibrat, 1931] and is known as Gibrat's law. Recent empirical work (xxx) has established that, at least conditioning on survival, small firms tend to grow faster than large firms, so there is a moderate degree of mean reversion.

2. Firm size is persistent, but the variance of the innovations are quite large. Gross reallocation of employment across firms exceeds in several orders of magnitude net reallocation (Davis, Haltiwanger and Schuh xxx) The variance of growth rates declines with size and age.

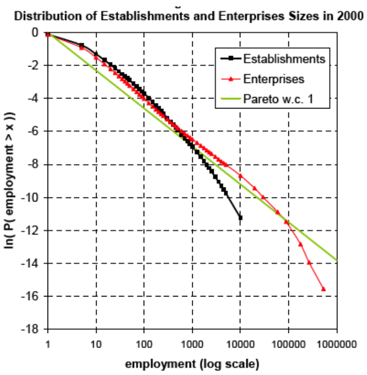

Figure 1: Pareto Size Distribution

3. Firm size increases with age. This is partly the result of selection, as small firms exhibit higher exit rates.

4. There is considerable degree of entry and exit, as documented in [Dunne *et al.*, 1988],[Dunne *e*
[Davis *et al.*, 1998],[Davis and Haltiwanger, 1992],[Evans, 1987b],[Leonard, 1988],
[Evans, 1987a],[Hall, 1987],[Geroski, 1995],[Caves, 1998]

5. Most of the firm level changes in employment respond to idiosyncratic shocks, i.e. they cannot be explained by aggregate, geographic or industry variables.

## 4.1 The model and the data

The model described above provides a simple structure to assess quantitatively the costs of different potential distortions. Starting with [Hopenhayn and Rogerson, 1993], most papers in the literature calibrate the model to the US manufacturing sector, under the assumption that there are no distortions. In static versions of the model, there is a one-to-one relationship between the size distribution of firms and the distribution of productivity shocks. To calibrate a dynamic version, [Hopenhayn and Rogerson, 1993] use data on firm employment dynamics as follows. Assume the $\ln e_{it}$ follows an AR1 process

$$\ln e_{it+1} = \bar{e}\,(1-\rho) + \rho \ln e_{it} + \varepsilon_{it+1}$$

where $\varepsilon_{it}$ is an iid process, normally distributed with mean zero and variance $\sigma^2$. Using equation (1) it follows that:

$$\ln n_{it+1} = \left(a + \frac{\bar{e}}{1-\eta}\right)(1-\rho) + \rho \ln_{it} + \frac{\varepsilon_{it+1}}{1-\eta}.$$

This implies that $\ln n_{it}$ also follows an AR1 process with the same persistence and normally distributed innovations with zero mean and variance $\sigma^2/\left(1-\eta\right)^2$. Parameters $\rho, a + \frac{\bar{e}}{1-\eta}$ and $\sigma^2/\left(1-\eta\right)^2$ can be estimated using panel data on firm employment. The parameter $\eta$ equals labor share. The distribution $G$ for the initial draw of firms can be inferred from the size distribution of entrants, up to an endogenous scale parameter $a$ that is monotonically decreasing in the equilibrium wage rate. This leaves three more parameters to calibrate, namely the cost of entry $c_e$, fixed cost $f$ and $\bar{e}$. There is clearly an identification problem between $a$ and $\bar{e}$ as they enter additively in the determination of average employment. Hence there is an extra degree of freedom and an arbitrary value of $a$ can be chosen, without any meaningful implications.

## 4.2 Implications for macroeconomics

As our aggregation exercise shows, firm level productivities are a key determinant of aggregate TFP. The data suggests that there is considerable variation in firm level productivities/demand and a high degree of resource flows across firms. Misallocation of resources could thus be a potentially important explanation of the wide differences in TFP across countries. Recent papers have tried to establish the relevance of this source of variation, suggesting it can be quite large. In particular, barriers to the reallocation of resources can also be a potentially important source of differences in TFP. Sections6 will discuss recent papers in this area.

The heterogeneity in firm growth rates can also have implications for transitional dynamics. In particular, the process of growth of firms with age leads to "time to build" in the aggregate and can have implications for the evolution of investment and productivity after large shocks hit the economy. This is discussed in Section 7.

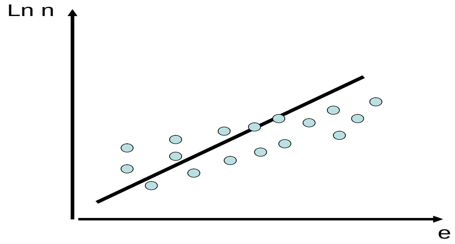

Figure 2: Wedges in marginal product

# 5 The distorted economy

This section analyzes the consequences of deviations from the optimal allocation of resources across productive units. Figure 2 provides a useful picture of the type of distortions that might occur:The solid line shows an optimal allocation, where $\ln n_i$ is a linear function of $e_i$. The dotted lines illustrate two types of distortions:

1. $n_i$ not equal for all firms with the same $e_i$, termed *uncorrelated distortions;*
2. average $\ln n_i(e) \neq a + \frac{1}{1-\eta}e$, termed *correlated distortions,* in the case of Figure 2 it is a distortion that results in reallocation of labor from more to less productive firms.

Both of these distortions result in losses of productivity as marginal product (or the marginal value of labor) is not equated across productive units. As an accounting device, it is useful to model these distortions as firm-specific implicit taxes/subsidies that create a wedge between its revenues and output:

$$\begin{aligned} r_i &= (1-\tau_i)\, y_i = (1-\tau_i)\, e_i n_i^\eta \\ &= \alpha \left(e_i (1-\tau_i)\right)^{\frac{1}{1-\eta}}, \end{aligned}$$

where $\alpha$ is a constant that depends only on the equilibrium wage. Equilibrium in this economy will be identical in terms of allocations to the equilibrium of an undistorted economy where the distribution of firm productivities is changed to $e_i(1-\tau_i)$. Total revenues are given by

$$r = N^\eta M^{1-\eta} \left( E\left[e_i (1-\tau_i)\right]^{\frac{\eta}{1-\eta}} \right)^{1-\eta} \tag{7}$$

and total output

$$y = \int y_i di = \int r_i (1 - \tau_i)^{-1} di$$

$$\frac{y}{r} = \frac{\int r_i (1 - \tau_i)^{-1} di}{\int r_i} = \frac{E (1 - \tau_i)^{-1} (e_i (1 - \tau_i))^{\frac{1}{1-\eta}}}{E (e_i (1 - \tau_i))^{\frac{1}{1-\eta}}}. \tag{8}$$

Using equations (7) and (8), it follows that

$$y = N^{\eta} M^{1-\eta} \frac{E (1 - \tau_i)^{-1} (e_i (1 - \tau_i))^{\frac{1}{1-\eta}}}{\left( E (e_i (1 - \tau_i))^{\frac{1}{1-\eta}} \right)^{\eta}}.$$

Interestingly, note that in the static case, if a planner were to choose optimally the entry of firms subject to the existing employment decisions, the choice would be independent of the distortions.[7] Indeed, it can also be shown that the equilibrium entry decisions would give rise to exactly the same number of firms. These results are not true in general for the sequence economy. Suppose the joint distribution of implicit taxes and efficiency for firms of age $s$ is $\tilde{\mu}_s (de, d\tau)$ and let

$$\tilde{r}_s = \int e^{\frac{1}{1-\eta}} (1 - \tau)^{\frac{1}{1-\eta}} d\tilde{\mu}_s (e, \tau) \tag{9}$$

be the average revenue-productivity. Hopenhayn (2010) shows that in a stationary equilibrium for the distorted economy, the equilibrium number of firms is determined by:

$$N - c_e m = \frac{\eta c_e}{(1 - \eta)} \frac{\sum_{s=0}^{\infty} \delta^s \tilde{r}_s}{\sum_{s=0}^{\infty} \beta^s \delta^s \tilde{r}_s}. \tag{10}$$

**Proposition 3** *(Hopenhayn, 2010) If $\beta = 1$, the equilibrium number of firms and the optimal number of firms are independent of distortions.*

The first part of this proposition follows directly from equation (10); the second is derived in Hopenhayn (2010). Equation (10) also suggests what are

[7] This would not be true if the cost of entry for creating new firms was not entirely denominated in units of labor, but also in terms of the homogenous consumption good. More generally, in an economy with intermediate goods the cost of entry needs to be produced with the same inputs as all intermediate goods.

the key elements in determining the effects of distortions on entry. Letting $m_0$ denote the number of firms in an undistorted economy and $r_s$ is defined by the right hand side of equation 9 when all $\tau's$ are zero.

$$\begin{aligned} \frac{N - c_e m}{N - c_e m_0} &= \left( \frac{\sum_{s=0}^{\infty} \delta^s \tilde{r}_s}{\sum_{s=0}^{\infty} \beta^s \delta^s \tilde{r}_s} \right) \Big/ \frac{\sum_{s=0}^{\infty} \delta^s r_s}{\sum_{s=0}^{\infty} \beta^s \delta^s r_s} \\ &= \frac{\sum_{s=0}^{\infty} \beta^s \delta^s r_s}{\sum_{s=0}^{\infty} \delta^s r_s} \Big/ \left( \frac{\sum_{s=0}^{\infty} \beta^s \delta^s \tilde{r}_s}{\sum_{s=0}^{\infty} \delta^s \tilde{r}_s} \right) \end{aligned}$$

**Proposition 4** *([?] and [?]) If distortions are age neutral, i.e $\tilde{r}_s / r_s$ is independent of $s$, then the equilibrium and optimal number of firms are independent of distortions.*

Again, the first part of the Proposition follows immediately from the above equation. Aside from these extreme cases, the net effect of distortions on entry depends on the specific age patterns. A sufficient condition for distortions to lead to more (less) entry is given below.

**Definition 5** *The sequence $\{\delta^s r_s\}$ dominates (is dominated by) the sequence $\{\delta^s \tilde{r}_s\}$ if and only if*

$$\frac{\sum_{s=0}^{t} \delta^s r_s}{\sum_{s=0}^{\infty} \delta^s r_s} \leq (\geq) \frac{\sum_{s=0}^{t} \delta^s \tilde{r}_s}{\sum_{s=0}^{\infty} \delta^s \tilde{r}_s}$$

*for all $t$.*

**Proposition 6** *([?]) Suppose $\{\delta^s r_s\}$ dominates (is dominated by) $\{\delta^s \tilde{r}_s\}$ Then $m \geq (\leq) m_0$.*

A sufficient condition for the sequence $\{\delta^s r_s\}$ to dominate (be dominated by) the sequence $\{\delta^s \tilde{r}_s\}$ is that $\frac{\sum_{s=0}^{t} \delta^s r_s}{\sum_{s=0}^{t} \delta^s \tilde{r}_s}$ be increasing (decreasing) in $t$. The following Corollary gives simpler necessary conditions.

**Corollary 7** *Suppose that*

$$\frac{\sum_{s=0}^{t} \delta^s r_s}{\sum_{s=0}^{t} \delta^s \tilde{r}_s} \leq (\geq) \frac{r_{t+1}}{\tilde{r}_{t+1}}.$$

*Then $m \geq (\leq) m_0$.*

Finally, an even stronger sufficient condition is that $r_s/\tilde{r}_s \leq (\geq)\ r_{s+1}/\tilde{r}_{s+1}$, which says that distortions are *relatively larger (smaller)* for older cohorts. An economy that benefits relatively established firms (e.g. as a consequence of their increased lobby power) will have less firms in equilibrium. Alternatively, an economy that tends to subsidize smaller firms will have the opposite effect, taking into account the well established fact that firm size increases with age.[8]

# 6 Quantitative analysis of distortions

The above framework has been used to asess the quantitative impact of distortions. The first paper to do that was [Hopenhayn and Rogerson, 1993], that studies the effect of layoff costs. A baseline model is calibrated using US firm dynamics data as indicated in section 4.1. Layoff costs are introduced to this benchmark as follows. Assume that there is a cost $f$ of firing a worker. The value of a firm with $n_0$ workers and productivity shock $e$ is given by the following Bellman equation:

$$v\left(e, n_0\right) = \max_n e n^\eta - wn - f \max(n_0 - n, 0) + \beta E\left[v\left(e', n\right)|e\right].$$

The solution to this problem is an $s - S$ type policy that can be easily characterized by two increasing functions $n_L\left(e\right) \leq n_H\left(e\right)$ with the following interpretation. Suppose a firm enters a period with employment $n_0$. I chooses employment $n$ in the current period according to the following rule:

1. If $n_0 < n_L\left(e\right)$, then $n = n_L\left(e\right)$
2. If $n_0 > n_H\left(e\right)$, then $n = n_H\left(e\right)$
3. If $n_L\left(e\right) \leq n_0 \leq n_H$

Figure 3 is an example of an $sS$ policy. The lower line corresponds to $n_L\left(e\right)$, the upper line to $n_H\left(e\right)$ and the middle line to the optimal employment with no distortions, that is always in between the other two. Firing costs put a wedge to the downward adjustment of employment, that explains

[8] [?] has a similar characterization for the equilibrium number of firms and provides a very interesting quantitative work showing the relevance of the entry and exit margins in evaluating the costs of distortions.

the partial nature of assumptions. The wedge on the upward adjustment to $n_L(e)$ is explained by the anticipated expected firing taxes and the resulting option value of delaying adjustment.

### 6.0.1 Firing costs and implicit wedges

The history of productivity shocks of a firm, through repeated application of this employment policy, determines the current employment of the firm. A stationary equilibrium implies a joint distribution of productivity/employment levels and consequently an aggregate level of TFP as discussed in the previous sections. Firms with employment below the undistorted level -the one given by the middle line in Figure 3- have a positive implicit $\tau$, while those with employment above the unconstrained level have negative $\tau$.

In order to get better insight on the nature of distortions that are generated from layoff costs, consider the following hypothetical example. Let there be three levels of productivity $e_1 < e_2 < e_3$ and suppose the corresponding employment thresholds are $n_L = \{5, 8, 14\}, n_H = \{9, 12, 20\}$ and the unconstrained employment levels $n^* = \{7, 10, 18\}$. Suppose the Markov process governing firms' productivities has the property that any level of productivity can be reached after some time with positive probability from any other level of productivity. This process generates a long run distribution with the following support: $\{e_{1,}9\}, \{e_2, 9\}, \{e_2, 12\}, \{e_3, 14\}$.

The explanation is quite simple: 1) eventually state $e_3$ is reached and employment increases to 14. Once it reaches this level, it can only approach state $e_1$ from above $n_H$ and so only $n = 9$ will be observed for firms with productivity $e_1$; 2) since state $e_3$ is approached only from below, employment $n = 14$ is the only compatible level in the long run for $e_3$; 3) state $e_2$ can be reached from state $e_1$ or $e_3$. In the first case, employment will be 9 and in the second case 12.

This implies that employment will be above the optimal level for $e_1$ and below for $e_2$, as if firms in $e_1$ faced a subsidy and those in $e_3$ a tax, hence a positive correlation of wedges. On the other hand, there will be firms with two implicit tax levels in state $e_2$, corresponding to a variance of wedges.

In order to get a quantitative order for these wedges consider the following

Table 1: Firing costs and tax gaps

| $e$ | $N_L$ | $N$ | $N_H$ | $N_H/N_L$ | tax gap |
|---|---|---|---|---|---|
| 1.5 | 4.5 | 6 | 20.7 | 4.6 | 25% |
| 2.2 | 29 | 59 | 170 | 5.9 | 27% |
| 3 | 243 | 483 | 1257 | 5.2 | 25% |
| 4 | 1677 | 3607 | 4454 | 2.7 | 15% |

variant of [Hopenhayn and Rogerson, 1993], with no entry and exit. Let

$$\begin{aligned} y_i &= e_i n_i^{0.85} \\ \ln e_{it} &= \bar{e}\,(1-\rho) + \rho \ln e_{it-1} + \varepsilon_{it} \end{aligned}$$

where $e_{it}$ is a lognormally distributed iid shock across firms and time with mean zero and variance $\sigma^2$,where $\rho = 0.92$ consistent with firm level employment in the US and a time unit of 5 years, $\sigma^2$ is chosen to match the standard deviation of $e_{it}$ in the US (from [Hsieh and Klenow, 2009]) , $\bar{e}$ generates an average size of firms with 50 workers in the undistorted economy similar to the US level. Take a firing cost $f$ equal to two years of wages, which is not an unreasonable mean value for countries with such type of regulations. Figure 3 shows precisely th $sS$ policy corresponding to the equilibrium of this economy. The solid lower and upper lines are the $N_L$ and $N_H$ boundaries. The middle line corresponds to the zero firing cost optimal employment level. The dashed lines show the restriction of the boundary decision rules to the support of the stationary distribution of productivity/employment states. As can be readily seen, there is a wide range of employment values between the two boundaries. Table 1 illustrates the range for a few values. The ratio of $N_H/N_L$ is extremely high, reaching almost a six-fold value. This employment range can be rationalized by taxes and subsidies to employment as described above. The tax gap, reflecting the difference between implicitly subsidy and tax rates for a given level of productivity, ranges from 15% to 27%.

These gaps might seem substantial, but what are their implications for aggregate TFP? Table xxx provides an answer to this question: the level of firing costs that we have considered (f=2 years) results in a 2.8% reduction in TFP. If $f = 5$ years, the TFP loss is 7.5% and if $f = 25$ years, it is 24.3%. This level of firing cost is close to prohibitive and firms respond by

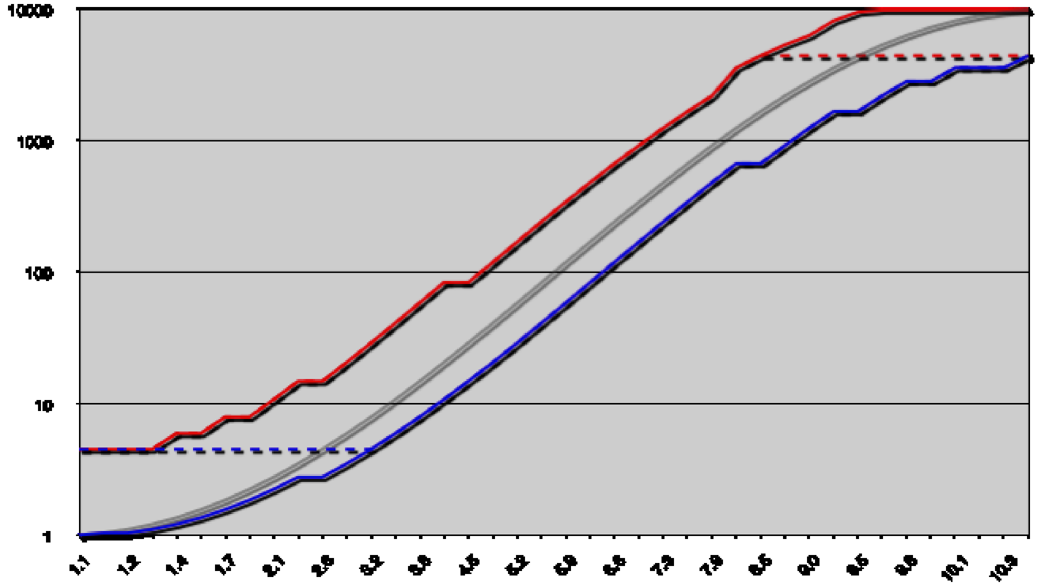

Figure 3: sS policy

not adjusting employment at all. With such degree of firing costs, it seems very likely that firms would negotiate and bargain with workers and induce quits in order to get around these barriers to adjustment. Hence a moderate range of firing costs seems a more plausible realistic scenario for employment rigidity and consequent TFP losses.

| Firing cost (years) | TFP loss | Gap (range of equivalent labor taxes) |
|---|---|---|
| 2 years | -2.8% | 32.9 |
| 5 years | -7.5% | 56.0 |
| 25 years | -24.3% | 97.6 |

Firing costs are an example of the very subtle connection that might appear between policies and implicit wedges. We discussed above the distinction between covariance and variance wedges. Figures 4 and 5 provide such kind of decomposition for the case of firing costs. More precisely, Figure 4 gives the average implicit tax rate for each level of productivity. The positive slope indicates positive covariance which increases substantially with firing costs. To provide some intuition for this result, consider the case where firing costs are infinite. Firms would then target a fixed employment level to maximize long run expected profits targeting the employment level for an average productivity shock (this would be exact without any discounting). As a consequence, there will be excessive employment for low productivity shocks and too little employment for high ones.

Figure 5 shows the variance of wedges for each level of productivity. The variance is larger for intermediate levels of productivity. This can be understood observing Figure 3. While the width of the bands does not seem to change much with productivity, the boundaries for the support of the long run joint distribution of productivity/employment -given by the dotted lines- get narrower for low and high shocks. This is partly due to the bounded support for productivity shocks used in the calculations, but also reflects the effect of mean reversion (recall that the AR1 coefficient used is 0.92.) The strength of this variance effect is also non monotonic in firing costs. This is also intuitive: for zero firing costs there is no variance since the optimal level of employment is targeted for each level of productivity; for very large firing costs employment hardly changes and is set at a constant level consistent with average long run productivity.

To get a quantitative idea of the role of variance and covariance in explaining the gap in TFP, we calculate the percentage gap closed if variance

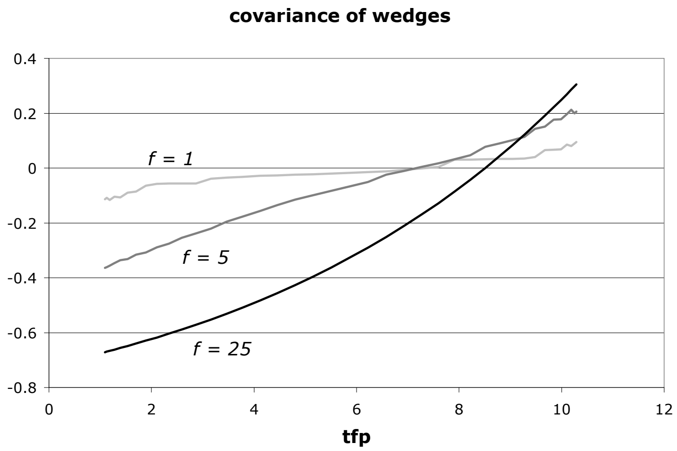

Figure 4: Covariance of wedges

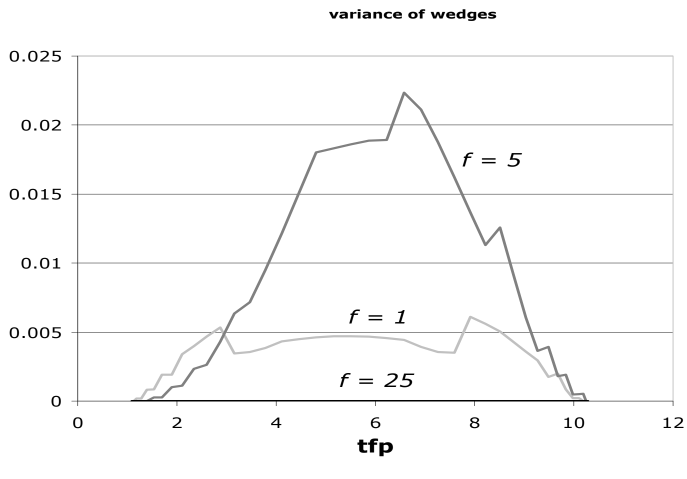

Figure 5: Variance of wedges

were set arbitrarily to zero by putting employment at the average level for each shock. The relative importance of variance shocks decreases with firing costs: 64% of the gap is closed when $f = 1$, 41% when $f = 5$ and zero when $f = 25$.

The model calibrated in this section has no entry and exit. This might affect somewhat the results. Though most of the employment adjustment takes place for incumbent firms, young firms are the ones that exhibit highest variance of innovations. Including entry in the model would change firm demographics, generating in steady state a cross section of different aged firms as we have described above. In addition, firms start small and tend to grow over time. This should imply that for younger vintages employment levels near the lower boundaries are more likely. Moreover, if we were to include higher variance of growth rates for younger firms, as occurs in the real world, the width of the bands would widen for youg firms, generating higher implicit wedges. This would appear as taxation to employment, as younger firms concentrate in the lower part of the sS band.

### 6.0.2 The potential effect of inter-firm distortions

In contrast to [Hopenhayn and Rogerson, 1993] where wedges arise as the result of a specific policy, the recent literature takes a more *agnostic* view and directly examines the effects on productivity of different distributions of wedges. The first paper to do is [Restuccia and Rogerson, 2008]. Expanding the above model to include capital, distortions occur both as departures in the output of a firm from its optimal value and the use of an inefficient capital/labor ratio.

**Restuccia-Rogerson** There is a fixed set (measure) of firms with distribution of productivities calibrated to match US size distribution, taken as a benchmark for an undistorted economy. In the distorted economy, profits for firm $i$ are given by: $\pi_i = (1 - \tau_i)\, y_i - wn_i - (1 + \tau_{ki})\, rk_i$, where $\tau_i$ and $\tau_{ki}$ are two wedges wedges are distributed according to a conditional density $\omega\left(\tau, \tau_k | e\right)$.[9] The exercise performed consists in taxing a subset of firms and subsidizing the remaining set so that steady state capital remains invariant. We consider here quantitative results for level taxes only. Table 2 gives a

[9] A more transparent interpretation of wedges would be obtained using $\pi_i = (1 - \tau_i)\, y_i - w\,(1 + \tau_{ni})\, n_i - r\,(1 + \tau_{ki})\, k_i$ and setting $(1 + \tau_{ki})^{\beta}\,(1 + \tau_{ni})^{\alpha} = 1$. This would make the tax on capital a pure input mix distortion.

Table 2: Relative TFP and Distortions

| % Estab. taxed | Uncorrelated | | Correlated | |
|---|---|---|---|---|
| | $\tau_t$ | | $\tau_t$ | |
| | 0.2 | 0.4 | 0.2 | 0.4 |
| 90 | 0.84 | 0.74 | 0.66 | 0.51 |
| 50 | 0.96 | 0.92 | 0.80 | 0.69 |
| 10 | 0.99 | 0.99 | 0.92 | 0.86 |

summary of the effects of distortions. The first two columns consider uncorrelated wedges and the remaining the correlated ones. Two observations follow: 1) correlated distortions have significantly stronger effects; 2) the impact on productivity is larger the higher is the fraction taxed. The latter can be explained by the fact that when a group of firms is taxed, the remaining firms are subsidized at a rate that makes total capital stock invariant. The larger the group taxed, the higher the subsidy has to be for the other group, thus creating a very large disparity in wedges. There is also an intuition for why correlated distortions have a bigger impact on TFP. In the extreme case where firms have near constant returns to scale, there is no effect of variance distortions as output can be totally reallocated within productivity groups. The same is not true with correlated distortions, as output is reallocated to firms with lower productivity. In particular, when the 90% most efficient firms are taxed the lowest 10% receive a 107% subsidy when $\tau = 0.2$ and 114% when $\tau = 0.4$.

The analysis done by Restuccia and Rogerson suggests that large interfirm distortions can have substantial aggregate TFP effects. But the distortions assumed are hypothetical. In contrast, [Hsieh and Klenow, 2009] derive these distortions from firm level data.

**Hsieh-Klenow** The setting they consider is one of monopolistic competition. As mentioned in section 2.2, this is not a major difference as for fixed number of firms the model is isomorphic to the Lucas' style model. The major novelty in their approach is the exercise of recovering wedges from firm level data. This can be explained easily in terms of the model developed in section 2. Measured TFP for each firm $i$ is simply $e_i = y_i/n_i^\eta$.This requires measuring inputs and outputs (or sales).[10]. Leaving aside measurement er-

[10] In case of [Hsieh and Klenow, 2009] capital is also included.

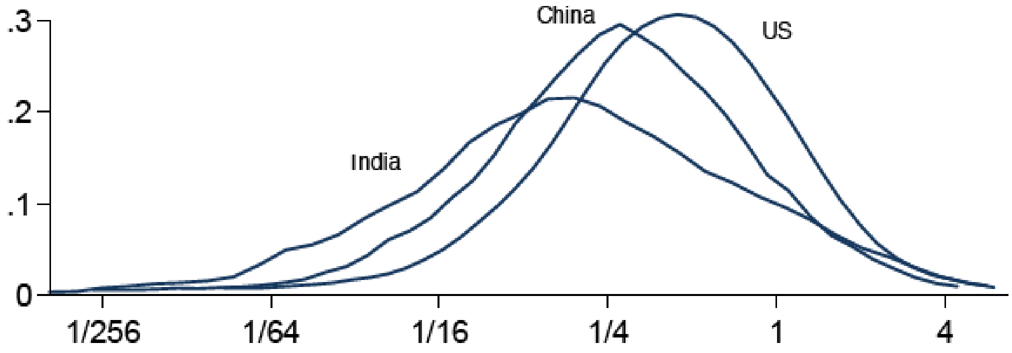

Figure 6: Distributions of Firm TFP

ror, one can calculate aggregate TFP in an efficient allocation using equation (3) as follows:

$$TFP_e = \left( E e_i^{\frac{1}{1-\eta}} \right)^{1-\eta} = \left( E \left( \frac{y_i}{n_i^{\eta}} \right)^{\frac{1}{1-\eta}} \right)^{1-\eta}.$$

In contrast, actual TFP is given by

$$TFP = y/N^{\eta} = \frac{y}{\left(\sum_i n_i\right)^{\eta}}$$

.

Hsieh and Klenow use firm level data from US, China and India to perform these calculations. Figure 6 gives the empirical distribution of firm level productivities $e_i$ as calculated by Hsieh and Klenow. The distribution for the US stochastically dominates that of China, which in turn stochastically dominates that of India -that also exhibits the highest levels of dispersion- except for very high levels of productivity.

As seen in Section 2, in an efficient allocation $y_i/n_i$ is equated across firms. The analogue for the output/labor ratio in Hsieh-Klenow is what in their paper is called TFPR (total factor productivity revenue.) The dispersion in

Table 3: Dispersion of LN TFPR

| | US (97) | China (98) | India (94) |
|---|---|---|---|
| SD | 0.49 | 0.74 | 0.67 |
| 75-25 | 0.53 | 0.97 | 0.81 |
| 90-10 | 1.19 | 1.87 | 1.60 |

ln TFPR across firms (absent measurement error) is an aggregate measure of the level of inter-firm distortions. Table 3 gives measures of dispersion of TFPR. All measures of dispersion, standard deviation, 75-25 percentiles and 90-10 percentiles show considerable higher dispersion for China and India than the US, indicating distortions are more important in these two countries.

In order to get a better idea of the nature of these distortions, we calculate the dispersion of implicit tax/subsidy rates would they imply. Assuming Cobb-Douglass production function with labor share $1-\alpha$ (as in [Hsieh and Klenow, 2009]) and considering only level taxes,

$$\ln TFPR_i - \ln TFPR_j = \ln\left(1-\tau_i\right)^{-1} - \ln\left(1-\tau_j\right)^{-1}$$

Table 4 converts the numbers in Table 3 into equivalent ratios of output taxes. For example, the value 2.6 for China represents the ratio $\left(1-\tau_{25}\right)/\left(1-\tau_{75}\right)$. This can be interpreted as follows: assuming the decile 25 corresponded to no taxes, decile 25 would have a subsidy of 160%, or assuming decile 25 had no subsidy, decile 75 would have a 61.5% tax on output. The comparison values for decile 90-10 are subsidies of 550% for decile 10 or taxes of approximately 85% for decile 90. These numbers are large, but they disregard measurement error, which could be quite large for several reasons. To get a relative expression, the table also gives relative ratios of taxes for China/US and India/US. For instance, for the 75-25 decile, China's distortion is 55% higher than in the US. If we considered US as the undistorted benchmark, the distortion for China would be equivalent to subsidizing 55% firms in percentile 25 (of the distribution of TFPR) if no tax were levied on percentile 75.

How much can these distortions explain of the current differences in TFP? To answer that question, Hsieh and Klenow perform the following counterfactual experiment. Take the case of China and the US, where the current gap is equal to $TFP_{china}/TFP_{us}$. Now suppose both countries eliminated all distortions, equating TFPR's across firms. The new ratio $TFP_{China}^{efficient}/TFP_{us}^{efficient}$

Table 4: Gap in Taxes

| | US (97) | China (98) | India (94) | China/US | India/US |
|---|---|---|---|---|---|
| 75-25 | 1.7 | 2.6 | 2.2 | 55% | 32% |
| 90-10 | 3.3 | 6.5 | 5 | 97% | 51% |

would still be less than one, since as we have seen in Figure 6 the distribution of firm level TFP in the US stochastically dominates that of China. A measure of how much closer is the TFP is to calculate what percentage of the gap was closed:

$$\frac{TFP_{China}^{efficient}/TFP_{us}^{efficient} - TFP_{china}/TFP_{us}}{TFP_{china}/TFP_{us}}$$

The same calculation was done for India. Values for China range from 30 to 50% and for India from 40 to 60%, depending on the years considered.

### 6.1 Distortions and the distribution of Productivities

The decomposition in [Hsieh and Klenow, 2009] separates the role of differences in the distribution of firm level productivities and interfirm misallocations as two separate sources of aggregate TFP differences. This might be a useful taxonomy, but misses the fact that these two sources are likely to be interrelated. Recent papers in Economic Development have emphasized this connection. [**?**], solve for a steadty state joint distribution of entrepreneurial talent and wealth in an economy where entrpreneurs must self-finance their investments. This joint distribution implies, both, a distribution of productivities and a distribution of wedges. Policies or institutions that relax the borrowing constraints should imply smalle wedges and a stronger selection effect by which low productivity entrepreneurs are driven out of production. [**?**] and [**?**] examine the link between borrowing constraints and technology selection. As borrowing constraints are relaxed, firms can adopt more productive technologies that to be profitable require a considerably larger scale and investment.

## 7 Firm demographics and TFP

In this section we explore a different application of firm microstructure to macroeconomics: the importance of firm heterogeneity and growth for aggregate adjustment dynamics. The mechanism described here is one of propagation The link that we explore in this section, is that the distribution of firm productivities depends on the age distribution of firms -what we called firm demographics- and is thus affected by the path of firm creation. Letting $\delta(a)$ denote the fraction of firms of age $a$, we can rewrite total productivity as a weighted average of age-cohort tfp's:

$$\begin{aligned} A &= \left(\sum_{a=0}^{\infty} \delta(a) E_a e_i^{\frac{1}{1-\eta}}\right)^{1-\eta} \qquad (11) \\ &= \left(\sum_{a=0}^{\infty} \delta(a) A_a^{\frac{1}{1-\eta}}\right)^{1-\eta} \\ A_a &= \left(E_a e_i^{\frac{1}{1-\eta}}\right)^{1-\eta}. \end{aligned}$$

How does TFP evolve as a cohort ages? In our model, there is a one-one mapping between the distribution of productivities and firm sizes and it is a well established fact that the size distribution of firms increases stochastically with the age of a cohort. This implies, through the lens of this model, a stochastic increase in productivities and thus a rise over time in the average productivity of a given cohort of entrants. There are two forces that suggest this is to be expected: learning by doing and selection. The first has been emphasized considerably in the IO literature. The idea of selection and its role for increasing firm size is also well established in the literature after Jovanovic's classic paper.

The average size of entrants is about 20% the average size of incumbent firms. Since $e_i/e_j = (n_i/n_j)^{-\alpha}$, using a value of $\alpha = 1/3$ this is consistent with a ratio in average productivity close to two. This does not take into account the increase in the variance of firm sizes with age, which applying Jensen's inequality to the term in brackets in the last line of equation (11) gives a further contribution to the average productivity of a cohort. These observations imply that, a shock to entry will lead to a decrease in productivity and a further rise.[11] In what follows, we examine two applications of this idea.

[11] This assumes that firm entry costs are properly accounted as part of the capital stock. Otherwise, the rise in the number of firms has the opposite effect.

Table 5: Life Cycle of an Industry

| **Stage** | **Mean duration** | **Mean annual net entry rate** | **Mean annual growth rate** |
|---|---|---|---|
| II | 9.7 years | 24.8% | 35% |
| III | 7.5 years | 0.2% | 12% |
| IV | 5.4 years | -9% | 8% |
| V | - | -0.5% | 1% |

## 7.1 Shakeout of firms

This section describes work in [Hopenhayn, 1993]. It is a fairly well established fact that as a new industry progresses through its life cycle, it experiences an increase and then a big fall in the number of firms as it approaches maturity. ([Gort and Klepper, 1982] document this shakeout to be in the order of 40% and describe 5 stages in the life cycle of an industry.) Table 5 documents some of the key factors extracted from the dataset used by Gort and Klepper. In the first stage, there are very few firms in the industry and output is very low. The second stage exhibits an exponential growth in the number of firms and a very high growth in total output, which slows down over the remaining stages. The third stage has little net entry and is where the peak number of firms is reached. The fourth stage is the shakeout, with a severe drop in the number of firms that can add to more than 50%. Note that the shakeout is not explained by a decay in the industry as total output continues to grow at a rate of 8% per year.

[Hopenhayn, 1993] provides the following explanation. For whatever reason, either because of a slowdown in demand growth or in technological change, Table 5 shows a slowdon in the rate of growth of the industry output over time. This leads to a slowdown in the entry of firms. As this happens and the large cohorts of firms that entered earlier become older, there is a change in the demographics of the industry, increasing the share of older firms. This results in a rise in average productivity and the size of firms, and thus a smaller number of firms are needed to clear the product market. [Hopenhayn, 1993] shows that this theory not only fits the facts qualitatively, but it also does a good job on the quantitative side.

To get an idea of the quantitative power of this effect, take a calibrated model as the one described in the paper of [Hopenhayn and Rogerson, 1993] . Consider demand growth as the source of expansion in the industry. Suppose

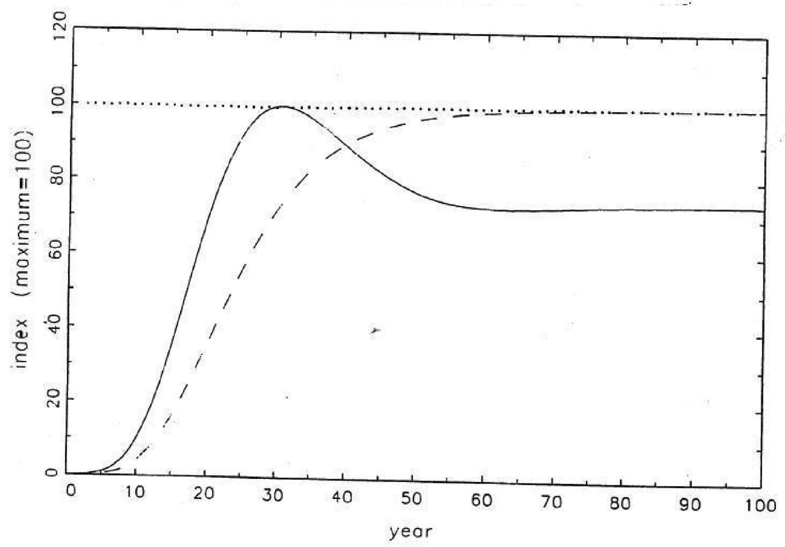

Figure 7: Shakeout of Firms

that demand grows at a decreasing rate following a Gompertz process, which is commonly used to model diffusion, and parametrize the model so it takes 40 years for the industry to reach 90% of its limiting total demand. Figure 5 depicts the evolution of total demand (and output) as the dashed line and the number of firms as the solid line. It can be seen that the number of firms peaks around period 30, then there is is a shakeout and the number of firms drops by 36% [Hopenhayn, 1993] uses data on price decreases to calibrate a process of cost reducing technological change which is taken as the driving force for industry expansion. The slowdown in entry is triggered by a slowdown in the rate of technological change coupled with a decrease in demand elasticity as the market expands.

## 7.2 Firm demographics and the productivity paradox

This section considers an application of the same idea to explain a productivity paradox during the second industrial revolution (1860-1900) studied by [Atkeson and Kehoe, 2007]. There was a 50 year lag between an increase

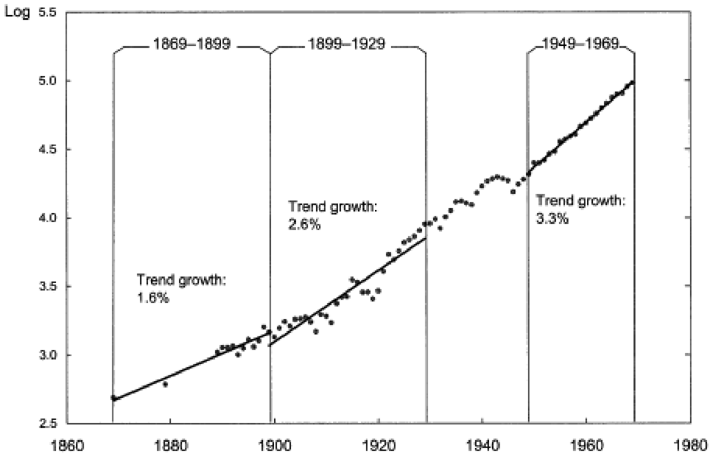

Figure 8: Productivity Paradox

in the growth rate of technological change that occurred starting 1860 and the subsequent growth rate of productivity. This was a period where many new technologies were introduced following the invention of electricity. Figure 8 from [Atkeson and Kehoe, 2007] gives account of the slow increase in productivity growth.

The authors link the productivity paradox with the slow diffusion of new technologies that had been documented in [Devine Jr, 1983]. This is illustrated in Figure 9, reproduced from their paper. Atkeson and Kehoe provide a model similar to the one described above to explain this slow diffusion process with the following elements: 1) it takes time for new firms to learn a new technology and 2) new technologies are embodied in new firms. From a modeling perspective, this is tantamount to assuming that the distribution of a cohort's productivity stochastically increases through time (as we did in Section 7.1) and that the technological frontier that is embodied in new firms grows at a constant rate. In the long run, the model delivers a stationary equilibrium along a balanced growth path that is very similar to the one described in Section 3. In what follows, we describe a similar model from Hopenhayn (1989, chapter 3).

The production function of a cohort of time $t$ entrants is given by $\gamma^t e_i n_i^\eta$, where $\gamma - 1 \geq 0$ is the rate of technological progress. Productivity shocks

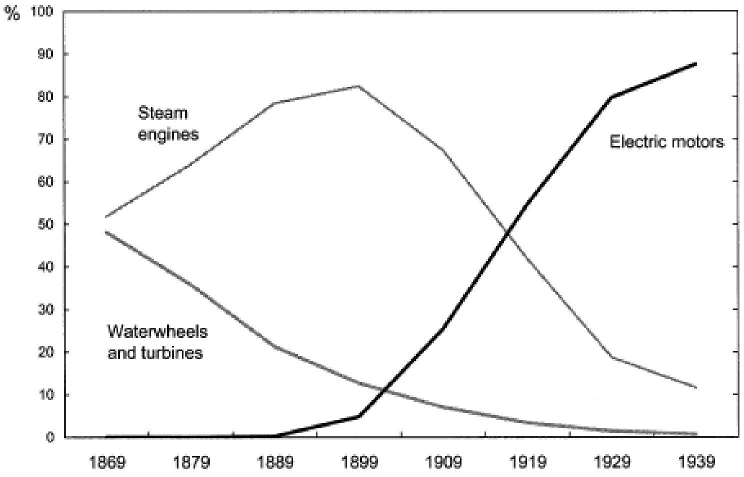

Figure 9: Diffusion of electric motors

$e_i$ at time of entry are drawn from a distribution with c.d.f. $G(e)$ and then follow a Markov process with conditional distribution $F\left(e'|e\right)$ as described in Section 3. In addition, firms face a fixed cost $f$ (in terms of labor) that is used to generate endogenous exit as in [Hopenhayn, 1992a]. In a balanced growth path, the wage rises at the rate of technological progress, so $w_t = \gamma^t w_0$. The value of a firm of age $\tau$ is given by:

$$v_\tau\left(e, \gamma^t w_0\right) = \max\left(0, \pi_\tau\left(e, \gamma^t w_0\right) + \beta E v_\tau\left(e', \gamma^{t+1} w_0 | e\right)\right)$$

where

$$\pi_\tau\left(e, w\right) = \max_n \gamma^\tau e n^\eta - wn - wf.$$

It is immediate that $\pi_\tau$ is homogeneous of degree one in $e, w$ and easy to show that so is $v_\tau$. So dividing through by $\gamma^t$

$$v_\tau\left(e, \gamma^t w_0\right) = \gamma^t v_0\left(\gamma^{-(t-\tau)} e, w_0\right) = \gamma^t \max\left(0, \pi_0\left(\gamma^{-(t-\tau)} e, w_0\right) + \beta \gamma^t E v_\tau\left(\gamma^{-(t-\tau)} e', \gamma w_0 | e\right)\right)$$

and thus

$$v_0\left(\gamma^{-(t-\tau)} e, w_0\right) = \max\left(0, \pi_0\left(\gamma^{-(t-\tau)} e, w_0\right) + \beta\gamma E v_\tau\left(\gamma^{-(t-\tau)} e'/\gamma, w_0 | e\right)\right).$$

Letting the state of a firm $\tilde{e} = \gamma^{-(t-\tau)} e$, this value function can be rewritten as:

$$v_0\left(\tilde{e}, w_0\right) = \max\left(0, \pi_0\left(\tilde{e}, w_0\right) + \beta\gamma E v_\tau\left(\tilde{e}/\gamma, w_0 | e\right)\right).$$

This is exactly the same as the value function for a firm in a stationary economy with no technological change with an appropriate shift in the conditional distribution of productivity $\tilde{F}\left(\tilde{e}'|\tilde{e}\right) = F(\gamma\tilde{e}'|\tilde{e})$. The equilibrium is solved exactly as in the model with no growth.

This shift introduces a *depreciating* element to a firm's productivity as it ages and falls further behind the technological frontier. There are now two forces working in opposite directions: the original property of the Markov process by which the distribution of a cohort's productivity increases with the age of  the cohort (what [Atkeson and Kehoe, 2007] call *learning*) and the counteracting force introduced by technological depreciation. These two forces determine the sequence of age-cohort productivities $\{A_a\}_{a=0}^{\infty}$ and in particular imply that $A_{a+1}/A_a < \gamma$. This is important for transitional dynamics. Start the economy in the steady state corresponding to a low $\gamma_0$ and consider a permanent increase increase to $\gamma > \gamma_0$. There will be a jump in the entry rate that will shift the demographics to lower ages. Though these new entrants embody the new technology, the average productivity of the cohort will does not fully reflect the higher productivity of the new technology, so average productivity will grow less than $\gamma$ (it could even fall!). As time goes by the economy converges (from below) to a balanced growth path where output and productivity grow at the new rate $\gamma$. In their calibrated model, [Atkeson and Kehoe, 2007] show that these forces can explain quite well the productivity paradox.

# 8 Final remarks

This paper examined the links between firm microstructure and aggregate productivity. Recent empirical work documenting the importance of productive heterogeneity and reallocation have increased the awareness among macroeconomists of the importance of looking *under the hood* of the aggregate production function. The development of parsimonious models of firm microstructure with very simple aggregation properties have contributed importantly to this very active research area. I have reviewed these basic models and some of the most important applications abstracting from many details and leaving aside other very important considerations and contributions. I will comment briefly on some important ommitted topics.

Our analysis has taken mostly as given the technological frontier or taken an exogenously given process of improvement. Research and Development and Innovation have been very important topics of research both in IO and Macroeconomics. As indicated by our analysis in Appendix 9, the area of *Economic Development, Technological Change and Growth* is the second largest overlap between IO and Macro research. Starting with [**?**], macroeconomists have been trying to understand better the forces behind economic growth. R&D has been and continues to be an important question in IO and there has been a great amount of research in this area. On the empirical side, [**?**] very influential paper developed the idea of measuring a firm's R&D stock as a determinant of productivity. On the theory side, there has been considerable attention on patent races and the timing of innovations (see [**?**]). From a regulation perspective, a large literature has also focused on patent design. The question about the incidence of competition and market structure on innnovation is a very important one that is at the center of both IO and Macro. Indeed, since [**?**], the idea of creative destruction as a major source of productivity growth has played a prominent role.[**?**] is the first paper to incoporate this idea into a general equilibrium macro model to understand the connection between innovation and growth. This is a very active area of research in the boundary of IO and Macro with great promise.

I have also abstracted from strategic considerations, that might play an important role in explaining productivity. This is an area of obvious relevance to IO, but one of considerable complexity. It has prompted severecriticisms to IO theory for its lack of general predictions and robustness that [Peltzman, 1991] well described as "theoretical chaos" There are two responses to this criticsm. One is to try to keep models at the simplest possible level, abstracting from complex strategic considerations and trying to focus on results that might have general application. This has been the direction followed by [**?**] that advance the idea of endogenous/strategic sunk costs as an important determinant of market structure. In the model presented above, exogenous sunk costs play an important role as a barrier to entry and can have a major impact on aggregate productivity. But if strategic sunk costs play an even more prominent role in detering entry, this should be a eustion of major interest to macroeconomists. This is definitely an area that should receive important contributions in the near future. Following [**?**], a second response to theoretical chaos has been to focus on Markov perfect equilibria of a dynamic game. This research is usually tied to structural estimation and has concentrated on somewhat narrow analysis of industries that do not easily aggregate into general insights of use in macroeconomics.

Needless to say, a very important area to analyze the importance of the firm microstructure for the aggregate economy is measurement. As an example, measuring firm level productivity is a complex task given the endogeneous nature of input choices. Starting with the work of [**?**], this question has received considerable attention. The availability of new longitudinal databases, such as the LBD and the role of the Bureau of the Census data centers in facilitating the access of this data to researchers is of considerable importance.

These are very exciting times for research at the boundary of IO and macro. It is an opportunity for IO economists interested in broadening perspectives to look into macro/development/international applications and more macroeconmists to look into IO. Central questions in economics such as the need for a better understanding of the determinants of productivity, innovation and growth are at the core of these two fields.

# References

[Adelman, 1958] I.G. Adelman. A stochastic analysis of the size distribution of firms. *Journal of the American Statistical Association*, pages 893–904, 1958.

[Alfaro *et al.*, 2007] L. Alfaro, A. Charlton, and F. Kanczuk. Firm-size distribution and cross-country income differences. *manuscript, Harvard Business School*, 2007.

[Alvarez and Lucas, 2007] F. Alvarez and R.E. Lucas. General equilibrium analysis of the Eaton–Kortum model of international trade. *Journal of Monetary Economics*, 54(6):1726–1768, 2007.

[Atkeson and Kehoe, 2007] A. Atkeson and P.J. Kehoe. Modeling the transition to a new economy: lessons from two technological revolutions. *American Economic Review*, 97(1):64–88, 2007.

[Bain, 1951] J.S. Bain. Relation of profit rate to industry concentration: American manufacturing, 1936-1940. *The Quarterly Journal of Economics*, 65(3):293–324, 1951.

[Bain, 1954] J.S. Bain. Economies of scale, concentration, and the condition of entry in twenty manufacturing industries. *The American Economic Review*, pages 15–39, 1954.

[Bain, 1956] J.S. Bain. *Barriers to new competition: their character and consequences in manufacturing industries.* Harvard Univ. Press, 1956.

[Bernard *et al.*, 2003] A.B. Bernard, J. Eaton, J.B. Jensen, and S. Kortum. Plants and productivity in international trade. *American Economic Review*, 93(4):1268–1290, 2003.

[Caves, 1998] R.E. Caves. Industrial organization and new findings on the turnover and mobility of firms. *Journal of economic literature*, 36(4):1947–1982, 1998.

[Davis and Haltiwanger, 1992] S.J. Davis and J. Haltiwanger. Gross job creation, gross job destruction, and employment reallocation. *The Quarterly Journal of Economics*, 107(3):819–863, 1992.

[Davis *et al.*, 1998] S.J. Davis, J.C. Haltiwanger, and S. Schuh. *Job creation and destruction.* The MIT Press, 1998.

[Devine Jr, 1983] W.D. Devine Jr. From shafts to wires: Historical perspective on electrification. *The Journal of Economic History*, 43(2):347–372, 1983.

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

[Dunne *et al.*, 1988] T. Dunne, M.J. Roberts, and L. Samuelson. Patterns of firm entry and exit in US manufacturing industries. *The RAND Journal of Economics*, 19(4):495–515, 1988.

[Dunne *et al.*, 1989a] T. Dunne, M.J. Roberts, and L. Samuelson. Plant turnover and gross employment flows in the US manufacturing sector. *Journal of Labor Economics*, 7(1):48–71, 1989.

[Dunne *et al.*, 1989b] T. Dunne, M.J. Roberts, and L. Samuelson. The growth and failure of US manufacturing plants. *The Quarterly Journal of Economics*, 104(4):671–698, 1989.

[Eaton and Kortum, 2002] J. Eaton and S. Kortum. Technology, geography, and trade. *Econometrica*, 70(5):1741–1779, 2002.

[Evans, 1987a] D.S. Evans. Tests of alternative theories of firm growth. *The Journal of Political Economy*, 95(4):657–674, 1987.

[Evans, 1987b] D.S. Evans. The relationship between firm growth, size, and age: estimates for 100 manufacturing industries. *The Journal of industrial economics*, 35(4):567–581, 1987.

[Geroski, 1995] PA Geroski. What do we know about entry? *International Journal of Industrial Organization*, 13(4):421–440, 1995.

[Gibrat, 1931] R. Gibrat. Les inégalités économiques. *Librairie du Recueil Sirey, Paris*, 1931.

[Gort and Klepper, 1982] M. Gort and S. Klepper. Time paths in the diffusion of product innovations. *The Economic Journal*, pages 630–653, 1982.

[Guner *et al.*, 2008] N. Guner, G. Ventura, and Y. Xu. Macroeconomic implications of size-dependent policies. *Review of Economic Dynamics*, 11(4):721–744, 2008.

[Hall, 1987] B.H. Hall. The relationship between firm size and firm growth in the US manufacturing sector. *The Journal of Industrial Economics*, 35(4):583–606, 1987.

[Hart and Prais, 1956] PE Hart and SJ Prais. The analysis of business concentration: a statistical approach. *Journal of the Royal Statistical Society. Series A (General)*, 119(2):150–191, 1956.

[Hopenhayn and Rogerson, 1993] H. Hopenhayn and R. Rogerson. Job turnover and policy evaluation: A general equilibrium analysis. *The Journal of Political Economy*, 101(5):915–938, 1993.

[Hopenhayn, 1992a] H.A. Hopenhayn. Entry, exit, and firm dynamics in long run equilibrium. *Econometrica*, 60(5):1127–1150, 1992.

[Hopenhayn, 1992b] H.A. Hopenhayn. Exit, selection, and the value of firms. *Journal of Economic Dynamics and Control*, 16(3-4):621–653, 1992.

[Hopenhayn, 1993] H.A. Hopenhayn. The shakeout. *Economics Working Papers*, 1993.

[Hsieh and Klenow, 2009] C.T. Hsieh and P.J. Klenow. Misallocation and Manufacturing TFP in China and India. *The Quarterly Journal of Economics*, 124(4):1403–1448, 2009.

[Jovanovic, 1982] B. Jovanovic. Selection and the Evolution of Industry. *Econometrica: Journal of the Econometric Society*, pages 649–670, 1982.

[Leonard, 1988] J.S. Leonard. In the wrong place at the wrong time: The extent of frictional and structural unemployment, 1988.

[Lucas Jr, 1978] R.E. Lucas Jr. On the size distribution of business firms. *The Bell Journal of Economics*, 9(2):508–523, 1978.

[Melitz, 2003] M.J. Melitz. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71(6):1695–1725, 2003.

[Peltzman, 1991] S. Peltzman. The Handbook of Industrial Organization: a review. *The Journal of Political Economy*, 99(1):201–217, 1991.

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

[Schmalensee, 1989] R. Schmalensee. Inter-industry studies of structure and performance. *Handbook of industrial organization*, 2:951–1009, 1989.

[Simon and Bonini, 1958] H.A. Simon and C.P. Bonini. The size distribution of business firms. *The American Economic Review*, 48(4):607–617, 1958.

[Sutton, 1995] J. Sutton. *Sunk costs and market structure: Price competition, advertising, and the evolution of concentration.* Mit Press, 1995.

[Sutton, 1997] J. Sutton. Gibrat's legacy. *Journal of economic Literature*, 35(1):40–59, 1997.

[Viner, 1932] J. Viner. Cost curves and supply curves. *Journal of Economics*, 3(1):23–46, 1932.

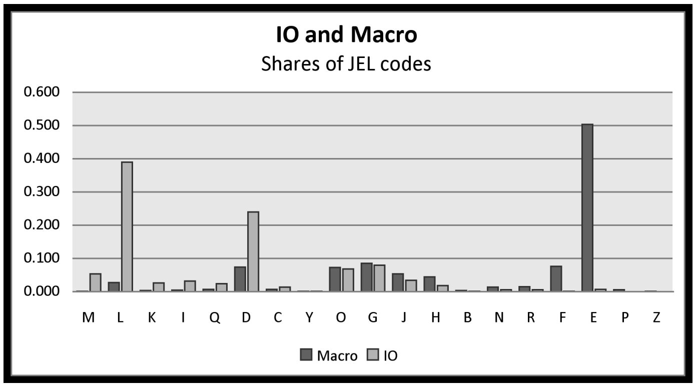

Figure 10: IO and Macro - JEL codes

# 9 Appendix: On the boundary of IO and Macro

I took a random sample of articles published between 1992 and 2010 in Rand Journal and Journal of Monetary Economics (JME), the two top field journals in IO and macro, respectively. We classified each of the papers in the sample –330 from RAND out of 761 published and 300 in JME out of 1234 – according to JEL codes, averaging between two and three codes per paper. Figure **??** provides a distribution of the relative shares of JEL codes in IO and Macro, ordered from left to right by the relative importance of IO. Excluding the small share codes, we can do the following classification:

1. Almost exclusively IO: Codes L (*Industrial Organization)* and M (*Business Administration and Business Economics; Marketing; Accounting)*
2. Almost exclusively Macro: Codes E (*Macroeconomics and Monetary Economics*) and F (*International Economics*)
3. Largely IO but important share of Macro: Code D (*Microeconomics*). The subcodes of more overlap are those related to consumer

Mixed. Ordered by importance: G (*Financial Economics*), O (*Economic Development, Technological Change, and Growth*), J (*Labor and Demographic Economics*) and H (*Health Economics*). The distribution shows pretty clearly two things: 1) that both fields seem to intersect with man other fields 2) the shared boundary between the two fields is pretty large. In order to get a finer

Table 6: Subfields intersection

| **JEL** | **Description** | **RAND** | **JME** |
|---|---|---|---|
| L110 | Production, Pricing and Market Structure; Size Distribution of Firms | 39 | 7 |
| D830 | Searcj; Learning; Information and Knowledge; Communication;Belief | 22 | 5 |
| G210 | Banks, Other Depository Institutions, Micro Financ Inst.; Mortgages | 5 | 23 |
| J240 | Human Capital; Skills; Occupational Choice; Labor Productivity | 5 | 5 |
| J310 | Wage Level and Struct; Wage Diff. by Skill; Training, Occupation, etc. | 5 | 7 |
| O330 | Technological Change: Choices and Consequences; Diffusion Proc. | 11 | 5 |
| D120 | Consumer Economics Empirical Analysis | 4 | 4 |
| D820 | Asymmetric and Private Information | 4 | 4 |

impression on the nature of the intersection it is useful to look at subfields within these JEL aggregates. These are ordered in Table 6 taking the minimum of the frequencies in Rand and in JME as a measure of intersection.